Inverse Problems and Imaging
Volume X, No. 0X, 200X, X–XX

Web site: http://www.aimSciences.org

IS SIFT SCALE INVARIANT?

Jean-Michel Morel

CMLA, ENS Cachan, 61 avenue du Président Wilson
94235 Cachan Cedex, France

Guoshen Yu

CMAP, Ecole Polytechnique
91128 Palaiseau Cedex, France

(Communicated by Professor Otmar Scherzer)

Abstract. This note is devoted to a mathematical exploration of whether Lowe's *Scale-Invariant Feature Transform* (SIFT) [21], a very successful image matching method, is similarity invariant as claimed. It is proved that the method is scale invariant only if the initial image blurs is exactly guessed. Yet, even a large error on the initial blur is quickly attenuated by this multiscale method, when the scale of analysis increases. In consequence, its scale invariance is almost perfect. The mathematical arguments are given under the assumption that the Gaussian smoothing performed by SIFT gives an aliasing free sampling of the image evolution. The validity of this main assumption is confirmed by a rigorous experimental procedure, and by a mathematical proof. These results explain why SIFT outperforms all other image feature extraction methods when it comes to scale invariance.

1. Introduction

Image comparison is a fundamental step in many computer vision and image processing applications. A typical image matching method first detects points of interest, then selects a region around each point, and finally associates with each region a descriptor. Correspondences between two images can then be established by matching the descriptors of both images. The images under comparison may have been taken under arbitrary viewpoints. Therefore the invariance to the viewpoint is crucial in image comparison. Many variations exist on the computation of invariant interest points, following the pioneering work of Harris and Stephens [13]. The Harris-Laplace and Hessian-Laplace region detectors [23, 26] are invariant to rotation and scale changes. Some moment-based region detectors [20, 2] including the Harris-Affine and Hessian-Affine region detectors [24, 26], an edge-based region detector [42], an intensity- based region detector [42], an entropy-based region detector [14], and two independently developed level line-based region detectors MSER ("maximally stable extremal region") [22] and LLD ("level line descriptor") [33, 34] are designed to be invariant to affine transformations. These two methods stem

2000 *Mathematics Subject Classification:* Primary: 68T10, 68T40; Secondary: 68T45, 93C85.
Key words and phrases: SIFT, scale invariance, Shannon interpolation, Gaussian blur, sampling theory, aliasing.
Research partially financed by the MISS project of Centre National d'Etudes Spatiales, the Office of Naval research under grant N00014-97-1-0839 and by the European Research Council, advanced grant "Twelve labors".

from the Monasse image registration method [29] that used well contrasted extremal regions to register images. MSER is the most efficient one and has shown better performance than other affine invariant detectors [28]. However, as pointed out in [21], none of these detectors is actually fully affine invariant: All of them start with initial feature scales and locations selected in a non-affine invariant manner. The difficulty comes from the scale change from an image to another: This change of scale is in fact an under-sampling, which means that the images differ by a blur.

In his milestone paper [21], Lowe has addressed this central problem and has proposed the so called scale-invariant feature transform (SIFT) descriptor, that is claimed to be invariant to image translations and rotations, to scale changes (blur), and robust to illumination changes. It is also surprisingly robust to large enough orientation changes of the viewpoint (up to 60 degrees when one of the images is in frontal view). Based on the scale-space theory [19], the SIFT procedure simulates all Gaussian blurs and normalizes local patches around scale covariant image key points that are Laplacian extrema. A number of SIFT variants and extensions, including PCA-SIFT [15] and gradient location-orientation histogram (GLOH) [27], that claim to have better robustness and distinctiveness with scaled-down complexity have been developed ever since [8, 18]. Demonstrated to be superior to other descriptors [25, 27], SIFT has been popularly applied for scene recognition [6, 30, 39, 44, 11, 40] and detection [9, 35], robot localization [3, 36, 32], image retrieval [12], motion tracking [43, 16], 3D modeling and reconstruction [38, 45], building panoramas [1, 4], or photo management [46, 17, 5]. Partially based on the SIFT similarity invariance proved in the present paper, recently, a variant of SIFT named affine-SIFT (ASIFT) has been mathematically proved to be fully invariant, and has been shown to give excellent performance even under very large viewpoint changes [31].

The initial goal of the SIFT method is to compare two images (or two image parts) that can be deduced from each other (or from a common one) by a rotation, a translation, and a zoom. In this method, following a classical paradigm, stable points of interest are supposed to lie at extrema of the Laplacian of the image in the image scale-space representation. The scale-space representation introduces a smoothing parameter σ. Images u_0 are smoothed at several scales to obtain $w(\sigma, x, y) := (G_\sigma * u_0)(x, y)$, where

$$G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

is the 2D-Gaussian function with integral 1 and standard deviation σ. The notation $*$ stands for the space 2-D convolution in (x, y). The description of the SIFT method involves sampling issues, which we shall discuss later.

Taking apart all sampling issues and several thresholds whose aim it is to eliminate unreliable features, the whole SIFT method can be summarized in one single sentence:

One sentence description *The SIFT method computes scale-space extrema (σ_i, x_i, y_i) of the Laplacian in (x, y) of $w(\sigma, x, y)$, and then samples for each one of these extrema a square image patch whose origin is (x_i, y_i), whose x-direction is one of the dominant gradients around (x_i, y_i), and whose sampling rate is proportional to (and usually smaller than) $\sqrt{\sigma_i^2 + \mathbf{c}^2}$.*

The constant $\mathbf{c} \approx 0.5$ is the tentative standard deviation of the initial image blur. The resulting samples of the digital patch at scale σ_i are encoded by their

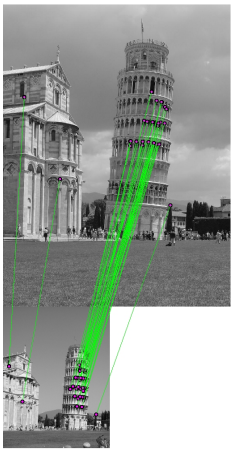


Figure 1. A result of the SIFT method, using an outliers elimination method [37]. Pairs of matching points are connected by segments.

gradient direction, which is invariant under nondecreasing contrast changes. This accounts for the robustness of the method to illumination changes. In addition, only local histograms of the direction of the gradient are kept, which accounts for the robustness of the final descriptor to changes of view angle (see Fig. 2).

The goal of this paper is to give the mathematical formalism to examine whether the method indeed is scale invariant, and to discuss whether its main assumption, that images are well-sampled under Gaussian blur, does not entail significant errors. We shall not propose a new variant or an extension of the SIFT method; on the contrary we intend to demonstrate that no other method will ever improve more than marginally the SIFT scale invariance (see Figs. 1 and 4 for striking examples). To the best of our knowledge, and in spite of the more than ten thousand papers quoting and using SIFT, the analysis presented here does not seem to have been done previously.

The paper is organized as follows. In Section 2, a simple formalism is introduced to obtain a condensed description of the SIFT shape encoding method. Using this formalism Section 3 proves mathematically that the SIFT method indeed computes translation, rotation and scale invariants. This proof is correct under the main assumption that the image initial blur is Gaussian, and that images with a Gaussian blur larger than 0.8 can be accurately retrieved by interpolation from their samples. Section 4 checks the validity of this crucial well sampling assumption through an experimental procedure and mathematical proofs.

2. Image operators formalizing SIFT

The analysis of the scale invariance is much easier on the continuous images whose samples form the digital image. The Shannon-Whittaker interpolation permits a perfect reconstruction of a continuous image from a discrete one, when the continuous image has been well-sampled [41]. Under the assumption that a Gaussian filtering can deliver well-sampled images up to a negligible error, (the validity of this assumption will be confirmed in Section 4), this section gives a formalized description of the SIFT procedure.

We denote by $u(\mathbf{x})$ a continuous image defined for every $\mathbf{x} = (x, y) \in \mathbb{R}^2$. All *continuous image* operators, including the sampling operator itself, will be written in capital letters A, B. Their composition will be, for the sake of simplicity, written as a mere juxtaposition AB. For any similarity transform A its application on u is defined as $Au(\mathbf{x}) := u(A\mathbf{x})$. For instance $H_\lambda u(\mathbf{x}) := u(\lambda\mathbf{x})$ denotes an expansion of u by a factor λ^{-1}. In the same way if R is a rotation, $Ru := u(R\mathbf{x})$ is the image rotation by R^{-1}.

2.1. SAMPLING AND INTERPOLATION. Digital (discrete) images are only defined for $\mathbf{k} = (k, l) \in \mathbf{Z}^2$ and are denoted in bold by $\mathbf{u}(\mathbf{k})$. For example the δ-sampled image $\mathbf{u} = \mathbf{S}_\delta u$ is defined on $\mathbf{Z}^2$ by

$$(\mathbf{S}_\delta u)(k, l) := u(k\delta, l\delta); \tag{1}$$

Conversely, the Shannon interpolate of a digital image is defined as follows [10]. Let $\mathbf{u}$ be a digital image, defined on $\mathbf{Z}^2$ and such that $\sum_{\mathbf{k}\in\mathbf{Z}^2} |\mathbf{u}(\mathbf{k})|^2 < \infty$ and $\sum_{\mathbf{k}\in\mathbf{Z}^2} |u(\mathbf{k})| < \infty$. These conditions are for example satisfied if the digital image has a finite number of non-zero samples. We call Shannon interpolate $\mathbf{Iu}$ of $\mathbf{u}$ the only $L^2(\mathbb{R}^2)$ function u which coincides with $\mathbf{u}$ on the samples $\mathbf{k}$ and with spectrum support contained in $(-\pi, \pi)^2$. $\mathbf{Iu}$ is defined by the Shannon-Whittaker formula

$$\mathbf{Iu}(x, y) := \sum_{(k,l)\in\mathbf{Z}^2} \mathbf{u}(k, l)\text{sinc}(x - k)\text{sinc}(y - l), \tag{2}$$

where $\text{sinc}\, x := \frac{\sin \pi x}{\pi x}$. The Shannon interpolation has the fundamental property $\mathbf{S}_1\mathbf{Iu} = \mathbf{u}$. Conversely, if u is L^2 and band-limited in $(-\pi, \pi)^2$, then

$$\mathbf{IS}_1 u = u. \tag{3}$$

In that ideal situation we say that u is *band-limited*. We shall also say that the digital image $\mathbf{u} = \mathbf{S}_1 u$ is *well-sampled* if it was obtained from a band-limited image u, and therefore permits to go back to u.

While the Shannon interpolation (2) allows a perfect reconstruction of a continuous image from a discrete one when the image is well-sampled, this interpolation $\mathbf{I}$ is unfortunately impractical since it assumes an infinite number of available samples. Instead, the DFT interpolation $\mathbf{I}_d$ that relies only on a limited number of samples in the discrete image rectangle is predominant in image processing. The DFT interpolation $\mathbf{I}_d$ will be defined in Section 4.1 where its approximation to the Shannon interpolation $\mathbf{I}$ will be checked. In the following we will thus proceed in this dual framework. For invariance proofs involving translations, rotations and zooms, the Shannon interpolation is necessary because we need to reason in a space of functions on $\mathbf{R^2}$ which is invariant by these geometric transforms. When it comes to image interpolation operations on the limited image domain, $\mathbf{I}_d$ will be invoked. It is well documented that the difference between these two interpolations causes negligible errors, but we shall anyway confirm this fact here.

