NICOP/N62909-12-1-7071 TR-01

Artificial Vision and
Intelligent Systems Laboratory

Università degli Studi
di Parma
Artificial Vision and
Intelligent Systems Lab

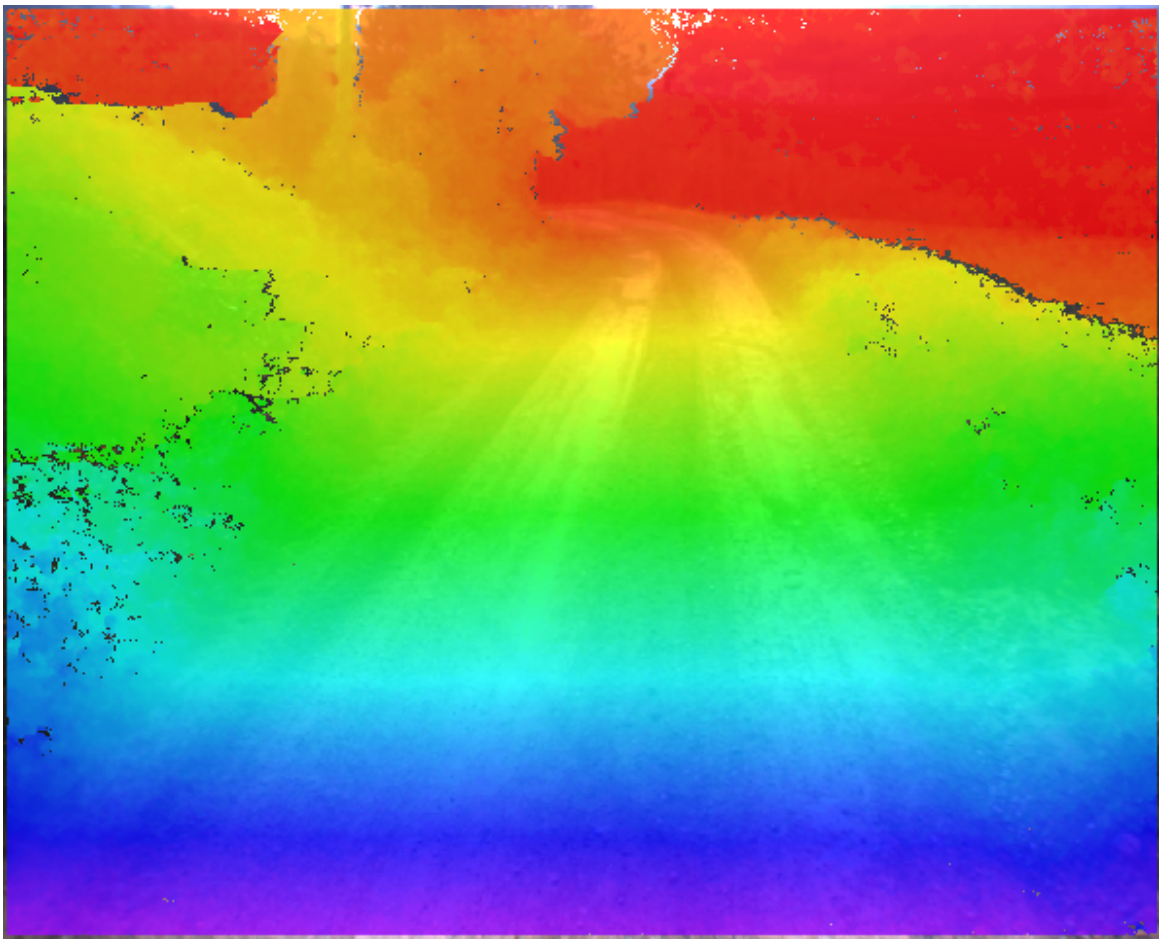
Naval International Cooperative Opportunities
in Science & Technology

Low Cost 3D Vision

Performance analysis of stereo reconstruction algorithms

Alberto Broggi, Mirko Felisa, Néstor Morales Hernández
and Paolo Zani

October 2012



Naval International Cooperative Opportunities in Science & Technology
NICOP/N62909-12-1-7071 TR-01
October 2012

Low Cost 3D Vision

Performance analysis of stereo reconstruction algorithms

Alberto Broggi, Mirko Felisa, Néstor Morales Hernández and Paolo Zani

Artificial Vision and
Intelligent Systems Laboratory (VisLab)
Università degli Studi di Parma
via G.P.Usberti 181/A
43124 Parma, ITALY

Technical Report

This work relates to Department of the Navy Grant N62909-12-1-7071 issued by Office of Naval Research Global.

Prepared for DEPARTMENT OF THE NAVY
Office of Naval Research Global
86 Blenheim Crescent
Ruislip, MIDDX, HA47HB,UK

Under N62909-12-1-7071

Abstract: Autonomous navigation features, including fully autonomous driving, require 360 degrees perception capabilities in a variety of conditions, in order to timely react to a constantly changing scenario. The vast majority of currently available unmanned vehicles use complex sets of active sensors, installed and wired with limited constraints on their position and size to map the environment.

An alternative, lower cost and more integrated solution is to exploit a stereo camera setup to perform dense 3D reconstruction of the vehicle surroundings. This report will provide a detailed analysis of a number of algorithms used to that end, and identify some possible strategies to improve their performance level.

Table of Contents

Figures and Tables

Figures

Tables

1 Introduction

Environment mapping targeted at enabling autonomous operation of a robotic platform has been widely studied over the years, leading to the creation of some prototype vehicles [1, 2, 3] which demonstrated that negotiating moderately complex and dynamic situations in real time was possible, albeit challenging. However, it was only with the development effort driven by the DARPA Challenges [4, 5] that the technology required to provide reliable operation both in off-road and urban scenarios proved to be within reach.

The vehicles that successfully took part to this series of events had to integrate planning and actuation capabilities with a sensing suite capable of coping with harsh environments, heavy traffic and wide temperature ranges, while keeping functional over extended amounts of time. Most competitors relied on high-end active sensors [6, 7, 8, 9], with some notable exceptions [10, 11]. At the time dense stereovision-based 3D mapping was still not feasible on commodity hardware in real time, but even sparse maps [12] proved to be good enough for navigation.

Since then the computational power of COTS [1] hardware has greatly increased, and a number of more advanced algorithms has become viable for autonomous driving applications. A quantitative and meaningful comparison of their performance level, however, is not an easy task, mainly because of the difficulty of producing ground truth information. Older data-sets were small, and either synthetic or taken in controlled environments [13], thus effectively limiting their usefulness as indicators of the actual algorithms ability to cope with outdoor scenarios. More recently the need for suitable metrics led to the definition of improved quality measures, which will be described and used in the following.

This technical report will compare the performance level of some state-of-the-art stereovision-based 3D mapping algorithms in automotive scenarios, using both evaluation sets available in literature and data specifically collected from a dedicated recording platform. One of the algorithms will also be analyzed in greater detail, and a number of variations will be tested in order to determine an improved configuration.

[1] Commercial Off-The-Shelf

2 Experimental setup

Providing reliable quantitative measures about the performance of stereo mapping algorithms in outdoor, uncontrolled scenarios is a difficult task, which can be tackled using different approaches.

One solution [14, 15] is to use an high-end LIDAR [2] unit [16] to directly map the area surrounding the vehicle: depth measurements usually have centimeter-level accuracy over the range 5-100 m, and produce reasonably dense maps, with up to 64 horizontal scanning planes. Another option is to exploit a prior over the data-set, such as the presence of free-space in front of a manually driven vehicle [17] to detect wrongly reconstructed points over an extended period of time. Finally, a virtual view synthesized from the reconstructed environment geometry can be compared with the actual data recorded by a third, suitably positioned physical camera [18, 19].

In this evaluation the algorithms have been tested using all the mentioned approaches, in order to get a better understanding of their actual behavior in real-world applications.

2.1 Dense LIDAR-based ground truth

As a reference, the test data-set available at [20] has been used, since it comes with manually refined ground truth data of about 200 scenes taken in urban and country settings. Ground truth for a given frame is obtained by registering 5 consecutive frames before and after the one selected and accumulating the resulting point clouds; ambiguous regions such as windows and fences are manually removed, and finally the corresponding disparity map is computed using calibration information (Fig. 1).

As performance metrics the percentage of bad pixels (i.e. those showing a measurement error exceeding 3 px), the average correspondence error in pixels and the image density have been gathered, much like it is done on [20].

However, the exact way of computing the values has been slightly changed, since the original metrics didn't look completely fair. In particular:

[2] Light Detection And Ranging, an optical remote sensing technology that can measure the distance to a target by illuminating the target with pulses from a laser

- in this work, only non-occluded, computed pixels have been considered. The original benchmark also gives statistics after linear interpolation of missing values, with the aim of making sparse and semi-dense methods comparable to dense ones; however, such an approximation is hardly optimal, and this reflects on unfairly worsened error metrics for non-dense algorithms.

- average errors have been computed considering only the values below the endpoint error, and not all the values, in order to get a better estimate of the behavior for relevant pixels.

- statistics for each frame are being considered, not just their average over an entire sequence. To better understand the collected data, it will be plotted in a graph with the independent variable (x-axis) representing the measured value, and the dependent one (y-axis) the percentage of frames falling below it. Better-performing algorithms are those with a lower x value for a given frame percentage (e.g. $y = 90\%$).

