HOW ACCURATE CAN BLOCK MATCHES BE IN STEREO VISION?

N. SABATER , J.-M. MOREL , AND A. ALMANSA

Abstract. This article explores the sub-pixel accuracy attainable for the disparity computed from a rectified stereo pair of images with small baseline. In this framework we consider translations as the local deformation model between patches in the images. A mathematical study shows first how discrete block-matching can be performed with arbitrary precision under Shannon-Whittaker conditions. This study leads to the specification of a block-matching algorithm which is able to refine disparities with sub-pixel accuracy. Moreover, a formula for the variance of the disparity error caused by the noise is introduced and proved. Several simulated and real experiments show a decent agreement between this theoretical error variance and the observed RMSE in stereo pairs with good SNR and low baseline. A practical consequence is that under realistic sampling and noise conditions in optical imaging, the disparity map in stereo-rectified images can be computed for the majority of pixels (but only for those pixels with meaningful matches) with a 1/20 pixel precision.

Key words. Block-matching, sub-pixel accuracy, noise error estimate.

1. Introduction. Stereo algorithms aim at reconstructing a 3D model from two or more images of the same scene acquired from different angles. Assuming for a sake of simplicity that the cameras are calibrated, and that the image pair has been stereo-rectified, our work will focus on the matching process. The matching of stereo images has been studied in depth for decades. We refer to [43] and [6] for a complete comparison of different methods.

Generally stereo matching methods are divided in two classes, the local algorithms and the global ones. Given two images of the same scene, the local methods compare a small block of pixels surrounding each pixel in the first image to the candidate blocks on the epipolar line in the second image. The blocks are usually compared by the normalized cross correlation (NCC) or the sum of squared differences (SSD). Having a minimum of the SSD does not guarantee at all that the match is correct. In general, only a significant proportion of the image can be reliably matched (about 40 to 80%). Block-matching methods can indeed produce wrong disparities near the intensity discontinuities in the images. This "fattening effect" is a classic problem in block-matching methods. It occurs when a salient image feature (typically an edge) lies *within* the comparison window φ but *away* from its center. This can produce a large error near points at which the disparity ε has a jump. Several papers have attempted with some success to alleviate this problem by using adaptive shape windows [21, 48, 22, 18], adaptive support-weight windows [53], a barycentric correction [12], or by feature matching methods [44]. If the images of the stereo pair u_1 and u_2 have little aliasing, [12] showed that the recovered disparity map obtained by minimizing a continuous quadratic distance between u_1 and u_2 has therefore two main error terms: the fattening error, and the error due to noise. There are other causes for mismatches. Some image parts are simply occluded in one of the images and cannot be observed in the other one. Often, image regions are too flat to reliably minimize the block distance, particularly the dark regions where noise dominates, and the saturated regions. The fractal structure of vegetation and the glossy surfaces are other causes of error, their aspect changing drastically even with a small change of viewpoint. We shall call such errors gross errors. Luckily, there are several techniques to avoid the gross errors: for example the coarse-to-fine scale refinement [17], the SIFT threshold [28] and the *a contrario* methods [31]. In this paper an *a contrario* rejection algorithm [41, 42] is used to detect and eliminate *a priori* all unreliable pixels. In geographic information systems, where high accuracy is particularly relevant, the reliable points

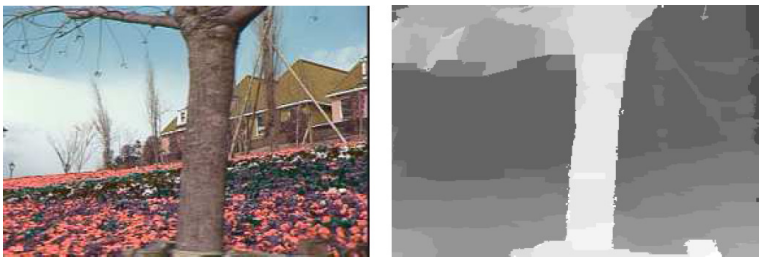

Fig. 1.1: Optimal disparity obtained with a global stereo functional and an efficient graph cuts algorithm (experiment taken from [5]). Such stereo methods depend on (at least) two parameters which it is difficult to estimate. Notice the obvious errors: the house and the sky have the same disparities. On the top-left the disparities of the tree branches and of the sky are mixed up. Global methods are not well adapted to a strong sub-pixel accuracy because they quantize the disparity to avoid a complexity blow up. In this experiment, there are only eight disparity levels.

correspond in general to the interior points of textured regions (roofs, lawn, terrains). The unreliable pixels, which include the pixels risking fattening, usually cover less than half the image.

Here lies, however, the main objection to a precision analysis in stereo vision: what is the use of giving sub-pixel error estimates, if a significant part of the matched points is simply mismatched? Block matching methods are nevertheless prone to an error analysis, precisely because there are techniques to rule out the risky pixels.

It would be nonetheless tempting to extend the error analysis to global methods. Global variational methods providing a dense disparity map are in significant progress, thanks to efficient new optimization techniques such as graph cuts [25], and clever uses of dynamic programming [33, 15]. Unfortunately, because of the global nature of the optimization process, these methods are not prone to a local precision analysis. The error can only be measured globally, and no formula can be derived for their local precision, because of the global nature of the optimization process. The dense results given by global methods can include many errors as illustrated in Fig. 1.1, taken from [5]. The left image is one of the matched images and the right image is the obtained disparity. Even though no ground truth is available for this classic example, the errors are obvious: parts of the sky seen through the tree branches inherit the disparity of the branches, and a wide stripe in the middle of the image mixes up house, sky and trees which get all the same disparity. Variational methods for stereovision depend necessarily on at least two parameters: Indeed, they all have at least three terms depending on the disparity: one is the cost for not attributing a disparity, namely the cost of the occlusions. The second term is the disparity regularity term penalizing oscillation and big jumps of the disparity. The third term is the fidelity term measuring the matching performance for a given disparity. The relative weighting of these three terms leaves two parameters to the user [25]. The optimal solution can change with a change in the parameter choice. Yet, the state of the theory does not seem to guarantee a reliable theory for the (crucial) estimation of these parameters, which can vary from an image pair to another. Another Achilles' heel of recent optimization techniques for stereo is their complexity, which blows up when the number of disparity

quanta increases.[1] This fact makes them poorly adapted to strong sub-pixel accuracy requirements.

These considerations explain why this paper will only analyze the stereo matching precision for a sub-pixel block-matching method.

1.1. Small Baseline and Sub-pixel Accuracy. Another aspect of the discussion is the geometric setting of the image pair. As we mentioned above, a large baseline increases considerably the risks of gross errors. Stereovision has nonetheless tended to consider pairs of images with a large baseline. The reason is to have a large disparity (measured in pixels) and therefore to get by triangulation a more accurate 3D reconstruction. But, as was pointed out in the seminal patent [17] and in [12], this argument against a small baseline becomes invalid if a reliable sub-pixel matching is possible. Particularly in satellite or aerial imaging, where the images are taken by two sweeps or snapshots from the same overflying object, the shadows can have moved, unless the snapshots are taken just after one another. The baseline proposed in [12], [17] is around 0.1 and even 0.05 instead of the more conventional ratios of the order of 1. It therefore requires a ten or twenty times sharper accuracy. With such small baselines the occlusion regions are very thin and the image aspect changes very little. Small baselines in conjunction with larger ones were considered in [34], a pragmatic study where different baselines were used to eliminate errors. However, its final sub-pixel results were computed with the large baseline samples.

To summarize, a low baseline is very desirable but requires a sharp sub-pixel accuracy. Yet, this question has seldom been tackled in the vast stereo literature. Birchfield and Tomasi [3] presented a dissimilarity measure which is insensitive to sampling and an algorithm which computes disparities at a pixel resolution. In order to get a sub-pixel disparity resolution they suggested to zoom in the image by a factor of about 10 to 15 before using their algorithm. Such an approach is sound but computationally very expensive. But there are clever solutions to refine the sub-pixel accuracy of a disparity map while adding little computation. For example, [47] has compared several sub-pixel registration algorithms where the surface maximum is computed in the sampling grid. These algorithms can be used for the sub-pixel refinement using an iterative gradient descent technique [29]. Similarly, the refinement can be computed by fitting a quadratic curve to the correlation at the discrete disparity [21]. The sub-pixel disparity is then reached with little additional computation. However, a systematic bias called *pixel-locking* has been observed when the disparities are computed at discrete values [45]. To avoid this effect [35] performs a linear interpolation and maximization of an enhanced NCC similarity measure. A more complex solution to avoid pixel-locking is the symmetric refinement strategy considered by [32]. Indeed, the images are treated symmetrically and a 2D surface fitting of the function cost is required to get sub-pixel coordinates for both matched points.

The sub-pixel accuracy is even more rare in global methods because of the complexity burden. [16] suggested a semi-global matching approach at a quarter pixel level which is computationally efficient. However the accuracy is hidden by the gross errors committed in the low-textured regions. Sub-pixel disparities in global methods can

[1] This shortcoming has been circumvented for instance in the case of graph-cuts algorithms for TV denoising in [8, 11, 20] where the optimization is solved for arbitrarily accurate gray-levels, by dyadic subdivision of the original problem into a sequence of binary ones. Such an approach has nevertheless not yet been extended to more complex and non-convex optimization problem such as those found in stereo.

be obtained in special cases by plane fitting [24]. (On a number of published results in the Middlebury evaluation benchmark it is assumed that the scenes are piecewise planar.) Some authors like Paparoditis and Dissard [23] even assume that the average matching precision is no more than a quarter of pixel. A sub-pixel refinement of the disparity can also be obtained by a direct anisotropic diffusion [2, 14].

The possibility of rigorous block-matching with sub-pixel accuracy by using a zero-padding factor 2 oversampling was first noticed in [46]. Its mathematical justification will be developed here. The first theoretical arguments towards high accuracy in stereovision were given in [12], who claimed that high precision matches can be obtained from a small baseline stereo pair of images if the Shannon-Whittaker conditions are met. However, this paper neither gave an accurate formula for the attainable precision (the given upper bound of the error due to noise is too pessimistic) nor demonstrated its practical feasibility. From an algorithmic point of view, there also differences with the work presented here. Their multiresolution algorithm (MARC) computes a disparity at each point and at each scale in order to use it as a initialization map in the next scale. Doing so, gross errors are propagated from the coarser scales to the finest one. A comparison of our algorithm with MARC has been done in [40].

Several contributions on the performance bounds on motion estimation have been published recently. Even if these contributions try to solve the registration problem instead of the stereo problem, the presented theoretical results are relevant. There are actually much more relevant to stereo than for global image registration, because these studies mainly focus on translations. A real image registration implies at least the compensation of a homography. The major contribution on this theme is due to Robinson and Milanfar [37] who studied the performance limits of image registration techniques and proposed a Cramer-Rao upper performance bound. Furthermore, they studied the bias inherent to the problem of image registration and in particular the bias of the gradient-based estimators. In order to minimize this bias, [38] proposed a methodology for designing filters based on image content which improves the estimator performance.

The Robinson and Milanfar approach is also followed and further refined in subsequent papers [52, 50, 26]. Kybic [26] addresses the problem of estimating uncertainty of image registration algorithms using the bootstrap resampling. The same paper also studies a fast registration accuracy estimation (FRAE) method whose performance is inferior to a bootstrap, but has a negligible computation overhead. Yetik and Nehorai [52] studied the Cramer-Rao bounds on a wide variety of geometric deformations including translation, rotation, shearing, rigid, more general affine and nonlinear transformations. But, as Xu et al. [50] pointed out, their Cramer-Rao lower bounds are functions of the unknown rotation angle and may be difficult to obtain in practice. For this reason Xu et al. assume that the prior information regarding the uncertainty model is known. In particular they address the rigid transformation case and study the Ziv-Zakai Bayesian bound.

The Fourier interpolation model which actually eliminates bias was ruled out in the previous registration bibliography since [37], who argue that it is not a good model for natural images. This is precisely the model we seek to numerically approximate here. Furthermore, in contrast to image registration, we use here much smaller windows sizes, which are better adapted to stereo problems. When using such small window sizes, and by being careful when dealing with discretization and the numerics of the Fourier interpolation model, we found that the bias in stereo block-matching is

negligible with respect to the variance due to noise in realistic noise conditions. For this fact we provide here theoretical and experimental evidence. This gives a new perspective to the problem of performance estimation, with practical results different from those obtained in the image registration context. From a theoretical point of view our work is still consistent with previous results. Our performance estimator is dominated by the noise error and lies between the upper bound given by Delon [12] and the Cramer-Rao lower bound. In fact, depending on image contents, the Cramer-Rao lower bound is off with respect to our estimate by a factor that varies between 1 and 2.