2.2. THE GAUSSIAN FILTERING. The Gaussian convolution that implements the scale-space plays a key role in the SIFT procedure. G_σ will denote the convolution operator on $\mathbb{R}^2$ with the Gaussian kernel $G_\sigma(x_1, x_2) = \frac{1}{2\pi\sigma^2} e^{-(x_1^2+x_2^2)/2\sigma^2}$, and the Gaussian kernel itself. Thus we simply write $G_\sigma u(x, y) := (G_\sigma * u)(x, y)$. G_σ satisfies the semigroup property

$$G_\sigma G_\beta = G_{\sqrt{\sigma^2+\beta^2}}. \tag{4}$$

The proof of the next formula is a mere change of variables in the integral defining the convolution.

$$G_\sigma H_\gamma u = H_\gamma G_{\sigma\gamma} u. \tag{5}$$

The discrete Gaussian convolution applied to a digital image is defined as a digital operator by

$$\mathbf{G}_\delta \mathbf{u} =: \mathbf{S}_1 G_\delta \mathbf{I}\mathbf{u}. \tag{6}$$

This discrete convolution is nothing but the continuous Gaussian convolution applied to the underlying continuous image. This definition maintains the Gaussian semi-group property used repeatedly in SIFT,

$$\mathbf{G}_\delta \mathbf{G}_\beta = \mathbf{G}_{\sqrt{\delta^2+\beta^2}}. \tag{7}$$

Indeed, using twice (6) and once (4) and (3),

$$\mathbf{G}_\delta \mathbf{G}_\beta \mathbf{u} = \mathbf{S}_1 G_\delta \mathbf{I}\mathbf{S}_1 G_\beta \mathbf{I}\mathbf{u} = \mathbf{S}_1 G_\delta G_\beta \mathbf{I}\mathbf{u} = \mathbf{S}_1 G_{\sqrt{\delta^2+\beta^2}} \mathbf{I}\mathbf{u} = \mathbf{G}_{\sqrt{\delta^2+\beta^2}} \mathbf{u}.$$

(The same formulas are adapted without alteration when replacing $\mathbf{I}$ by $\mathbf{I}_d$.)

The SIFT method makes the following assumption, whose validity will be confirmed both experimentally and mathematically in Section 4.

Assumption 1. *For every σ larger than 0.8 and every Shannon interpolated digital image u_0, the Gaussian blurred image $G_\sigma u_0$ satisfies the Shannon inversion formula up to a negligible error, namely $I\mathbf{S}_1 G_\sigma u_0 \simeq G_\sigma u_0$.*

2.3. FORMALIZED SCALE INVARIANT FEATURES TRANSFORM. The Assumption 1 that the Gaussian pre-filtering leads to a nearly aliasing-free subsampling allows a perfect reconstruction of continuous images from discrete ones with Shannon-Whittaker interpolation. The main steps of the SIFT method can therefore be formalized in a continuous setting as follows.

1. **Geometry:** there is an underlying infinite resolution planar image $u_0(\mathbf{x})$ that has undergone a similarity Au_0 (modeling the composition of a rotation, a translation, and a homothety) before sampling.
2. **Sampling and blur:** the camera blur is assumed to be a Gaussian with standard deviation $\mathbf{c}$. The initial digital image is therefore $\mathbf{u} = \mathbf{S}_1 G_\mathbf{c} A u_0$;
3. **Sampled scale space:** the SIFT method computes enough samples of the scale space function $u(\sigma, \cdot) = G_\sigma G_\mathbf{c} A u_0$ to detect accurately "key points" $(\sigma, \mathbf{x})$, defined as scale and space local extrema of $\Delta u(\sigma, \cdot)$.
4. **Covariant resampling:** a 32×32 grid centered at the key point is used to sample $u(\sigma, \cdot)$ around each key point $(\sigma, \mathbf{x})$. The grid mesh is proportional to $\sqrt{\sigma^2 + \mathbf{c}^2}$. The directions of the sampling grid axes are fixed by a dominant direction of $\nabla u(\sigma, \cdot)$ in a neighborhood of the key point, whose size is also proportional to the key point scale σ. This yields for each key point a rotation, translation and scale invariant square sampled subimage samples in which the four parameters of A have been eliminated (see Fig. 3);
5. **Illumination invariance:** the final SIFT descriptors keep mainly the orientation of the samples gradient, to gain invariance with respect to light conditions.

Steps 1 to 5 are the main steps of the method. We have omitted all details that are not relevant in the discussion to follow. Let them be mentioned briefly. The Laplacian extrema are kept only if they are larger than a fixed threshold that

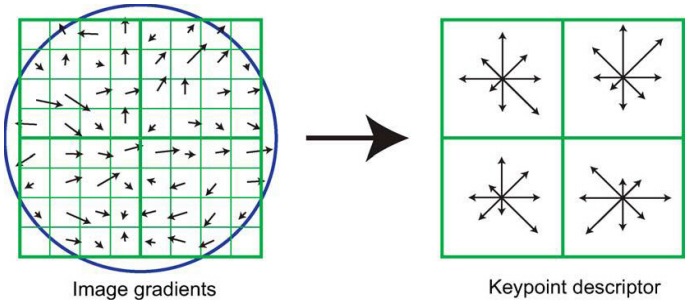


FIGURE 2. Each key-point is associated *a square image patch whose size is proportional to the scale and whose side direction is given by the assigned direction.* Example of a 2×2 descriptor array of orientation histograms (right) computed from an 8×8 set of samples (left). The orientation histograms are quantized into 8 directions and the length of each arrow corresponds to the magnitude of the histogram entry.

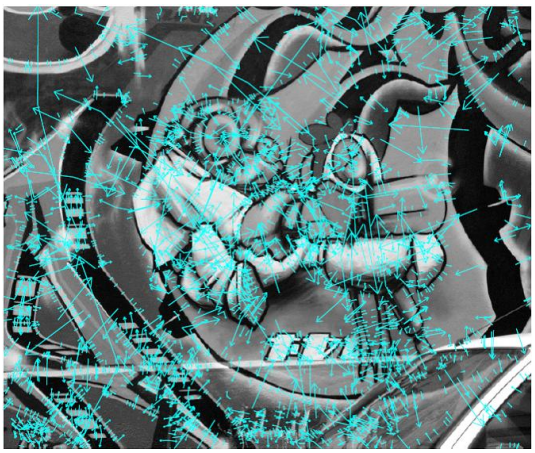


FIGURE 3. SIFT key points. The arrow starting point, length and the orientation signify respectively the key point position, scale, and dominant orientation. These features are claimed to be covariant to any image similarity.

eliminates small features mainly due to noise. This threshold is not scale invariant. The ratio of the eigenvalues of the Hessian of the Laplacian must be close enough to 1 to ensure a good key point localization. (Typically, straight edge points have only one large Hessian eigenvalue, are poorly localized, and are therefore ruled out by this second threshold, which is scale invariant.)

The SIFT method assumes that the initial image satisfies $\mathbf{c} = 0.5$ (meaning that it is the result of a convolution with a Gaussian with standard deviation $\mathbf{c}$). This implies a slight under-sampling which is compensated by a complementary Gaussian blur applied to the image. Following Assumption 1 it increases the initial blur to 0.8. In accordance with this choice, a 2×2 subsampling in the SIFT scale-space computations can be made only when a $2 \times 0.8 = 1.6$ Gaussian blur has been reached.

Postponing the verification of the SIFT main Assumption 1 to Section 4, the next Section proves that SIFT has an almost perfect scale invariance, which is the main result of the present paper.

3. Scale and SIFT: consistency of the method

Let $\mathcal{T}$, R, H and G be respectively an arbitrary image translation, an arbitrary image rotation, an arbitrary image homothety, and an arbitrary Gaussian convolution, all applied to continuous images. We say that there is strong commutation if we can exchange the order of application of two of these operators. We say that there is weak commutation between two of these operators if we have (e.g.) $R\mathcal{T} = \mathcal{T}'R$, meaning that given R and $\mathcal{T}$ there is $\mathcal{T}'$ such that the former relation occurs. The next lemma is straightforward.

Lemma 1. *All of the aforementioned operators weakly commute. In addition, R and G commute strongly.*

In this Section, in conformity with the SIFT model of Section 2, the digital image is a frontal view of an infinite resolution ideal image u_0. In that case, $A = H\mathcal{T}R$ is the composition of a rotation R, a translation $\mathcal{T}$ and a homothety H. Thus the digital image is $\mathbf{u} = \mathbf{S}_1 G_\delta H\mathcal{T}Ru_0$, for some H, $\mathcal{T}$, R. Assuming that the image is not aliased boils down, by the experimental results of Section 4, to assuming $\delta \geq 0.8$. The next Lemma shows that SIFT is rotation- and translation-invariant.

Lemma 2. *For any rotation R and any translation $\mathcal{T}$, the SIFT descriptors of $\mathbf{S}_1 G_\delta H\mathcal{T}Ru_0$ are identical to those of $\mathbf{S}_1 G_\delta Hu_0$.*

Proof. Using the weak commutation of translations and rotations with all other operators (Lemma 1), it is easily checked that the SIFT method is rotation and translation invariant: The SIFT descriptors of a rotated or translated image are identical to those of the original. Indeed, the set of scale space Laplacian extrema is covariant to translations and rotations. Then the normalization process for each SIFT descriptor situates the origin at each extremum in turn, thus canceling the translation, and the local sampling grid defining the SIFT patch has axes given by peaks in its gradient direction histogram. Such peaks are translation invariant and rotation covariant. Thus, the normalization of the direction also cancels the rotation. □

Lemma 3. *Let $\mathbf{u}$ and $\mathbf{v}$ be two digital images that are frontal snapshots of the same continuous flat image u_0, $\mathbf{u} = \mathbf{S}_1 G_\beta H_\lambda u_0$ and $\mathbf{v} := \mathbf{S}_1 G_\delta H_\mu u_0$, taken at different distances, with different Gaussian blurs and possibly different sampling rates. Let $w(\sigma, \mathbf{x}) := (G_\sigma u_0)(\mathbf{x})$ denote the scale space of u_0. Then the scale spaces of $\mathbf{u}$ and $\mathbf{v}$ are*

$$u(\sigma, \mathbf{x}) = w(\lambda\sqrt{\sigma^2 + \beta^2}, \lambda\mathbf{x}) \quad \text{and} \quad v(\sigma, \mathbf{x}) = w(\mu\sqrt{\sigma^2 + \delta^2}, \mu\mathbf{x}).$$

If $(s_0, \mathbf{x}_0)$ is a key point of w satisfying $s_0 \geq \max(\lambda\beta, \mu\delta)$, then it corresponds to a key point of u at the scale σ_1 such that $\lambda\sqrt{\sigma_1^2 + \beta^2} = s_0$, whose SIFT descriptor is sampled with mesh $\sqrt{\sigma_1^2 + \mathbf{c}^2}$, where $\mathbf{c}$ is the tentative standard deviation of the initial image blur as described in Section 2.3. In the same way $(s_0, \mathbf{x}_0)$ corresponds to a key point of v at scale σ_2 such that $s_0 = \mu\sqrt{\sigma_2^2 + \delta^2}$, whose SIFT descriptor is sampled with mesh $\sqrt{\sigma_2^2 + \mathbf{c}^2}$.