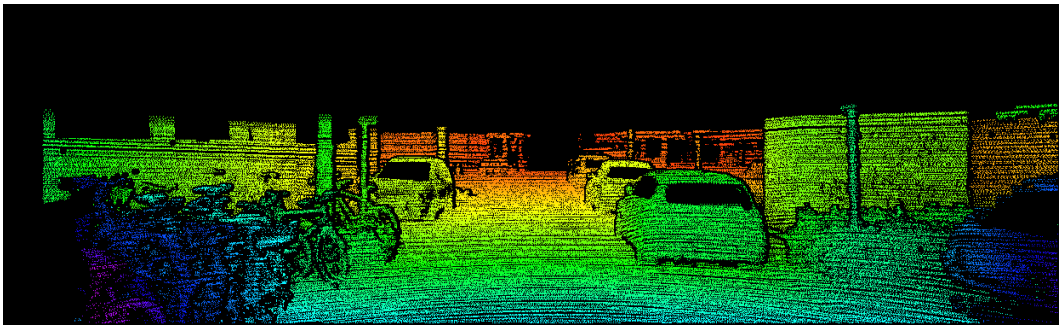
2.2 False correspondences estimation

This benchmark is an adaptation of one of the techniques described in [17]: when driving manually, a safety distance of about 1 s is usually kept from a leading vehicle; this means that a (speed-dependent) volume of free space is present at all times in front of the ego-vehicle, and any reconstructed point falling within said area must be considered as an erroneous estimate, as shown in Fig. 2. The false correspondences percentage $m_{fc} = 100 \times N_{fc}/N$ is then the ratio of points inside the object-free volume with respect to the total number of 3D points.

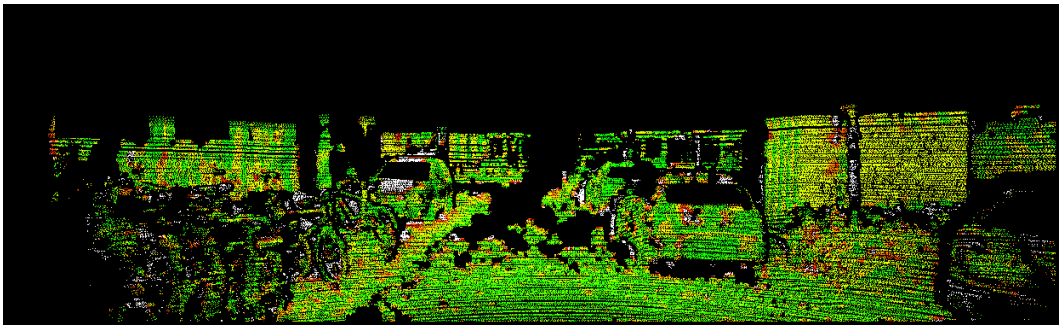
2.3 Normalized cross correlation

The approaches introduced so far have some limitations: LIDAR-based ground truth still takes time to be produced, so it cannot be provided for large datasets, while leading vehicle measurement can be easily performed even on long sequences, but it is an indirect performance metric, albeit a relevant one. As an alternative, the use of a third camera [19], as illustrated in Fig. 3, allows to directly compare a reconstructed view with the actual images without any manual intervention. The computed disparity map is used to transform image pixels taken from the reference camera into control camera coordinates, thus effectively creating a virtual image (Fig. 4-a) that can be compared with that recorded by the control camera (Fig. 4-b) to produce a cross correlation map (Fig. 4-c). However, care must be taken to ensure that the results are

(a)



(b)



(c)

Figure 1. LIDAR-based ground truth example. a) The original image, b) the LIDAR-generated ground-truth disparity map, and c) the error map, with colors spanning from green (error = 0 px) to red (error = 3 px). White pixels correspond to areas with a reconstruction error greater than 3 px.

meaningful: camera calibration is a source of error which must be kept to a minimum, and the occasional presence of obstacles within the object-free volume must be handled as well. The chosen metric is the Normalized Cross Correlation (NCC), computed as:

$$NCC(I_v, I_c) = \frac{1}{|\Omega|} \sum_{(x,y)\in\Omega} \frac{[I_c(x,y) - \mu_c][I_v(x,y) - \mu_v]}{\sigma_c \sigma_v} \tag{1}$$

where Ω is the subset of all pixels having a valid disparity; $\mu_c, \mu_v, \sigma_c, \sigma_v$ are the mean and standard deviation of the control and virtual images respectively.

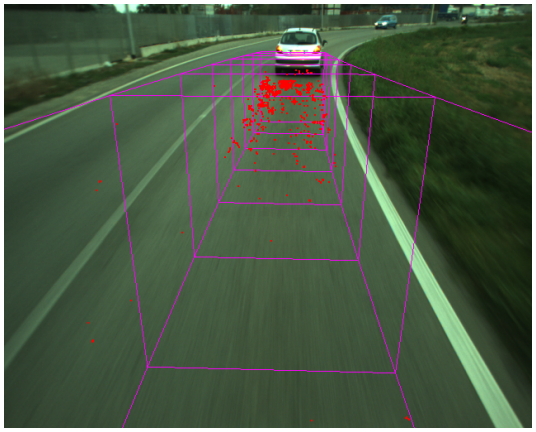


Figure 2. False correspondences estimation example. In pink, the object-free volume which is present in front of the vehicle and, in red, the points falling within said area, produced by false matches.

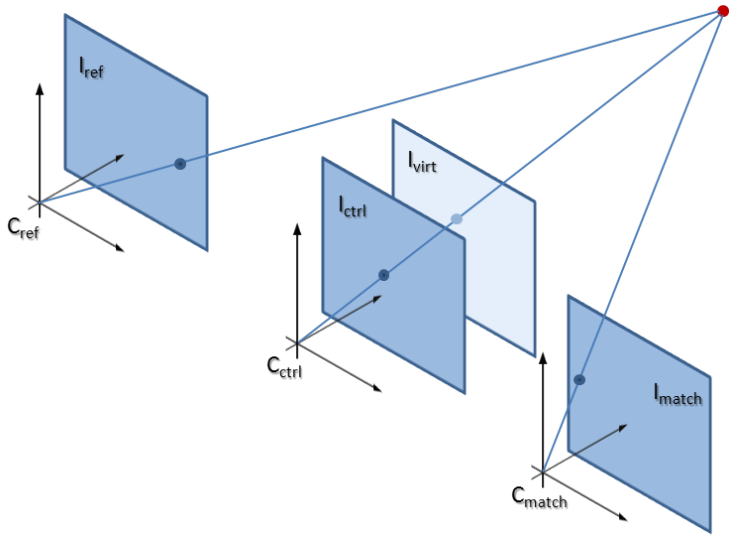


Figure 3. Trinocular camera setup for stereo evaluation. The reference and match cameras (C_{ref} and C_{match} respectively) are used to produce a virtual image I_{virt} in the control camera's reference system that can be compared to the actual recorded image I_{ctrl}.

(a) (b) (c)

Figure 4. NCC metric example. The reference image is used to produce a virtual image a), which is compared to the control camera's view b), to produce a cross-correlation map c).

It is worth noticing that in [18] it is suggested a configuration with the reference and match camera lying 30 cm apart, and the control camera at 50 cm from the reference camera; however, the recording platform used in this work uses much shorter distances (24 and 12 cm respectively), as illustrated in Sec. 2.4, since it has been equipped with a pre-calibrated trinocular camera. Moreover, in this study the control camera has been kept in between the reference and match cameras. This layout makes the virtual and control images look more similar, and could potentially yield to improved NCC scores due to the lower disparity ranges encountered. However, factory calibration is still significantly better than what can be currently achieved with lab equipment, and the resulting accuracy improvement has been deemed more relevant than the quantization introduced by the disparity range reduction.

2.4 Recording platform setup

The tests described in Sec. 2.2 and Sec. 2.3 have been carried out on data acquired using the vehicle depicted in Fig. 5, which has been equipped with a forward-looking Point Grey Bumblebee® XB3-13S2C color camera with 3.5 mm optics working at a resolution of 1280×960 pixels, mounted on top of the van above the windshield. The imaging system is synchronized to a SICK LD-MRS-400001 4-plane LIDAR unit running at 12.5 Hz, with an angular resolution of 0.125 ° and a field of view of 85 °, integrated in the front bull-bar. GPS and INS information are provided by a Topcon AGI-3 unit, and are used to predict the vehicle trajectory.

(a) (b) (c)

Figure 5. The recording platform. Data has been collected using one of the electric vans a) which had been set up in 2010 to take part to the VisLab Intercontinental Autonomous Challenge (VIAC) [21]. The imaging unit b) is a Point Grey Bumblebee®s XB3-13S2C, synchronized to a SICK LD-MRS-400001 LIDAR c) working at a frequency of 12.5 Hz

2.4.1 LIDAR-to-camera calibration

In order to obtain meaningful results for the test described in Sec. 2.2, the LIDAR unit has been used to detect the occasional presence of close preceding vehicles within the defined free space area. To perform this operation it is necessary to measure the relative positioning of the stereo rig and the laser-scanner, which is challenging, given the relatively small amount of data that the available 4-plane LIDAR produces.