The subpixel accuracy in image registration has also been studied and related to the underlying image interpolation models. [36, 9, 1] addressed the problem by phase correlation. [36] proposed a method based on linear weighting the main phase correlation peak and the difference between its two neighboring sidepeaks. [9] proposed high order statistics in order to attain subpixel accuracy when registering in noisy images with a low SNR. [1] showed that the continuous and discrete phase difference is a 2D sawtooth signal. They count the number of cycles of the phase difference and give the noninteger fraction of the last cycle as the subpixel shift. Meanwhile [39] studied the oscillation artifacts caused by interpolation of the correlation and proposed high degree B-spline interpolation. This interpolation is close to the interpolation (sinc) used in this paper.

Contributions of the paper. This paper analyzes the attainable precision in stereo with a block-matching method and proposes a theoretical sub-pixel accuracy estimate whose principal term is caused by the noise. In particular, a formula for the variance of the disparity error at each image point $\mathbf{x}$ is proved. This formula depends on the noise variance in both images and on the image gray levels within the block around $\mathbf{x}$.

In our theoretical analysis we also study i) the conditions under which the discrete block-matching distance is equal to the continuous one and ii) how the block-matching distance should be sampled in order to satisfy the Shannon-Whittaker conditions. Under ideal Shannon conditions, this permits to avoid the block matching errors due to wrong interpolation and wrong sampling.

The theoretical error formula due to noise is matched experimentally to the real error, thanks to a method discarding the potentially wrong matches. While this *a contrario* method developed in [41, 42] is not the object of this paper, the existence of algorithms filtering out all risky pixels is crucial for this study. Otherwise the precision could never be checked. So we will assume that the reliable pixels are selected *a priori* on each image pair.

Simulated pairs and real examples including benchmark data will examine to which extent the theoretical error bounds are reached.

Furthermore, on several realistic simulations and on benchmark examples, theory and practice will confirm a 1/20 pixel accuracy. From the experiments

one observes that that the predicted theoretical error variance at reliable pixels has the same order of magnitude as the observed error, the difference between RMSE and predicted error being of about 20%. The formula for the main disparity error term due to noise given is new, exact, and actually far more accurate than the upper bound proposed in [12] for the same error, (which was more than 10 times larger).

Plan of the paper. Section 2 describes the theoretical assumptions and the accurate interpolation techniques permitting to compute sub-pixel disparities. Section

3 proves a formula for the theoretical noise error. Section 4 gives all details on the algorithms and a complexity analysis.

Section 5 shows the obtained results for several stereo pairs with simulated and real ground truths, and compares the practical error with their theoretical prediction.

2. Preliminaries on Sub-Pixel Interpolation. We denote by $\mathbf{x} = (x, y)$ an image point in the continuous image domain, and by $u_1(\mathbf{x}) = u_1(x, y)$ and $u_2(\mathbf{x})$ the images of a stereo rectified image pair. The epipolar direction will always be the x axis. The underlying depth map can be deduced from the disparity function $\varepsilon(\mathbf{x})$ giving the shift of an observed physical point $\mathbf{x}$ from the left image u_1 in the right image u_2. The physical disparity $\varepsilon(\mathbf{x})$ may be quite irregular (especially in urban environments), and therefore is not well-sampled (in the Shannon-Nyquist sense) with respect to the sampling rate of the stereo pair. Therefore, it cannot be recovered at all points, but only essentially at points $\mathbf{x}$ around which the depth map is continuous and with small derivative. Around such points, a deformation model holds:

$$\begin{aligned} u_1(\mathbf{x}) &= u(x + \varepsilon(\mathbf{x}), y) + n_1(\mathbf{x}) \\ u_2(\mathbf{x}) &= u(\mathbf{x}) + n_2(\mathbf{x}), \end{aligned} \tag{2.1}$$

where n_i are Gaussian noises and $u(\mathbf{x})$ is the ideal bandlimited image that would be observed instead of $u_2(\mathbf{x})$ if there was no noise. The deformation model (2.1) is *a priori* valid when the angle of the 3D surface at $\mathbf{x}$ with respect to the camera changes moderately, which is systematically true for small (0.02 to 0.15) baseline stereo systems. The restriction brought by (2.1) is moderate. Indeed, the trend in stereo vision is to have multiple views of the 3D object to be reconstructed and therefore many pairs with small baseline.

Consider the two images $u_1(\mathbf{x})$ and $u_2(\mathbf{x})$ defined on a square $[0, a]^2$ and a window function $\varphi(\mathbf{x})$. Block-matching is the search for a value of μ minimizing the continuous quadratic distance

$$e_{\mathbf{x}_0}(\mu) := \int_{[0,a]^2} \varphi(\mathbf{x} - \mathbf{x}_0)\big(u_1(\mathbf{x}) - u_2(\mathbf{x} + (\mu, 0))\big)^2 d\mathbf{x}. \tag{2.2}$$

This section will prove a discrete formula for the quadratic distance which is faithful to the continuous image interpolates. Thanks to it, an accurate sub-pixel matching becomes possible. Without loss of generality, all considered images u, u_1, etc. are defined on the $[0, a]^2$ and are supposed to be square integrable. Thus, the Fourier series decomposition applies

$$u(x, y) = \sum_{k,l \in \mathbb{Z}} \tilde{u}_{k,l} e^{\frac{2i\pi(kx+ly)}{a}}, \tag{2.3}$$

where the $\tilde{u}_{k,l}$ are the Fourier series coefficients (or shortly the Fourier coefficients) of u. By the classic Fourier series isometry, for any two square integrable functions $u(\mathbf{x})$ and $v(\mathbf{x})$ on $[0, a]^2$,

$$\int_{[0,a]^2} u(\mathbf{x})\overline{v}(\mathbf{x}) d\mathbf{x} = a^2 \sum_{k,l \in \mathbb{Z}} \tilde{u}_{k,l} \overline{\tilde{v}_{k,l}}. \tag{2.4}$$

The digital images are usually given by their N^2 samples $u(\mathbf{m})$ for $\mathbf{m}$ in the grid

$$\mathbb{Z}^1_a = [0, a]^2 \cap \Big((\frac{a}{2N}, \frac{a}{2N}) + \frac{a}{N}\mathbb{Z}^2 \Big).$$

Similarly, the over-sampling grid with four times more samples is denoted by

$$\mathbb{Z}_a^{1/2} = [0,a]^2 \cap \left((\frac{a}{4N}, \frac{a}{4N}) + \frac{a}{2N}\mathbb{Z}^2 \right). \tag{2.5}$$

N is always an even integer. In all that follows we shall assume that the images obtained by a stereo vision system are band-limited. This assumption is classical and realistic, the aliasing in good quality CCD cameras being moderate. Another classical assumption in image processing which we adopt is the (forced) a-periodicity assumption. Under this assumption a band-limited image becomes a trigonometric polynomial. This periodicity assumption is not natural, but it only entails a minor distortion near the boundary of the image domain $[0,a]^2$. The payoff for the *band-limited + periodic assumption* is that the image can be interpolated, and its Fourier coefficients computed from discrete samples. Indeed, given N^2 samples $u_{\mathbf{m}}$ for $\mathbf{m}$ in $\mathbb{Z}_a^1$, there is a unique trigonometric polynomial in the form

$$u(x,y) = \sum_{k,l=-N/2}^{N/2-1} \tilde{u}_{k,l} e^{\frac{2i\pi(kx+ly)}{a}} \tag{2.6}$$

such that $u(\mathbf{m}) = u_{\mathbf{m}}$. We shall call such polynomials *N-degree trigonometric polynomials*. The coefficients $\tilde{u}_{k,l}$ are again the *Fourier coefficients* of u in the Fourier basis $e^{\frac{2i\pi(kx+ly)}{a}}$, $k,l \in \mathbb{Z}$. The map $u_{\mathbf{m}} \to u_{k,l}$ is nothing but the *2D Discrete Fourier Transform (DFT)*, and the map $(u_{\mathbf{m}}) \to N(\tilde{u}_{k,l})$ is an isometry from $\mathbb{C}^{N^2}$ to itself. The function $u(x,y)$ is therefore usually called the *DFT interpolate of the samples* $u_{\mathbf{m}}$. In consequence, there is an isometry between the set of N-degree trigonometric polynomials endowed with the $L^2([0,a]^2)$ norm, and $\mathbb{C}^{N^2}$ endowed with the usual Euclidean norm:

$$\int_{[0,a]^2} |u(x,y)|^2 d\mathbf{x} = a^2 \sum_{k,l=-N/2}^{N/2-1} |\tilde{u}_{k,l}|^2 = \frac{a^2}{N^2} \sum_{\mathbf{m}\in\mathbb{Z}_a^1} |u(\mathbf{y}+\mathbf{m})|^2, \tag{2.7}$$

where the relation is valid with a grid with arbitrary origin $\mathbf{y}$. If $u(\mathbf{x})$ and $v(\mathbf{x})$ are two N-degree trigonometric polynomials, we therefore also have

$$\int_{[0,a]^2} u(\mathbf{x})\overline{v}(\mathbf{x}) d\mathbf{x} = a^2 \sum_{k,l=-N/2}^{N/2-1} \tilde{u}_{k,l}\overline{\tilde{v}_{k,l}} = \frac{a^2}{N^2} \sum_{\mathbf{m}\in\mathbb{Z}_a^1} u(\mathbf{y}+\mathbf{m})\overline{v(\mathbf{y}+\mathbf{m})}, \tag{2.8}$$

where $\overline{v}$ is the complex conjugate of v. Taking four times more samples, it follows from (2.8) that

$$\int_{[0,a]^2} u(\mathbf{x})\overline{v}(\mathbf{x}) d\mathbf{x} = a^2 \sum_{k,l=-N}^{N-1} \tilde{u}_{k,l}\overline{\tilde{v}_{k,l}} = \frac{a^2}{4N^2} \sum_{\mathbf{m}\in\mathbb{Z}_a^{1/2}} u(\mathbf{m})\overline{v(\mathbf{m})}. \tag{2.9}$$

which is also valid if $u(\mathbf{x})$ and $v(\mathbf{x})$ are $2N$-degree trigonometric polynomials in $\mathbf{x}$.

This last fact has a first important consequence in block-matching leading to the next

PROPOSITION 2.1 (**Equality of the discrete and the continuous quadratic distance**). *Let $u_1(\mathbf{x})$ and $u_2(\mathbf{x})$ be two N-degree trigonometric polynomials on $[0,a]^2$ and let $\varphi(\mathbf{x})$ be a window function which we assume to be a $2N$-degree trigonometric polynomial. Then*

$$e_{\mathbf{x}_0}(\mu) = e^d_{\mathbf{x}_0}(\mu), \quad \textit{where} \tag{2.10}$$

$$e^d_{\mathbf{x}_0}(\mu) := \frac{a^2}{4N^2} \sum_{\mathbf{m}\in\mathbb{Z}_a^{1/2}} \varphi(\mathbf{m}-\mathbf{x}_0)\big(u_1(\mathbf{m}) - u_2(\mathbf{m}+(\mu,0))\big)^2. \tag{2.11}$$

Proof. Since $\big(u_1(\mathbf{x}) - u_2(\mathbf{x}+(\mu,0))\big)^2$ and $\varphi(\mathbf{x}-\mathbf{x}_0)$ are $2N$-degree trigonometric polynomials in $\mathbf{x}$, by (2.9) the discrete scalar product defining $e^d_{\mathbf{x}_0}(\mu)$ equals the continuous scalar product defining $e_{\mathbf{x}_0}(\mu)$.

Thus *the continuous block distance is a finite sum of discrete samples.* □

The block distance function $\mu \to e_{\mathbf{x}_0}(\mu)$, whose minimization is our main objective, can itself easily be sampled. By (2.11) it is a $2N$-degree trigonometric polynomial with respect to μ. This proves:

PROPOSITION 2.2 (**Sub-pixel correlation requires $\times$2 zoom**). *Let $u_1(\mathbf{x})$ and $u_2(\mathbf{x})$ be two N-degree trigonometric polynomials. Then the quadratic distance $e^d_{\mathbf{x}_0}(\mu)$ is well-sampled provided it has at least $2N$ successive samples. Thus the computation of $e^d_{\mathbf{x}_0}(\mu)$ at half samples $\mu \in \frac{a\mathbb{Z}}{2}$ (via zero-padding of u_1 and u_2) allows the exact reconstruction of $e^d_{\mathbf{x}_0}(\mu)$ for any real μ by DFT interpolation.* (The bottom line argument here is that the block matching is a squared distance, and the bandwidth of u^2 is the double of the bandwidth of u by the Fourier convolution theorem.) Remark that the last proposition does not require any assumption on the window function $\varphi(\mathbf{x})$. Prop. 2.2, which opens the way to rigorous block-matching with sub-pixel accuracy, has been noticed in [46]. It is also used in the MARC method [17] used by the French space agency (CNES). But Prop. 2.1 and the simple proof of Prop. 2.2 are new.