Proof. The interpolated initial images are by (3)

$$u := \mathbf{IS}_1 G_\beta H_\lambda u_0 = \mathbf{G}_\beta H_\lambda u_0 \text{ and } v := \mathbf{IS}_1 G_\delta H_\mu u_0 = G_\delta H_\mu u_0.$$

Computing the scale-space of these images amounts to convolve them for every $\sigma > 0$ with G_σ, which yields, using the Gaussian semigroup property (4) and the commutation relation (5):

$$u(\sigma, \cdot) = G_\sigma G_\beta H_\lambda u_0 = G_{\sqrt{\sigma^2+\beta^2}} H_\lambda u_0 = H_\lambda G_{\lambda\sqrt{\sigma^2+\beta^2}} u_0.$$

By the same calculation, this function is compared by SIFT with

$$v(\sigma, \cdot) = H_\mu G_{\mu\sqrt{\sigma^2+\delta^2}} u_0$$

. Let us set $w(s, \mathbf{x}) := (G_s u_0)(\mathbf{x})$. Then the scale spaces compared by SIFT are

$$u(\sigma, \mathbf{x}) = w(\lambda\sqrt{\sigma^2+\beta^2}, \lambda\mathbf{x}) \text{ and } v(\sigma, \mathbf{x}) = w(\mu\sqrt{\sigma^2+\delta^2}, \mu\mathbf{x}).$$

Let us consider an extremal point $(s_0, \mathbf{x}_0)$ of the Laplacian of the scale space function w. If $s_0 \geq \max(\lambda\beta, \mu\delta)$, an extremal point occurs at scales σ_1 for (the Laplacian of) $u(\sigma, \mathbf{x})$ and σ_2 for (the Laplacian of) $v(\sigma, \mathbf{x})$ satisfying

$$s_0 = \lambda\sqrt{\sigma_1^2+\beta^2} = \mu\sqrt{\sigma_2^2+\delta^2}. \tag{8}$$

We recall that each SIFT descriptor at a key point $(\sigma_1, \mathbf{x}_1)$ is computed from space samples of $\mathbf{x} \to u(\sigma, \mathbf{x})$. The origin of the local grid is $\mathbf{x}_1$, the intrinsic axes are fixed by one of the dominant directions of the gradient of $u(\sigma_1, \cdot)$ around $\mathbf{x}_1$, in a circular neighborhood whose size is proportional to σ_1. The SIFT descriptor sampling rate around the key point is proportional to $\sqrt{\sigma_1^2+\mathbf{c}^2}$ for $u(\sigma_1, \mathbf{x})$, and to $\sqrt{\sigma_2^2+\mathbf{c}^2}$ for $u(\sigma_2, \mathbf{x})$, as described in Section 2.3. □

Theorem 1. *Let* $\mathbf{u}$ *and* $\mathbf{v}$ *be two digital images that are frontal snapshots of the same continuous flat image* u_0*,* $\mathbf{u} = \mathbf{S}_1 G_\beta H_\lambda \mathcal{T} R u_0$ *and* $\mathbf{v} := \mathbf{S}_1 G_\delta H_\mu u_0$*, taken at different distances, with different Gaussian blurs and possibly different sampling rates, and up to a camera translation and rotation around its optical axis. Without loss of generality, assume* $\lambda \leq \mu$*. Then if the initial blurs are identical for both images (if* $\beta = \delta = \mathbf{c}$*), then each SIFT descriptor of* $\mathbf{u}$ *is identical to a SIFT descriptor of* $\mathbf{v}$*. If* $\beta \neq \delta$ *(or* $\beta = \delta \neq \mathbf{c}$*), the SIFT descriptors of* $\mathbf{u}$ *and* $\mathbf{v}$ *become (quickly) similar when their scales grow, namely as soon as* $\frac{\sigma_1}{\max(\mathbf{c},\beta)} \gg 1$ *and* $\frac{\sigma_2}{\max(\mathbf{c},\delta)} \gg 1$*, where* σ_1 *and* σ_2 *are respectively the scales of the key points in the two images.*

Proof. By the result of Lemma 2, we can neglect the effect of translations and rotations. Therefore assume without loss of generality that the images under comparison are as in Lemma 3. Consider a key point $(s_0, \mathbf{x}_0)$ of w with scale $s_0 \geq \max(\lambda\beta, \mu\delta)$. Following Lemma 3, there is a corresponding key point $(\sigma_1, \frac{\mathbf{x}_0}{\lambda})$ for $\mathbf{u}$ whose sampling rate is fixed by the method to $\sqrt{\sigma_1^2+\mathbf{c}^2}$ and a corresponding key point $(\sigma_2, \frac{\mathbf{x}_0}{\mu})$ whose sampling rate is fixed by the method to $\sqrt{\sigma_2^2+\mathbf{c}^2}$ for $\mathbf{v}$. To have a common reference for these sampling rates, it is convenient to refer to the corresponding sampling rates for $w(s_0, \mathbf{x})$, which are $\lambda\sqrt{\sigma_1^2+\mathbf{c}^2}$ for the SIFT descriptors of $\mathbf{u}$ at scale σ_1, and $\mu\sqrt{\sigma_2^2+\mathbf{c}^2}$ for the descriptors of $\mathbf{v}$ at scale σ_2. Thus the SIFT descriptors of $\mathbf{u}$ and $\mathbf{v}$ for $\mathbf{x}_0$ will be identical if and only if $\lambda\sqrt{\sigma_1^2+\mathbf{c}^2} = \mu\sqrt{\sigma_2^2+\mathbf{c}^2}$. Since

we have $\lambda\sqrt{\sigma_1^2+\beta^2}=\mu\sqrt{\sigma_2^2+\delta^2}$, the SIFT descriptors of $\mathbf{u}$ and $\mathbf{v}$ are identical if and only if

$$\lambda\sqrt{\sigma_1^2+\beta^2}=\mu\sqrt{\sigma_2^2+\delta^2}\Rightarrow\lambda\sqrt{\sigma_1^2+\mathbf{c}^2}=\mu\sqrt{\sigma_2^2+\mathbf{c}^2}. \tag{9}$$

In other terms $\lambda\sqrt{\sigma_1^2+\mathbf{c}^2}=\mu\sqrt{\sigma_2^2+\mathbf{c}^2}$ if and only if

$$\lambda^2\beta^2-\mu^2\delta^2=(\lambda^2-\mu^2)\mathbf{c}^2. \tag{10}$$

Since λ and μ correspond to camera distances to the observed object u_0, their values are arbitrary. Thus in general the only way to get (10) is to have $\beta=\delta=\mathbf{c}$, which means that the blurs of both images have been guessed correctly.

The second statement is straightforward: if σ_1 and σ_2 are large enough with respect to β, δ and $\mathbf{c}$, the relation $\lambda\sqrt{\sigma_1^2+\beta^2}=\mu\sqrt{\sigma_2^2+\delta^2}$, implies $\lambda\sqrt{\sigma_1^2+\mathbf{c}^2}\approx\mu\sqrt{\sigma_2^2+\mathbf{c}^2}$. □

The almost perfect scale invariance of SIFT stated in Theorem 1 is illustrated by the striking example of Fig. 4. The 25 SIFT key points of a very small image $\mathbf{u}$ are compared to the 60 key points obtained by zooming in $\mathbf{u}$ by a 32 factor: The resulting digital image is $\mathbf{v}:=\mathbf{S}_{\frac{1}{32}}\mathbf{I}_d\mathbf{u}$, again obtained by zero-padding. For better observability, both images are displayed with the same size by enlarging the pixels of $\mathbf{u}$. Almost each key point (18 out of 25) of $\mathbf{u}$ finds its counterpart in $\mathbf{v}$. 18 matches are detected between the descriptors as shown on the right. Let us check how this extreme example is covered by Theorem 1. We compare an initial image $\mathbf{u}=\mathbf{S}_1G_\delta\mathbf{I}_d\mathbf{u}_0$ (with $\delta=\mathbf{c}$) with its zoomed in version $\mathbf{v}=\mathbf{S}_{\frac{1}{32}}G_\delta\mathbf{I}_d\mathbf{u}_0$. But we have by (5)

$$\mathbf{v}=\mathbf{S}_{\frac{1}{32}}G_\delta\mathbf{I}_d\mathbf{u}_0=\mathbf{S}_1H_{\frac{1}{32}}G_\delta\mathbf{I}_d\mathbf{u}_0=\mathbf{S}_1G_{32\delta}H_{\frac{1}{32}}\mathbf{I}_d\mathbf{u}_0.$$

Here the numerical application of the relations in the above proof give: We want (9) to hold approximately, where $\mu=1$, $\lambda=\frac{1}{32}$, $\beta=32\delta$. Thus we want $\frac{1}{32}\sqrt{\sigma_1^2+(32\delta)^2}=\sqrt{\sigma_2^2+\delta^2}$ to imply $\frac{1}{32}\sqrt{\sigma_1^2+\mathbf{c}^2}\approx\sqrt{\sigma_2^2+\mathbf{c}^2}$ which means $\sqrt{(\frac{\sigma_1}{32})^2+\mathbf{c}^2}=\sqrt{\sigma_2^2+\mathbf{c}^2}$ to imply $\sqrt{(\frac{\sigma_1}{32})^2+(\frac{\mathbf{c}}{32})^2}\approx\sqrt{\sigma_2^2+\mathbf{c}^2}$. This is true only if σ_1 is significantly larger than 32, which is true, since σ_1 is the scale of the SIFT descriptors in the image $\mathbf{v}$, which has been zoomed in by a 32 factor.

By the second part of Theorem 1 the reliability of the SIFT matching increases with scale. This fact is illustrated in Fig. 5. Starting from a high resolution image $\mathbf{u}_0$, two images $\mathbf{u}$ and $\mathbf{v}$ are obtained by simulated zoom out, $\mathbf{u}=\mathbf{S}_1G_\beta H_\lambda\mathbf{I}_d\mathbf{u}_0=\mathbf{S}_\lambda G_{\lambda\beta}\mathbf{I}_d\mathbf{u}_0$ and $\mathbf{v}=\mathbf{S}_\mu G_{\mu\delta}\mathbf{I}_d\mathbf{u}_0$, with $\lambda=2$, $\mu=4$, $\beta=\delta=0.8$. Pairs of SIFT descriptors of $\mathbf{u}$ and $\mathbf{v}$ in correspondence, established by a SIFT matching, are compared using an Euclidean distance $\mathbf{d}$. The scale rate σ_1/σ_2 as well as the distance d between the matched key points are plotted against σ_2 in Fig. 5. That $\sigma_1/\sigma_2\approx2$ for all key points confirms that the SIFT matching process is reliable. As stated by the theorem, the rate σ_1/σ_2 goes to $\mu/\lambda=2$ when σ_2 increases, and the distance $\mathbf{d}$ goes down. However, as also apparent in the numerical result, when the scale is small ($\sigma_2<1$), σ_1/σ_2 is very different from 2 and $\mathbf{d}$ is large.

