Robust Confidence Intervals in Stereo Matching using Possibility Theory

Roman Malinowski, Emmanuelle Sarrazin, Loïc Dumas, Emmanuel Philippe Dubois, Sébastien Destercke

Robust Confidence Intervals in Stereo Matching using Possibility Theory

Roman Malinowski
CNES, CS, UTC
18 avenue E. Belin, Toulouse, France
roman.malinowski@utc.fr

Emmanuelle Sarrazin
Centre National d'Etudes Spatiales (CNES)
18 avenue E. Belin, Toulouse, France
emmanuelle.sarrazin@cnes.fr

Loïc Dumas
CS
5 rue Brindejonc des Moulinais, Toulouse, France
loic.dumas@csgroup.eu

Emmanuel Dubois
Centre National d'Etudes Spatiales (CNES)
18 avenue E. Belin, Toulouse, France
emmanuel.dubois@cnes.fr

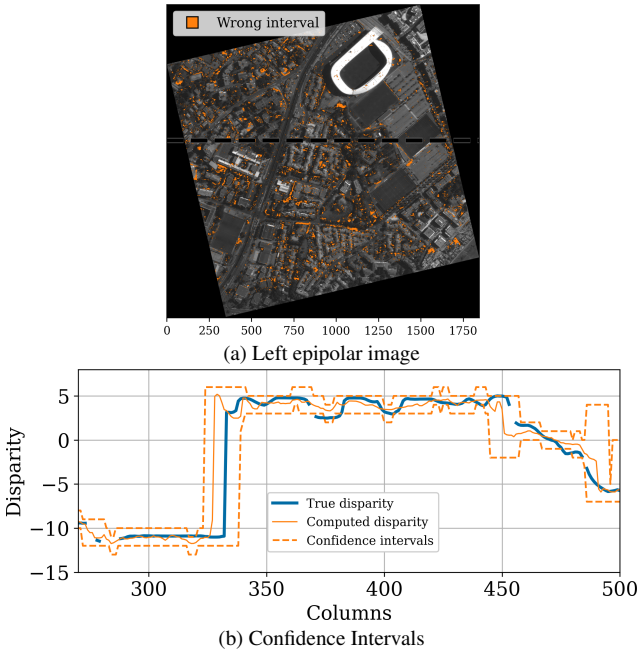
Sébastien Destercke
UTC
Université de Technologie de Compiègne (UTC)
Avenue de Landshut, Compiègne, France
sebastien.destercke@utc.fr

Abstract

We propose a method for estimating disparity confidence intervals in stereo matching problems. Confidence intervals provide complementary information to usual confidence measures. To the best of our knowledge, this is the first method creating disparity confidence intervals based on the cost volume. This method relies on possibility distributions to interpret the epistemic uncertainty of the cost volume. Our method has the benefit of having a white-box nature, differing in this respect from current state-of-the-art deep neural networks approaches. The accuracy and size of confidence intervals are validated using the Middlebury stereo datasets as well as a dataset of satellite images. This contribution is freely available on GitHub.

1. Introduction

Stereo matching is used as a mean to estimate the depth of a scene in numerous applications, ranging from autonomous driving to Earth observation [11, 19]. With the growing availability of satellite imagery [22], many stereo algorithms have been proposed to perform 3D reconstruction from remote sensing images [10, 26, 32, 36]. All those algorithms contain a dense matching step, which consists in determining the displacement of every pixel, called disparity, between the different images. Such algorithms usually start by computing the similarity or sets of features between pixels in the form of a cost volume, from which the disparity can then be deduced [18, 30]. Point clouds are retrieved



(a) Left epipolar image

(b) Confidence Intervals

Figure 1. Example of intervals on a image of the city of Montpellier, France. Fig. 1a presents the left image, colored pixels indicate wrong interval locations. Fig. 1b contains confidence intervals along a section of the dashed line in Fig. 1a

from the disparity, which can themselves be converted into a mesh or a digital surface model.

Estimating the confidence in the disparity estimation is crucial in many applications, and can lead to the improvement of overall results [17, 27]. In the context of 3D reconstruction from satellite imagery, it can even be propagated to be provided as the confidence related to the final 3D product. As such, it has become an important research topic [16, 25]. There are two aspects to the uncertainty regarding disparity computation: how confident we are in the disparity prediction, and what would be the magnitude of the potential error. Although state-of-the-art confidence measures reliably indicate how likely a predicted disparity is to be correct, they do not indicate the extent of the potential error. Those two notions are linked, but are not the same: it could be that a prediction is made with high confidence but would have a large associated error if wrong (meaning there would be a great gap between the predicted and the true disparity in case of an error). Similarly, a prediction could be made with low confidence, but the set of possible disparities is restricted to few values close to the predicted disparity. Estimating the magnitude of the error and providing sets of possible disparities bring additional information that could help users to improve the disparity map, similarly to current work with classical confidence measures [24, 33].

In this article, we present a method for creating robust disparity confidence intervals on the disparity estimation. The intervals will be propagated in future applications to produce confidence intervals on digital surface models, but this lies outside the scope of this paper and thus will not be covered here. We design our method so that it can be fully integrated in classical 3D pipelines [10, 26, 32, 36] using a cost-volume based stereo matching algorithm depicted in Scharstein *et al*. [30]. To the best of our knowledge, this is the first approach providing disparity confidence intervals for stereo matching problems. For each pixel of the reference image, we give a lower and upper displacement of its position in the target image as in Fig. 1. We aim for an accuracy of 90%, using robust uncertainty models called *possibility distributions*. We think this additional information about the magnitude of the error gives a deeper understanding of the uncertainty in stereo matching algorithms. No training is required to produce confidence intervals, also sparing the method from classical criticisms regarding black-box aspects, in the sense that all processing can be followed and monitored. Additionally, we detail precautions that must be taken when applying some post-processing steps [30] to the disparity map so that it stays consistent with the confidence intervals. Our method for creating confidence intervals can be summarized as follows:

1. Computation of the matching cost volume and confidence measure
2. Transformation of matching cost curves into possibility distributions
3. Deduction of disparity intervals by taking α-cuts on the possibility distributions
4. Filtering of intervals while maintaining consistency with the disparity map
5. Statistical regularization of intervals in low-confidence zones

Section 2 contains an overview of current work regarding stereo algorithms, confidence measures, and uncertainty models used in this paper. Section 3 details the method for constructing confidence intervals. Finally, the confidence intervals robustness and size are validated in Section 4 using images from the Middlebury dataset and from a dataset of high-resolution optical satellite images of various landscapes around the city of Montpellier. The code is freely available on GitHub: `https://github.com/CNES/Pandora`.