Interpolating the noise too. Sub-pixel block-matching requires the interpolation of noisy images, whose noise is by nature discrete. Thus, following Shannon's classical observation, the noise itself will also be interpolated as a band-limited function. In the periodic framework it therefore becomes a trigonometric polynomial. We shall need some estimates on the interpolated noise, in particular its variance and the variance of its derivative at each point. Assume that $(n_{\mathbf{m}}), \mathbf{m} \in \mathbb{Z}_a^1$ are N^2 independent $\mathcal{N}(0,\sigma^2)$ noise samples, implying that $(n_{\mathbf{m}})$ is a Gaussian vector. Since the DFT is an isometry, the noise Fourier coefficients $N(\tilde{n}_{\mathbf{k}})$ also form a Gaussian vector with diagonal covariance matrix $\sigma^2 Id$. By (2.8), the mapping $(n_{\mathbf{m}})_{\mathbf{m}\in\mathbb{Z}_a^1} \to (n(\mathbf{x}+\mathbf{m}))_{\mathbf{m}\in\mathbb{Z}_a^1}$ is an isometry from $\mathbb{C}^{N^2}$ to itself. It follows that $n(\mathbf{x})$ is $\mathcal{N}(0,\sigma^2)$ for every $\mathbf{x}$.

An estimate of $\mathrm{Var}(n_x(\mathbf{x}))$ will also be needed, where $n_x(\mathbf{x}) = \frac{\partial n}{\partial x}(x,y)$.

$$\mathrm{Var}(n_x(\mathbf{x})) = \mathrm{Var}\left(\sum_{k,l=-N/2}^{N/2-1} \tilde{n}_{k,l} \frac{2ik\pi}{a} e^{2i\frac{k\pi x + l\pi y}{a}}\right) =$$

$$= \frac{4\pi^2\sigma^2 N}{N^2a^2} \sum_{k=-N/2}^{N/2-1} k^2 \simeq \frac{4\pi^2\sigma^2}{a^2N}\frac{N^3}{12} = \frac{\pi^2N^2}{3a^2}\sigma^2.$$

Since $n(\mathbf{x})$ is a normal law, $n(\mathbf{x})^2$ is a χ^2 law of order 1. Thus its variance is $2\sigma^4$. Finally we shall need to evaluate $\mathrm{Var}(n_1(\mathbf{x})n_2(\mathbf{x}))$, where n_i are two independent interpolated white noises of the above kind. Thus $n_1(\mathbf{x})n_2(\mathbf{x})$ is the product of two normal laws. The expectation of the product is zero and the variance is therefore $\mathrm{Var}(n_1n_2) = \mathbb{E}(n_1n_2)^2 = \mathbb{E}n_1^2\mathbb{E}n_2^2 = (\mathbb{E}n^2)^2 = \mathrm{Var}(n)^2 = \sigma^4$. In summary:

LEMMA 2.3. *Let $(n_{\mathbf{m}})_{\mathbf{m}\in\mathbb{Z}^1_a}$ be N^2 independent white Gaussian noise samples with variance σ^2. Then the DFT interpolate $n(\mathbf{x})$ on $[0,a]^2$ is $\mathcal{N}(0,\sigma^2)$ for every $\mathbf{x}$. If n_1 and n_2 are two independent noises like n, one has*

$$\mathrm{Var}(n^2(\mathbf{x}))=2\sigma^4, \tag{2.12}$$

$$\mathrm{Var}(n_x(\mathbf{x}))\simeq\frac{\pi^2N^2}{3a^2}\sigma^2, \tag{2.13}$$

$$\mathrm{Var}(n_1(\mathbf{x})n_2(\mathbf{x}))=\sigma^4. \tag{2.14}$$

LEMMA 2.4. *Take $a = N$ and let $n(\mathbf{x})$ be the DFT interpolate on $[0,N]^2$ of a white noise with variance σ^2 on $\mathbb{Z}^1_N$, as defined above. Let $\varphi(\mathbf{x})$ be a $2N$-degree trigonometric polynomial on $[0,N]^2$. Then*

$$\mathrm{Var}\left(\int_{[0,N]^2}\varphi(\mathbf{x})n(\mathbf{x})n_x(\mathbf{x})d\mathbf{x}\right)\leqslant\frac{\sigma^4}{2}\int_{[0,N]^2}\varphi_x(\mathbf{x})^2d\mathbf{x}, \tag{2.15}$$

and the expectation of this random variable is null. Let $g(\mathbf{x})$ be any square integrable function on $[0,N]^2$ and let g_N be its least square approximation by a N-degree trigonometric polynomial. Then

$$\mathrm{Var}\left(\int g(\mathbf{x})n(\mathbf{x})d\mathbf{x}\right)=\sigma^2\int_{[0,N]^2}g_N(\mathbf{x})^2d\mathbf{x}\leqslant\sigma^2\int_{[0,N]^2}g(\mathbf{x})^2d\mathbf{x}. \tag{2.16}$$

The proof of this lemma is in appendix.

3. Block-Matching Errors Due to Noise. Consider the digital images of a stereo pair and their DFT interpolates $u_1(\mathbf{x})$, $u_2(\mathbf{x})$ satisfying (2.1). Block matching amounts to look for every $\mathbf{x}_0$ for the estimated disparity at $\mathbf{x}_0$ minimizing

$$e_{\mathbf{x}_0}(\mu)=\int_{[0,N]^2}\varphi(\mathbf{x}-\mathbf{x}_0)\big(u_1(\mathbf{x})-u_2(\mathbf{x}+(\mu,0))\big)^2d\mathbf{x}. \tag{3.1}$$

where $\varphi(\mathbf{x}-\mathbf{x}_0)$ is a soft window function centered at $\mathbf{x}_0$. For a sake of compactness in notation, $\varphi_{\mathbf{x}_0}(\mathbf{x})$ stands for $\varphi(\mathbf{x}-\mathbf{x}_0)$, $\int_{\varphi_{\mathbf{x}_0}}u(\mathbf{x})$ will be an abbreviation for $\int\varphi(\mathbf{x}-\mathbf{x}_0)u(\mathbf{x})d\mathbf{x}$; we will write $u(\mathbf{x}+\mu)$ for $u(\mathbf{x}+(\mu,0))$ and ε for $\varepsilon(\mathbf{x})$. The minimization problem (3.1) rewrites

$$\min_\mu\int_{\varphi_{\mathbf{x}_0}}\big(u(\mathbf{x}+\varepsilon(\mathbf{x}))+n_1(\mathbf{x})-u(\mathbf{x}+\mu)-n_2(\mathbf{x}+\mu)\big)^2d\mathbf{x}.$$

Differentiating this energy with respect to μ implies that any local minimum $\mu=\mu(\mathbf{x}_0)$ satisfies

$$\int_{\varphi_{\mathbf{x}_0}} \Big(u(\mathbf{x}+\varepsilon(\mathbf{x}))+n_1(\mathbf{x})-u(\mathbf{x}+\mu)-n_2(\mathbf{x}+\mu)\Big)\times\Big(u_x(\mathbf{x}+\mu)+(n_2)_x(\mathbf{x}+\mu)\Big)\,d\mathbf{x}=0. \tag{3.2}$$

One has by Taylor-Lagrange formula $u_x(\mathbf{x}+\mu)=(u(\mathbf{x}+\varepsilon))_x+O_1(\mu-\varepsilon)$, with

$$O_1(\mu-\varepsilon)\le|\mu-\varepsilon|\max|u(\mathbf{x}+\varepsilon)_{xx}| \tag{3.3}$$

and $u(\mathbf{x}+\varepsilon(\mathbf{x}))-u(\mathbf{x}+\mu)=(u(\mathbf{x}+\varepsilon))_x(\varepsilon-\mu)+O_2((\varepsilon-\mu)^2)$, where

$$|O_2((\varepsilon-\mu)^2)|\le\frac{1}{2}\max|(u(\mathbf{x}+\varepsilon))_{xx}|(\varepsilon-\mu)^2\,.$$

Thus equation (3.2) yields

$$\int_{\varphi_{\mathbf{x}_0}}\Big((u(\mathbf{x}+\varepsilon))_x(\varepsilon-\mu)+O_2((\varepsilon-\mu)^2)+n_1(\mathbf{x})-n_2(\mathbf{x}+\mu)\Big)\times \Big((u(\mathbf{x}+\varepsilon))_x+O_1(\mu-\varepsilon)+(n_2)_x(\mathbf{x}+\mu)\Big)\,d\mathbf{x}=0. \tag{3.4}$$

and therefore

$$\mu\int_{\varphi_{\mathbf{x}_0}}(u(\mathbf{x}+\varepsilon))_x^2d\mathbf{x}=\int_{\varphi_{\mathbf{x}_0}}(u(\mathbf{x}+\varepsilon))_x^2\,\varepsilon(\mathbf{x})\,d\mathbf{x}+\tilde{\mathcal{A}}+\tilde{\mathcal{B}}+\mathcal{O}_1+\mathcal{O}_2, \tag{3.5}$$

where

$$\tilde{\mathcal{A}}=\int_{\varphi_{\mathbf{x}_0}}(u(\mathbf{x}+\varepsilon))_x\big(n_1(\mathbf{x})-n_2(\mathbf{x}+\mu)\big)d\mathbf{x}; \tag{3.6}$$

$$\tilde{\mathcal{B}}=\int_{\varphi_{\mathbf{x}_0}}\big(n_1(\mathbf{x})-n_2(\mathbf{x}+\mu)\big)(n_2)_x(\mathbf{x}+\mu)d\mathbf{x}; \tag{3.7}$$

$$\begin{aligned}\mathcal{O}_1=&\int_{\varphi_{\mathbf{x}_0}}(u(\mathbf{x}+\varepsilon))_x(\varepsilon-\mu)(n_2)_x(\mathbf{x}+\mu)d\mathbf{x}\\&+\int_{\varphi_{\mathbf{x}_0}}O_1(\mu-\varepsilon)\big(n_1(\mathbf{x})-n_2(\mathbf{x}+\mu)\big)d\mathbf{x};\end{aligned} \tag{3.8}$$

$$\begin{aligned}\mathcal{O}_2=&\int_{\varphi_{\mathbf{x}_0}}O_2(\varepsilon-\mu)^2(u(\mathbf{x}+\varepsilon))_xd\mathbf{x}\\&+\int_{\varphi_{\mathbf{x}_0}}O_2(\varepsilon-\mu)^2[O_1(\mu-\varepsilon)+(n_2)_x(\mathbf{x}+\mu)]d\mathbf{x}\\&+\int_{\varphi_{\mathbf{x}_0}}O_1(\mu-\varepsilon)(u(\mathbf{x}+\varepsilon))_x(\varepsilon-\mu)d\mathbf{x}.\end{aligned} \tag{3.9}$$

Denote by $\bar{\varepsilon}$ the average of ε weighted by $\varphi(\mathbf{x}-\mathbf{x}_0)$. By the Taylor-Lagrange theorem we have

$$\tilde{\mathcal{A}}=\mathcal{A}+\mathcal{O}_{\mathcal{A}}\,,$$

where

$$\mathcal{A} = \int_{\varphi_{\mathbf{x}_0}} (u(\mathbf{x}+\varepsilon))_x \big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\overline{\varepsilon})\big) d\mathbf{x} \tag{3.10}$$

and

$$\mathcal{O}_{\mathcal{A}} = (\overline{\varepsilon} - \mu) \int_{\varphi_{\mathbf{x}_0}} (u(\mathbf{x}+\varepsilon))_x (n_2)_x (\mathbf{x} + \tilde{\varepsilon}(\mathbf{x})) d\mathbf{x}, \tag{3.11}$$

where $\tilde{\varepsilon}(\mathbf{x})$ satisfies $\tilde{\varepsilon}(\mathbf{x}) \in [\min(\mu, \overline{\varepsilon}), \max(\mu, \overline{\varepsilon})]$. In the same way,

$$\tilde{\mathcal{B}} = \int_{\varphi_{\mathbf{x}_0}} \big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\mu)\big)(n_2)_x(\mathbf{x}+\mu) d\mathbf{x}\,.$$

so that $\tilde{\mathcal{B}} = \mathcal{B} + \mathcal{O}_{\mathcal{B}}$, where

$$\mathcal{B} = \int_{\varphi_{\mathbf{x}_0}} \big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\overline{\varepsilon})\big)(n_2)_x(\mathbf{x}+\overline{\varepsilon}) d\mathbf{x} \tag{3.12}$$

and

$$\mathcal{O}_{\mathcal{B}} = (\mu - \overline{\varepsilon}) \int_{\varphi_{\mathbf{x}_0}} n_1(\mathbf{x})(n_2)_{xx}(\mathbf{x} + \tilde{\varepsilon}(\mathbf{x})) - (n_2)_x(\mathbf{x} + \tilde{\varepsilon}(\mathbf{x})) d\mathbf{x}. \tag{3.13}$$

The terms $\mathcal{A}$ and $\mathcal{B}$ are stochastic and we must estimate their expectation and variance. The terms $\mathcal{O}_1$, $\mathcal{O}_2$, $\mathcal{O}_{\mathcal{A}}$, $\mathcal{O}_{\mathcal{B}}$ are higher order terms with respect to $\varepsilon - \mu$ and therefore negligible if $\varepsilon - \mu$ is small. The next lemma computes the variances of the main error terms caused by the noise.