4. The right Gaussian blur to achieve well-sampling

This section first checks the close relation between the impractical Shannon interpolation, with which we had proved the scale invariance of SIFT, and the DFT interpolation that is always applied in practice for image resampling. The rest of the section checks the SIFT main Assumption 1. Assumption 1 is that a sufficiently

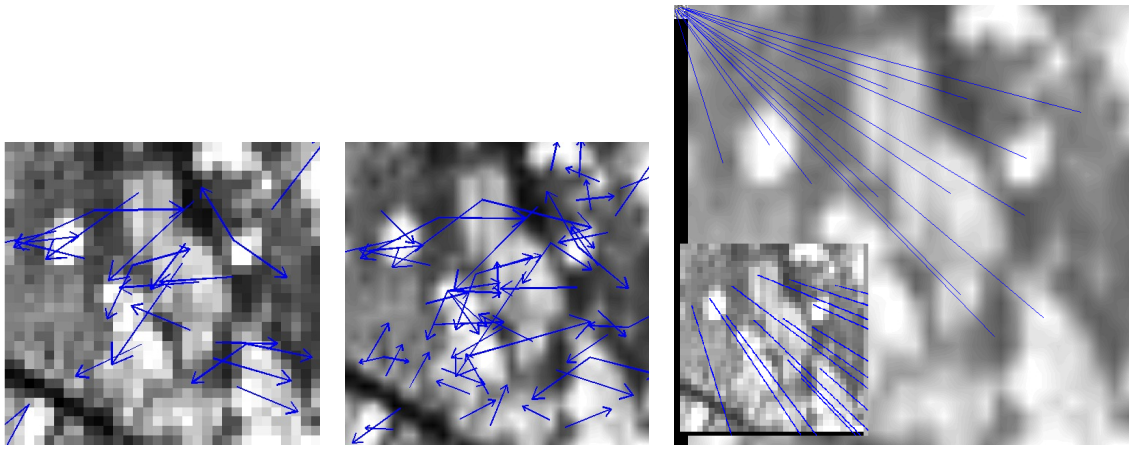


FIGURE 4. Scale invariance of SIFT, an illustration of Theorem 1. Left: a very small digital image $\mathbf{u}$ with its 25 key points. For the conventions to represent key points and matches, see the comments in Fig. 3. Middle: this image is over sampled by a 32 factor to $\mathbf{S}_{\frac{1}{32}}\mathbf{I}_d\mathbf{u}$. It has 60 key points. Right: 18 matches found between $\mathbf{u}$ and $\mathbf{S}_{\frac{1}{32}}\mathbf{I}_d\mathbf{u}$. A zoom of the small image $\mathbf{u}$ on the up-left corner is shown in the bottom left. It can be observed that all the matches are correct.

broad Gaussian blur may lead to well-sampled images. In principle, a Gaussian blur cannot lead to well-sampled images, because it is not *stricto sensu* band limited. Thus, after Gaussian blur, the image is no more well sampled in the sense defined above. However, it will be shown that Assumption 1 is approximatively true. Section 4.2 defines an experimental procedure which checks that a Gaussian blur works. This procedure will fix the minimum standard deviation of the Gaussian blur ensuring well-sampling, up to a minor error. This error will be formally computed and shown to be very small in Section 4.3.

4.1. SHANNON INTERPOLATION VS DFT INTERPOLATION. The Shannon interpolation, assuming infinitely many samples, is impractical. The image processing interpolation is slightly different. A digital image is usually given by its samples on a rectangle (for simplicity we take it to be a square $\{0,\dots N-1\}^2$). These samples are (implicitly) extended by N-periodicity into a periodic image on $\mathbf{R^2}$, and it is assumed that the resulting underlying image is band-limited with spectrum in $[-\pi,\pi]^2$. Needless to say, the periodicity assumption is enforced and contradicts in some extent the band-limited assumption. Even if the samples came from a really band-limited image, the loss of the samples outside the sampling square is irreversible. Thus, the band limited image and periodic image which is finally interpolated from the samples is different from the Shannon interpolate, but easier to compute. Indeed, given a digital image $\mathbf{u}(\mathbf{k})$ with $\mathbf{k}=(k,l)\in\{0,\dots,N-1\}^2$ its band-limited N-periodic interpolate is nothing but the trigonometric polynomial

$$\mathbf{I}_d u(\mathbf{x}) = \sum_{\mathbf{m}=(m,n)\in[-N/2,N/2-1]^2} \tilde{u}_{\mathbf{m}} e^{\frac{2i\pi\mathbf{m}.\mathbf{x}}{N}}, \tag{11}$$

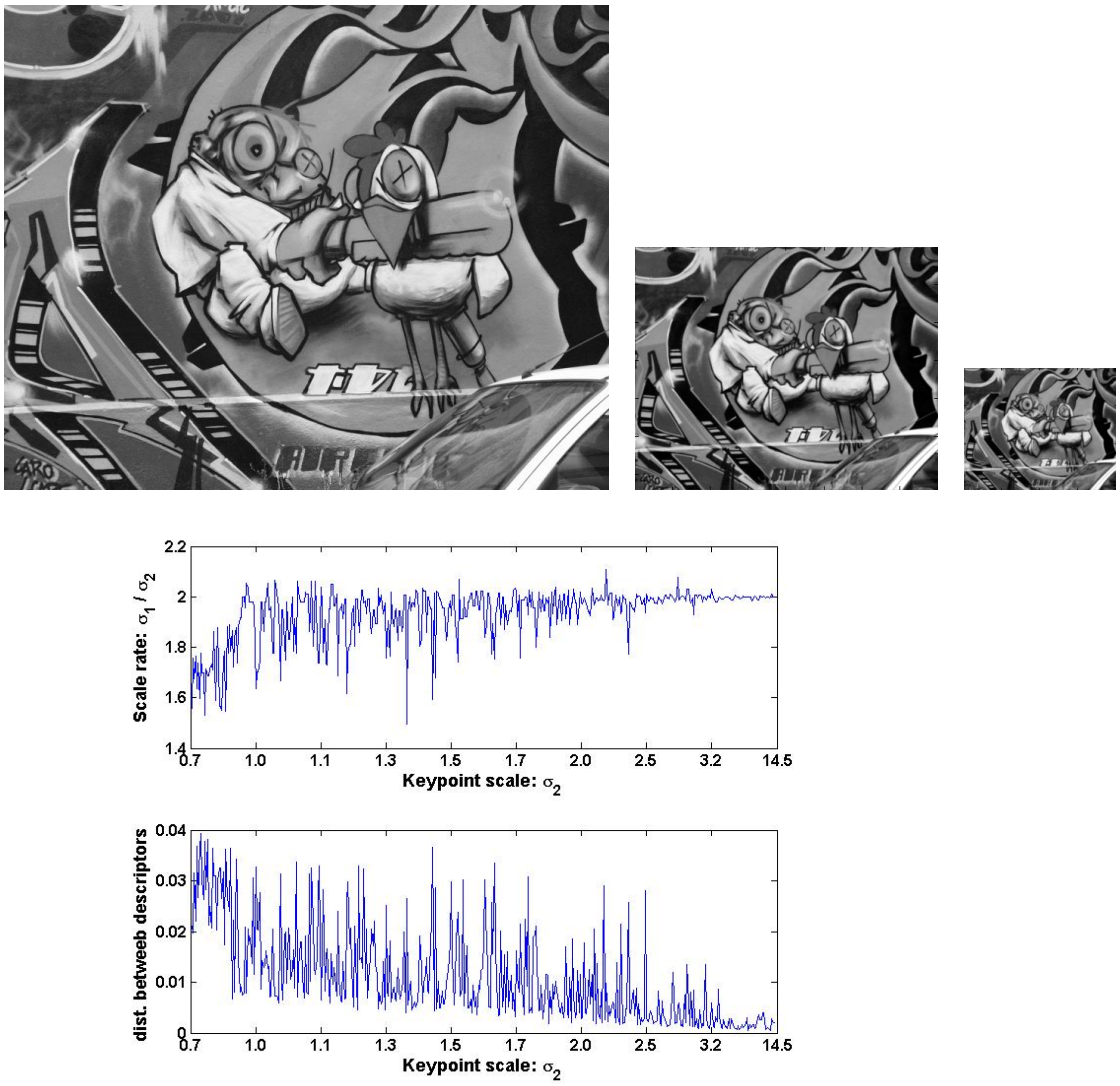
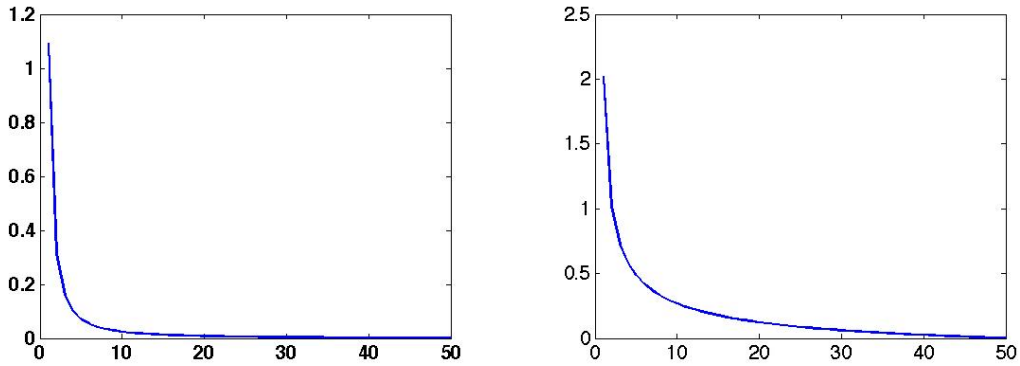


FIGURE 5. Top (from left to right): $\mathbf{u}_0$, $\mathbf{u}$, $\mathbf{v}$. Middle: Rate of scales σ_1/σ_2 of matched key points in $\mathbf{u}$ and $\mathbf{v}$ against σ_2. Bottom: Distance between matched descriptors of $\mathbf{u}$ and $\mathbf{v}$ against σ_2.

where $\tilde{u}_{\mathbf{m}}$ with $\mathbf{m} = (m, n)$ are the Discrete Fourier Transform (DFT) coefficients of the N^2 samples $u(\mathbf{k})$.

It is not the object of this paper to investigate in depth the interpolation error caused by ignoring the samples outside the image domain. But, still, a fast numerical experiment was made to get some hints about this *DFT interpolation error*. The classic digital test image Lena was inserted in a larger digital image $\mathbf{u}_z$ by surrounding its samples with a frame of black (zero) samples. The width of the



(a) **(b)**

FIGURE 6. **(a).** The root mean square error $\mathrm{RMS}(\mathbf{v}_z, \mathbf{v}_{z+1})$ between successive interpolation. **(b).** The root mean square error $\mathrm{RMS}(\mathbf{v}_z, \mathbf{v}_{50})$ between the current and the last interpolations.

frame evolved from $z = 0$ to $z = 50$. For each image $\mathbf{u}_z$, a 2×2 zoom-in by zero-padding [1] was performed. Let us call $\mathbf{v}_z$ the digital image which is the restriction of the zoomed in image to the new samples (excluding the known samples, and also all samples in the added frame).