The calibration procedure starts with an initial rough alignment step; after that, easily recognizable LIDAR points are manually associated to the corresponding image pixel. The accuracy of each association is constrained by several factors, such as the LIDAR angular resolution, and the ambiguity arising by the limited number of scanning planes hitting each surface; however, a large number of samples can be quickly collected over different frames, and used together in a non-linear Maximum-Likelihood minimization framework, using as a cost function:

$$\arg\min_{[\mathrm{R}|\mathrm{t}]} \sum_i \|p_i - \mathrm{K}[\mathrm{R}|\mathrm{t}]w_i\|^2 \tag{2}$$

with [R|t] being the unknown rototranslation matrix to estimate, p a given 2D (undistorted) image pixel and w the corresponding world point. Assuming to operate under a pin-hole camera model, the camera projection matrix K is defined as:

$$\mathrm{K} = \begin{bmatrix} k & 0 & u_0 \\ 0 & k & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

with k being the pixel focal length and u_0, v_0 the camera optical center. As a non-linear solver, the Levenberg-Marquardt [22] approach has been chosen, given its robustness an relatively fast convergence times.

2.4.2 Camera-to-camera calibration

In order to perform the test described in Sec. 2.3 with the hardware setup in use the relative positioning between all of the cameras has to be computed, since Point Grey Bumblebee® XB3-13S2C cameras only provide combined rectification and dedistorsion look-up tables for left-right and center-right baselines. Let P_{mw} and P_{ms} be two homogeneous disparity points on the right (match, see Fig. 3) camera in the wide and short reference systems respectively:

$$P_{mw} = (u_{rw} - d_{rw}, v_{rw}, -d_{rw}, 1) \qquad P_{ms} = (u_{rs} - d_{rs}, v_{rs}, -d_{rs}, 1) \tag{4}$$

then there exist a 3D homography matrix H so that $P_{ms} = \mathrm{H}P_{mw}$, in the form:

$$\mathrm{H} = \mathrm{Q_s^{-1}} \begin{bmatrix} \mathrm{R} & \mathrm{t} \\ 0 & 1 \end{bmatrix} \mathrm{Q_w} \tag{5}$$

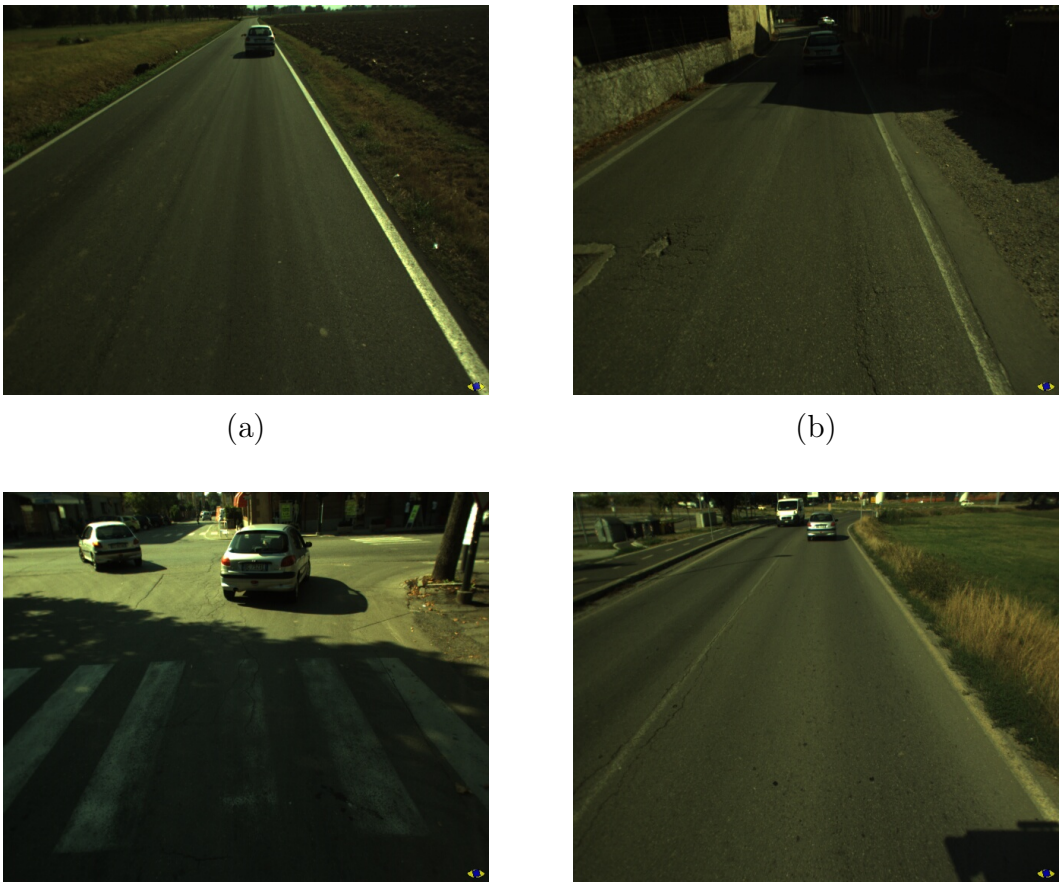
with

$$\mathrm{Q_s^{-1}} = \begin{bmatrix} k_s & 0 & u_{0s} & 0 \\ 0 & k_s & v_{0s} & 0 \\ 0 & 0 & 0 & k_s b_s \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad \mathrm{Q_w} = \begin{bmatrix} \frac{1}{k_w} & 0 & 0 & -\frac{u_{0w}}{k_w} \\ 0 & \frac{1}{k_w} & 0 & -\frac{v_{0w}}{k_w} \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{1}{k_w b_w} & 0 \end{bmatrix} \tag{6}$$

The $k_{\{w,s\}}$ terms in Eq. 6 represent the focal lengths of the rectified images in the wide and short baselines respectively, $u_{0\{w,s\}}$, $v_{0\{w,s\}}$ are the corresponding optical centers and $b_{\{w,s\}}$ the baseline lengths. The R and t terms in Eq. 5, instead, represent the rotation and translation components that align the two baselines. The unknown rotation R is very close to the identity since the three cameras are almost physically aligned, and a linear solver has proved enough to directly estimate the terms of the H matrix, using feature-based correspondences to generate the needed pixelwise associations. Once H is known, the luminance value $I(P_{rw})$ of each point with a known disparity value is used to build the virtual image, by projecting it into coordinates (u_{rs}, v_{rs}), exploiting Eq. 4

2.5 Test data-set

A test sequence has been recorded in a mixed suburban and country environment, as shown in Fig. 6. The data-set has been acquired along a 15 Km loop around the University of Parma campus surroundings; the recording session took place at around 13:14 on a sunny September day, and the scenarios include narrow country roads (Fig. 6-a), small downtowns (Fig. 6-c), intersections (Fig. 6-e) and motorways (Fig. 6-g).



(a) (b)

(c) (d)

Figure 6. Samples of the recorded data.

3 Algorithms

In order to have a broad range of performance statistics, three different dense reconstruction algorithm implementations have been tested. The first two are both based on the so-called Semi-Global Matching approach (in short, SGM) first presented in [23], albeit exploiting different metrics for cost volume initialization, while the latter [24] matches sparse features in the left and right images to restrict the search range of a local window-based approach.

3.1 Semi-Global Matching

The Semi-Global Matching approach aims at identifying the disparity map D that minimizes the energy function

$$E(D) = E_{data}(D) + E_{smooth}(D) \tag{7}$$

with $E_{data}(D)$ representing the pixel-wise matching cost and $E_{smooth}(D)$ a smoothness constraint.

In particular, the $E_{data}(D)$ term is the sum of all pixel matching costs C for the disparities of D:

$$E_{data}(D) = \sum_{\mathrm{p}} C(\mathrm{p}, D_{\mathrm{p}}) \tag{8}$$

while the E_{smooth} term adds a small penalty P_1 to all pixels q in the neighborhood Np of p, for which the disparity varies from p by one, and a higher penalty P_2 if the difference is greater:

$$\begin{aligned} E_{smooth} = & \sum_{\mathrm{q} \in N\mathrm{p}} P_1 \mathrm{T}[|D_{\mathrm{p}} - D_{\mathrm{q}}| = 1] + \\ & \sum_{\mathrm{q} \in N\mathrm{p}} P_2 \mathrm{T}[|D_{\mathrm{p}} - D_{\mathrm{q}}| > 1] \end{aligned} \tag{9}$$

with

$$\mathrm{T}[x] = \begin{cases} 1 \text{ if } x \text{ is true} \\ 0 \text{ otherwise} \end{cases} \tag{10}$$

Global optimization of $E(D)$ is a complex task (i.e. NP-complete), and currently intractable in real time; however, good results can be obtained by applying a dynamic programming strategy that computes the values of $E(D)$

along 1D paths from 8 directions towards each pixel. The costs L'_{r} of each path r are aggregated as described in Eq. 12 for each pixel p and disparity d:

$$\begin{aligned} L'_{\mathrm{r}}(\mathrm{p},d) = C(\mathrm{p},d) + \min(&L'_{\mathrm{r}}(\mathrm{p}-\mathrm{r},d),\\ &L'_{\mathrm{r}}(\mathrm{p}-\mathrm{r},d-1)+P_1, L'_{\mathrm{r}}(\mathrm{p}-\mathrm{r},d+1)+P_1,\\ &\min_i L'_{\mathrm{r}}(\mathrm{p}-\mathrm{r},i)+P_2) \end{aligned} \tag{11}$$

The final disparity value for each pixel is then determined by a winner-takes-all strategy applied to the values of L'_{r}.

