J Math Imaging Vis
DOI 10.1007/s10851-016-0657-5

An Analysis of the Factors Affecting Keypoint Stability in Scale-Space

Ives Rey-Otero[1] · Jean-Michel Morel[1] · Mauricio Delbracio[2]

Received: 23 November 2015 / Accepted: 12 April 2016

Abstract The most popular image matching algorithm SIFT, introduced by D. Lowe a decade ago, has proven to be sufficiently scale invariant to be used in numerous applications. In practice, however, scale invariance may be weakened by various sources of error inherent to the SIFT implementation affecting the stability and accuracy of keypoint detection. The density of the sampling of the Gaussian scale-space and the level of blur in the input image are two of these sources. This article presents a numerical analysis of their impact on the extracted keypoints stability. Such an analysis has both methodological and practical implications, on how to compare feature detectors and on how to improve SIFT. We show that even with a significantly oversampled scale-space numerical errors prevent from achieving perfect stability. Usual strategies to filter out unstable detections (e.g., poorly contrasted extrema) are shown to be inefficient. We also prove that the effect of the error in the assumption on the initial blur is asymmetric and that the method is strongly degraded in the presence of aliasing or without a correct assumption on the camera blur. This analysis leads to a series of practical recommendations.

Keywords Scale-space · Keypoint detectors · Scale invariance · SIFT · Gaussian convolution

1 Introduction

SIFT [19,20] is a popular image matching method extensively used in image processing and computer vision applications. SIFT relies on the extraction of keypoints and the computation of local invariant feature descriptors. The scale invariance property is crucial. The matching of SIFT features is used in various applications such as image stitching [4], 3D reconstruction [31], and camera calibration [35].

SIFT was proved to be theoretically scale invariant [24]. Indeed, SIFT keypoints are covariant, being the extrema of the image Gaussian scale-space [16,41]. In practice, however, the computation of the SIFT keypoints is affected in many ways, which in turn limits the scale invariance.

The literature on SIFT focuses on variants, alternatives, and accelerations [1–7,10,12–15,17,21,23,25–27,32,34,36–38,40,42,43]. A majority of them use the scale-space keypoints as defined in the SIFT method. The huge amount of citations of SIFT indicates that it has become a standard and a reference in many applications. In contrast, there are almost no articles discussing the scale-space settings in the SIFT method and trying to compare SIFT with itself. By this comparison, we mean the question of comparing the scale invariance claim in SIFT with its empirical invariance, and the influence of the SIFT scale-space and keypoint detection parameters on its own performance. On this strict subject D. Lowe's paper [20] remains the principal reference, and it seems that very few of its claims on the parameter choices of the method have undergone a serious scrutiny. This paper intends to fill in the gap for the main claim of the SIFT method, namely the scale invariance of its keypoint detector, and incidentally on its translation invariance. This is investigated by means of a strict image simulation framework allowing us to control the main image and scale-space sampling parameters: initial blur, scale and space sampling rates, and noise level. We show that even in a particularly favorable scenario, many of the detected SIFT keypoints are unstable. We prove that the scale-space sampling has an influence on the scale invariance and that finely sampling the

✉ Ives Rey-Otero
ives.rey-otero@cmla.ens-cachan.fr

[1] CMLA, ENS-Cachan, Cachan, France

[2] ECE, Duke University, Durham, USA

Published online: 30 April 2016

Gaussian scale-space improves the detection of scale-space extrema. We quantify how the empirical invariance is affected by image aliasing and other errors due to wrong assumptions on the input image blur level.

Also, we verify the importance of the quadratic interpolation proposed in SIFT for refining the precision of the localized extrema. This is a fundamental step for the overall algorithm stability by filtering out unstable discrete extrema. On the other hand, we show that the contrast threshold proposed in SIFT is ineffective to remove the unstable detections.

Some of the conclusions of this paper were announced in [30]. The present article incorporates a more thorough rigorous analysis of the scale-space extrema and their stability. We reach this by separating the mathematical definition of the scale-space from the numerical implementation. We also add an analysis of the difference of Gaussians (DoG) scale-space operator and a discussion on how fine the scale-space should be sampled to fulfill the SIFT invariance claim.