To check the impact of the exterior samples on the interpolation inside, the root mean square error $\mathrm{RMS}(\mathbf{v}_z, \mathbf{v}_{z+1})$ was computed. As plotted in Fig. 6-(a), $\mathrm{RMS}(\mathbf{v}_z, \mathbf{v}_{z+1})$ converges to zero when the width of the black frame added outside the image increases. Assuming therefore $\mathbf{v}_{50}$ to be very close to the final Shannon interpolate (which fixes all exterior samples to zero), Fig. 6-(b) shows the root mean square error $\mathrm{RMS}(\mathbf{v}_z, \mathbf{v}_{50})$ as a function of z. The initial error (about 2 for an image ranging from 0 to 255) is not negligible at all. However, for $z = 20$ this error becomes negligible. This simple experiment (which should be completed by a theoretical study in the spirit of Section 4) indicates that *the perturbations of image interpolation due to the ignorance of exterior samples only affect significantly the image samples near the boundary. The samples at distance larger than 20 to the image boundary can be indifferently computed by DFT or Shannon interpolation.* This justifies our dual use of the interpolation operators $\mathbf{I}$ and $\mathbf{I}_d$: They accurately coincide away from the image boundary.

4.2. EXPERIMENTAL VERIFICATION OF ASSUMPTION 1. To understand the right Gaussian blur to achieve a well-sampled image, we shall distinguish two types of blur. The *absolute* blur with standard deviation $\mathbf{c}_a$ is the one that must be applied to an ideal infinite resolution (blur free) image to create an approximately band-limited image before 1-sampling. The *relative* blur $\sigma = \mathbf{c}_r(t)$ is the one that must be applied to a well-sampled image before a sub-sampling by a factor of t. In the

[1] The zero-padding permits to compute the values on a finer grid of the digital image with (11). It proceeds as follows: a) apply the $N \times N$ DFT to the samples $\mathbf{u}(k,l)$, $(k,l) \in \{0,\dots,N-1\}^2$ to get the discrete Fourier coefficients, $\tilde{u}_{\mathbf{m}}$, $\mathbf{m} \in \{-N/2, N/2-1\}^2$. b) complete these DFT coefficients to get a matrix $\tilde{u}_{\mathbf{m}}$, $\mathbf{m} \in \{-N, N-1\}^2$, the new DFT coefficients being zeros; c) Apply the inverse $2N \times 2N$ DFT to get $4N^2$ samples $v_{\mathbf{k}}$, $\mathbf{k} \in \{0,\dots,2N\}^2$. Then $v_{2\mathbf{k}} = u_{\mathbf{k}}$ for $\mathbf{k} \in \{0,\dots,N-1\}^2$.

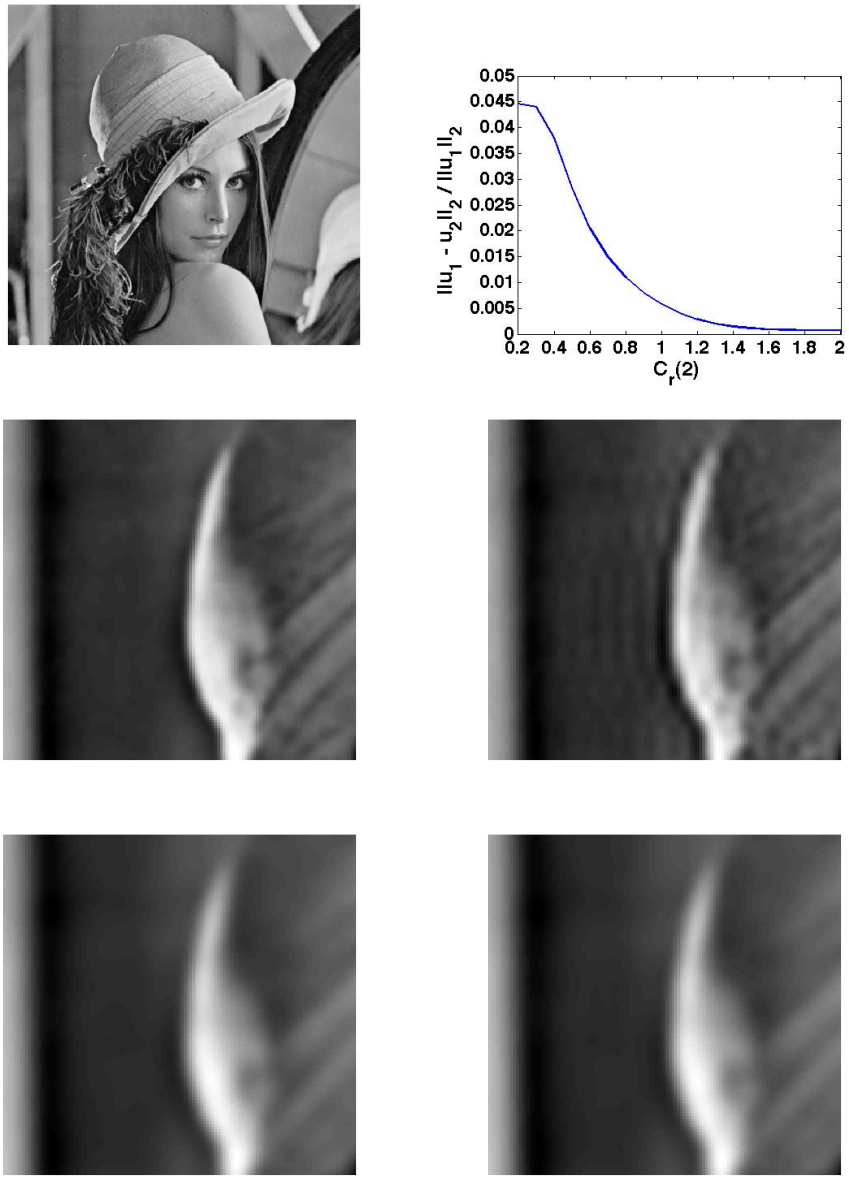


FIGURE 7. Top left: the classic test image $\mathbf{u}$. Top right: relative error $\epsilon(\mathbf{u}_1, \mathbf{u}_2) = \|\mathbf{u}_1 - \mathbf{u}_2\|_2 / \|\mathbf{u}_1\|_2$ vs $\mathbf{c}_r(4)$, the error introduced by a 2×2 sampling after a Gaussian blur with $\sigma = \mathbf{c}_r(2)$. Middle (from left to right): $\mathbf{u}_1$ and $\mathbf{u}_2$ (zoomed) with $\mathbf{c}_r(2) = 0.8$, $\epsilon(\mathbf{u}_1, \mathbf{u}_2) = 1.1e-02$. The aliasing error is large and conspicuous ringing artifacts appear on the right image (refer to the electronic copy for better visibility). Bottom (from left to right): $\mathbf{u}_1$ and $\mathbf{u}_2$ (zoomed) with $\mathbf{c}_r(2) = 1.6$. $\epsilon(\mathbf{u}_1, \mathbf{u}_2) = 9.2e-04$. These images are almost identical, which confirms that a Gaussian blur with standard deviation 1.6 is sufficient before 2×2-sub-sampling.

case of Gaussian blur, because of the semi-group formula (4), the relation between the absolute and relative blur is

$$t^2\mathbf{c}_a^2 = \mathbf{c}_r^2(t) + \mathbf{c}_a^2,$$

which yields

$$\mathbf{c}_r(t) = \mathbf{c}_a\sqrt{t^2-1}. \tag{12}$$

In consequence, if $t \gg 1$, then $\mathbf{c}_r(t) \approx \mathbf{c}_a t$.

Two experiments have been designed to calculate the anti-aliasing absolute Gaussian blur $\mathbf{c}_a$ ensuring that an image is approximately well-sampled. The first experiment finds the appropriate relative blur $\mathbf{c}_r(t)$ to achieve well-sampled images and then calculates the absolute blur $\mathbf{c}_a$ using (12). $\mathbf{c}_r(t)$. It compares for several values of $\mathbf{c}_r(t)$ the digital images

$$\mathbf{u}_1 := \mathbf{G}_{\mathbf{c}_r(t)}\mathbf{u} = \mathbf{S}_1 G_{\mathbf{c}_r(t)}\mathbf{I}_d\mathbf{u} \quad \text{and} \quad \mathbf{u}_2 := (\mathbf{S}_{1/t}\mathbf{I}_d)\mathbf{S}_t G_{\mathbf{c}_r(t)}\mathbf{I}_d\mathbf{u},$$

where $\mathbf{u}$ is an initial digital image that is well-sampled, $\mathbf{S}_t$ is a t-sub-sampling operator (t is an integer), $\mathbf{S}_{1/t}$ a t-over-sampling operator, and $\mathbf{I}_d$ the DFT interpolation operator. The discrete convolution by a Gaussian is defined in (6). Since t is an integer, the t-sub-sampling is trivial. The DFT over-sampling $\mathbf{S}_{1/t}\mathbf{I}_d$ with an integer zoom factor t is computed by zero-padding as explained in footnote 1. If the anti-aliasing filter size $\mathbf{c}_r(t)$ is too small, $\mathbf{u}_1$ and $\mathbf{u}_2$ can be very different. The right value of $\mathbf{c}_r(t)$ should be the smallest value permitting that, systematically, $\mathbf{u}_1 \approx \mathbf{u}_2$.

Fig. 7 shows $\mathbf{u}_1$ and $\mathbf{u}_2$ with $t = 2$ and plots their relative error $\epsilon(\mathbf{u}_1, \mathbf{u}_2) = \|\mathbf{u}_1 - \mathbf{u}_2\|_2/\|\mathbf{u}_1\|_2$. An anti-aliasing filter with $\mathbf{c}_r(2) = 0.8$ is clearly not broad enough: $\mathbf{u}_2$ presents strong ringing artifacts. The ringing artifact is instead hardly noticeable with $\mathbf{c}_r(2) = 1.6$. The value $\mathbf{c}_r(2) \approx 1.6$ is a good visual candidate, and this choice is confirmed by the curve showing that $\epsilon(\mathbf{u}_1, \mathbf{u}_2)$ decays rapidly until $\mathbf{c}_r(2)$ gets close to 1.6, and is stable and small thereafter. By (12), this value of $\mathbf{c}_r$ yields $\mathbf{c}_a \approx 0.8$. This value has been confirmed by experiments on ten digital images. A doubt can be cast on this experiment, however: Its result slightly depends on the assumption that the initial blur on $\mathbf{u}$ is equal to $\mathbf{c}_a$.

In a second experiment, $\mathbf{c}_a$ has been evaluated directly by using a binary image $\mathbf{u}_0$, thus containing a minimal blur. As illustrated in Fig. 8, $\mathbf{u}_0$ is obtained by binarizing the classic test image Lena (Fig. 7), the threshold being the image median value. Since $\mathbf{u}_0$ is now close to be blur-free, we can compare for several values of $\mathbf{c}_a$ and for $t = 2$, which is large enough, the digital images

$$\mathbf{u}_1 := \mathbf{G}_{t\mathbf{c}_a}\mathbf{u} = \mathbf{S}_1 G_{t\mathbf{c}_a}\mathbf{I}_d\mathbf{u} \quad \text{and} \quad \mathbf{u}_2 := (\mathbf{S}_{1/t}\mathbf{I}_d)\mathbf{S}_t G_{t\mathbf{c}_a}\mathbf{I}_d\mathbf{u},$$

As shown in Fig. 8, $\mathbf{c}_a = 0.8$ is the smallest value ensuring no visual ringing in $\mathbf{u}_2$. Under this value, for example for $\mathbf{c}_a = 0.4$, clear ringing artifacts are present in $\mathbf{u}_2$. That $\mathbf{c}_a = 0.8$ is the correct value is confirmed by the $\epsilon(\mathbf{u}_1, \mathbf{u}_2)$ curve showing that the relative error decays rapidly until $\mathbf{c}_a$ goes down to 0.8, and is stable and small thereafter. The result, confirmed in ten experiments with different initial images, is consistent with the value obtained in the first experimental setting.