2. Related Works

2.1. Stereo matching

Our method is designed for classical 3D reconstruction pipelines from remote sensing imagery [10, 26, 32, 36]. Those pipelines retrieve stereo information by means of dense matching algorithms, mainly falling into two main categories: classical approaches following the steps established by Scharstein *et al*. [30], and full deep-learning approaches. Classical approaches usually contain the following steps: matching cost computation, cost aggregation, disparity computation and disparity refinement. On the other hand, deep-learning approaches provide strong results (see [18] for details), but are prone to generalization issues when using images differing from the training dataset, especially in the context of satellite imagery [20]. Obtaining ground truth data of various landscape can prove difficult, we therefore focus here on classical approaches as we aim to produce a general and robust method for creating confidence intervals. We consider two similarity functions, the Census cost function [37], and the MC-CNN cost function [38] learned using convolutional neural networks. Those similarity functions are *minimitive*, meaning that a low value indicates a strong similarity of the compared patches. Both handcrafted and learned similarity functions are considered here to highlight the generic nature of our method for creating intervals. The cost volume obtained using those functions is regularized using the semi-global matching (SGM) methods [9, 15], used in other state-of-the-art methods [3].

Numerous confidence measures have been proposed regarding the disparity [16], handcrafted using the properties from the reference image, the cost volume or the disparity map itself. Learning-based methods constitute the majority of state of the art confidence measures [25]. We can mention for instance deep learning methods estimating confi-

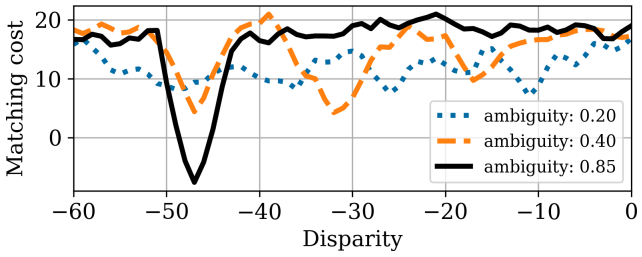


Figure 2. Example of three MC-CNN cost curves with different confidences from ambiguity.

dence using CNNs on the cost volume [21] or on the disparity map [24], and random forests on handcrafted confidence measures [12]. We refer to [16, 25] for more in-depth details. Estimating the confidence supporting the disparity map can lead to new strategies to improve the overall results [8, 17, 34]. In this work, we use use a confidence measure computed from the matching cost volume, called *ambiguity* [27].

In [4], authors estimate the magnitude of the absolute symmetric error using a MLP. Their work demonstrate the interest in estimating the magnitude of the error as a complement to confident measures, and we push this idea even further. Indeed, estimating the absolute error provides valuable information on the magnitude of the error, but does not indicate *where* the correct disparity should be, i.e. if it is probably higher or lower than the predicted one.

We therefore propose to answer this question using confidence intervals on the disparity. Our method also differs as it can be plugged on a vast range of cost-volume based stereo algorithms with *winner-takes-all* strategy. Additionally, our method does not rely on the accuracy of the algorithm used for the disparity prediction. In contrast, the network in [4] specifically takes as input 4 disparities at various resolutions estimated by a 3D CNN [13], limiting its applicability. To the best of our knowledge, this is the only work that estimates the disparity error in a manner similar to ours. The restricted setting in which they work also means that we cannot compare our two approaches.

2.2. Possibility distributions

To create confidence intervals, we consider using *possibility distributions*, closely related *fuzzy sets* [6], as uncertainty models. This allows to correctly represent *epistemic* uncertainty, i.e., due to the partial nature of available information [1]. In stereo matching problems, errors are mostly due to epistemic uncertainty. Indeed, similarity functions evaluate how much two patches are alike based on given or learned features, and there exists some uncertainty regarding how well this similarity indicates a match between corresponding pixels. Using possibility distributions aims to address the downsides of probability distributions regarding epistemic uncertainty [35]. Possibility distributions are well-suited to model an expert's opinion on the uncertainty of an imprecise observation, for instance in the context of groundwater contamination [1, 2]. In the context of stereo matching, a regularized cost curve using SGM can be seen as an expert evaluating if two patches are homologous or not, based on their features and global properties of the cost volume. The use of possibility distributions allows to benefit from the advanced state of knowledge of IP towards robust estimation of uncertainty.

Formally, a possibility distribution is defined as a mapping $\pi : \Omega \to [0, 1]$ verifying:

$$\exists \omega \in \Omega, \pi(\omega) = 1 \tag{1}$$

where Ω is the set of possible observed states. π represents the degree of possibility of an event ω, $\pi(\omega) = 1$ meaning that ω is fully possible, and $\pi(\omega) = 0$ meaning that ω is impossible. Possibility distributions can be used to define the envelope of a convex set of probability distributions $\mathbb{P}$ [7]:

$$\mathbb{P} = \{P : 2^{\Omega} \to [0, 1] \mid P(E) \leqslant \sup_{\omega \in E} \pi(\omega)\} \tag{2}$$

where P is a probability distribution on the power set 2^{Ω} of Ω.