The remainder of the paper is organized as follows. Section 2 presents the SIFT algorithm and details how to implement the Gaussian scale-space for the requirements of the present work. Section 3 examines the SIFT theoretical scale invariance. With that aim in view, we explicit the camera model consistent with the SIFT method. Section 4 details how input images are simulated to be rigorously consistent with SIFT camera model. Section 5 explores the extraction of SIFT keypoints at each stage of the algorithm focusing on the impact of the scale-space sampling on detections. Section 6 looks at the impact of image aliasing and of errors in the estimation of camera blur. To confirm some of the most obvious findings of this analysis, Sect. 7 briefly examines performance on a matching scenario. Section 8 is the conclusion which contains a list of practical recommendations.

2 The SIFT Method and Its Exact Implementation

In this section, we briefly review the SIFT method and fix the adjustments that are required to make it ideally precise. This ideal SIFT will be used in the next sections to explore the limits of the SIFT method to detect scale-space extrema.

2.1 SIFT Overview

SIFT derives from scale invariance properties of the Gaussian scale-space [16,41]. The Gaussian scale-space of an initial image u is the 3D function defined as the convolution on $\mathbb{R}^2$ with the isotropic Gaussian function of integral equal to one:

$$v : (\sigma, \mathbf{x}) \mapsto (G_\sigma * u)(\mathbf{x}) = \int_{\mathbb{R}^2} G_\sigma(\mathbf{x}')u(\mathbf{x}-\mathbf{x}')d\mathbf{x}',$$

where the Gaussian kernel is parameterized by its standard deviation $\sigma > 0$ (the scale),

$$G_\sigma(\mathbf{x}) = \frac{1}{2\pi\sigma^2} e^{-\frac{\|\mathbf{x}\|^2}{2\sigma^2}}.$$

We shall shorten the notation by also letting G_σ denote the convolution operator, thus simply writing $G_\sigma u(\mathbf{x}) := (G_\sigma * u)(\mathbf{x})$. In this framework, the Gaussian kernel acts as an approximation of the optical blur introduced in the camera (represented by its point spread function). Among other important properties [16], the Gaussian approximation is convenient because it satisfies the semi-group property

$$G_\sigma G_\gamma u(\mathbf{x}) = G_{\sqrt{\sigma^2+\gamma^2}} u(\mathbf{x}). \quad (1)$$

In particular, this permits to simulate distant snapshots from closer ones. Thus, the scale-space can be seen as a stack of images, each one corresponding to a different zoom factor. Matching two images with SIFT consists in matching keypoints extracted from these two stacks.

SIFT keypoints are defined as the 3D extrema of the difference of Gaussians (DoG) scale-space. Let v be the Gaussian scale-space and $\kappa > 1$, the DoG is the 3D function

$$w : (\sigma, \mathbf{x}) \mapsto v(\kappa\sigma, \mathbf{x}) - v(\sigma, \mathbf{x}).$$

When $\kappa \to 1$, the DoG operator acts as a good approximation of the normalized Laplacian of the scale-space [16,20],

$$v(\kappa\sigma, \mathbf{x}) - v(\sigma, \mathbf{x}) \approx (\kappa - 1)\sigma^2 \Delta v(\sigma, \mathbf{x}).$$

Continuous 3D extrema of the digital DoG are calculated in two successive steps. First, the DoG scale-space is scanned for localizing discrete extrema. This is done by comparing each voxel to its 26 neighbors. Since the location of the discrete extrema is constrained to the scale-space sampling grid, SIFT refines the position and scale of each candidate keypoint using a local interpolation model. Given a detected discrete extremum $(\sigma, \mathbf{x})$ of the digital DoG space, we denote by $\boldsymbol{\omega}_{\sigma,\mathbf{x}}(\boldsymbol{\alpha})$ the quadratic function at sample point $(\sigma, \mathbf{x})$ given by

$$\boldsymbol{\omega}_{\sigma,\mathbf{x}}(\boldsymbol{\alpha}) = \mathbf{w}_{\sigma,\mathbf{x}} + \boldsymbol{\alpha}^T g_{\sigma,\mathbf{x}} + \frac{1}{2}\boldsymbol{\alpha}^T H_{\sigma,\mathbf{x}}\boldsymbol{\alpha}, \quad (2)$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3) \in [-1/2, 1/2]^3$; $g_{\sigma,\mathbf{x}}$ and $H_{\sigma,\mathbf{x}}$ denote the 3D gradient and Hessian at $(\sigma, \mathbf{x})$ computed with a finite difference scheme. This quadratic function can be interpreted as an approximation of the second-order Taylor expansion of the underlying continuous function (where its derivatives are approximated by finite differences).