3.1.1 Census cost metric

Instead of using mutual information as the pixel-wise matching function, as it is done in the original work [23], the Hamming distance of the Census transform of a 5×5 window cropped around each pixel has been implemented, since it provides similar results [25] while reducing the overall computational burden.

The Census transform of a window W taken from an image I and centered around a pixel p is defined as:

$$census(I,p) = \bigotimes_{\overline{p} \in W} \xi(I(p), I(\overline{p})) \tag{12}$$

where $\bigotimes$ denotes concatenation, and $\xi(I_p, I_{\overline{p}})$ is defined as:

$$\xi(I_p, I_{\overline{p}}) = \begin{cases} 1 \text{ if } I_p < I_{\overline{p}} \\ 0 \text{ otherwise} \end{cases} \tag{13}$$

Each position $C(p,d)$ of the cost volume is then initialized with the number of differing bits between the corresponding transformed values of the left and right images.

3.1.2 Birchfield-Tomasi cost metric

The freely available OpenCV SGM implementation [26] (BT-SGM in the following) uses the Birchfield-Tomasi pixel dissimilarity metric [27] to initialize the cost volume.

Let I_L and I_R be the 1-D functions representing the intensity values along a given scan-line in the left and right images respectively, and $\hat{I}_L(x_L)$, $\hat{I}_R(x_R)$ the linearly interpolated functions between the sample points around x_L and

x_R. It is then possible to determine how well the intensity at x_L fits into the linearly interpolated region surrounding x_R

$$\overline{d}(x_L, x_R, I_L, I_R) = \min_{x_R - \frac{1}{2} \leq x \leq x_R + \frac{1}{2}} \left| I_L(x_L) - \hat{I}_R(x) \right| \tag{14}$$

and symmetrically:

$$\overline{d}(x_R, x_L, I_R, I_L) = \min_{x_L - \frac{1}{2} \leq x \leq x_L + \frac{1}{2}} \left| I_R(x_R) - \hat{I}_L(x) \right| \tag{15}$$

The dissimilarity between the pixels at x_L and x_R then becomes:

$$d(x_L, x_R) = \min\{\overline{d}(x_L, x_R, I_L, I_R), \overline{d}(x_R, x_L, I_R, I_L)\} \tag{16}$$

and is used to initialize the cost volume for any given pixel and disparity value.

3.2 Efficient Large-Scale Stereo Matching

This method, proposed in [24] and referred to as ELAS in the following is particularly suited for handling the high disparity ranges which can arise by using large baselines or very high resolutions images.

It exploits sparse, robustly matched control points to generate a 2D mesh via Delaunay triangulation, which in turn is leveraged to create a prior that is used to reduce the disparity search range for the remaining pixels. Said prior is formed by computing a piecewise linear function induced by the support point disparities and the triangulated mesh.

3.3 Additional filters

A number of pre- and post-processing filters, and their combinations, have been tested in order to determine which would be the most effective in improving the quality metrics discussed in Chap. 2:

- Gaussian filter – the following 3×3 Gaussian smoothing mask is applied to both gray-scale input images:

$$\frac{1}{16} \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \tag{17}$$

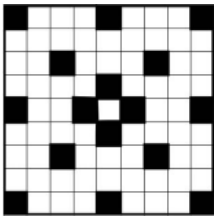


Figure 7. Sparse Census pattern.

- Sparse Census mask – following [28], the sparse pattern displayed in Fig. 7 is used to compute the Census transform of the input images:

- Ternarized Census – in order to improve the amount of information about the local image structure encoded in the resulting images, the Census transform function $\xi(I_p, I_{\overline{p}})$ has been modified to return the following bit strings:

$$\xi(I_p, I_{\overline{p}}) = \begin{cases} 00 \text{ if } I_p - I_{\overline{p}} < -1 \\ 11 \text{ if } I_p - I_{\overline{p}} > 1 \\ 01 \text{ otherwise} \end{cases} \tag{18}$$

- Hamming scores aggregation – as suggested in [28], a window W of size 5×5 centered around each pixel is used to preprocess each score $C(p, d)$ in the input cube:

$$C(p, d) = \frac{1}{25} \sum_{\overline{p} \in W} C(\overline{p}, d) \tag{19}$$

- Uniqueness constraint – The ratio between the first and second minima of the aggregated cost function for a given pixel is used to determine whether a match is reliable or not: higher ratios correspond to a strong minimum, which is more likely to be correct.

- Adaptive mean – an 8×8 adaptive mean filter [24] is applied over the resulting disparity map D:

$$D(p) = \frac{\sum_{\overline{p} \in W} D(\overline{p}) \max\{4 - |D(p) - D(\overline{p})|, 0\}}{\sum_{\overline{p} \in W} \max\{4 - |D(p) - D(\overline{p})|, 0\}} \tag{20}$$

- Despeckle filter – Small disparity image patches with values very different from their neighborhood are usually likely to correspond to wrong associations, so the strategy proposed in [29] is used to identify and remove them.

- Gap filter – constant interpolation along 1D horizontal and vertical paths in the disparity image is performed in order to fill small (≤ 3 px) areas with missing disparity values [24].

Let p_L and p_R be the first two pixels with valid disparity values before and after the current gap; the filling value then becomes:

$$d = \begin{cases} \frac{D(p_L)+D(p_R)}{2} \text{ if } |D(p_L) - D(p_R)| < 3 \\ \min\{D(p_L), D(p_R)\} \text{ otherwise} \end{cases} \quad (21)$$

Each filter has been tested individually against a Census-SGM baseline configuration, and three promising setups have been selected. Each setup has then been compared against other approaches, which also share some of the same filtering strategies, as detailed in Tab. 1.

Table 1. Algorithm configurations. After individual testing the most promising filter setups have been evaluated for the Census-SGM case, while for the BT-SGM and ELAS algorithms the setups suggested in [15] have been followed.

	Census-SGM			BT-SGM	ELAS
	Config 1	Config 2	Config 3		
Gaussian filter	✓	✓	✓	-	-
Sparse Census mask	-	-	-	-	-
Ternarized Census	-	-	-	-	-
Hamming scores aggregation	-	-	-	-	-
Uniqueness constraint	10	20	20	10	15
Adaptive mean	✓	✓	✓	-	✓
Despeckle filter	✓	✓	✓	✓	✓
Gap filter	✓	✓	-	-	✓
Other parameters	P1=10, P2=50, L/R check			see [30]	see [30]

4 Benchmarks

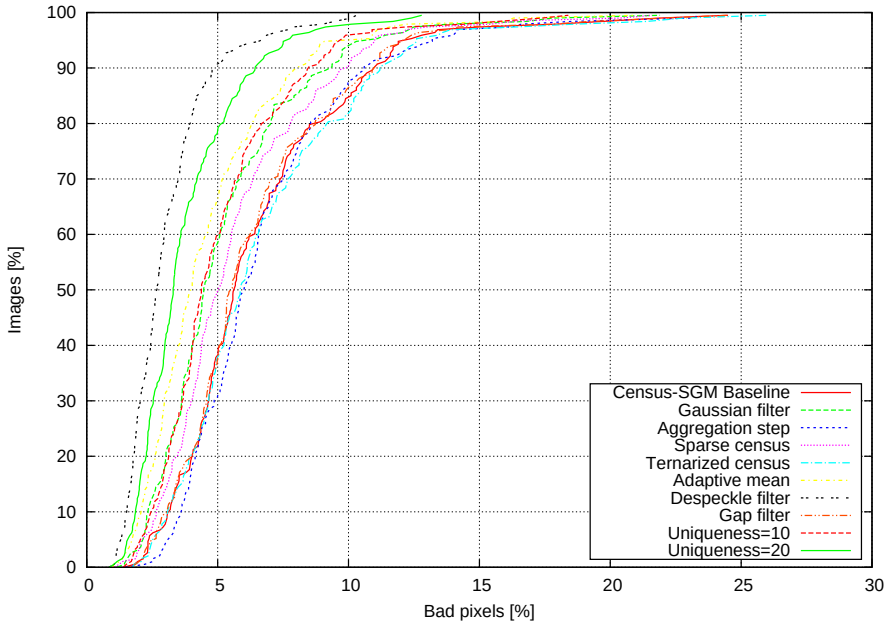
As explained in Sec. 2.1, full performance graphs are presented in this chapter for the tests performed; for the sake of brevity, the following notations will be used:

- LGT – LIDAR-based ground truth evaluation (Sec. 2.1).
- NFC – Number of false correspondences (Sec. 2.2).
- NCC – Normalized cross correlation (Sec. 2.3).