Alongside possibility distributions are often defined α-cuts C_{α}^{π}, which will be used for constructing the confidence intervals:

$$C_{\alpha}^{\pi} = \{\omega \in \Omega \mid \pi(\omega) \geqslant \alpha\} \tag{3}$$

α-cuts are the maximal sets whose possibility is at least equal to α for every $\omega \in \Omega$. From a probabilistic point of view, α-cuts are composed of all $\omega \in \Omega$ for which there is a $P \in \mathbb{P}$ from Eq. (2) such that $P(\omega) \geqslant \alpha$:

$$C_{\alpha}^{\pi} = \{\omega \mid \exists P \in \mathbb{P},\ P(\omega) \geqslant \alpha\} \tag{4}$$

3. Creation of Disparity Confidence Intervals

In the following, we consider that the images have been resampled in epipolar geometry so that the displacement of a pixel between left I_L and right I_R images can only occur horizontally in a given disparity range $\mathcal{D} = [d_{min}, d_{max}]$. In this setting, a pixel of the left image $p = (i, j)$ with a disparity $d \in \mathcal{D}$ is matched to the pixel $q = (i, j + d)$ of the right image with a cost $C_V(i, j, d)$.

3.1. From Cost Curves to Possibility Distributions

We use possibility distributions to model the epistemic uncertainty associated with similarity functions, in the same way that an expert would state an opinion for every pixel on which disparities are more likely to be the correct ones.

In order to construct a possibility distribution consistent with a cost curve, it is first necessary to normalize the cost

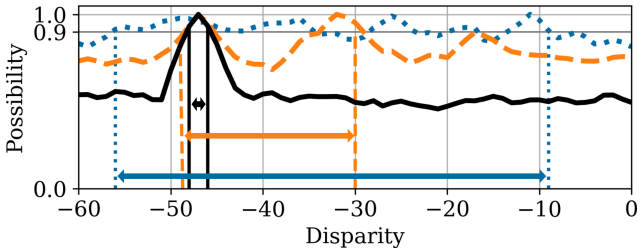


Figure 3. The possibility distributions obtained from the cost curves of Fig. 2. The arrows and vertical lines indicate the disparity intervals obtained with $\alpha = 0.9$.

curve to ensure its values lies in $[0, 1]$. Normalization can be done using the minimum and maximum values of the matching cost volume after SGM regularization. Given a pixel $p = (i, j)$, its cost curve is normalized as follows:

$$f_{i,j}^{norm}(d) = \frac{C_V(i,j,d) - \max C_V}{\min C_V - \max C_V} \tag{5}$$

The $\min$ and $\max$ operators of standard normalization are reversed as the possibility needs to be maximal when the cost function is minimal. $f_{i,j}^{norm}$ is not yet a possibility distribution as it does not necessarily verify Eq. (1). As noticed in Eq. (1), we can interpret it as a form of inconsistency, in the sense that there is no disparity value such that the two patches of pixels perfectly match. Assuming $p = (i, j)$ has an homologous pixel in the right image, we can restore consistency through normalisation [23] , resulting in a possibility distribution $\pi_{i,j}(d) : \mathcal{D} \rightarrow [0, 1]$:

$$\pi_{i,j}(d) = f_{i,j}^{norm}(d) + 1 - \max_{d \in \mathcal{D}} f_{i,j}^{norm}(d) \tag{6}$$

Examples of cost curved transformed into possibility distributions are presented in Figs. 2 and 3. Another way of verifying Eq. (1) would have been to obtain $f_{i,j}^{norm}$ using the $\min$ and $\max$ operators on the cost curve instead of the whole cost volume. Doing so would have artificially accentuated the differences between the matching costs [23]. Using Eq. (6) instead keeps the curvature of the cost curve.

3.2. Deducing Intervals from Alpha-cuts

Having defined possibility distributions, we now look to define a set of possible disparities verifying our 90% confidence objective. Every possibility distribution defines a set of probability distributions $\mathbb{P}$ using Eq. (2). Disparities d for which every probability measure in $\mathbb{P}$ evaluated on d are lower than 0.9 are deemed to be unlikely. According to Eq. (4), considering disparities d for which there is a probability $P \in \mathbb{P}$ such that $P(d) \geqslant 0.9$ is equivalent to consider the α-cut $C_\alpha^{\pi_{i,j}}$ with $\alpha = 0.9$. Different values of α reflect different levels of confidence. We compared different values of α in the ablation study presented in Tab. 2.

In general, α-cuts are sets and not intervals. We are able to define a single confidence interval $I_\alpha(i, j)$ from an α-cut by taking its extrema:

$$I_\alpha(i,j) = [\min C_\alpha^{\pi_{i,j}}, \max C_\alpha^{\pi_{i,j}}] \tag{7}$$

Switching from $C_\alpha^{\pi_{i,j}}$ to the interval $I_\alpha(i, j)$ reduces the amount of information available by adding disparities with low possibilities to our considered set of disparities. However, as only two interval bounds need to be considered instead of every disparity of $C_\alpha^{\pi_{i,j}}$, our solution consumes less memory, facilitates further processing and is easier to understand for users. The level of confidence of $I_\alpha(i, j)$ is also guaranteed to be at least equal to that of $C_\alpha^{\pi_{i,j}}$ as $C_\alpha^{\pi_{i,j}} \subseteq I_\alpha(i, j)$. Examples of confidence intervals on disparities are presented in Fig. 3.

3.3. Refinement and Filtering with Intervals

In most stereo pipelines, the output disparity map deduced from the matching cost volume is being post-processed. Namely, sub-pixel refinement and filtering steps are usually applied to improve overall results. The confidence intervals need to be processed accordingly to maintain their coherence with the disparity map.

Sub-pixel refinement is taking into account by slightly extending the confidence intervals in the case where the predicted disparity $d_{i,j}$ is one of the interval bounds. For clarity, we refer to the lower and upper bounds of an interval I as $\underline{I}$ and $\overline{I}$ respectively. Confidence intervals are modified as follows:

$$\text{if } d_{i,j} = \min C_\alpha^{\pi_{i,j}}, \underline{I}_\alpha(i,j) = \min C_\alpha^{\pi_{i,j}} - 1 \tag{8}$$

$$\text{if } d_{i,j} = \max C_\alpha^{\pi_{i,j}}, \overline{I}_\alpha(i,j) = \max C_\alpha^{\pi_{i,j}} + 1 \tag{9}$$

This interval extension is coherent with different methods of interpolation, like parabolic-fit for instance. V-fit refinement [14] is used in our experiments.