LEMMA 3.1. *Consider the main error terms*

$$\mathcal{A} = \int_{\varphi_{\mathbf{x}_0}} (u(\mathbf{x} + \varepsilon(\mathbf{x})))_x \big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\overline{\varepsilon})\big) d\mathbf{x}$$

and

$$\mathcal{B} = \int_{\varphi_{\mathbf{x}_0}} \big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\overline{\varepsilon})\big)(n_2)_x(\mathbf{x}+\overline{\varepsilon}) d\mathbf{x}$$

as defined above. One has $\mathbb{E}\mathcal{A} = \mathbb{E}\mathcal{B} = 0$ *and*

$$\begin{aligned}
\mathrm{Var}(\mathcal{A}) &= 2\sigma^2 \int [\varphi(\mathbf{x} - \mathbf{x}_0) u(\mathbf{x}+\varepsilon)_x]_N^2 \, d\mathbf{x} \\
&\leq 2\sigma^2 \int \varphi(\mathbf{x} - \mathbf{x}_0)^2 (u(\mathbf{x}+\varepsilon))_x^2; \\
\mathrm{Var}(\mathcal{B}) &\leq \frac{2\pi^2\sigma^4}{3} \int \varphi(\mathbf{x} - \mathbf{x}_0)^2 d\mathbf{x} + \sigma^4 \int \varphi_x(\mathbf{x} - \mathbf{x}_0)^2 d\mathbf{x}.
\end{aligned}$$

See the appendix for the proof of this lemma.

THEOREM 3.2. **(Main disparity formula and noise error estimate)** *Let* $u_1(\mathbf{x})$ *and* $u_2(\mathbf{x})$ *be two images related by the geometrical deformation model described*

in (2.1) *and consider an optimal disparity* $\mu(\mathbf{x}_0)$ *obtained as any global minimizer of* $e_{\mathbf{x}_0}(\mu)$ *(defined by* (2.2)*). Then*

$$\mu(\mathbf{x}_0) = \frac{\int_{\varphi_{\mathbf{x}_0}} [u(\mathbf{x}+\varepsilon(\mathbf{x}))]_x^2 \varepsilon(\mathbf{x}) d\mathbf{x}}{\int_{\varphi_{\mathbf{x}_0}} [u(\mathbf{x}+\varepsilon(\mathbf{x}))]_x^2 d\mathbf{x}} + \mathcal{E}_{\mathbf{x}_0} + \mathcal{F}_{\mathbf{x}_0} + \mathcal{O}_{\mathbf{x}_0}\,, \tag{3.14}$$

where

$$\mathcal{E}_{\mathbf{x}_0} = \frac{\int_{\varphi_{\mathbf{x}_0}} (u(\mathbf{x}+\varepsilon(\mathbf{x})))_x \big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\overline{\varepsilon})\big) d\mathbf{x}}{\int_{\varphi_{\mathbf{x}_0}} [u(\mathbf{x}+\varepsilon(\mathbf{x}))]_x^2 d\mathbf{x}}$$

is the dominant noise term,

$$\mathcal{F}_{\mathbf{x}_0} = \frac{\int_{\varphi_{\mathbf{x}_0}} \big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\overline{\varepsilon})\big)(n_2)_x(\mathbf{x}+\overline{\varepsilon}) d\mathbf{x}}{\int_{\varphi_{\mathbf{x}_0}} [u(\mathbf{x}+\varepsilon(\mathbf{x}))]_x^2 d\mathbf{x}}$$

and

$$\mathcal{O}_{\mathbf{x}_0} = \frac{\mathcal{O}_1 + \mathcal{O}_2 + \mathcal{O}_{\mathcal{A}} + \mathcal{O}_{\mathcal{B}}}{\int_{\varphi_{\mathbf{x}_0}} [u(\mathbf{x}+\varepsilon(\mathbf{x}))]_x^2 d\mathbf{x}},$$

is made of smaller order terms in $\overline{\varepsilon} - \mu$*. In addition the variances of the main error terms due to noise satisfy*

$$\mathrm{Var}(\mathcal{E}_{\mathbf{x}_0}) = 2\sigma^2 \frac{\int [\varphi(\mathbf{x}-\mathbf{x}_0) u(\mathbf{x}+\varepsilon)_x]_N^2 \, d\mathbf{x}}{\Big(\int \varphi(\mathbf{x}-\mathbf{x}_0) u(\mathbf{x}+\varepsilon)_x^2 d\mathbf{x}\Big)^2}; \tag{3.15}$$

$$\mathrm{Var}(\mathcal{F}_{\mathbf{x}_0}) \leqslant \frac{\frac{2\pi^2}{3}\sigma^4 \int \varphi(\mathbf{x}-\mathbf{x}_0)^2 d\mathbf{x} + \sigma^4 \int \varphi_x(\mathbf{x}-\mathbf{x}_0)^2 d\mathbf{x}}{\Big(\int \varphi(\mathbf{x}-\mathbf{x}_0) u(\mathbf{x}+\varepsilon)_x^2 d\mathbf{x}\Big)^2}. \tag{3.16}$$

Proof: All above results are a direct consequence of the decomposition formula for the disparity, (3.5), the definitions of the error terms (3.11), (3.13) and (3.6), completed with the variance estimates in Lemma 3.1. □

Remark In all treated examples, it will be observed that $\mathrm{Var}(\mathcal{B}) \ll \mathrm{Var}(\mathcal{A})$, which by Lemma 3.1 will be true if

$$\sigma^2 \left[\frac{2\pi^2}{3} \int \varphi(\mathbf{x}-\mathbf{x}_0)^2 + \int \varphi_x(\mathbf{x}-\mathbf{x}_0)^2\right] \ll 2 \int [\varphi(\mathbf{x}-\mathbf{x}_0) u(\mathbf{x}+\overline{\varepsilon})_x]_N^2, \tag{3.17}$$

therefore if the noise variance σ is small enough with respect to the average image contrast of u_x in the block.

Discussion about the meaning of the theorem The expressions "dominant noise term" and "made of smaller terms" in the statement of Theorem 3.2 are mere interpretations of the terms of the formula. The first term in (3.14) gives a deterministic estimate of the correlation-maximizing μ, as weighted average of the real disparity ε in the block. This estimate is exact in the noiseless case at points around which ε is constant. Otherwise the difference between this deterministic term and the real $\varepsilon(\mathbf{x}_0)$ is the "fattening error", as pointed out in [12].

The variance of the second main term $\mathcal{E}_{\mathbf{x}_0}$ given by (3.15) will experimentally prove very close to the empirically observed RMSE of the measured disparity. The error term $\mathcal{O}_{\mathbf{x}_0}$, being formally an $O(\bar{\varepsilon} - \mu)$, will be small if $\mu = \mu(\mathbf{x}_0)$ is close to $\varepsilon(\mathbf{x}_0)$. Yet, this last fact is neither assumed nor ensured by the theorem. The theorem only assumes that μ is one global minimizer of $e_{\mathbf{x}_0}(\mu)$. Thus, Theorem 3.2 will be useful only if we can by some *oracle* select the points $\mathbf{x}_0$ at which $\mu(\mathbf{x}_0)$ is very close to $\varepsilon(\mathbf{x}_0)$ and therefore make $\mathcal{O}_{\mathbf{x}_0}$ small. On the other hand, if we happen to dispose of an *oracle* discarding the wrong matches, Theorem 3.2 gives a precise estimate of the dominant noise error term. Section 5 will compare the orders of magnitude of the overall error term and of its prediction (3.15).

4. The Numerical Implementation. The numerical implementation aiming at sharp sub-pixel accuracy and optimality must be very careful. This section details the main steps, namely the choice of the soft window φ, and the discrete correlation algorithm.

4.1. Choice of the Function φ. Section 2 showed that the minimization of $e_{\mathbf{x}_0}(\mu)$ only requires its knowledge for $\mu \in \mathbb{Z}_a^{1/2}$. The other values of $e_{\mathbf{x}_0}(\mu)$ are obtained by DFT interpolation. The 2-over-sampling of u_1 and u_2 is easy by zero-padding. The one-dimensional interpolation of $e_{\mathbf{x}_0}$ is done by a numerical approximation to the DFT interpolation.

Concerning the window function φ we would like it to be simultaneously:

- *of small spatial support*, say a few pixels, in order to both reduce computations, and to make the distance as local as possible, thus avoiding fattening (*a.k.a.* adhesion) effects; and
- *sufficiently regular* (a trigonometric polynomial of degree no larger than $2N$), in order to preserve the equality between the discrete and continuous quadratic distances (see proposition 2.1). This is required to make precise continuous computations possible on discrete samples, which is essential to make the link between accuracy the results in Theorem 3.2 (which is necessarily stated in teh continuous domain) with any discrete computer algorithm.

Since no function can have compact support both in the spatial and frequency domain

both requirements are apparently contradictory, but there is a sensible solution in this case.

Let us concentrate first on the small spatial support, then on the spectral support. At the end we explain how to construct a window function φ that conciliates both criteria.

φ with small spatial support. Here we relax slightly the band-limitedness assumption in favor of a small spatial support, to reduce computations and to better localize the result.

A prolate window function φ is optimal [27], in the sense that for a given spatial support $[-b, b]^2$ of size $(2b)^2$

it concentrates its Fourier coefficients as much as possible within $[-N, N-1]^2$.

$$\varphi_{\text{prolate}} = \underset{\varphi \in L^2([0,a]),\, \text{supp}(\varphi) \subseteq [-b,b]^2,\, \int |\varphi| = 1}{\arg\max} \frac{\sum_{\mathbf{k} \in ([-N,N-1] \cap \mathbb{Z})^2} |\tilde{\varphi}_{\mathbf{k}}|^2}{\sum_{\mathbf{k} \in \mathbb{Z}^2} |\tilde{\varphi}_{\mathbf{k}}|^2} \tag{4.1}$$

For instance for a spatial support $b = 1.5\frac{a}{2N}$ of 3×3 half-pixels the prolate function concentrates more than 99.8% of its L^2 energy within its central $(2N)^2$

Fourier coefficients. Typical correlation window sizes are larger (about 13×13 or even 17×17 half-pixels, in order to ensure a sufficiently dense set of meaningful matches [41, 42]), which leaves many more degrees of freedom, and quasi 100% energy concentration within the low-frequency band $[-N, N-1]^2$. The additional degrees of freedom may be used to satisfy further constraint as in [4]. This parameter choice makes the discrete correlation e^d almost equal (up to less than 0.2% error) to the continuous one e, in agreement with (2.10). The cost is just a 2 over-sampling, as specified in formula (2.11).

The fact that doubling the sampling rate was necessary to obtain accurate results had already been observed in [46], but their use of cubic interpolation and step window functions for φ limited the accuracy of their results. Exact interpolation and prolate functions have to be used to attain the twentieth of pixel. This is a crucial point: Otherwise, the resulting error is considerably higher, as shown in Section 5.2.

φ with compact spectral support and small *discrete* spatial support. The previous choice of φ is computationally convenient but it has the disadvantage that it only approximately satisfies the hypothesis of proposition 2.1. Thus the equality between the computed e^d and the continuous e is only approximate (up to about 0.2% error), and so are all our error estimates, which are based on the continuous version e.