Fig. 9 illustrates the same experiment on a Gaussian white noise image that has constant spectrum energy over all frequencies. This is not at all an unrealistic image, since textured images can have parts with a flat spectrum which is not so much different from white noise. The critical value $\mathbf{c}_a = 0.8$ is reconfirmed. Under this value, for example for $\mathbf{c}_a = 0.4$, u_2 is visually very different from u_1, while the two very close (numerically and visually as well) with $\mathbf{c}_a = 0.8$. The relative error

for a white noise, about 4% cannot be considered negligible. But it is a worst case, no digital image having such a flat spectrum. The experimental value of this error will be retrieved by a formal calculation, and shown to be independent of the image size.

4.3. Mathematical confirmation of Assumption 1. This section confirms by exact computations the experimental estimates made on simulated Gaussian white noise. The spectrum of natural digital images has a spectrum usually decaying much faster than a white noise. Thus, the aliasing error can be safely estimated on white noise, which has a flat spectrum. Sampling operations shown to be safe for white noise will therefore also be safe for natural images. We will first calculate the aliasing error in the general case, and then show that after an anti-aliasing Gaussian filtering with a relative blur $\sigma = 2 \times 0.8 = 1.6$, the aliasing error created by a 2×2 subsampling on a Gaussian white noise is negligible.

Consider a digital image $\mathbf{u}(\mathbf{k})$ with $\mathbf{k} = (k, l) \in \{0, \dots, N-1\}^2$ and its DFT interpolate

$$u(\mathbf{x}) = (\mathbf{I}_d \mathbf{u})(\mathbf{x}) = \sum_{\mathbf{m}=(m,n)\in[-N/2,N/2-1]^2} \tilde{u}_{\mathbf{m}} e^{\frac{2i\pi \mathbf{m}.\mathbf{x}}{N}}, \quad \mathbf{x} \in \mathbf{R^2}, \tag{13}$$

where $\tilde{u}_{\mathbf{m}}$ with $\mathbf{m} = (m, n)$ are the Discrete Fourier Transform (DFT) coefficients of the N^2 samples $u(\mathbf{k})$, and the Fourier Series coefficients of u as well. Recall the classic isometries between samples, DFT coefficients, and the L^2 norm of the image:

$$||u||^2_{L^2([0,N]^2)} = ||\mathbf{u}||^2_{l^2(\{0,\dots,N-1\}^2)} = N^2 ||\tilde{u}||^2_{l^2(\{-N/2,\dots,N/2-1\}^2)}. \tag{14}$$

The scale space convolves u with a Gaussian $G_\sigma(\mathbf{x}) = \frac{1}{2\pi\sigma^2} e^{-\frac{\mathbf{x}^2}{2\sigma^2}}$. By computing its DFT, an easy calculation shows that the result of the convolution of u with G_σ is

$$v(\mathbf{x}) := (G_\sigma * u)(\mathbf{x}) = \sum_{\mathbf{m}\in[-\frac{N}{2},\frac{N}{2}-1]^2} \tilde{u}_{\mathbf{m}} \hat{G}_\sigma \left(\frac{2\mathbf{m}\pi}{N}\right) e^{\frac{2i\pi \mathbf{m}.\mathbf{x}}{N}} \tag{15}$$

with $\hat{G}_\sigma(\xi) = e^{\frac{-\sigma^2\xi^2}{2}}$. Indeed,

$$(G_\sigma * e^{\frac{2i\pi \mathbf{m}.\mathbf{x}}{N}})(\mathbf{x}) = \int_{\mathbb{R}^2} G_\sigma(\mathbf{y}) e^{\frac{2i\pi \mathbf{m}.(\mathbf{x}-\mathbf{y})}{N}} d\mathbf{y} = e^{\frac{2i\pi \mathbf{m}.\mathbf{x}}{N}} \int_{\mathbb{R}^2} G_\sigma(\mathbf{y}) e^{-\frac{2i\pi \mathbf{m}.\mathbf{y}}{N}} d\mathbf{y} = e^{\frac{2i\pi \mathbf{m}.\mathbf{x}}{N}} \hat{G}_\sigma \left(\frac{2\pi\mathbf{m}}{N}\right).$$

The convolved digital image is

$$v(\mathbf{k}) = \sum_{\mathbf{m}\in[-\frac{N}{2},\frac{N}{2}-1]^2} \tilde{v}_{\mathbf{m}} e^{\frac{2i\pi \mathbf{m}.\mathbf{k}}{N}}. \tag{16}$$

Subsampling this image by a factor of 2×2 and interpolating the remaining $N^2/4$ samples with zero-padding interpolation yields a new trigonometric polynomial

$$w(\mathbf{x}) = \sum_{\mathbf{m}\in[-N/4,N/4-1]^2} \tilde{w}_{\mathbf{m}} e^{\frac{2i\pi \mathbf{m}.\mathbf{x}}{N}}. \tag{17}$$

Let us compute $\tilde{w}_{\mathbf{m}}$.

Lemma 4. *Let $\mathbf{w} = \mathbf{S}_2\mathbf{v}$ be the $\frac{N}{2} \times \frac{N}{2}$ image obtained by a 2×2 subsampling of a digital $N \times N$ image $\mathbf{v}$. Then $\tilde{w}_{m,n}$, the DFT of $\mathbf{w}$, satisfies for $m, n = -\frac{N}{4}, ..., \frac{N}{4}-1$,*

$$\tilde{w}_{m,n} = \sum_{(\varepsilon_1,\varepsilon_2)\in\{0,1,-1\}} \tilde{v}_{m+\varepsilon_1\frac{N}{2},n+\varepsilon_2\frac{N}{2}}. \tag{18}$$

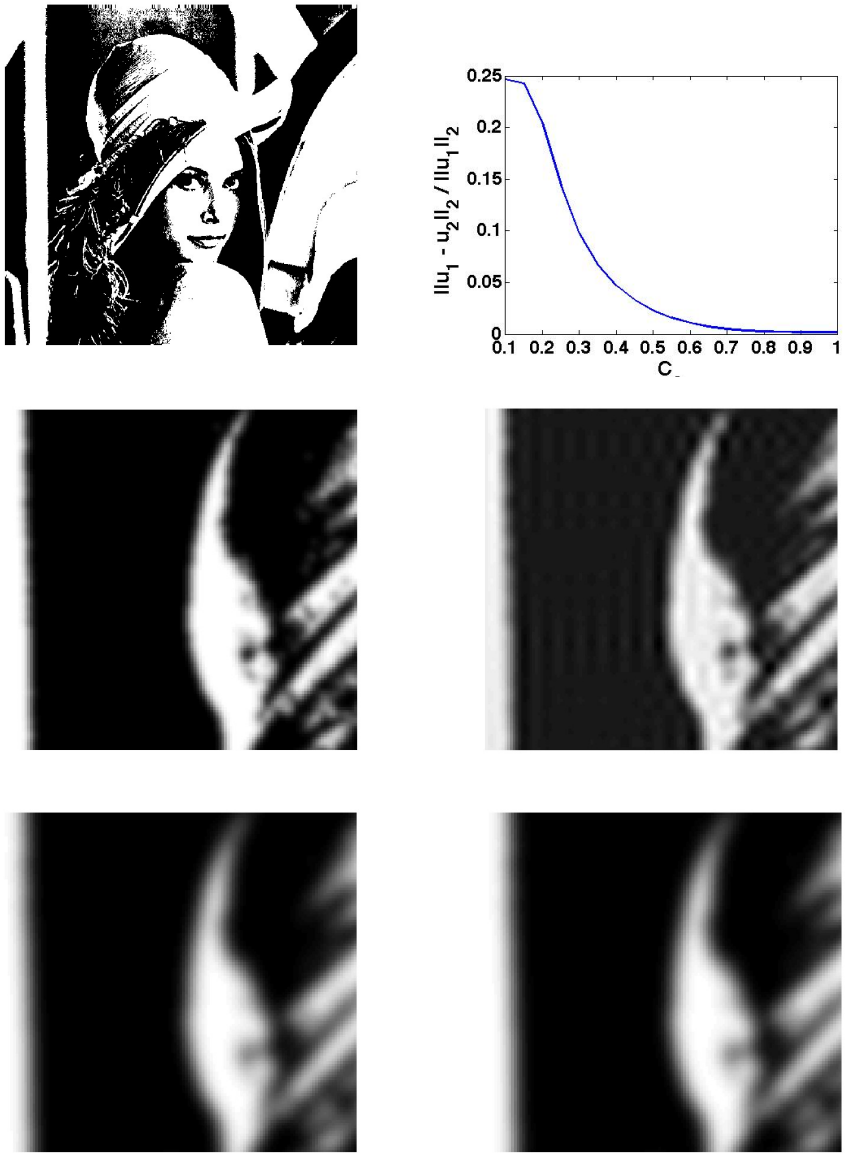


FIGURE 8. Top left: $\mathbf{u}$ Binarized Lena (gray-levels 50 and 0). Top right: $\epsilon(\mathbf{u}_1, \mathbf{u}_2) = \|\mathbf{u}_1 - \mathbf{u}_2\|_2 / \|\mathbf{u}_1\|_2$ vs $\mathbf{c}_a$, the error introduced by a 2×2 sampling after a Gaussian blur with $\sigma = 2\mathbf{c}_a$. Middle (from left to right): $\mathbf{u}_1$ and $\mathbf{u}_2$ (zoomed) with $\mathbf{c}_a = 0.4$. $\epsilon(\mathbf{u}_1, \mathbf{u}_2) = 4.7e-02$. The aliasing error is large and conspicuous ringing artifacts appear on the right image (refer to the electronic copy for better visibility). Bottom (from left to right): $\mathbf{u}_1$ and $\mathbf{u}_2$ (zoomed) with $\mathbf{c}_a = 0.8$. $\epsilon(\mathbf{u}_1, \mathbf{u}_2) = 2.5e-03$. These last two images are almost identical, while strong aliasing artifacts are visible with $\mathbf{c}_a = 0.4$.