Similarly, multiple filtering of the disparity map can be considered. We use a median filter in our experiments, which is a popular method in many stereo algorithms [9, 30]. Using a median filter modifies the disparity map and might create inconsistencies with the confidence intervals. However, it is possible to demonstrate that for every set of pixels $\{p(i, j)\}$ and confidence intervals $\{I_\alpha(i, j)\}$ verifying $\forall (i, j), p(i, j) \in I_\alpha(i, j)$ then:

$$median(\{\underline{I}_\alpha(i,j)\}) \leqslant median(\{p(i,j)\}) \tag{10}$$

$$median(\{p(i,j)\}) \leqslant median(\{\overline{I}_\alpha(i,j)\}) \tag{11}$$

Previous equations mean that the median filter can be applied independently to the disparity map and the confidence interval bounds while still maintaining their coherence.

3.4. Regularization in Low Confidence Areas

Despite running post-processing steps, confidence interval performances heavily depend on the quality of the similarity function used. Near surface discontinuities, SGM algorithm sometimes struggle to correctly detect disparity changes due to the continuity constraint. A shift between the predicted and the true disparity can be observed, which induces biases in the cost curve and challenges the interpretation of the cost curve as an expert's opinion. To overcome this limitation, confidence intervals are processed with a more pessimistic approach in those areas.

Low confidence areas are detected using confidence measures. As such, our approach is complementary to classical confidence estimation approaches. We use the confidence from ambiguity measure [27] as it presents the advantage of being both explainable and efficient for this task. This confidence measure aims to represent the difficulty to single out a disparity value as the correct disparity. The higher the confidence from ambiguity, the easier it is easy to identify the correct disparity, whereas a value near 0 indicates that numerous patches present the same locally minimal similarity. Examples of cost curves with different values of ambiguities can be found in Fig. 2. The ambiguity measure is computed pixel-wise, and thus can present strong spatial variations in low confidence zones. To compensate this effect, we first smooth the ambiguity map using a 1×5 min convolution kernel, and pixels whose ambiguity is under a threshold τ are considered to be in low confidence zones:

$$amb_{smooth}(i,j) = \min_{-2 \leqslant k \leqslant 2} amb(i, j+k) \tag{12}$$

$$\text{Low confidence if } \; amb_{smooth}(i,j) \leqslant \tau \tag{13}$$

We fix empirically $\tau = 0.6$ for our experiments with ambiguity, as it does not seem to depend on the similarity function used.

Low confidence areas usually correspond to regions which possess a strong disparity variation. The intervals in low confidence areas might be extended to the minimal and maximal bounds found in the area, but experiments show this approach is too pessimistic. Indeed, a single unnecessary large confidence interval would penalize the whole area. Instead, we advocate for a statistical approach by extending the intervals using quantiles instead of the the minimal and maximal bounds. This approach has the advantage of being more robust to outliers.

For each low confidence pixel, we determine the maximal set A of contiguous low confidence pixels within l lines above and bellow, as presented for $l = 2$ in Fig. 4. Experiments show that $l = 2$ allows the set A to contain enough pixels to be statistically relevant, while maintaining a relatively low computation time. Using values higher than 2 does not improve the results. Once A has been established,

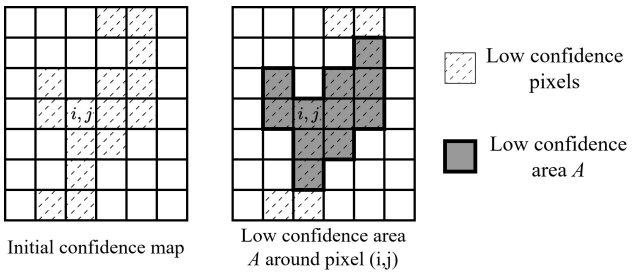
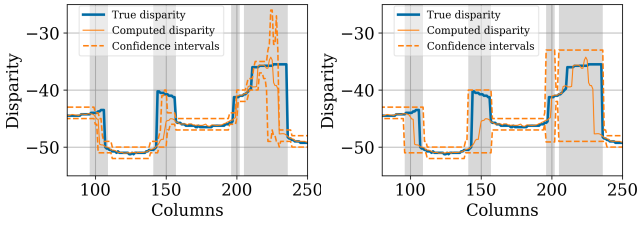


Figure 4. Low confidence areas with $l = 2$.

the lower interval bound of the low confidence pixel is set to the 10^{th} quantile of the lower interval bounds in A. The same procedure is applied to the upper bounds with the 90^{th} quantile. Examples of intervals with and without regularization along a row of a scene are presented in Fig. 5. When no regularization is applied, errors occur for intervals in low confidence areas.



(a) Intervals without regularization (b) Interval with regularization

Figure 5. Intervals without (Fig. 5a) and with (Fig. 5b) regularization from Middlebury's *cones*. CENSUS cost function is used. Areas with low confidence are indicated in gray.

4. Evaluation

We define different metrics in order to simultaneously evaluate intervals reliability and size. The metrics are adapted to our future objective of propagating the disparity intervals into height intervals for 3D reconstruction from satellite imagery. As such, intervals must be both reliable and small. The metrics are first assessed globally by considering every intervals for each dataset. Then, they are measured separately for high confidence and low confidence areas. The distinction between both regions is performed using a threshold on the ambiguity measure, as in Sec. 3.4.