Alternatively we can take any spatial support $[-b, b]^2$, arbitrarily choose the values of φ at half-pixels

$$\varphi_f^d(\mathbf{n}) = \begin{cases} f(\mathbf{n}) & \text{if } \mathbf{n} \in [-b,b]^2 \cap \overline{\mathbb{Z}}_a^{1/2} \\ 0 & \text{otherwise } (i.e.\ \mathbf{n} \in \overline{\mathbb{Z}}_a^{1/2}) \end{cases}$$

and define φ_f as the $2N$-degree trigonometric polynomial interpolating those samples.[2] Such a construction ensures the equality between discrete and continuous distances e^d and e, and the small (discrete) spatial support allows to make computations fast. However, it has the disadvantage that the continuous φ_f may have a large spatial support thus loosing localization of the result and potentially introducing fattening effects. This is especially true if φ_f^d is chosen as a box-function, thus leading to ringing artifacts when calculating the interpolated φ_f. However if we choose φ_f^d to decay smoothly to 0 near the borders of $[-b, b]^2$ then those ringing artifacts will be minimized. This is the idea of the combined solution explored next.

The final compromise. In order to conciliate both criteria we shall choose the half-pixel samples f within the spatial support $[-b, b]^2$ of φ_f^d so as to minimize the L^2 energy of φ_f outside this spatial support:

$$\varphi = \text{argmin}_{f \in \mathbb{R}^m} \frac{\int_{[0,a]^2 \setminus [-b,b]^2} |\varphi_f(\mathbf{x})|^2 d\mathbf{x}}{\int_{[0,a]^2} |\varphi_f(\mathbf{x})|^2 d\mathbf{x}}$$

where $dm = \sharp([-b, b]^2 \cap \overline{\mathbb{Z}}_a^{1/2})$. The construction is similar to the prolate window described in the previous paragraph, but inverting the roles of the Fourier and spatial domains and adding zero-interpolation constraints at half-pixel locations outside

[2] To be consistent with the common choice of a 0-centered window function φ_f, we used here a 0-centered grid

$$\overline{\mathbb{Z}}_a^{1/2} = \left[-\frac{a}{2}, \frac{a}{2}\right]^2 \cap \left(\left(\frac{a}{4N}, \frac{a}{4N}\right) + \frac{a}{2N}\mathbb{Z}^2\right)$$

instead of the one defined in equation (2.5).

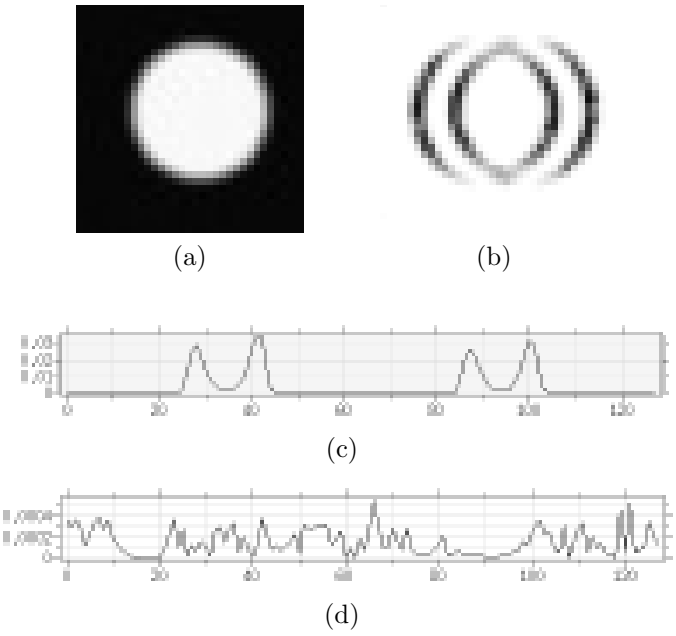

Fig. 4.1: (a) Reference image. The secondary image is a trivial DFT translation of the reference. (b) Unsigned error image. Small errors appear close to the edges of the disk due to the lack of samples in the interpolation. (c) Plot of a horizontal line of the error image. The error peak is approximately of $3 * 10^{-2}$ pixels. (d) Plot of the same line error when using $2N$ samples for interpolating e^d. The error peaks close to the disk boundaries disappear and the error is smaller than $5 * 10^{-4}$ for all the pixels of the line image. However, this is more expensive computationally. The use of a function φ as explained in Section 4.1 alleviates the numerical error due to the lack of samples. Indeed, the error doubles when a step window function φ is used, all other parameters being equal (peak of the order of $6 * 10^{-2}$).

$[-b, b]^2$. This way we can obtain a window function φ for which the equality $e = e^d$ is exactly true, and which has a small discrete spatial support (3×3 half-pixels), and a concentration of 99.8% of the L^2 energy of the continuous φ within this discrete spatial support, which is sufficient to avoid fattening effects beyond the size of the discrete window $[-b, b]^2$. In practice, we can obtain almost the same window and energy concentration by choosing φ_f, the vector f being the half-integer samples of φ_{prolate} within $[-b, b]^2$.

Numerical Error. In practice, the Shannon hypotheses are not completely satisfied in the interpolation of e^d. Indeed, not all of the $2N$ samples will be used for complexity reasons in this 1D interpolation. A slight accuracy loss in pixels close to edges of the image can therefore be observed in the toy example of a translated disk (see Fig. 4.1). In this example, we have compared the committed error when interpolating the truncated e^d with some samples and the complete e^d with $2N$ samples. The small error committed with the truncated e^d will be neglected in the sequel, because it is much smaller than the noise error.

4.2. The discrete Correlation Algorithm. Since the quadratic distance $e_{\mathbf{x}_0}(\mu)$ may present several local minima, the algorithm for accurately finding the minimizing μ is composed of two steps:

1. Localization of the "correct" local minimum along the epipolar line of $\mathbf{x}_0$ within an interval of length less than one pixel and elimination of errors by an "oracle" ruling out false matches happening by chance.
2. Fine localization of the selected local minimum up to the desired or attainable accuracy.

The oracle in step 1 is not the subject of this paper. It uses the *a contrario* method presented in [41, 42]. Here we suppose that the set of reliable matches in the images are known as well as the disparities with one pixel accuracy. So, the algorithm presented in this paper aims to refine such disparities.

Our algorithm is actually very similar to the "classic" correlation algorithm but using the theoretical results in Section 2 makes the difference. Indeed, Propositions (2.1) and (2.2) and a good choice of the window φ are the secrets to obtain very accurate disparities. In particular, the critical aspect is how $g = e_{\mathbf{x}_0}(\mu)$ is interpolated. The common approach, which consists in sampling g for integer disparities and interpolating these samples, provides a wrong result because of insufficient sampling rate. But DFT interpolation of a set of half-integer samples of g provides an *exact* interpolation, as shown in Section 2 (cf. Proposition 2.2). This is why the images are previously zoomed by a factor of 2 (see line 1 in Algorithm 1).

On the other hand, the spatial extent of the DFT has to be limited in order to save computational time. Here we used a DFT interpolation within an interval of length $L = 8$ around the initial search point μ_0 (see line 3 in Algorithm 1). Furthermore g is computed with a $W \times W$ ($W = 17$ in our case) half-pixels window size for φ. (see subsection 4.1 for the choice of φ.) The complete pseudo-code of our algorithm is:

Algorithm 1: PSEUDO-CODE

input : Images u_1 and u_2. Set of meaningful points $\mathbf{x}_0$ and its integer disparities μ_0.
output: Subpixel disparities for the meaningful points.

1 Zoom x2 of the images u_1 and u_2 (zero-padding);
2 **foreach** *meaningful point* $\mathbf{x}_0$ *and its associated* μ_0 **do**
3 Compute $e_{\mathbf{x}_0}(\mu)$ using the zoomed images (Eq. (2.11)) for $\mu \in [\mu_0 - \frac{L}{2}, \mu_0 + \frac{L}{2}]$;
4 Compute $\tilde{e}_{\mathbf{x}_0}(\mu)$ = DFT interpolation (x32) of $e_{\mathbf{x}_0}(\mu)$;
5 Compute $\tilde{\mu_0} = \text{argmin}_\mu \tilde{e}_{\mathbf{x}_0}(\mu)$;
6 Refine $\tilde{\mu_0}$ with the analytical minimum of the parabola fitting the current $\tilde{\mu_0}$ and its two closest points.
7 **end**
8 Return the set of disparities $\tilde{\mu_0}$;

Let us now analyze the complexity of our algorithm:

Initialization Computation of the half-integer samples $e_{\mathbf{x}_0}(\mu)$ for $\mu \in \mu_0 + \frac{1}{2}(\mathbb{Z} \cap [-\frac{L}{2}, \frac{L}{2}])$ requires

1. 2x zoom by zero-padding to obtain $u_1(\mathbf{m})$ and $u_2(\mathbf{m})$ at half integer locations $\mathbf{m} \in \mathbb{Z}_a^{1/2}$. This is done only once globally for the whole image ($(16 + 20 \log_2 N)$ flops/pixel).

2. Computation of $e_{\mathbf{x}_0}(\mu)$ using equation (2.11)
for each of the L samples: ($L \times 2W^2$ flops/pixel)

Total: $2LW^2 + 16 + 20\log_2 N^2$ flops/pixel

Evaluation of $e_{\mathbf{x}_0}(\mu)$ for a new value of $\mu \in \mu_0 + [-0.5, 0.5]$ requires a 1D Fourier translation (DFT interpolation) of length L, *i.e.* $2L\log_2 L$ flops/pixel/iteration. Note that we pay no penalty for each new interpolation and just a small initialization penalty with respect to the inexact version based on integer disparity sampling. On the other hand, an equally exact but brute-force solution based on image-interpolation instead of quadratic distance interpolation would transfer the burden of the initialization cost to each new evaluation of the distance function:

Initialization 2x zoom by zero-padding for u_1 and u_2 ($(16 + 20\log_2 N)$ flops/pixel).

Evaluation of $e(\mathbf{x}_0, \mu)$ for a new value of $\mu \in \mu_0 + [-0.5, 0.5]$ requires:

1. Non-integer translation of a $W \times W$ patch of the zoomed u_1 by the computationally least expensive among 1D sinc interpolation of size L or Fourier translation of size $M = L + W$.
($\min(LW, 2M\log_2 M)W$ flops/pixel)
2. Computation of $e_{\mathbf{x}_0}(\mu)$ using equation (2.11).
($2W^2$ flops/pixel)

Total: $(2W + \min(LW, (L+W)\log_2(L+W)))W$ flops/pixel/iteration

So, if the optimum search takes K iterations then the algorithm takes $16 + 20\log_2 N + 2LW^2 + K \times [2L\log_2 L]$ whereas the brute force approach would take $16 + 20\log_2 N + K \times [(2W + \min(LW, (L+W)\log_2(L+W)))W]$ The previous mathematical analysis shows that the proposed method is as accurate as the brute force method, but for typical values of $W = 17$, $L = 8$ and $N = 1024$ it computes each iteration **50 times faster** at the cost of a longer initialization. For typical values of K (5 to 7) this still means a global speedup of a factor **3**.

5. Results and Evaluation. Three experiments were performed to evaluate the attainable disparity error under realistic noise conditions. The standing problem of such evaluations is to obtain a reliable ground truth. So much more so for high sub-pixel accuracy, since there is no validated highly accurate benchmark data. Two ways were found to go around this problem. The first was to simulate urban aerial stereo pairs with realistic depth map and noise. Several simulated translations were also applied to Brodatz textures, thus avoiding the adhesion problem and focusing on the noise factor. Finally, images from the Middlebury dataset[3] were tested. In that case the noise was estimated, and we shall see that the quantized manual ground truth can be improved by cross-validation. In all cases, the resulting performance is evaluated by the Root Mean Squared Error (RMSE) measured on all pixels $\mathbf{q}$ in the set M of reliable match points,

$$RMSE = \left(\frac{\sum_{\mathbf{q} \in M} \left(\mu(\mathbf{q}) - \varepsilon(\mathbf{q})\right)^2}{\#M} \right)^{\frac{1}{2}},$$

where $\mu(\mathbf{q})$ is the computed disparity and $\varepsilon(\mathbf{q})$ is the ground truth disparity. This RMSE will be compared to the theoretical prediction of the noise error given by (3.15).

For the simulated cases the influence of noise in the matching process is studied with several image signal to noise ratios $SNR = \frac{\| u \|_2}{\sigma_n}$, where σ_n is the standard

[3] www.vision.middlebury.edu/stereo/

deviation of the noise. In each case σ_n is known and the predicted noise error is given by formula (3.15).

A main feature of the experimental setting is the use of a blind *a contrario* rejection method that *does not use the ground truth*. Thus, the percentage of wrong matches is also given, bad matches being those where the computed disparity differs by more than one pixel from the ground truth.

As explained in the introduction, the accuracy of matches can only be evaluated on pixels not exposed to fattening effect, which therefore lie away from disparity edges. These edges being *a priori* unknown, security zones were computed by dilating the strong grey level edges by the same window used for block-matching. The other pixels were matched only if they passed an *a contrario* test to ensure that the match is meaningful. As shown in [41, 42], the conjunction of both safety filters usually keeps more than half the pixels and ensures that the matched pixels are right with very high probability. For all experiments the sub-pixel refinement step goes up to $\frac{1}{64}$ pixel.