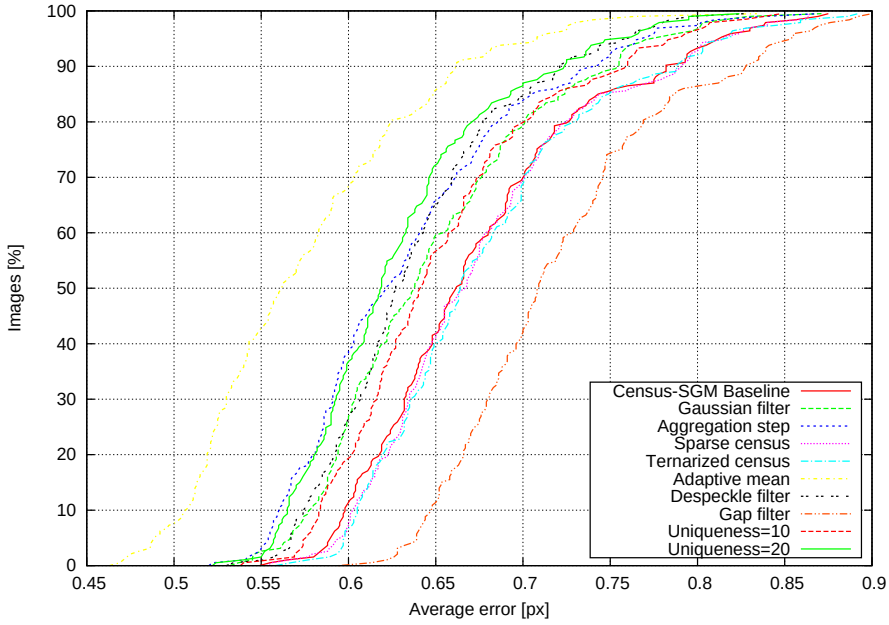
4.1 Isolated filters

LGT evaluation results for each single filter presented in Sec. 3.3 are plotted in Fig. 8. Biggest improvements can be obtained through the use of the despeckle filter; the uniqueness constraint with a strict threshold is also quite effective at removing spurious values, albeit at the cost of a reduced density. Adaptive mean consistently reduces the average reconstruction error, even if on a relatively small scale (around 0.1 px). At the level of sub-pixel error the gap filter produces worse results, which can be explained by the fact that the constant value interpolation that it performs is not accurate enough to capture pixel to pixel variations in the disparity values. Fig. 9 shows the results for the NFC test: the despeckle and uniqueness filters still show clear improvements, as it does the adaptive mean, reinforcing the idea that their combined use is likely to boost the reconstruction performance. Results for the NCC test are plotted in Fig. 10: unfortunately, the scores obtained from the different filters are almost overlapping. This behavior is not easily explained, although some factors are likely to contribute to it:

- NCC scores depend on the luminance of the image points being compared, so wrong reconstructions are not evenly weighted;
- the relatively small baseline in use reduces the measurable effects of reconstruction errors;
- the reconstruction quality is always quite good when using the algorithms described in Chap. 3 irrespective of the filters applied, and the test scores might be dominated by other error sources, such as the calibration.



(a)



(b)

Figure 8. Isolated filters LGT performance. a) Bad pixels percentage and b) average error using LGT.

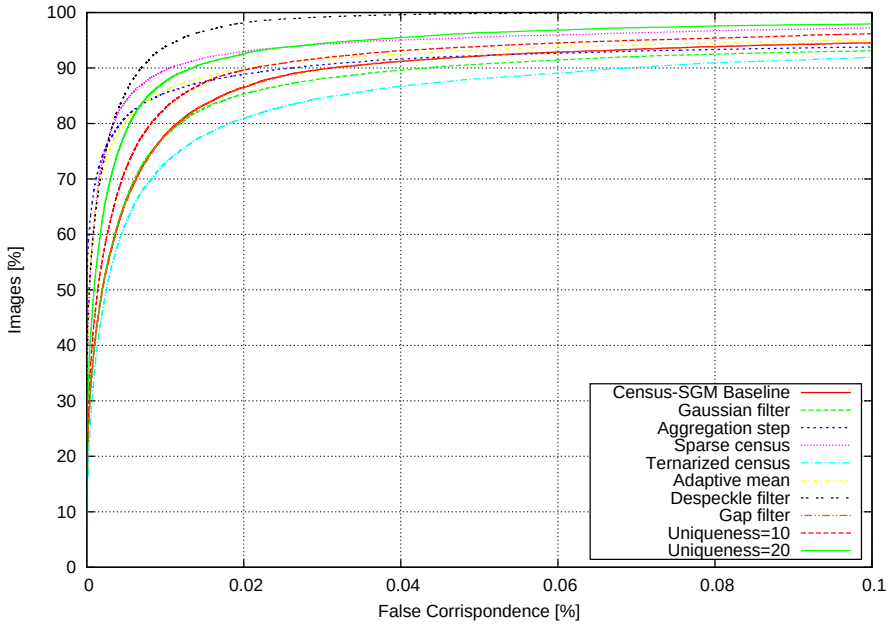


Figure 9. Isolated filters NFC performance.

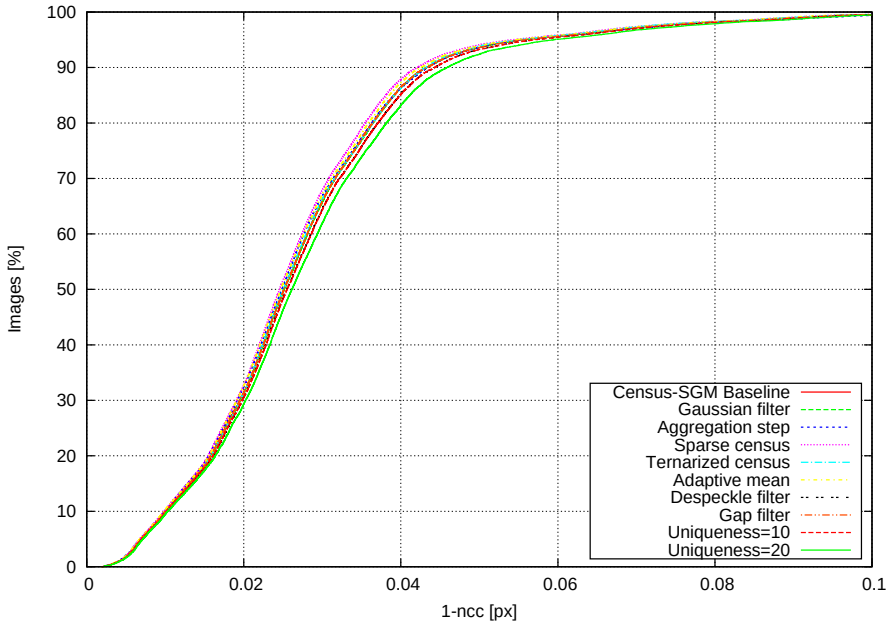
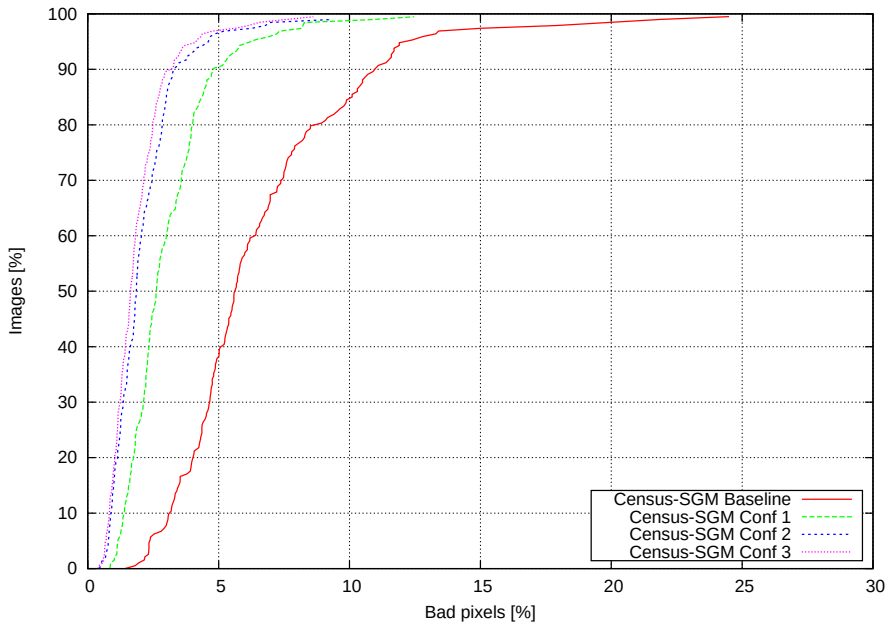


Figure 10. Isolated filters NCC performance.