4.1. Evaluation Metrics

The confidence intervals are evaluated following different criteria:

- Their global *accuracy*. An interval is considered accurate if it contains the true disparity. The accuracy is computed as:

$$Acc = \frac{\#accutate\ intervals}{\#intervals} \tag{14}$$

	Global				High confidence areas				Low confidence areas					
	Accuracy ↑		Relative Size ↓		Accuracy ↑		Relative Size ↓		Accuracy ↑		Relative Size ↓		Overestimation ↓	
Dataset	CENSUS	MCCNN	CENSUS	MCCNN	CENSUS	MCCNN	CENSUS	MCCNN	CENSUS	MCCNN	CENSUS	MCCNN	CENSUS	MCCNN
2003	**0.973**	0.954	**0.033**	**0.033**	**0.983**	0.968	**0.033**	**0.033**	**0.942**	0.89	**0.183**	0.233	**0.165**	0.182
2005	0.963	**0.971**	**0.026**	0.038	0.969	**0.973**	**0.026**	**0.026**	0.95	**0.969**	**0.218**	0.256	**0.152**	0.228
2006	**0.989**	**0.989**	**0.026**	0.038	**0.993**	0.992	**0.026**	**0.026**	0.98	**0.985**	0.569	**0.397**	**0.109**	0.268
2014	0.957	**0.983**	0.063	**0.029**	0.912	**0.972**	**0.007**	0.013	0.991	**0.996**	**0.872**	0.993	**0.12**	0.339
2021	0.936	**0.991**	**0.594**	1.0	0.818	**0.969**	**0.012**	0.026	0.987	**0.999**	**0.859**	1.0	**0.168**	0.314
Rural	0.904	**0.975**	**0.100**	0.250	0.867	**0.967**	**0.083**	0.222	0.952	**0.998**	**0.286**	1.0	**0.360**	0.713
Urban	0.926	**0.986**	**0.100**	0.263	0.894	**0.981**	**0.091**	0.238	0.965	**0.999**	**0.286**	1.0	**0.367**	0.722

Table 1. Accuracy Acc, Relative size S_{rel} and Relative overestimation O_{rel} for different Middlebury and satellite datasets. "Global" column consider every intervals in the dataset, while "High confidence areas" and "Low confidence areas" separate the intervals based on the confidence measure (Sec. 3.4). Two cost functions are compared: CENSUS and MC-CNN. The best results for each dataset appear in bold font.

where # refers to the number of elements of a set.

- The *relative size* of the intervals compared to the disparity range. The relative size is computed over a scene or a whole dataset as:

$$S_{rel} = median\left(\frac{\overline{I} - \underline{I}}{d_{max} - d_{min}}\right) \tag{15}$$

This criterion is important as one could achieve 100% accuracy by simply setting every interval to $[d_{min}, d_{max}]$.

- S_{rel} is not adapted in low confidence areas as we purposely extended the intervals (see Sec. 3.4). Thus, we define an additional criterion only for low confidence areas, called *relative overestimation*:

$$O_{rel} = median\left(1 - \frac{\Delta|d - \hat{d}|}{\overline{I} - \underline{I}}\right) \tag{16}$$

where $\Delta|d-\hat{d}|$ is the maximal difference between the true disparity and the predicted disparity over the low confidence area. It is therefore the size of the optimal interval in the area. Figure 6 allows to visualize $\Delta|d - \hat{d}|$ and $\overline{I} - \underline{I}$. O_{rel} is the median of intervals overestimation over low confidence areas.

In Eq. (15) and Eq. (16), the median is used to evaluate the sizes of the intervals instead of the mean in order to gain statistical robustness. It is noteworthy that using the mean yields very similar results.

In the absence of any other method for creating disparity confidence intervals, we compare the accuracy and relative size of our method with a "naive" approach that serves as a baseline. This approach, referred to as *baseline* in Tab. 2, consists in simply normalizing every cost curve with its maximum and minimum values, and defining the disparity

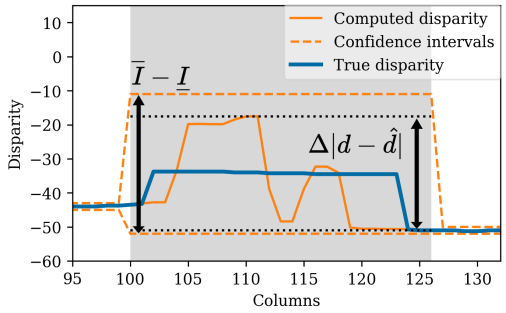


Figure 6. A low confidence area in gray, with a representation of $\Delta|d - \hat{d}|$ and $\overline{I} - \underline{I}$ from Eq. (16).

interval as the minimal and maximal disparities for which the cost is greater that 90%. We compare this to our method using possibilities without regularization and with α-levels taking values in $[50\%, 80\%, 90\%, 98\%]$. We also present our method with the regularization step and an α-level of 90%, referred to as "90 w/ reg". This is the same method used in Tab. 1. This ablation study enables us to observe the impact of different α-levels on the accuracy of the intervals, and proves the necessity of the regularization step.

Section 4.3 discusses the performance of the method with regard to the criterion from Eq. (14). We aim to validate at least 90% accuracy on the intervals over every scene. Section 4.4 evaluates criteria of Eq. (15) and Eq. (16). In high confidence areas, the relative size S_{rel} needs to be as small as possible. We consider a relative size of around 25% as a satisfying objective. In low confidence areas, a relative overestimation O_{rel} of about 30% seems a feasible objective, while still providing enough information for a later propagation into elevation intervals. The numeri-

cal objectives are given for information purposes, as users needs may vary depending on the application.

	Baseline	50	80	90	90 w/ reg	98
2003	0.503	**0.997**	0.985	0.982	0.973	0.98
2005	0.592	**0.978**	0.958	0.950	0.963	0.944
2006	0.626	0.986	0.982	0.979	**0.989**	0.976
2014	0.039	0.543	0.519	0.494	**0.957**	0.474
2021	0.03	0.549	0.478	0.442	**0.936**	0.420

Table 2. Ablation study. Evaluation of the accuracy from Eq. (14) with different methods using the CENSUS cost function. $50, 80, 90, 98$ refer to different value of α. 90 /w reg refers to an α value of 90% and a regularization step from Sec. 3.4.