5.1. Simulated Stereo Pair. In order to provide the quantitative error when doing stereo sub-pixel matching, a secondary image was simulated from a reference image and a ground truth provided by IGN (French National Geographic). In this case the resulting couple of images has a low baseline ($B/H = 0.045$) and a 25 cm/pixel resolution. Figure 5.1 shows the reference stereo image, its ground truth, the mask of matched points, the sparse disparity map obtained by the sub-pixel block-matching algorithm, and the theoretical predicted error variance caused by the noise predicted by Formula (3.15) at each point. In this Figure the results obtained by Graph-Cuts [25] with the public code of Kolmogorov and Zabih are shown as well. This last algorithm is not well adapted to image pairs where a sub-pixel accuracy is needed. The range of disparity being [-2, 2] for this image, the obtained disparity map with Graph-Cuts is piecewise planar with only four labels. The image could be previously zoomed to get sub-pixel labels, as Birchfield and Tomasi suggested, but the ensuing complexity is unbearable with large images.

After the simulation of the secondary image a Gaussian noise was added independently to both images. Table 5.1 gives the RMSE on the disparity committed by the sub-pixel algorithm for decreasing SNRs. This table also gives the theoretical disparity RMSE, as predicted from the noise variance, the percentage of matched pixels, and the percentage of wrong matches. The case without noise (SNR $= \infty$) shows the limit of the sub-pixel accuracy, with a 0.023 numerical error (see Section 4.1). In presence of noise the observed RMSE differs by less than 0.008 pixel from its prediction. Thus, when the RMSE increases, the gap between predicted and real accuracy becomes very small, but this better estimate is obtained on fewer pixels.

5.2. Matching Textured Images. This experiment simulates the ideal case of two textured images (Figure 5.2-(a)) obtained from each other by a 2.5 pixels translation using zero-padding. An independent Gaussian noise was added independently to both images. Again, the observed RMSE has the same order of magnitude as the predicted noise error. For several textured images and signals the results were similar (see Table 5.2).

The very same test was led with cubic interpolation as proposed in [46] instead of the exact DFT interpolating method. The match of two textured images *without noise* had a RMSE of 0.24 instead of 0.0053. This test confirms that adopting the right interpolation is crucial for a sub-pixel stereo technology.

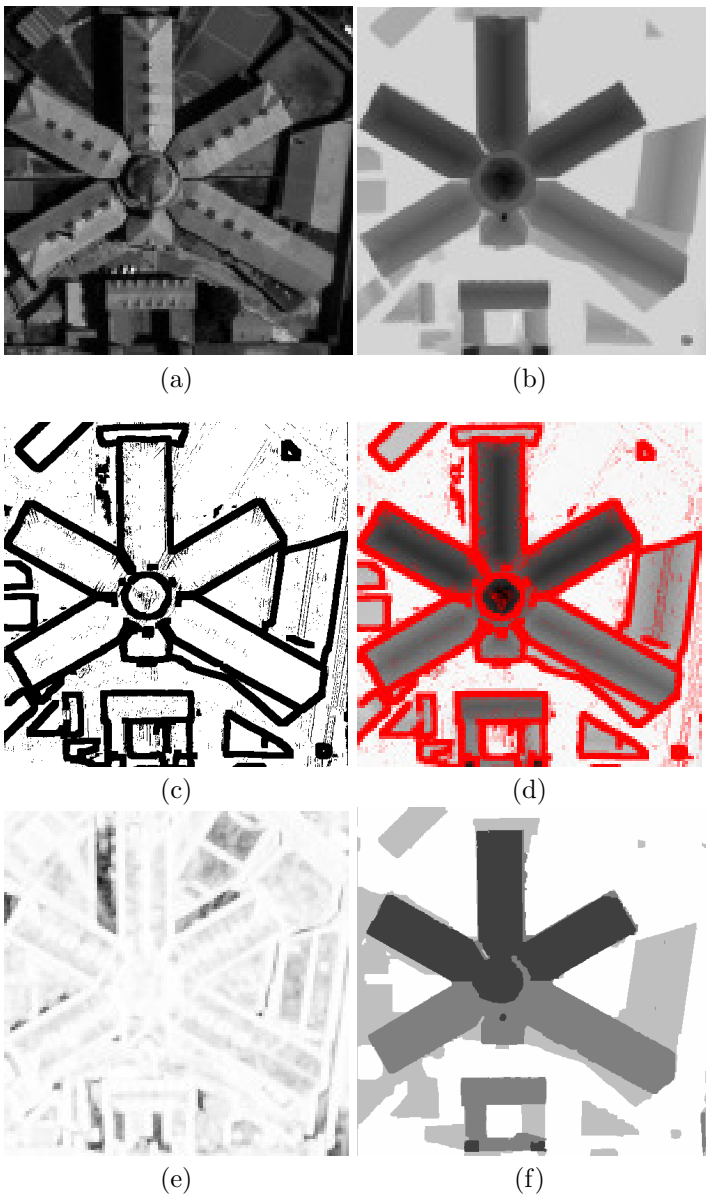

Fig. 5.1: Results for the simulated stereo pair. (a) Reference aerial image. (b) Ground truth. (c) Mask of matched points (white points (70.6%)). (d) Sparse disparity map. (e) Noise error prediction at each point. The darker the pixel, the higher the predicted error. Low gradient regions (e.g. shadows) have a larger predicted error. (f) Disparity map obtained by Graph-Cuts. Original stereo-pair and ground-truth were kindly provided by CNES, who holds the copyright. Simulated stereo-pair courtesy of Gabriele Facciolo [13].

SNR	RMSE (predicted)	RMSE (observed)	Matched points percentage	Wrong matches percentage
∞	**0**	**0.023**	70.6%	0.00%
357.32	**0.029**	**0.033**	63.3%	0.00%
178.66	**0.041**	**0.049**	54.2%	0.01%
125.06	**0.052**	**0.058**	41.5%	0.02%

Table 5.1: Qualitative results for the simulated stereo pair (Fig. 5.1). From left to right: image signal to noise ratio; RMSE (in pixels) predicted by Formula (3.15); RMSE to ground truth (in pixels); percentage of matched points and percentage of bad matches.

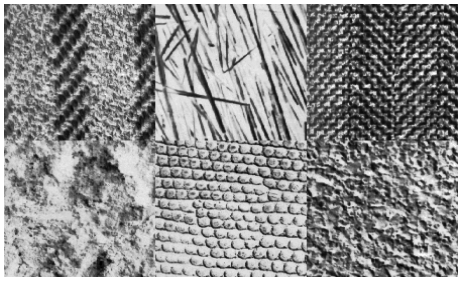

Fig. 5.2: Brodatz texture.

Table 5.3 summarizes the orders of magnitude of the terms in our main error formula (3.5) for the images in figs 5.1 and 5.2. For these images we know exactly the ground truth and the standard deviation σ of the added noise. The standard deviations of the main error terms in Theorem 3.2, $\mathcal{E}_{\mathbf{x}_0}$ and $\mathcal{E}_{\mathbf{x}_0} + \mathcal{F}_{\mathbf{x}_0}$ and $\mathcal{O}_{\mathbf{x}}$ were computed (R_E, R_{E+F} and R_O respectively) where $\mathrm{Var}(\mathcal{E}_{\mathbf{x}_0} + \mathcal{F}_{\mathbf{x}_0})$ was bounded from above by $\mathrm{Var}(\mathcal{E}_{\mathbf{x}_0}) + \mathrm{Var}(F_{\mathbf{x}_0}) + 2(\mathrm{Var}\,\mathcal{E}_{\mathbf{x}_0}\,\mathrm{Var}\,\mathcal{F}_{\mathbf{x}_0})^{1/2}$ in the computation of R_{E+F}. This table confirms that the formula (3.5) scales correctly the orders of magnitude, and that the main error term is due to the noise and not to the adhesion.

5.3. Simulated 3D scene. In this experiment a virtual 3D scene with perfect planes was created. Then a texture was projected on each plane. Finally two high resolution simulated snapshots of the scene were used as input for the sub-pixel algorithm (with a 0.04 B/H ratio). In this simulated example the ground truth is known with a 0.001 pixel precision. Figure 5.3 shows the used images, the ground truth and the result and Table 5.4 summarizes the qualitative results of this experiment, in particular the predicted and obtained error when adding independent white noises to the images.

5.4. Middlebury Images. The last experiments were done on the Middlebury classic dataset, which also publishes a hand-made ground truth. The first test was made on Sawtooth, a piecewise planar stereo pair. Table 5.5 gives (column R_0) a 20/100 pixel distance to the ground truth. Dequantizing the Middlebury ground truth (column R_1) improves slightly this distance to the ground truth to roughly 16/100. Still, with comparable noise level, this distance is thrice the error in the simulated experiments! *Yet a closer analysis of the results shows that the real error is*

SNR	RMSE (predicted)	RMSE (observed)	Matches percentage	Wrong matches percentage
∞	**0**	**0.0053**	100%	0.0%
96.38	**0.0048**	**0.0073**	99.8%	0.0%
48.19	**0.0096**	**0.0109**	99.8%	0.0%
32.12	**0.0141**	**0.0160**	98.7%	0.0%
24.09	**0.0192**	**0.0203**	87.1%	0.0%

Table 5.2: Quantitative results for textures (Fig. 5.2). From left to right: image signal to noise ratio; RMSE (theoretical prediction (3.15)); observed RMSE (in pixels); percentage of matched points, percentage of wrong matches.

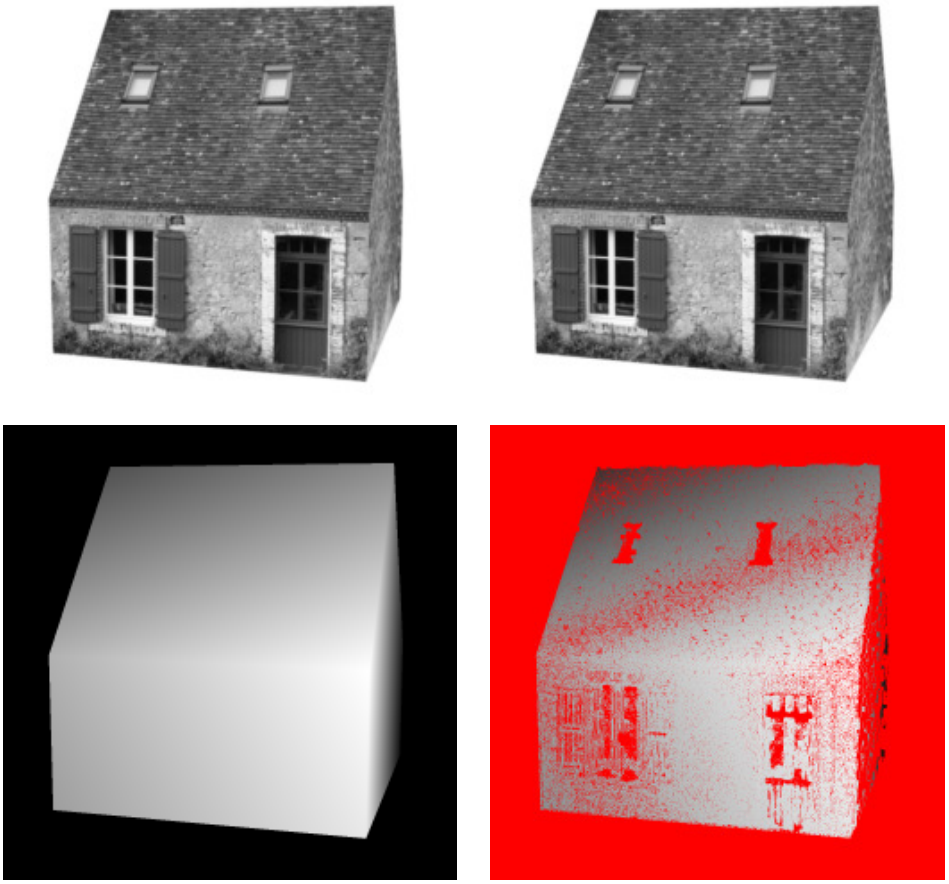

Fig. 5.3: Top: Two high precision simulated snapshots of a 3D scene. Bottom: ground truth and disparity map. Ground truth and synthetic stereo-pair courtesy of Lionel Moisan [30] and CNES, who holds the copyright of these images.