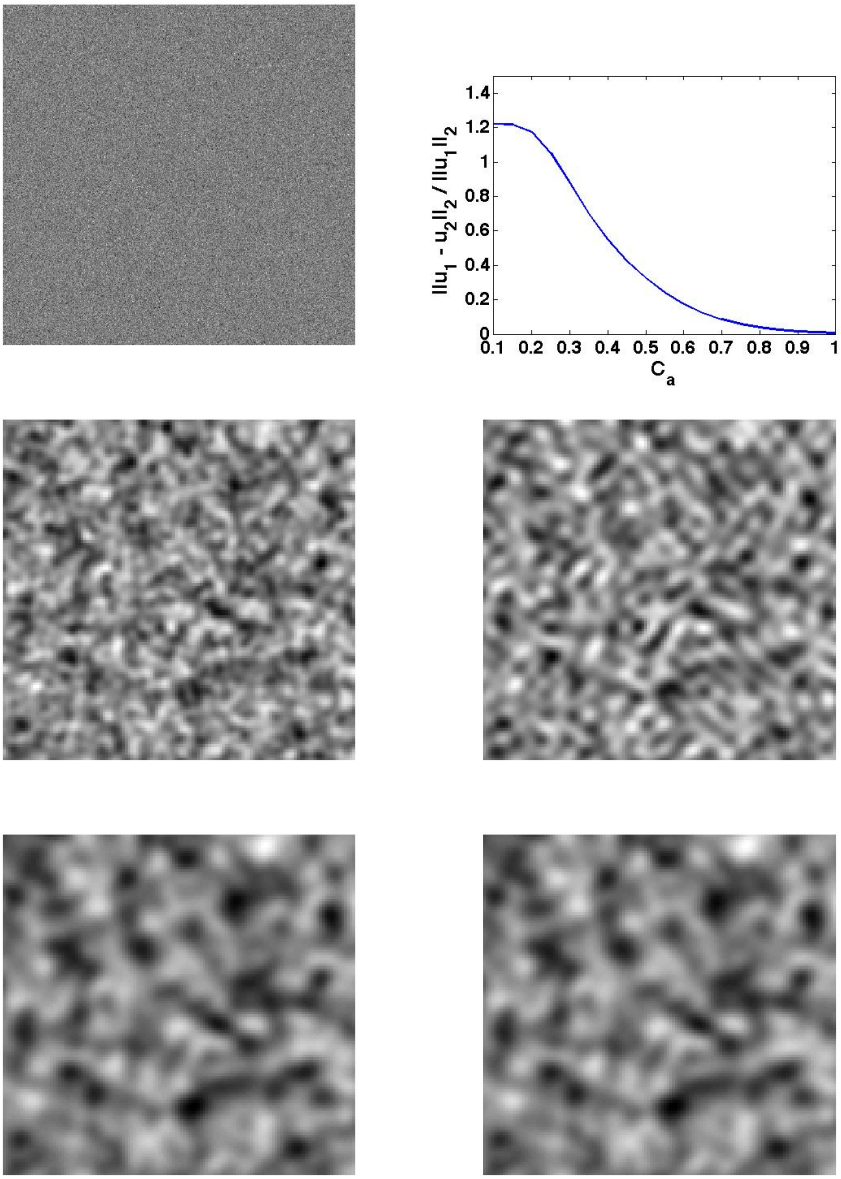


FIGURE 9. Top left: $\mathbf{u}$ Gaussian white noise. Top right: $\epsilon(\mathbf{u}_1, \mathbf{u}_2) = \|\mathbf{u}_1 - \mathbf{u}_2\|_2 / \|\mathbf{u}_1\|_2$ vs $\mathbf{c}_a$, the error introduced by a 2×2 sampling after a Gaussian blur with $\sigma = 2\mathbf{c}_a$. Middle (from left to right): $\mathbf{u}_1$ and $\mathbf{u}_2$ (zoomed) with $\mathbf{c}_a = 0.4$. $\epsilon(\mathbf{u}_1, \mathbf{u}_2) = 0.549$ and the images are also visually very dissimilar. Bottom (from left to right): $\mathbf{u}_1$ and $\mathbf{u}_2$ (zoomed) with $\mathbf{c}_a = 0.8$. $\epsilon(\mathbf{u}_1, \mathbf{u}_2) = 0.0389$, and the images are visually identical.

Notice that $\tilde{v}_{m,n}$ is supported by $[-\frac{N}{2}, \frac{N}{2}-1]^2$ and that $\tilde{w}_{m,n}$ is supported by $[-\frac{N}{4}, \frac{N}{4}-1]^2$. For each (m,n), only at most four coefficients out of the nine $v_{m+\varepsilon_1 \frac{N}{2}, n+\varepsilon_2 \frac{N}{2}}$ are non-zero, because if for example $m,n \leq 0$, $\tilde{v}_{m-\frac{N}{2},n}$, $\tilde{v}_{m-\frac{N}{2},n-\frac{N}{2}}$, $\tilde{v}_{m,n-\frac{N}{2}}$, $\tilde{v}_{m+\frac{N}{2},n-\frac{N}{2}}$, and $\tilde{v}_{m-\frac{N}{2},n+\frac{N}{2}}$ are all null. To make apparent the four non-zero coefficients, we take the convention to $N \times N$ periodize $(m,n) \to \tilde{v}_{m,n}$. With this convention, (18) rewrites

$$(19)\ \ \forall (m,n) \in \left[-\frac{N}{4}, \frac{N}{4}-1\right]^2,\ \tilde{w}_{m,n} = \tilde{v}_{m,n} + \tilde{v}_{m+\frac{N}{2},n+\frac{N}{2}} + \tilde{v}_{m+\frac{N}{2},n} + \tilde{v}_{m,n+\frac{N}{2}}.$$

The digital image $\mathbf{w}$ is twice smaller than $\mathbf{v}$. To compare it with $\mathbf{v}$ we must go back to the underlying trigonometric polynomials $v(\mathbf{x})$ and $w(\mathbf{x})$ such that $\mathbf{w} = \mathbf{S}_2 w$ and $\mathbf{v} = \mathbf{S}_1 v$, and we must compare their corresponding samples. Thus set $\mathbf{w}^1 = \mathbf{S}_1 w$. The image $\mathbf{w}^1$ is obtained by zero padding from $\tilde{\mathbf{w}}$ by setting $\tilde{w}^1_{m,n} = 0$ if $(m,n) \in \{-\frac{N}{2}, ..., \frac{N}{2}-1\}^2 \backslash \{-\frac{N}{4}, \dots, \frac{N}{4}-1\}^2$ and $\tilde{w}^1_{m,n} = \tilde{w}_{m,n}$ if $(m,n) \in \{-\frac{N}{4}, \dots, \frac{N}{4}-1\}^2$. Then using (16), (17) and (14), we have

(20)
$$||v-w||^2_{L^2([0,N]^2)} = ||\mathbf{v}-\mathbf{w}_1||^2_{l^2(\{0,\dots,N-1\}^2)} = N^2 ||\tilde{v}-\tilde{w}||^2_{l^2(\{-\frac{N}{2},\dots,\frac{N}{2}-1\})} = e_h(v)+e_a(v).$$

Using the above N-periodicity convention for the Fourier coefficients $\tilde{v}(m,n)$, we have proved the following proposition.

Proposition 1. *Let u be a digital image and $v := G_\sigma * u$ be an image obtained after a Gaussian convolution with standard deviation σ. Then the variance of the error incurred in doing a 2×2-sub-sampling followed by a zero-padding interpolation is*

$$||v-w||_2^2 = e_h(v) + e_a(v), \tag{21}$$

where

$$e_h(v) = N^2 \left[\sum_{(m,n) \in [-N/2, N/2-1]^2 \backslash [-N/4, N/4-1]^2} |\tilde{v}(m,n)|^2 \right] \tag{22}$$

$$= N^2 \left[\sum_{(m,n) \in [-N/4, N/4-1]^2} \left(|\tilde{v}(m, n+N/2)|^2 + |\tilde{v}(m+N/2, n)|^2 + |\tilde{v}(m+N/2, n+N/2)|^2 \right) \right] \tag{23}$$

is the error due to the elimination of the high frequencies and

(24)
$$e_a(v) = N^2 \left[\sum_{(m,n) \in [-N/4, N/4-1]^2} |\tilde{v}(m, n+N/2) + \tilde{v}(m+N/2, n) + \tilde{v}(m+N/2, n+N/2)|^2 \right]$$

is the error introduced by the spectrum aliasing, with

$$\tilde{v}(\mathbf{m}) = \tilde{u}(\mathbf{m}) \hat{G}_\sigma \left(\frac{2\mathbf{m}\pi}{N} \right). \tag{25}$$

Following Shannon [41] we call white noise image with variance τ^2 a continuous image

$$v(\mathbf{x}) = \sum_{\mathbf{m} \in [-N/2, N/2-1]^2} \tilde{v}_{\mathbf{m}} e^{\frac{2i\mathbf{m}\mathbf{x}}{N}}$$

whose samples $v(k,l)$, $(k,l) \in \{0, \dots, N-1\}^2$ are i.i.d. $\mathcal{N}(0, \tau^2)$ Gaussian variables. Using Proposition 1, the next Corollary shows that with a Gaussian anti-aliasing filtering with $\sigma = 2 \times 0.8 = 1.6$, the aliasing error created by a 2×2 subsampling on a Gaussian white noise image is very small.

Corollary 1. *Let u be a Gaussian white noise image and $v := u * G_\sigma$ be an image obtained after a Gaussian convolution with standard deviation σ. Then the relative error incurred in doing a 2×2 subsampling followed by a zero-padding 2×2 interpolation is*

$$\epsilon(\sigma) := \frac{\sqrt{\mathbf{E}||v-w||^2}}{\sqrt{\mathbf{E}||v||^2}} = \sqrt{2\frac{\sum_{\mathbf{m}\in[-N/2,N/2-1]^2\setminus[-N/4,N/4-1]^2} e^{-4\pi^2\sigma^2\frac{|\mathbf{m}|^2}{N^2}}}{\sum_{\mathbf{m}\in[-N/2,N/2-1]^2} e^{-4\pi^2\sigma^2\frac{|\mathbf{m}|^2}{N^2}}}}; \tag{26}$$

$$\epsilon(\sigma) \approx \sqrt{2\frac{\int_{[-\frac{1}{2},\frac{1}{2}]^2\setminus[-\frac{1}{4},\frac{1}{4}]^2} e^{-4\pi^2\sigma^2\mathbf{x}^2}d\mathbf{x}}{\int_{[-\frac{1}{2},\frac{1}{2}]^2} e^{-4\pi^2\sigma^2\mathbf{x}^2}d\mathbf{x}}}. \tag{27}$$

In particular, $\epsilon(1.6)\approx 0.0389$.

The aim of (27) is to point out that for white noise, the relative error incurred by a Gaussian sub-sampling is independent from the size N of the image.