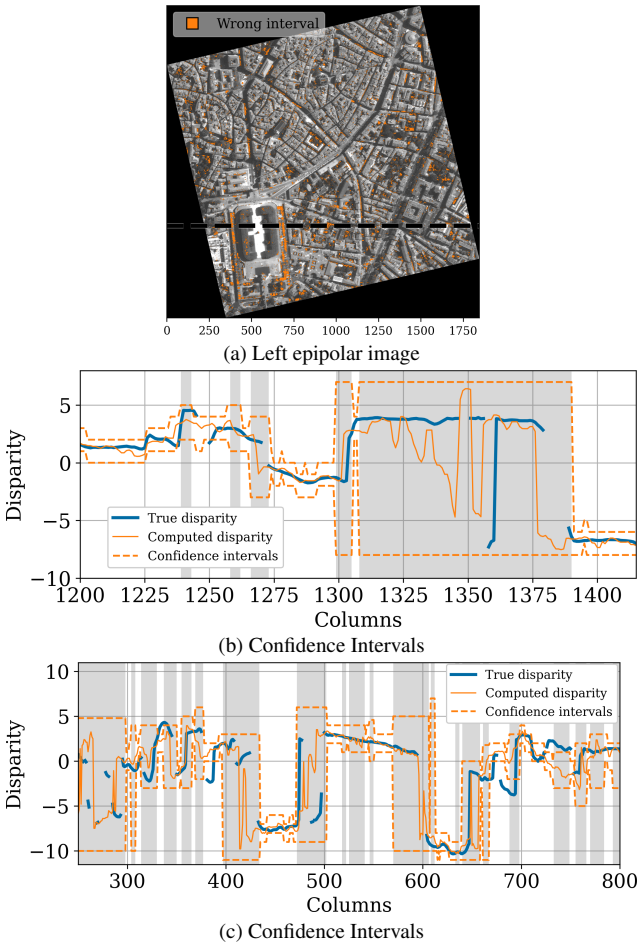
4.2. Reference Datasets

We used 83 scenes from Middlebury $2003, 2005, 2006, 2014$ and 2021 datasets for evaluation [28–31]. We use quarter-size and third-size versions of the data for $2003, 2005$ and 2006 datasets and full resolution for 2014 and 2021 datasets. We use the disparity range indicated in the calibration files. Each year contains respectively 2, 6, 21, 23 and 24 pairs of images with different shapes, and the size of the disparity intervals ranges between 60 and 1110. We also use 120 1845×1845 pairs of epipolar images generated using [5] from satellite images of the region of Montpellier, France, with a resolution of 50cm/pixel. The disparity range for those images is between 20 and 50 pixels, depending on the scene. During the evaluation, the dataset is split into two categories: *urban*, for images containing mostly buildings, and *rural*, for image mostly composed of forests and fields. The ground truth disparity was retrieved using LiDAR data.

4.3. Accuracy Results

First, intervals accuracy can be analyzed on the specific example of Fig. 7a. Inaccurate intervals are colored in the left image. Figures 7b and 7c present confidence interval values along a row, as well as the disparity estimation and the true disparity. Low confidence areas are indicated by the gray sections. Figs. 7b and 7c contain both high confidence areas with small intervals, and low confidence areas with important disparity variations. We observe a low confidence area between columns 1300 and 1390 where the computed disparity is far from the true disparity, but the confidence intervals remain correct.

Evaluating the accuracy statistics on each dataset yields strong results. Scores per year for intervals computed using the CENSUS and MC-CNN cost functions are presented in Tab. 1. CENSUS-based intervals have an accuracy always superior to 90% over each dataset. They verify the 90% accuracy objective on 80 of the 83 scenes from Middlebury, and on all 120 satellite images. MC-CNN-based intervals have an accuracy superior to 95% on every datasets. They validate the 90% accuracy objective on all 83 scenes from Middlebury and 120 satellite images. Those strong performances come nonetheless with large interval size, as detailed in section Sec. 4.4.



(a) Left epipolar image

(b) Confidence Intervals

(c) Confidence Intervals

Figure 7. Fig. 7a is the left image of the city of Montpellier where colored pixels indicate wrong interval location. Figs. 7b and 7c present detailed confidence intervals, computed disparity, and true disparity along the black dashed line from figure Fig. 7a. Areas with low confidence are indicated in gray.

An ablation study is carried out in Tab. 2, highlighting the importance of the regularization process. Without the regularization, the accuracy drops for datasets with many low confidence areas. The value of α has a small impact when compared to the regularization step. We also observe that the naive approach of the baseline produces very inaccurate intervals in comparison to our method.

4.4. Discussions on the Size of the Intervals

Although the intervals are very accurate, confidence intervals with unnecessarily large sizes need to be avoided. In Figs. 7b and 7c, confidence intervals have a small relative size in high confidence areas, and a larger size in low confidence areas. In those areas, the intervals are properly adjusted to contain the predicted disparity and the true disparity without overestimating the error. Similar observations can be made in Fig. 8b.