SNR	R_E	R_{E+F}	R_O
∞	0	0	0
357.32	0.029	0.030	0.0015
178.66	0.041	0.044	0.0027
125.06	0.052	0.053	0.0033
∞	0	0	0
96.38	0.0048	0.0051	0.0024
48.19	0.0096	0.0103	0.0030
32.12	0.0141	0.0145	0.0039
24.09	0.0192	0.0193	0.0042

Table 5.3: Order of magnitude of the terms in formula (3.5). Top of the table: simulated stereo pair (Fig. 5.1). Bottom of the table: Textures (Fig. 5.2). From left to right: Signal to noise ratio; R_E predicted noise error computed from $\mathcal{E}_{\mathbf{x}_0}$; R_{E+F} error from $\mathcal{E}_{\mathbf{x}_0} + \mathcal{F}_{\mathbf{x}_0}$. R_O: variance of the explicit computation of O using the ground truth ε. The contribution of $\mathcal{F}_{\mathbf{x}_0}$ in R_{E+F} is negligible and R_O is of the order of 0.003 pixel.

SNR	RMSE (observed)	RMSE predicted	Accepted matches	Wrong matches
∞	0.04	0	73.8%	0.0%
189	0.07	0.058	59.3%	0.0%
126	0.11	0.077	33.1%	0.00%
63	0.14	0.134	12.9%	0.008%

Table 5.4: Comparative results for the experiment of Section 5.3

close to 9/100 *pixel.* Indeed, we know that the manual ground truth in Middlebury is not sub-pixel accurate: As explained in the Middlebury web site, the ground truth is a quantized estimation of the affine motion of each planar, manually labeled, image object. *A more faithful ground truth can actually be recovered from the image pair itself.* Indeed, assuming that the data set was accurately piecewise planar permits to compute the error between the sub-pixel matching result and its own plane-fit. **The standard deviation of this error goes down to** 9/100 **respectively (see column** R_2**).** On the other hand, an independent error estimate of the obtained disparity map (not relying on the ground truth) can be obtained by *cross-validating* the disparity measurements applied to several different stereo pairs of the same 3D scene. Indeed, the Middlebury data set provides nine ortho-rectified views at uniform intervals, so that disparity maps taking the central view as reference image are related to one another by a scaling constant which depends on the baseline ratios. *The RMSE error between the scaled disparity maps (see column* R_3*) turns out to be in full agreement with the piecewise planar check (column* R_2*).* The predicted noise error was computed by using an estimation of the standard deviation of the noise of the image given by the accurate noise estimator in [7].

The above Sawtooth test demonstrates the (relatively) poor accuracy of the ground truth. In consequence, for the current four pairs of images in the Middlebury

	w.r.t. ground truth		cross-validation		RMSE
	R_0	R_1	R_2	R_3	(predicted)
Sawtooth	0.213	0.162	0.09	0.090	0.076

Table 5.5: From left to right: RMSE to "official" ground truth. RMSE to the plane-fit of the official ground truth. RMSE to the plane-fit of the sub-pixel correlation results. RMSE of cross-validation with 7 additional views. Finally, predicted disparity error due to noise (using an accurate noise estimate on the pictures themselves).

benchmark, we decided to cross-validate the sub-pixel correlation results by using all the images of each scene in the dataset (5 images for Tsukuba and 9 images for Venus, Teddy and Cones). Table 5.6 compares the obtained RMSE by cross-validation with the predicted theoretical noise error.

	w.r.t. ground truth	cross-validation	Predicted Noise Error
Tsukuba	0.357	0.080	0.069
Venus	0.225	0.101	0.042
Teddy	0.424	0.093	0.072
Cones	0.319	0.082	0.066

Table 5.6: Quantitative results for the Tsukuba, Venus, Teddy and Cones images. The first column corresponds to the RMSE to the ground truth computed in the mask of valid points. The second column is the RMSE by cross-validation of the 5 or 9 images in the dataset. Finally, the noise error (3.15) predicted by the theory appears in the last column.

Comparison of existing algorithms in our mask. For the sake of evidence of the ground truth imperfection we have also checked that classic stereo algorithms provide disparity maps that are closer to each other (and to our result) than to the ground truth. The tested algorithms are actually at the top of the Middlebury evaluation table: AdaptingBP [24], CoopRegion [49], SubPixDoubleBP [51], CSemiGlob [19] and GC+SegmBorder [10].

Table 5.7 gives the quadratic errors when comparing two by two the considered algorithms. The values in the diagonal of the tables (gray) are the RMSE with respect to the ground truth. All of these error values have been computed in our same valid point mask. The distance between any two solutions is for most of the cases smaller than the distance of these solutions to the ground truth. The only exception of the 6 tested algorithms is GC+SegmBorder.

5.5. Conclusion, and a few objections. The empirical sub-pixel accuracy in stereo vision can attain its predicted limit, which only depends on the noise at regular disparity points. The experiments on realistically simulated pairs and real benchmark images show a 1/20 pixel accuracy to be attained by block-matching, for more than half the image pixels, under realistic noise conditions. These image pixels were not found *a posteriori*, they were specified *a priori* by an autonomous algorithm, meaning that the small predicted accuracy can be predicted *a priori* for these points. The two ingredients, namely a sub-pixel accuracy and a strict *a priori* mismatch control, could make stereo-vision (with high SNR and low baseline) into potential competitor

IMAGES	Our algorithm	AdaptingBP	CoopRegion	SubPixDoubleBP	CSemiGlob	GC+SegmBorder	ALGORITHMS
Tsukuba	0.357	0.281	0.258	0.245	0.216	0.223	Our algorithm
		0.297	0.251	0.289	0.253	0.300	AdaptingBP
			0.337	0.241	0.214	0.241	CoopRegion
				0.264	0.253	0.272	SubPixDoubleBP
					0.275	0.272	CSemiGlob
						0.207	GC+SegmBorder
Venus	0.225	0.231	0.130	0.143	0.161	0.239	Our algorithm
		0.215	0.146	0.163	0.176	0.223	AdaptingBP
			0.131	0.122	0.129	0.142	CoopRegion
				0.162	0.148	0.174	SubPixDoubleBP
					0.192	0.212	CSemiGlob
						0.142	GC+SegmBorder
Teddy	0.424	0.341	0.336	0.346	0.352	0.511	Our algorithm
		0.421	0.330	0.312	0.303	0.531	AdaptingBP
			0.354	0.262	0.317	0.509	CoopRegion
				0.385	0.291	0.519	SubPixDoubleBP
					0.411	0.542	CSemiGlob
						0.481	GC+SegmBorder
Cones	0.319	0.281	0.267	0.245	0.253	0.421	Our algorithm
		0.331	0.262	0.301	0.319	0.483	AdaptingBP
			0.272	0.252	0.262	0.446	CoopRegion
				0.349	0.234	0.487	SubPixDoubleBP
					0.365	0.510	CSemiGlob
						0.400	GC+SegmBorder

Table 5.7: Comparison of several stereo algorithms in the top classification. Values on the diagonals (gray) are the values with respect to the ground truth. In general the distance between two solutions is smaller than the distance to the ground truth, with the exception of GC+SegmBorder.

with laser range scanners. Indeed, at a 70cm distance, the best triangulation scanners achieve a 20μ accuracy, only on Lambertian objects and very controlled environment and this requires multiple scans. On the other hand a well-calibrated stereo camera system with 4000 pixels width, situated at the same distance of the object can achieve a 70/4000cm disparity accuracy at the well matched points if the accuracy is pixelian. This yields a $1.7510^{-2}cm = 175\mu$ accuracy for pixelian stereo systems. But if the stereo disparity attains a 0.10 pixels accuracy we get a 17.5μ accuracy, which is exactly comparable to the best current triangulation scanner accuracy.

The above Middlebury experiments also showed that the ground truth provided as a benchmark reference is actually less accurate than the attainable level of 1/20 pixel (see [51] for similar conclusions). This study also suggests that a rigorous methodology

to create reliable ground truths is needed. Such ground truths should be built up by several different automatic devices used repeatedly on the same objects, so as to provide a cross-validated estimate of their own accuracy.

A first objection to the sub-pixel algorithm presented herewith is that it only delivers a non dense disparity map, while the users usually desire a 3D rendering of the objects and therefore a dense map. There is no contradiction, however, with this goal in the present study. Global interpolation methods can be used to complete the reliable pixels, knowing of course that the accuracy of the new disparities will not be guaranteed. It is clear that stereo algorithms giving a final dense map must be more complex than just block-matching. But our conclusion is that a block-matching is necessary and fruitful if accuracy is at stake.

A second objection is that the study does not take into account the actual trend of working not with a stereo pair, but rather with a whole set of images of a 3D scene taken from many viewpoints. The present study can and must be extended to this general setting with some advantage. For example if we dispose of five successive snapshots with low baseline, a fusion of the four obtained disparities should decrease the error variance by a four factor, and therefore go beyond the accuracy estimated here.

A third objection is that the translation model (2.1) which is the basis of all computations here, is geometrically too simplistic. It is in essence only true for the parts of the observed objects facing the cameras, and in the low baseline framework. However, it so happens that for the low baseline case, the translation dominates the other perspective deformations, as has been systematically observed in experiments. In the case of slanted surfaces with respect to the cameras, the present theory can be used anyway. In fact, local matching is reduced to an (approximate) translation, after a local affine transform on one of the images. This is doable but goes beyond the scope of the present study.

Acknowledgments. The authors acknowledge financial support from the French Space Agency (CNES), ECOS Sud project U06E01, ANR FREEDOM and Callisto projects, European Research Council (advanced grant Twelve Labours) and Office of Naval research (grant N00014-97-1-0839). The stereo pairs in Figures 5.1 and 5.3 are copyrighted by CNES, and provided with contributions from Lionel Moisan and Gabriele Facciolo.

Appendix A: Proof of Lemma 2.4. Integrating by parts in x we have

$$V := \mathrm{Var}\left(\int \varphi(\mathbf{x}) n(\mathbf{x}) n_x(\mathbf{x}) d\mathbf{x}\right) = \mathrm{Var}\left(\frac{1}{2}\int \varphi_x(\mathbf{x}) n(\mathbf{x})^2 d\mathbf{x}\right).$$

Since $n(\mathbf{x})^2$ and $\varphi(\mathbf{x})$ are $2N$-degree trigonometric polynomials, (2.9) can be used with $a = N$:

$$V = \frac{1}{4}\,\mathrm{Var}\left(\frac{1}{4}\sum_{\mathbf{m}\in\mathbb{Z}_N^{1/2}} \varphi_x(\mathbf{m}) n(\mathbf{m})^2\right).$$

This sum can be split into

$$V = \frac{1}{4^3}\,\mathrm{Var}(S_1+S_2+S_3+S_4) \leqslant \frac{1}{4^2}\sum_{i=1}^{4}\mathrm{Var}(S_i)\,, \tag{5.1}$$

where $S_i = \sum_{\mathbf{m}\in A_i} \varphi_x(\mathbf{m}) n(\mathbf{m})^2$, $A_i = [0,N]^2 \cap (a_i + \mathbb{Z}^2)$, $a_1 = (1/4, 1/4)$, $a_2 = (1/4, 3/4)$, $a_3 = (3/4, 1/4)$, and $a_4 = (3/4, 3/4)$. We shall evaluate for example

$$\mathrm{Var}(S_1) = \mathrm{Var}\left(\sum_{\mathbf{m}\in A_1} \varphi_x(\mathbf{m}) n(\mathbf{m})^2 \right).$$

The samples $n(\mathbf{m})$, $\mathbf{m} \in A_1$ being independent, $\mathrm{Var}(S_1) = \sum_{\mathbf{m}\in A_1} \varphi_x(\mathbf{m})^2 \mathrm{Var}(n(\mathbf{m})^2)$ which by Lemma 2.3 yields $\mathrm{Var}(S_1) = 2\sigma^4 \sum_{\mathbf{m}\in A_1} \varphi_x(\mathbf{m})^2$. Thus, from (5.1) follows that $V \leqslant \frac{2\sigma^4}{4^2} \sum_{\mathbf{m}\in \mathbb{Z}_N^{1/2}} \varphi_x(\mathbf{m})^2$ which, using again (2.9) with $a = N$, yields

$$V \leqslant \frac{4 \times 2\sigma^4}{4^2} \int \varphi_x^2(\mathbf{x}) d\mathbf{x} = \frac{\sigma^4}{2} \int \varphi_x^2(\mathbf{x}) d\mathbf{x}.$$

Also,

$$\mathbb{E} \int \varphi(\mathbf{x}) n(\mathbf{x}) n_x(\mathbf{x}) d\mathbf{x} = -\frac{1}{2} \mathbb{E} \int \varphi_x(\mathbf{x}) n(\mathbf{x})^2 d\mathbf{x} =$$

$$= -\frac{1}{2} \int \varphi_x(\mathbf{x}) \mathbb{E} n(\mathbf{x})^2 d\mathbf{x} = -\frac{\sigma^2}{2} \int \varphi_x(\mathbf{x}) d\mathbf{x} = 0.$$