Proof. The DFT coefficients $\tilde{u}(\mathbf{m})$ of a Gaussian white noise image u with standard deviation τ are independent Gaussian variables with standard deviation $\frac{\tau}{N}$. Using (21) and (25), we therefore have

$\mathbf{E}||v-w||^2 = \mathbf{E}e_h(v) + \mathbf{E}e_a(v)$ and

$$\mathbf{E}e_a(v) = N^2 \sum_{(m,n)\in\{-N/4,N/4-1\}^2} \mathbf{E}\left|\tilde{v}(m,n+\frac{N}{2}) + \tilde{v}(m+\frac{N}{2},n) + \tilde{v}(m+\frac{N}{2},n+\frac{N}{2})\right|^2$$

$$= N^2 \sum_{(m,n)\in\{-N/4,N/4-1\}^2} \tau^2\left(\hat{G}_\sigma\left(\frac{2\pi}{N}(m,n+\frac{N}{2})\right)^2 + \hat{G}_\sigma\left(\frac{2\pi}{N}(m+\frac{N}{2},n)\right)^2 + \hat{G}_\sigma\left(\frac{2\pi}{N}(m+\frac{N}{2},n+\frac{N}{2})\right)^2\right).$$

Using (22) we deduce that

$$\mathbf{E}||v-w||^2 = \mathbf{E}e_h(v)+\mathbf{E}e_a(v) = 2\mathbf{E}e_a(v) = 2N^2\tau^2 \sum_{\mathbf{m}\in\{-N/2,\dots N/2-1\}^2\setminus\{-N/4,\dots,N/4-1\}^2} \hat{G}_\sigma\left(\frac{2\pi\mathbf{m}}{N}\right)^2$$

$$= 2N^2\tau^2 \sum_{\mathbf{m}\in\{-N/2,\dots N/2-1\}^2\setminus\{-N/4,\dots,N/4-1\}^2} e^{-4\pi^2\sigma^2\frac{|\mathbf{m}|^2}{N^2}}$$

On the other hand,

$$\mathbf{E}||v||^2 = \mathbf{E}\sum_{(k,l)\in\{0,\dots,N-1\}^2} |v(k,l)|^2 = N^2\mathbf{E}||\tilde{v}||^2_{l^2(\{-N/2,\dots,N/2-1\}^2)} =$$

$$= N^2 \sum_{\mathbf{m}\in\{-N/2,\dots,N/2-1\}^2} \mathbf{E}|\tilde{u}_\mathbf{m}|^2\hat{G}_\sigma(\frac{2\mathbf{m}\pi}{N})^2 = N^2\tau^2 \sum_{\mathbf{m}\in\{-N/2,\dots,N/2-1\}^2} e^{-4\pi^2\sigma^2\frac{|\mathbf{m}|^2}{N^2}}.$$

Thus

$$\frac{\sqrt{\mathbf{E}||v-w||^2}}{\sqrt{\mathbf{E}||v||^2}} = \sqrt{2\frac{\sum_{(m,n)\in\{-N/2,\dots N/2-1\}^2\setminus\{-N/4,\dots,N/4-1\}^2} e^{-4\pi^2\sigma^2\frac{|\mathbf{m}|^2}{N^2}}}{\sum_{(m,n)\in\{-N/2,\dots N/2-1\}^2} e^{-4\pi^2\sigma^2\frac{|\mathbf{m}|^2}{N^2}}}}$$

$$\approx \sqrt{2\frac{\int_{[-\frac{1}{2},\frac{1}{2}]^2\setminus[-\frac{1}{4},\frac{1}{4}]^2} e^{-4\pi^2\sigma^2\mathbf{x}^2}d\mathbf{x}}{\int_{[-\frac{1}{2},\frac{1}{2}]^2} e^{-4\pi^2\sigma^2\mathbf{x}^2}d\mathbf{x}}}.$$

The last equivalent is computed by noticing that the discrete sums are Riemann sums for the integrals on the right. For $\sigma = 1.6$, this value is approximately 0.0389. □

The error $\epsilon(1.6) \approx 0.0389$ calculated from (26) was numerically verified with the experiments on Gaussian white noise. It is actually not negligible, but must be considered as a worst case which is never approximated by real images. Indeed, the power spectra of natural images having a much faster decay than Gaussian white noise, the value $\epsilon(1.6)$ is smaller on natural images: For Lena, Peppers, Baboon and Barbara as shown in Figure 10, $\epsilon(1.6)$ is respectively $9.2e-04$, $1.2e-03$, $1.9e-03$, $2.0e-03$.

Lena **Peppers** **Baboon** **Barbara**

FIGURE 10. Standard test images. From left to right: Lena, Peppers, Baboon, and Barbara.

5. Conclusion

Our overall conclusion is that no substantial improvement of the SIFT method can be ever hoped, as far as translation, rotation and scale invariance are concerned. As pointed out by several benchmarks, the robustness and repeatability of the SIFT descriptors outperforms other methods. However, such benchmarks mix three very different criteria that, in our opinion, should have been discussed separately. The first one is the formal invariance of each method when all thresholds have been eliminated. This formal invariance has been proved here for SIFT when the initial blurs of both images are equal to a known value $\mathbf{c}$, and it has been proved to be approximately true even with images having undergone very different blurs, like in the surprising experiment of fig. 5. The second criterion is the practical validity of the sampling method used in SIFT, that has been again checked in the present note. The last criterion is the clever fixing of several thresholds in the SIFT method ensuring robustness, repeatability, and a low false alarm rate. This one has been extensively tested and confirmed in previous benchmark papers (see also the recent and complete report [7]). We think, however, that the success of SIFT in these benchmarks is primarily due to its full scale invariance.

References

[1] A. Agarwala, M. Agrawala, M. Cohen, D. Salesin, and R. Szeliski. Photographing long scenes with multi-viewpoint panoramas. *International Conference on Computer Graphics and Interactive Techniques*, pages 853–861, 2006.

[2] A. Baumberg. Reliable feature matching across widely separated views. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 1:774–781, 2000.

[3] M. Bennewitz, C. Stachniss, W. Burgard, and S. Behnke. Metric localization with scale-Invariant visual features using a single perspective camera. *European Robotics Symposium*, page 195, 2006.

[4] M. Brown and D. Lowe. Recognising panorama. In *Proc. the 9th Int. Conf. Computer Vision, October*, pages 1218–1225, 2003.

[5] E.Y. Chang. EXTENT: fusing context, content, and semantic ontology for photo annotation. *Proceedings of the 2nd International Workshop on Computer Vision Meets Databases*, pages 5–11, 2005.

[6] Q. Fan, K. Barnard, A. Amir, A. Efrat, and M. Lin. Matching slides to presentation videos using SIFT and scene background matching. *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 239–248, 2006.

[7] L. Février. A wide-baseline matching library for Zeno. *Internship report, ENS, Paris, France, www.di.ens.fr/~fevrier/papers/2007-InternsipReportILM.pdf*, 2007.

[8] J.J. Foo and R. Sinha. Pruning SIFT for scalable near-duplicate image matching. *Proceedings of the Eighteenth Conference on Australasian Database*, 63:63–71, 2007.

[9] G. Fritz, C. Seifert, M. Kumar, and L. Paletta. Building detection from mobile imagery using informative SIFT descriptors. *Lecture Notes in Computer Science*, pages 629–638, 2005.

[10] C. Gasquet and P. Witomski. *Fourier Analysis and Applications: Filtering, Numerical Computation, Wavelets*. Springer Verlag, 1999.

[11] I. Gordon and D.G. Lowe. What and where: 3D object recognition with accurate pose. *Lecture Notes in Computer Science*, 4170:67, 2006.

[12] J.S. Hare and P.H. Lewis. Salient regions for query by image content. *Image and Video Retrieval: Third International Conference, CIVR*, pages 317–325, 2004.

[13] C. Harris and M. Stephens. A combined corner and edge detector. *Alvey Vision Conference*, 15:50, 1988.

[14] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *European Conference on Computer Vision*, pages 228–241, 2004.

[15] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2:506–513, 2004.

[16] J. Kim, S.M. Seitz, and M. Agrawala. Video-based document tracking: unifying your physical and electronic desktops. *Proc. of the 17th Annual ACM Symposium on User interface Software and Technology*, 24(27):99–107, 2004.

[17] B.N. Lee, W.Y. Chen, and E.Y. Chang. Fotofiti: web service for photo management. *Proceedings of the 14th Annual ACM International Conference on Multimedia*, pages 485–486, 2006.

[18] H. Lejsek, F.H. Ásmundsson, B.T. Jónsson, and L. Amsaleg. Scalability of local image descriptors: a comparative study. *Proceedings of the 14th Annual ACM International Conference on Multimedia*, pages 589–598, 2006.

[19] T. Lindeberg. Scale-space theory: a basic tool for analyzing structures at different scales. *Journal of Applied Statistics*, 21(1):225–270, 1994.

[20] T. Lindeberg and J. Garding. Shape-adapted smoothing in estimation of 3-D depth cues from affine distortions of local 2-D brightness structure. *Proc. ECCV*, pages 389–400, 1994.

[21] D.G Lowe. Distinctive image features from scale-invariant key points. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[22] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.

[23] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. *Proc. ICCV*, 1:525–531, 2001.

[24] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. *Proc. ECCV*, 1:128–142, 2002.

[25] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 257–263, June 2003.

[26] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[27] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. PAMI*, pages 1615–1630, 2005.

[28] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L.V. Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1):43–72, 2005.

[29] P. Monasse. Contrast invariant image registration. *Proc. of the International Conf. on Acoustics, Speech and Signal Processing, Phoenix, Arizona*, 6:3221–3224, 1999.

[30] P. Moreels and P. Perona. Common-frame model for object recognition. *Neural Information Processing Systems*, pages 953–960, 2004.

[31] J.M. Morel and G. Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2):438–469, 2009.

[32] A. Murarka, J. Modayil, and B. Kuipers. Building local safety maps for a wheelchair robot using vision and lasers. In *Proceedings of the The 3rd Canadian Conference on Computer and Robot Vision*. IEEE Computer Society Washington, DC, USA, 2006.

[33] P. Musé, F. Sur, F. Cao, and Y. Gousseau. Unsupervised thresholds for shape matching. *Proc. of the International Conference on Image Processing*, 2:647–650, 2003.

[34] P. Musé, F. Sur, F. Cao, Y. Gousseau, and J.M. Morel. An a contrario decision method for shape element recognition. *International Journal of Computer Vision*, 69(3):295–315, 2006.

[35] A. Negre, H. Tran, N. Gourier, D. Hall, A. Lux, and JL Crowley. Comparative study of people detection in surveillance scenes. *Structural, Syntactic and Statistical Pattern Recognition, Proceedings Lecture Notes in Computer Science*, 4109:100–108, 2006.

[36] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2161–2168, 2006.

[37] J. Rabin, Y. Gousseau, and J. Delon. A statistical approach to the matching of local features. *SIAM Journal on Imaging Sciences*, 2(3):931–958, 2009.

[38] F. Riggi, M. Toews, and T. Arbel. Fundamental matrix estimation via TIP-transfer of invariant parameters. *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)-Volume 02*, pages 21–24, 2006.

[39] J. Ruiz-del Solar, P. Loncomilla, and C. Devia. A new approach for fingerprint verification based on wide baseline matching using local interest points and descriptors. *Lecture Notes in Computer Science*, 4872:586–599, 2007.

[40] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional SIFT descriptor and its application to action recognition. *Proceedings of the 15th International Conference on Multimedia*, pages 357–360, 2007.

[41] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:623–656, 1948.

[42] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affine invariant regions. *International Journal of Computer Vision*, 59(1):61–85, 2004.

[43] L. Vacchetti, V. Lepetit, and P. Fua. Stable real-time 3D tracking using online and offline information. *IEEE Trans PAMI*, pages 1385–1391, 2004.

[44] M. Veloso, F. von Hundelshausen, and PE Rybski. Learning visual object definitions by observing human activities. In *Proc. of the IEEE-RAS Int. Conf. on Humanoid Robots,*, pages 148–153, 2005.

[45] M. Vergauwen and L. Van Gool. Web-based 3D reconstruction service. *Machine Vision and Applications*, 17(6):411–426, 2005.

[46] K. Yanai. Image collector III: a web image-gathering system with bag-of-keypoints. *Proc. of the 16th Int. Conf. on World Wide Web*, pages 1295–1296, 2007.