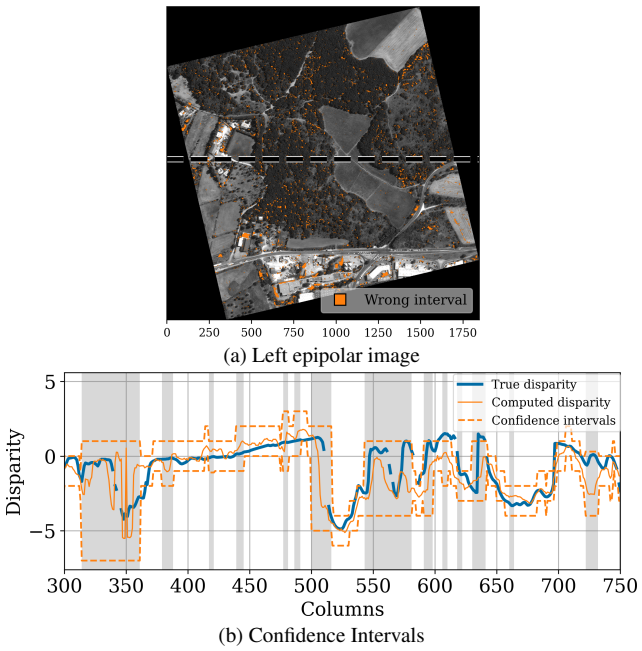
Detailed statistics of intervals relative size and relative overestimation are presented in Tab. 1 alongside accuracy results. For MC-CNN-based intervals, the 5% relative size criterion is validated for each dataset in high confidence areas, with a relative size on all datasets of only 1.5%. The global relative size is below 4% for years $2001, 2003, 2005, 2006$ and 2014. S_{rel} is maximal for the 2021 dataset due to the high proportion of low confidence intervals on this dataset. Intervals in low confidence areas are only overestimated by around 30%, meaning that large intervals are unavoidable on this dataset. This can also be explained by the complexity and high resolution of 2021 (and 2014) scenes, which leads to larger confidence intervals in general. It results in poorer performances of the disparity prediction as a majority of pixels have low confidence. Intervals computed on satellite images have a relative size around 25%, which is relatively low as most scenes have a disparity range of around 20 pixels. They however tend to overestimate the intervals in low confidence areas: around 30% when using the CENSUS cost function, and 70% when using the MC-CNN cost function.

Intervals computed using the CENSUS cost function validate the 25% relative size objective in high confidence areas. In low confidence areas, their relative overestimation is around 13%, meaning that they are close to the ideal intervals. They outperform MC-CNN based intervals on all datasets when comparing their relative size in high confidence, and on all datasets regarding the relative overestimation criterion. This is probably due to the SGM regularization, which uses different weight for the CENSUS and MC-CNN cost functions. CENSUS based weights seem to produce curves with a more pronounced/narrow peak near the minima, resulting in smaller intervals, and thus better relative sizes and over-estimation metrics.

5. Conclusion and Perspectives

To the best of our knowledge, we present the first method for creating confidence intervals on the disparity in stereo matching problems. Our method is designed to work with any stereo algorithm computing a 3D cost volume. Matching cost functions are transformed into possibility distributions and then interpreted as an expert's opinion. We rely on the advanced theoretical background of possibility distributions to compute robust uncertainty estimations. Confidence intervals are deduced from the α-cuts of possibility distributions, and regularized in low confidence areas. Post-processing steps handling is also taken into account to maintain consistency between the predicted disparity map and the confidence intervals. As we have not found existing accurate methods for comparison, we assess the intervals based on accuracy, relative size, and overestimation. Criteria are evaluated on the Middlebury datasets. 90% accuracy objective is achieved while maintaining a small relative size in high confidence area and without overestimating the intervals size in low confidence areas. The accuracy of the confidence intervals does not depend on the performance of the disparity estimation. All of our contributions are available on our GitHub repository. This work aims to motivate further research in detecting, locating and quantifying the magnitude of the error in disparity maps. We demonstrate in this paper that possibility distributions can model and process epistemic uncertainty in a understandable and explainable way.

Future work will include propagating the disparity confidence intervals into elevation confidence intervals for 3D reconstruction. Those intervals can then be provided as a complementary product alongside digital surface models, often used in many Earth Observation applications.



(a) Left epipolar image

(b) Confidence Intervals

Figure 8. Fig. 8a is the left image of a rural area near Montpellier. Colored pixels indicate wrong interval locations. Figure 8b details confidence intervals, computed disparity, and true disparity along the black dashed line from figure Fig. 8a. Areas with low confidence are indicated in gray.

References

[1] Cédric Baudrit, Dominique Guyonnet, and Didier Dubois. Joint propagation of variability and imprecision in assessing the risk of groundwater contamination. *Journal of Contaminant Hydrology*, 93(1-4):72–84, 2007. 3

[2] András Bárdossy, Axel Bronsterts, and Bruno Merz. l-, 2- and 3-dimensional modeling of water movement in the unsaturated soil matrix using a fuzzy approach. *Advances in Water Resources*, 1995. 3

[3] Mohamed Ali Chebbi, Ewelina Rupnik, Marc Pierrot-Deseilligny, and Paul Lopes. DeepSim-Nets: Deep Similarity Networks for Stereo Image Matching. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2097–2105, Vancouver, BC, Canada, 2023. IEEE. 2

[4] Liyan Chen, Weihan Wang, and Philippos Mordohai. Learning the Distribution of Errors in Stereo Matching for Joint Disparity and Uncertainty Estimation, 2023. arXiv:2304.00152 [cs]. 3

[5] Myriam Cournet, Emmanuelle Sarrazin, Loïc Dumas, Julien Michel, Jonathan Guinet, David Youssefi, Véronique Defonte, and Quentin Fardet. Ground Truth Generation and Disparity Estimation for Optical Satellite Imagery. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2020:127–134, 2020. 7

[6] Didier Dubois and Henri Prade. Random sets and fuzzy interval analysis. *Fuzzy Sets and Systems*, 42(1):87–101, 1991. 3

[7] Didier Dubois and Henri Prade. When upper probabilities are possibility measures. *Fuzzy Sets and Systems*, 49(1):65–74, 1992. 3

[8] Loïc Dumas, Véronique Defonte, Yoann Steux, and Emmanuelle Sarrazin. Improving Pairwise DSM With 3SGM: a Semantic Segmentation for SGM Using an Automatically Refined Neural Network. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-2-2022:167–175, 2022. 3

[9] Gabriele Facciolo, Carlo de Franchis, and Enric Meinhardt. MGM: A Significantly More Global Matching for Stereovision. In *Procedings of the British Machine Vision Conference 2015*, pages 90.1–90.12, Swansea, 2015. British Machine Vision Association. 2, 4