The second part of the lemma is easier. By the Fourier series isometry (2.4),

$$\int_{[0,N]^2} g(\mathbf{x}) n(\mathbf{x}) d\mathbf{x} = N^2 \sum_{k,l\in\mathbb{Z}} \tilde{g}_{k,l} \tilde{n}_{k,l} =$$

$$= N^2 \sum_{-\frac{N}{2} \leqslant k,l \leqslant \frac{N}{2}-1} \tilde{g}_{k,l} \tilde{n}_{k,l}.$$

Indeed, n being a degree N-trigonometric polynomial, $\tilde{n}_{k,l} = 0$ for $(k,l) \notin [-N/2, N/2-1]^2$. Since the $\tilde{n}_{k,l}$ are independent with variance $\frac{\sigma^2}{N^2}$, we obtain the announced result by taking the variance of the last finite sum:

$$\mathrm{Var}\left(\int_{[0,N]^2} g(\mathbf{x}) n(\mathbf{x}) d\mathbf{x} \right) = \sigma^2 N^2 \sum_{-\frac{N}{2} \leq k,l \leqslant \frac{N}{2}-1} |\tilde{g}_{k,l}|^2.$$

By (2.7), this yields

$$\mathrm{Var}\left(\int_{[0,N]^2} g(\mathbf{x}) n(\mathbf{x}) d\mathbf{x} \right) = \sigma^2 \int_{[0,N]^2} g_N(\mathbf{x})^2 d\mathbf{x},$$

where

$$g_N(\mathbf{x}) := \sum_{-N/2 \leqslant k,l \leqslant N/2-1} \tilde{g}_{k,l} e^{\frac{2i\pi(kx+ly)}{a}}$$

is the degree N-trigonometric polynomial best approximating g for the quadratic distance. □

Appendix B: Proof of Lemma 3.1. Notice that $n_1(\mathbf{x})$ and $n_2(\mathbf{x}+\overline{\varepsilon})$ are independent Gaussian noises with variance σ^2. Thus their difference is again a Gaussian noise with variance $2\sigma^2$. It therefore follows from (2.16) in the second part of Lemma 2.4 that

$$\mathrm{Var}(\mathcal{A}) = 2\sigma^2 \int [\varphi(\mathbf{x}-\mathbf{x}_0)u(\mathbf{x}+\varepsilon)_x)]_N^2 \, d\mathbf{x} \leq 2\sigma^2 \int \varphi(\mathbf{x}-\mathbf{x}_0)^2 (u(\mathbf{x}+\varepsilon)_x)^2 d\mathbf{x}.$$

The noises n_1 and n_2 being independent, by the second part of Lemma 2.4, by the second relation in Lemma 2.3 and by (2.15) in the first part of Lemma 2.4,

$$\mathrm{Var}(\mathcal{B}) \leq 2\left[\mathrm{Var}(\int_{\varphi_{\mathbf{x}_0}} n_1(\mathbf{x})(n_2)_x(\mathbf{x}+\overline{\varepsilon})d\mathbf{x}) + \mathrm{Var}(\int_{\varphi_{\mathbf{x}_0}} n_2(\mathbf{x}+\overline{\varepsilon})(n_2)_x(\mathbf{x}+\overline{\varepsilon})d\mathbf{x})\right]$$

$$\leq 2\left[\sigma^2 \times \frac{\pi^2\sigma^2}{3}\int \varphi^2(\mathbf{x}-\mathbf{x}_0)d\mathbf{x} + \frac{\sigma^4}{2}\int \varphi_x(\mathbf{x}-\mathbf{x}_0)^2 d\mathbf{x}\right]$$

$$= \frac{2\pi^2\sigma^4}{3}\int \varphi(\mathbf{x}-\mathbf{x}_0)^2 d\mathbf{x} + \sigma^4 \int \varphi_x(\mathbf{x}-\mathbf{x}_0)^2 d\mathbf{x}.$$

□

REFERENCES

[1] M. Balci and H. Foroosh. Subpixel estimation of shifts directly in the fourier domain. *Image Processing, IEEE Transactions on*, 15(7):1965 –1972, 2006.

[2] A. Banno and K. Ikeuchi. Disparity map refinement and 3d surface smoothing via directed anisotropic diffusion. In *3-D Digital Imaging and Modeling (3DIM)*, 2009.

[3] S. Birchfield and C. Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *Trans. PAMI*, 20(4):401–406, 1998.

[4] G. Blanchet, A. Buades, B. Coll, J.-M. Morel, and B. Rougé. Fattening free correlation algorithms.

[5] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.

[6] M. Z. Brown, D. Burschka, and G. D. Hager. Advances in computational stereo. *Trans. PAMI*, 25(8):993–1008, 2003.

[7] T. Buades, Y. Lou, J.-M. Morel, and Z. Tang. A note on multi-image denoising. 2008.

[8] A. Chambolle. Total variation minimization and a class of binary mrf models. In Anand Rangarajan, Baba Vemuri, and Alan L. Yuille, editors, *Energy Minimization Methods in Computer Vision and Pattern Recognition*, volume 3757 of *Lecture Notes in Computer Science*, chapter 10, pages 136–152. Springer-Verlag, Berlin/Heidelberg, 2005.

[9] L. Chen and K.-H. Yap. An effective technique for subpixel image registration under noisy conditions. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 38(4):881 –887, 2008.

[10] W. Chen, M. Zhang, and Z. Xiong. Segmentation-based stereo matching with occlusion handling via region border constraints. 2009.

[11] J. Darbon and M. Sigelle. A fast and exact algorithm for total variation minimization. In *Pattern Recognition and Image Analysis*, pages 351–359. 2005.

[12] J. Delon and B. Rougé. Small baseline stereovision. *Journal of Mathematical Imaging and Vision*, 28(3):209–223, 2007.

[13] G. Facciolo. Variational adhesion correction with image based regularization for digital elevation models. Master's thesis, Universidad de la República (Uruguay), August 2005.

[14] G. Facciolo, A. Lecumberry, F.and Almansa, A. Pardo, V. Caselles, and B. Rougé. Constrained anisotropic diffusion and some applications. In *British Machine Vision Conference (BMVC 2006)*, Edinburgh, September 2006.

[15] S. Forstmann, Y. Kanou, J. Ohya, S. Thuering, and A. Schmitt. Real-time stereo by using dynamic programming. In *IEEE CVPR Workshop*, volume 3, pages 29–36, Washington, 2004.

[16] S. Gehrig and U. Franke. Improving stereo sub-pixel accuracy for long range stereo. In *ICCV VRML workshop*, 2007.

[17] A. Giros, B. Rougé, and H. Vadon. Appariement fin d'images stéréoscopiques et instrument dédié avec un faible coefficient stéréoscopique. French Patent N.0403143, 2004.

[18] M. Gong, R. Yang, L. Wang, and M. Gong. A performance study on different cost aggregation approaches used in real-time stereo matching. *International Journal of Computer Vision*, 75(2):283–296, 2007.

[19] H. Hirschmuller. Stereo vision in structured environments by consistent semi-global matching. *CVPR*, pages 2386–2393, 2006.

[20] H. Ishikawa. Exact optimization for markov random fields with convex priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1333–1336, 2003.

[21] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *Trans. PAMI*, 16(9):920–932, 1994.

[22] S. B. Kang, R. Szeliski, and J. Chai. Handling occlusions in dense multi-view stereo. *CVPR*, I:103–110, 2001.

[23] M. Kasser and Y. Egels. *Digital Photogrammetry*. Taylor & Francis, 2002.

[24] A. Klaus, M. Sormann, and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. *ICPR*, pages 15–18, 2006.

[25] V. Kolmogorov and R. Zabih. Graph cut algorithms for binocular stereo with occlusions. In *Handbook of Mathematical Models in Computer Vision*. Springer-Verlag, 2005.

[26] J. Kybic. Bootstrap resampling for image registration uncertainty estimation without ground truth. *Image Processing, IEEE Transactions on*, 19(1):64 –73, 2010.

[27] H.J. Landau and H.O. Pollak. Fourier spheroidal wave functions, Fourier analysis and uncertainty (III): the dimension of the space of essentially time and bandlimited signals. *Bell System Technical Journal*, 41(1295-1336), 1962.

[28] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, vol. 60(2):91–110, 2004.

[29] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *In Seventh International Joint Conference on Artificial Intelligence (IJCAI-81)*, pages 674–679, 1981.

[30] L. Moisan. Simulation of high-precision stereo pairs using geometric integration, 2010.

[31] P. Musé, F. Sur, F. Cao, J.-L. Lisani, and J.-M. Morel. A Theory of Shape Identification. Lecture Notes in Mathematics. Springer, 2008.

[32] D. Nehab, S. Rusinkiewicz, and J. Davis. Improved sub-pixel stereo correspondences through symmetric refinement. In *ICCV*, volume 1, pages 557–563, Washington, DC, USA, 2005. IEEE Computer Society.

[33] Y. Ohta and T. Kanade. Stereo by intra- and inter-scanline search using dynamic programming. *Trans. PAMI*, 7(2):139–154, 1985.

[34] M. Okutomi and T. Kanade. A multiple-baseline stereo. *Trans. PAMI*, 15(4):353–363, 1993.

[35] E. Z. Psarakis and G. D. Evangelidis. An enhanced correlation-based method for stereo correspondence with sub-pixel accuracy. In *ICCV*, volume 1, pages 907–912, Washington, DC, USA, 2005. IEEE Computer Society.

[36] J. Ren, J. Jiang, and T. Vlachos. High-accuracy sub-pixel motion estimation from noisy images in fourier domain. *Image Processing, IEEE Transactions on*, 19(5):1379 –1384, May 2010.

[37] D. Robinson and P. Milanfar. Fundamental performance limits in image registration. *Image Processing, IEEE Transactions on*, 13(9):1185 –1199, 2004.

[38] D. Robinson and P. Milanfar. Bias minimizing filter design for gradient-based image registration. *SIGNAL PROCESSING-IMAGE COMMUNICATION*, 20:554–568, 2005.

[39] G.K. Rohde, A. Aldroubi, and D.M. Healy. Interpolation artifacts in sub-pixel image registration. *Image Processing, IEEE Transactions on*, 18(2):333 –345, 2009.

[40] N. Sabater, L. Moisan, G. Blanchet, A. Almansa, and J.-M. Morel. Review of low-baseline stereo algorithms and benchmarks. In *Proceedings of SPIE, vol. 7830 (Image and Signal Processing for Remote Sensing XVI)*, 2010.

[41] N. Sabater, J.-M. Morel, and A. Almansa. Rejecting wrong matches in stereovision, CMLA Preprint 2008-28. 2008.

[42] N. Sabater, J.-M. Morel, A. Almansa, and G. Blanchet. Discarding moving objects in quasi-simultaneous stereovision. In *IEEE International Conference on Image Processing, ICIP*, 2010.

[43] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspon-

dence algorithms. *IJCV*, (47):7–42, 2002.

[44] C. Schmid and A. Zisserman. The geometry and matching of lines and curves over multiple views. *IJCV*, 40(3):199–234, 2000.

[45] M. Shimizu and M. Okutomi. Precise sub-pixel estimation on area-based matching. *International Conference on Computer Vision*, 1:90–97, 2001.

[46] R. Szeliski and D. Scharstein. Sampling the disparity space image. *Trans. PAMI*, 26(3):419–425, 2004.

[47] Q. Tian and M.N. Huhns. Algorithms for subpixel registration. *Computer Vision, Graphics and Image Processing*, 35(2):220–233, 1986.

[48] O. Veksler. Fast variable window for stereo correspondence using integral images. *CVPR*, 1:556–561, 2003.

[49] Z.-F. Wang and Z.-G. Zheng. A region based stereo matching algorithm using cooperative optimization. 2008.

[50] M. Xu, H. Chen, and P.K. Varshney. Ziv-zakai bounds on image registration. *Signal Processing, IEEE Transactions on*, 57(5):1745 –1755, May 2009.

[51] Q. Yang, R. Yang, J. Davis, and D. Nister. Spatial-depth super resolution for range images. In *CVPR '07.*, pages 1–8, 2007.

[52] I.S. Yetik and A. Nehorai. Performance bounds on image registration. *Signal Processing, IEEE Transactions on*, 54(5):1737 – 1749, May 2006.

[53] S. Yoon, K.-J.and Kweon. Adaptive support-weight approach for correspondence search. *Trans. PAMI*, 28(4):650–656, 2006.