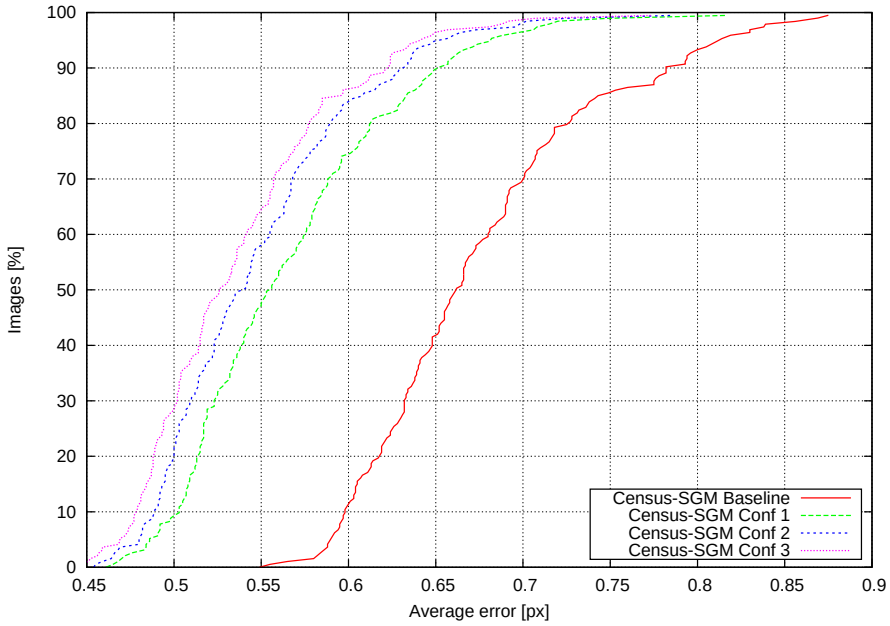
4.2 Composite filters

By combining different filters it is possible to obtain even better performances than when using them separately. Fig. 11 plots the results for the three Census-SGM configurations described in Tab. 1 under the LGT test. Looking at the 90th percentile of Fig. 11-a it can be observed an improvement of around 6% in the number of pixels exceeding the endpoint error for configurations 2 and 3; the average pixels error, instead, decreases by around 0.175 px for the same two configurations (Fig. 11-b). These improvements, however, come at the cost of a decreased disparity map density, as it is apparent in Fig. 11-c: configuration 3, at the 90th percentile, has a density of around 58%, while configuration 2 scores better, at about 65%, which can still be considered acceptable for autonomous driving tasks; for comparison, the baseline method has a density close to 78%, with 12% bad pixels. Configurations 2 and 3 also produce the best results in the NFC test, effectively reducing the number of wrong reconstructions falling within the vehicle trajectory to a negligible amount. The NCC test, instead, seems to indicate an opposite behavior across the tested configurations, but as explained in Sec. 4.1 this data is likely to be very loosely related to the configuration in use.

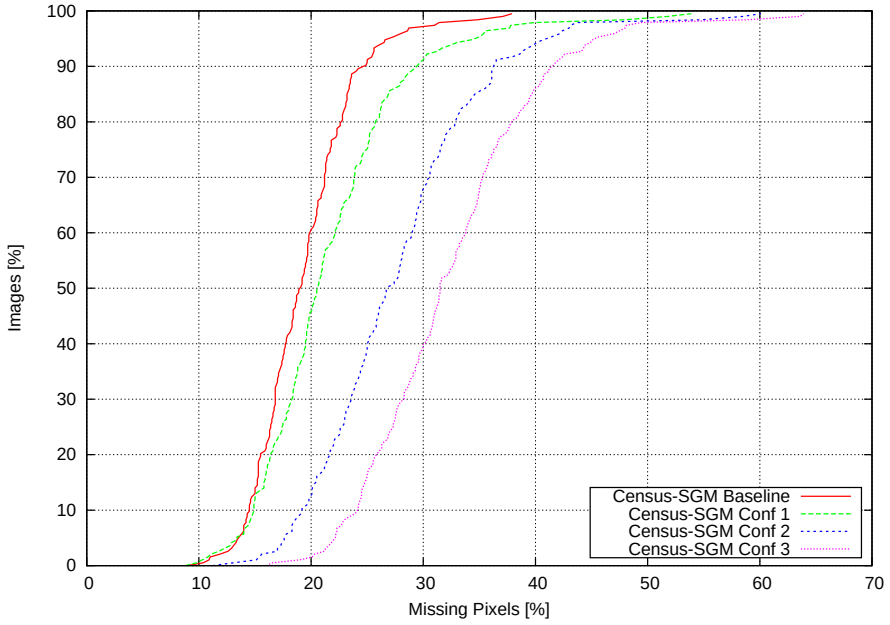


(a)

Figure 11. (cont.) Composite filters LGT performance: a) bad pixels percentage



(b)



(c)

Figure 11. b) average error, c) output density using LGT.

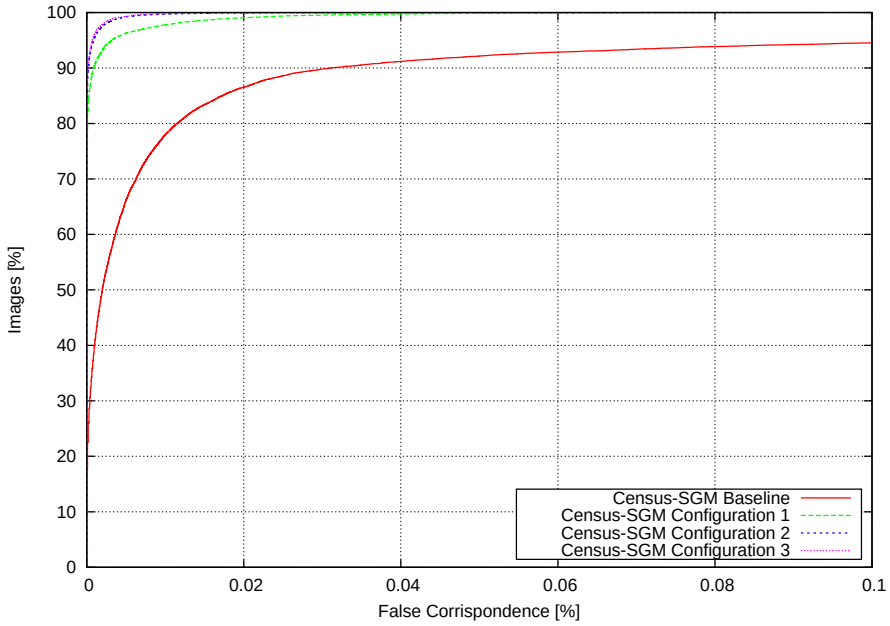


Figure 12. Composite filters NFC performance.

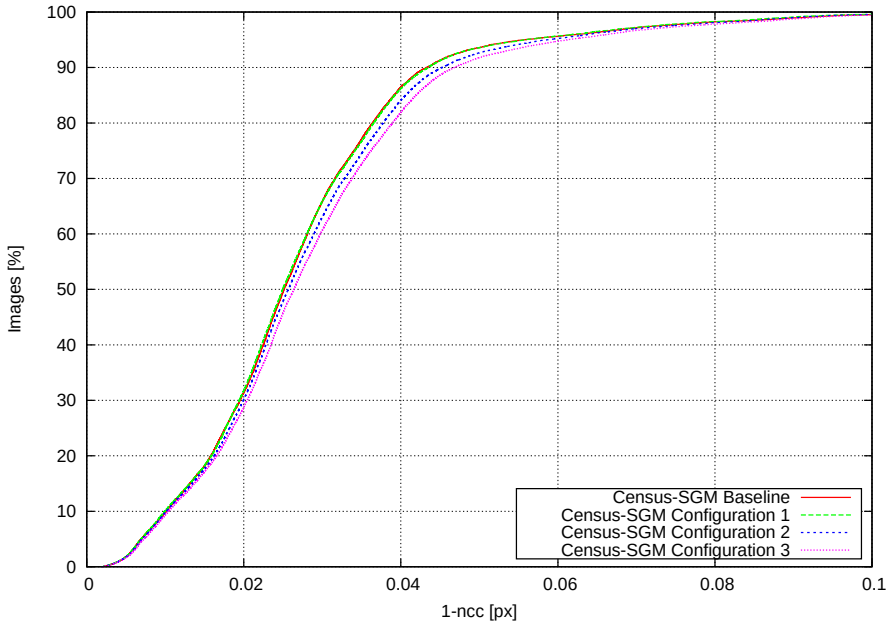
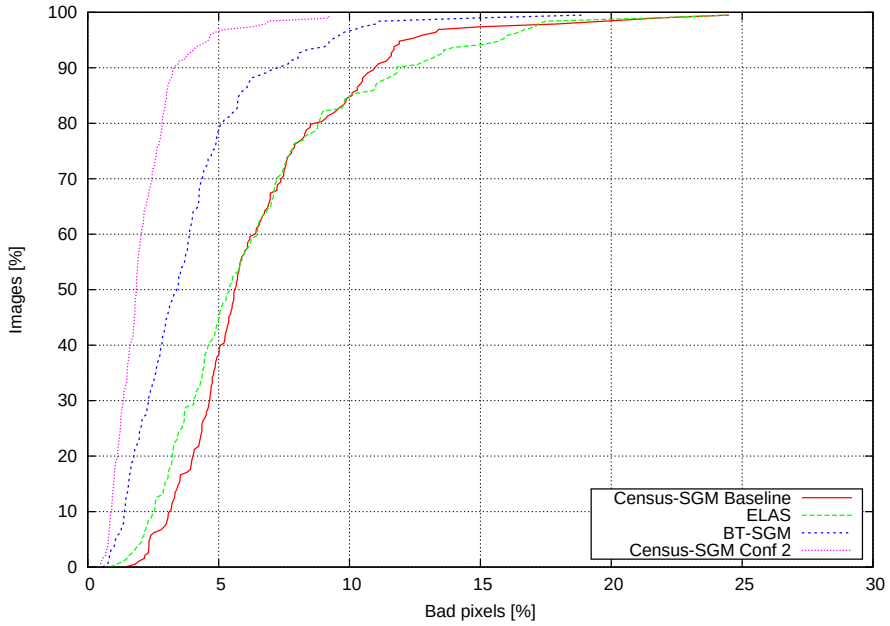


Figure 13. Composite filters NCC performance.