[10] Carlo de Franchis, Enric Meinhardt, Julien Michel, Jean-Michel Morel, and Gabriele Facciolo. An automatic and modular stereo pipeline for pushbroom images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3:49–56, 2014. 1, 2

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, Providence, RI, 2012. IEEE. 1

[12] Rafael Gouveia, Aristotle Spyropoulos, and Philippos Mordohai. Confidence Estimation for Superpixel-Based Stereo Matching. In *2015 International Conference on 3D Vision*, pages 180–188, Lyon, France, 2015. IEEE. 3

[13] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-Wise Correlation Stereo Network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3268–3277, Long Beach, CA, USA, 2019. IEEE. 3

[14] Istvan Haller, Cosmin D. Pantilie, Florin Oniga, and Sergiu Nedevschi. Real-time semi-global dense stereo solution with improved sub-pixel accuracy. In *2010 IEEE Intelligent Vehicles Symposium*, pages 369–376, La Jolla, CA, USA, 2010. IEEE. 4

[15] Heiko Hirschmüller. Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 807–814, San Diego, CA, USA, 2005. IEEE. 2

[16] Xiaoyan Hu and Philippos Mordohai. A Quantitative Evaluation of Confidence Measures for Stereo Vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34 (11):2121–2133, 2012. 2, 3

[17] Mark Höllmann, Max Mehltretter, and Christian Heipke. Geometry-Based Regularization For Dense Image Matching Via Uncertainty-Driven Depth Propagation. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-2-2020:151–159, 2020. 2, 3

[18] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A Survey on Deep Learning Techniques for Stereo-Based Depth Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4): 1738–1764, 2022. 1, 2

[19] Renaud Marti, Simon Gascoin, Etienne Berthier, M. de Pinel, Thomas Houet, and Dominique Laffly. Mapping snow depth in open alpine terrain from stereo satellite imagery. *The Cryosphere*, 10(4):1361–1380, 2016. 1

[20] Roger Marí, Thibaud Ehret, and Gabriele Facciolo. Disparity Estimation Networks for Aerial and High-Resolution Satellite Images: A Review. *Image Processing On Line*, 12:501–526, 2022. 2

[21] Max Mehltretter and Christian Heipke. CNN-Based Cost Volume Analysis as Confidence Measure for Dense Matching. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2070–2079, Seoul, Korea (South), 2019. IEEE. 3

[22] Olivier Melet, David Youssefi, Céline L'Helguen, Julien Michel, Emmanuelle Sarrazin, Florie Languille, and Laurent Lebègue. CO3D Mission Digital Surface Model Production Pipeline. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2020, 2020. 1

[23] Mourad Oussalah. On the normalization of subnormal possibility distributions: New investigations. *International Journal of General Systems*, 31(3):277–301, 2002. 4

[24] Matteo Poggi and Stefano Mattoccia. Learning a General-Purpose Confidence Measure Based on O(1) Features and a Smarter Aggregation Strategy for Semi Global Matching. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 509–518, Stanford, CA, USA, 2016. IEEE. 2, 3

[25] Matteo Poggi, Seungryong Kim, Fabio Tosi, Sunok Kim, Filippo Aleotti, Dongbo Min, Kwanghoon Sohn, and Stefano

Mattoccia. On the confidence of stereo matching in a deep-learning era: a quantitative evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 2, 3

[26] Ewelina Rupnik, Mehdi Daakir, and Marc Pierrot Deseilligny. MicMac – a free, open-source solution for photogrammetry. *Open Geospatial Data, Software and Standards*, 2(1): 14, 2017. 1, 2

[27] Emmanuelle Sarrazin, Myriam Cournet, Loïc Dumas, Véronique Defonte, Quentin Fardet, Y. Steux, N. Jimenez Diaz, E. Dubois, D. Youssefi, and F. Buffe. Ambiguity Concept In Stereo Matching Pipeline. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2021: 383–390, 2021. 2, 3, 5

[28] Daniel Scharstein and Chris Pal. Learning Conditional Random Fields for Stereo. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, MN, USA, 2007. IEEE. 7

[29] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, pages I–195–I–202, Madison, WI, USA, 2003. IEEE Comput. Soc.

[30] Daniel Scharstein, Richard Szeliski, and Ramin Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, pages 131–140, Kauai, HI, USA, 2001. IEEE Comput. Soc. 1, 2, 4

[31] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. In *Pattern Recognition*, pages 31–42. Springer International Publishing, Cham, 2014. Series Title: Lecture Notes in Computer Science. 7

[32] David E. Shean, Oleg Alexandrov, Zachary M. Moratto, Benjamin E. Smith, Ian R. Joughin, Claire Porter, and Paul Morin. An automated, open-source pipeline for mass production of digital elevation models (DEMs) from very-high-resolution commercial stereo satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 116:101–117, 2016. 1, 2

[33] Aristotle Spyropoulos, Nikos Komodakis, and Philippos Mordohai. Learning to Detect Ground Control Points for Improving the Accuracy of Stereo Matching. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1621–1628, Columbus, OH, USA, 2014. IEEE. 2

[34] Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Leveraging Confident Points for Accurate Depth Refinement on Embedded Systems. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 158–167, Long Beach, CA, USA, 2019. IEEE. 3

[35] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Springer US, Boston, MA, 1991. 3

[36] David Youssefi, Julien Michel, Emmanuelle Sarrazin, Fabrice Buffe, Myriam Cournet, Jean-Marc Delvit, Celine L'Helguen, Olivier Melet, Aurelie Emilien, and Julien Bosman. CARS: A Photogrammetry Pipeline Using Dask Graphs to Construct A Global 3D Model. In *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 453–456, Waikoloa, HI, USA, 2020. IEEE. 1, 2

[37] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *Computer Vision — ECCV '94*, pages 151–158. Springer Berlin Heidelberg, Berlin, Heidelberg, 1994. Series Title: Lecture Notes in Computer Science. 2

[38] Jure Žbontar and Yann LeCun. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *Journal of Machine Learning Research 17*, 2016. arXiv: 1510.05970. 2