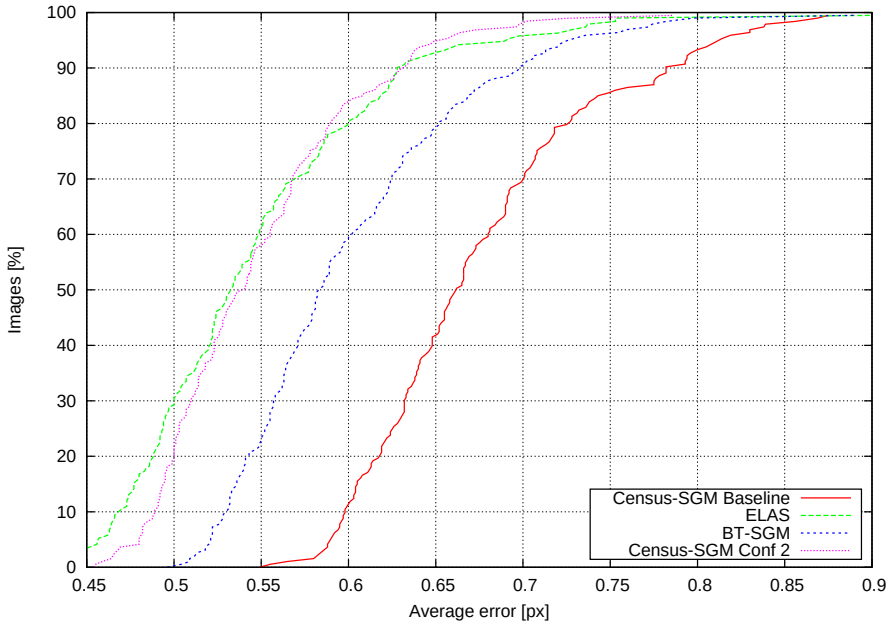
4.3 Algorithms comparison

Census-SGM configuration 2 has been selected as the best compromise between reconstruction quality and map density, and the following graphs illustrate its performance compared to that of the other two approaches described in Chap. 3. LGT evaluation (Fig. 14) shows that the bad pixel percentage is cut by around 7.5% at the 90th percentile with respect to the baseline configuration, and by about 4.5% if compared to the BT-SGM algorithm. The average error is also reduced by 0.15 px, when using Census-SGM configuration 2, making it in line with the values obtained by ELAS. The missing pixels percentage increases to around 35%, which is 12% more than the baseline setup; however, a substantial portion of the additional unreconstructed points is due to the improved error suppression capabilities of the algorithm, and as such is expected behavior.

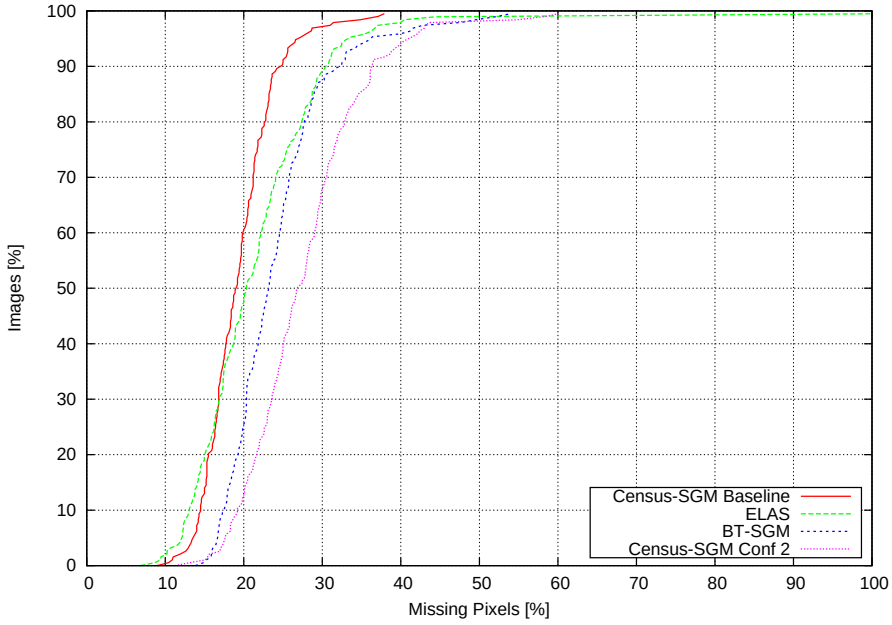


(a)

Figure 14. (cont.) Algorithms LGT performance: a) bad pixels percentage



(b)



(c)

Figure 14. b) average error, c) output density using LGT.

NFC evaluation produces results which are in line with the one obtained with the LGT test, which means that Census-SGM configuration 2 is measurably and consistently better than the alternative approaches, and as such the winning algorithm in this comparison.

NCC scores for the ELAS and BT-SGM algorithms are quite close, and better than the Census-SGM baseline configuration (as expected), but the placement of Census-SGM configuration 2 looks suspicious (i.e. worse than the Census-SGM baseline configuration). For this reason, data coming from this test will have to undergo further investigation before it can be trusted as a reliable indicator of an algorithm's performance.

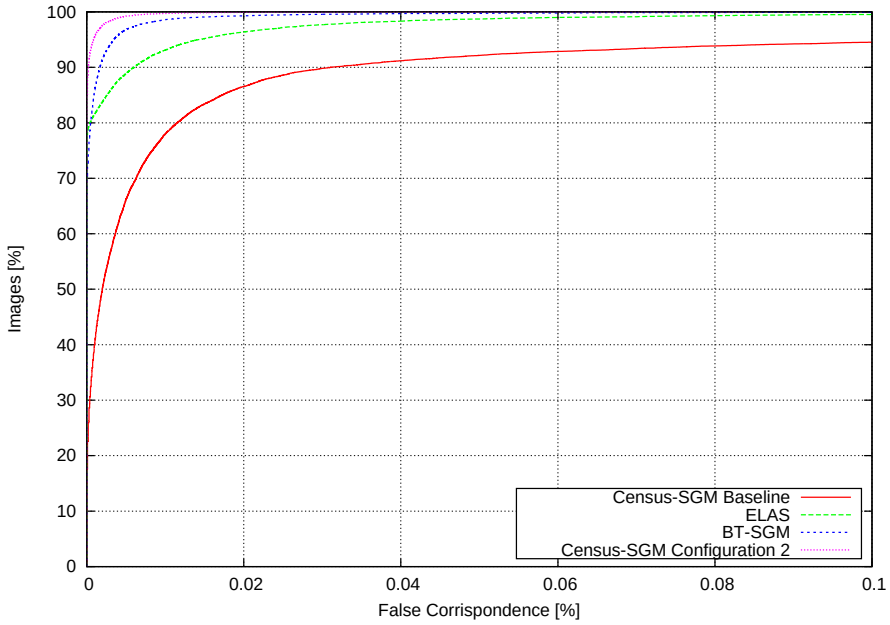


Figure 15. Algorithms NFC performance.

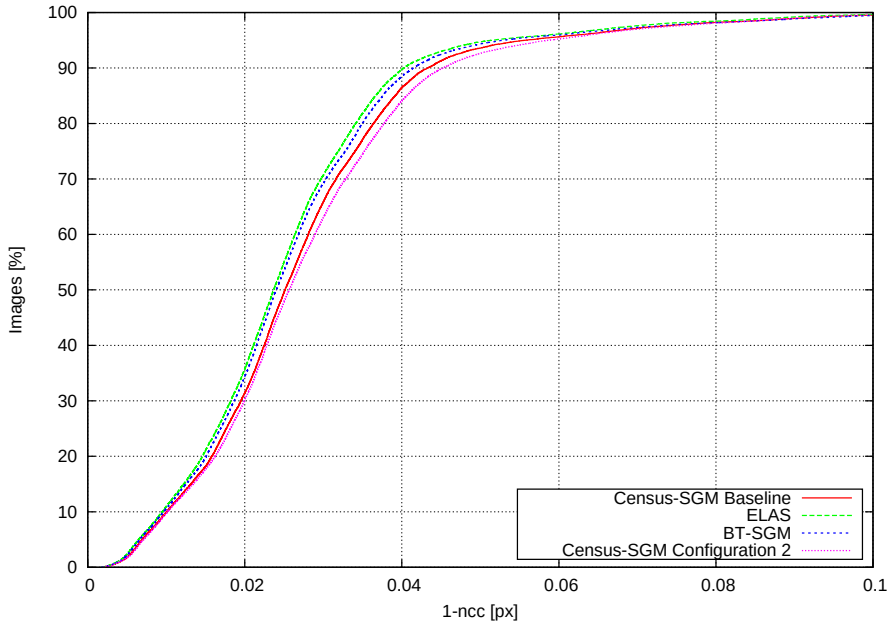


Figure 16. Algorithms NCC performance.

5 Conclusions

The tests conducted so far have quantitatively confirmed how targeted filtering strategies can substantially reduce the amount of wrong pixels computed during stereo reconstruction, while also improving the results accuracy.

In particular, the Census-SGM configuration 2 described in Sec. 3.3 reduces the number of bad pixels by 7.5% at the 90th percentile, while also improving the average error by 0.15 px, making it the candidate of choice among those tested.

This evaluation also served to validate the strategies employed. While LIDAR-based ground truth (see Sec. 2.1) is very effective at producing reliable statistics, it remains quite expensive to carry out, both in terms of the equipment required and of the manual post-processing that has to be performed to produce a single frame.

As an alternative, the use of a prior on the vehicle movement (Sec. 2.2) can be successfully exploited to identify a portion of the wrongly reconstructed points. The advantage of this approach is that it can be effectively used to evaluate the behavior of an algorithm on big data-sets, thus covering a broad range of environmental conditions. On the other hand, the portion of space that can be checked is limited to the area in front of a moving vehicle, which often times is the most critical, but nonetheless can introduce a bias in the resulting statistics.

The use of a third camera for evaluation (Sec. 2.3) is conceptually appealing, but in practice has shown to produce poor results. Further testing will be needed to assess its real effectiveness in real-world scenarios.

References

[1] M. Maurer, R. Behringer, S. Furst, F. Thomanek, and E. D. Dickmanns. A compact vision system for road vehicle guidance. In Proceedings of the International Conference on Pattern Recognition (ICPR '96) Volume III, ICPR '96, pages 313–317, Washington, DC, USA, 1996. IEEE Computer Society.

[2] Dean Pomerleau and Todd Jochem. Rapidly adapting machine vision for automated vehicle steering. IEEE Expert: Intelligent Systems and Their Applications, 11(2):19–27, April 1996.

[3] Alberto Broggi, Massimo Bertozzi, Alessandra Fascioli, and Gianni Conte. Automatic Vehicle Guidance: the Experience of the ARGO Vehicle. World Scientific, Singapore, April 1999. ISBN 9810237200.

[4] Martin Buehler, Karl Iagnemma, and Sanjiv Singh. The 2005 DARPA Grand Challenge: The Great Robot Race. Springer Publishing Company, Incorporated, 1st edition, 2007.

[5] The DARPA Urban Challenge: Autonomous Vehicles in City Traffic (Springer Tracts in Advanced Robotics). Springer, 1 edition, November 2009.

[6] Chris Urmson, Joshua Anhalt, Drew Bagnell, Christopher Baker, Robert Bittner, M. N. Clark, John Dolan, Dave Duggins, Tugrul Galatali, Chris Geyer, Michele Gittleman, Sam Harbaugh, Martial Hebert, Thomas M. Howard, Sascha Kolski, Alonzo Kelly, Maxim Likhachev, Matt McNaughton, Nick Miller, Kevin Peterson, Brian Pilnick, Raj Rajkumar, Paul Rybski, Bryan Salesky, Young-Woo Seo, Sanjiv Singh, Jarrod Snider, Anthony Stentz, William "Red" Whittaker, Ziv Wolkowicki, Jason Ziglar, Hong Bae, Thomas Brown, Daniel Demitrish, Bakhtiar Litkouhi, Jim Nickolaou, Varsha Sadekar, Wende Zhang, Joshua Struble, Michael Taylor, Michael Darms, and Dave Ferguson. Autonomous driving in urban environments: Boss and the urban challenge. Journal of Field Robotics: Special Issues on the 2007 DARPA Urban Challenge, 25(8):425–466, August 2008.

[7] Michael Montemerlo, Jan Becker, Suhrid Bhat, Hendrik Dahlkamp, Dmitri Dolgov, Scott Ettinger, Dirk Haehnel, Tim Hilden, Gabe Hoffmann, Burkhard Huhnke, Doug Johnston, Stefan Klumpp, Dirk Langer, Anthony Levandowski, Jesse Levinson, Julien Marcil, David Orenstein, Johannes Paefgen, Isaac Penny, Anna Petrovskaya, Mike Pflueger, Ganymed Stanek, David Stavens, Antone Vogt, and Sebastian Thrun. Junior: The stanford entry in the urban challenge. Journal of Field Robotics: Special Issues on the 2007 DARPA Urban Challenge, 25(9):569–597, September 2008.

[8] Andrew Bacha, Cheryl Bauman, Ruel Faruque, Michael Fleming, Chris Terwelp, Charles Reinholtz, Dennis Hong, Al Wicks, Thomas Alberi, David Anderson, Stephen Cacciola, Patrick Currier, Aaron Dalton, Jesse Farmer, Jesse Hurdus, Shawn Kimmel, Peter King, Andrew Taylor, David Van Covern, and Mike Webster. Odin: Team victortango's entry in the darpa urban challenge. Journal of Field Robotics: Special Issues on the 2007 DARPA Urban Challenge, 25(8):467–492, August 2008.

[9] Sören Kammel, Julius Ziegler, Benjamin Pitzer, Moritz Werling, Tobias Gindele, Daniel Jagzent, Joachim Schröder, Michael Thuy, Matthias Goebl, Felix von Hundelshausen,

Oliver Pink, Christian Frese, and Christoph Stiller. Team annieway's autonomous system for the 2007 darpa urban challenge. Journal of Field Robotics: Special Issues on the 2007 DARPA Urban Challenge, 25(9):615–639, September 2008.

[10] Alberto Broggi, Claudio Caraffi, Pier Paolo Porta, and Paolo Zani. The Single Frame Stereo Vision System for Reliable Obstacle Detection used during the 2005 Darpa Grand Challenge on TerraMax. In Procs. IEEE Intl. Conf. on Intelligent Transportation Systems 2006, pages 745–752, Toronto, Canada, September 2006.

[11] Alberto Broggi, Andrea Cappalunga, Claudio Caraffi, Stefano Cattani, Stefano Ghidoni, Paolo Grisleri, Pier Paolo Porta, Matteo Posterli, and Paolo Zani. TerraMax Vision at the Urban Challenge 2007. IEEE Trans. on Intelligent Transportation Systems, 11(1):194–205, March 2010.

[12] Mirko Felisa and Paolo Zani. Incremental Disparity Space Image computation for automotive applications. In Procs. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems, St.Louis, Missouri, USA, October 2009.

[13] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision, 47:7–42, 2001.

[14] Sandino Morales and Reinhard Klette. Ground truth evaluation of stereo algorithms for real world applications. In ACCV Workshops (2), pages 152–162, 2010.

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In Computer Vision and Pattern Recognition (CVPR), Providence, USA, June 2012.

[16] `http://velodynelidar.com/lidar/hdlproducts/hdl64e.aspx`.

[17] Pascal Steingrube, Stefan K. Gehrig, and Uwe Franke. Performance evaluation of stereo algorithms for automotive applications. In Proceedings of the 7th International Conference on Computer Vision Systems: Computer Vision Systems, ICVS '09, pages 285–294, Berlin, Heidelberg, 2009. Springer-Verlag.

[18] Sandino Morales, Simon Hermann, and Rehinard Klette. Real-world stereo-analysis evaluation. Technical Report MItech-TR-77, The University of Auckland, New Zealand, 2011.

[19] Sandino Morales and Reinhard Klette. A third eye for performance evaluation in stereo sequence analysis. In Proceedings of the 13th International Conference on Computer Analysis of Images and Patterns, CAIP '09, pages 1078–1086, Berlin, Heidelberg, 2009. Springer-Verlag.

[20] `http://www.cvlibs.net/datasets/kitti`.

[21] Massimo Bertozzi, Luca Bombini, Alberto Broggi, Michele Buzzoni, Elena Cardarelli, Stefano Cattani, Pietro Cerri, Stefano Debattisti, Rean Isabella Fedriga, Mirko Felisa, Luca Gatti, Alessandro Giacomazzo, Paolo Grisleri, Maria Chiara Laghi, Luca Mazzei, Paolo Medici, Matteo Panciroli, Pier Paolo Porta, and Paolo Zani. The VisLab Intercontinental Autonomous Challenge: 13,000 km, 3 months, no driver. In Procs. 17^{th} World Congress on ITS, Busan, South Korea, October 2010.

[22] K. Levenberg. A method for the solution of certain nonlinear problems in least squares. Q. Appl. Math, 2:164–168, 1944.

[23] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02, CVPR '05, pages 807–814, Washington, DC, USA, 2005. IEEE Computer Society.

[24] Andreas Geiger, Martin Roser, and Raquel Urtasun. Efficient large-scale stereo matching. In Proceedings of the 10th Asian conference on Computer vision - Volume Part I, ACCV'10, pages 25–38, Berlin, Heidelberg, 2011. Springer-Verlag.

[25] Heiko Hirschmuller and Daniel Scharstein. Evaluation of stereo matching costs on images with radiometric differences. IEEE Trans. Pattern Anal. Mach. Intell., 31(9):1582–1599, September 2009.

[26] `http://opencv.willowgarage.com`.

[27] Stan Birchfield and Carlo Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. IEEE Trans. Pattern Anal. Mach. Intell., 20(4):401–406, April 1998.

[28] C.D. Pantilie and S. Nedevschi. Sort-sgm: Subpixel optimized real-time semiglobal matching for intelligent vehicles. Vehicular Technology, IEEE Transactions on, 61(3):1032 –1042, march 2012.

[29] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. IEEE Trans. Pattern Anal. Mach. Intell., 30(2):328–341, February 2008.

[30] `http://www.cvlibs.net/datasets/kitti/eval_stereo_flow.php?benchmark=stereo`.