To refine the position of a discrete extremum $(\sigma_0, \mathbf{x}_0)$ SIFT proceeds as follows.

1. Initialize $(\sigma, \mathbf{x}) = (\sigma_0, \mathbf{x}_0)$.
2. Find the extrema of $\boldsymbol{\omega}_{\sigma,\mathbf{x}}$ by solving $\nabla \boldsymbol{\omega}_{\sigma,\mathbf{x}}(\boldsymbol{\alpha}) = 0$. This yields $\boldsymbol{\alpha}^* = -\left(H_{\sigma,\mathbf{x}}\right)^{-1} g_{\sigma,\mathbf{x}}$ and a refined DoG value $\boldsymbol{\omega}_{\sigma,\mathbf{x}}(\boldsymbol{\alpha}^*)$. The corresponding keypoint coordinates are updated accordingly.
3. If $\|\boldsymbol{\alpha}^*\|_\infty < M_{\text{offset}} = 0.6$ the extremum is accepted. Otherwise, go back to Step 1 and recompute the quadratic model at the closest point in the scale-space discrete grid.

This process is repeated up to N_{interp} times (in SIFT, $N_{\text{interp}} = 5$) or until the interpolation is validated. If after five iterations the result is not yet validated, the candidate keypoint is discarded.

Low contrast detections are filtered out by discarding keypoints with a small DoG value. Keypoints lying on edges are also discarded since their location is not precise due to their intrinsic translation invariant nature.

A reference keypoint orientation is computed based on the dominant gradient orientation in the keypoint surrounding. This orientation along with the keypoint coordinates are used to extract a covariant patch. Finally, the gradient orientation distribution in this patch is encoded into a 128 elements feature, the so-called SIFT descriptor. We shall not discuss further the constitution of the descriptor and refer to the abundant literature [6,14,22,25,34,39]. Numerous variations of the SIFT method exist, each variant substituting one or more elements of the SIFT algorithmic chain either to lower the computational cost [3,5] or to improve the localization accuracy [8,18,44]. For a detailed description of the SIFT method we refer the reader to [28].

2.2 The Gaussian Scale-Space and Its Implementation

Let us assume that the input image has Gaussian blur level c. The construction of the digital scale-space begins with the computation of a *seed* image. For that purpose, the input image is oversampled by a factor $1/\delta_{\min}$ and filtered by a Gaussian kernel $G_{\sqrt{\sigma_{\min}{}^2 - c^2}}$ to reach the minimal level of blur $\sigma_{\min}$ and inter-pixel distance $\delta_{\min}$. The scale-space set is split into subsets where images share a common inter-pixel distance. Since in the original SIFT algorithm the sampling rate is iteratively decreased by a factor of two, these subsets are called *octaves*. We shall denote by n_{spo} the number of scales per octave.

The subsequent images are computed iteratively from the *seed* image using the semi-group property (1) to simulate the blurs following a geometric progression

$$\sigma_s = \sigma_{\min} 2^{s/n_{\text{spo}}}, \quad s = 1, \ldots, n_{\text{spo}} - 1.$$

The digital Gaussian scale-space architecture is unequivocally defined by four parameters: the number of octaves n_{oct}, the minimal blur level $\sigma_{\min}$ in the scale-space, the number of scales per octave n_{spo}, and the initial oversampling factor $\delta_{\min}$. The standard values proposed in SIFT [19] are $n_{\text{spo}} = 3$, $\delta_{\min} = 1/2$ and $\sigma_{\min} = 0.8$. By increasing n_{spo} the scale dimension can be sampled arbitrarily finely. In the same way, by considering a small $\delta_{\min}$ value, the 2D spatial position can be sampled finely.

From this digital Gaussian scale-space the difference of Gaussian scale-space (DoG) is computed. A DoG image at scale σ is computed by subtracting from the image with blur level $\kappa\sigma$ the image with blur level σ (with $\kappa > 1$). Originally, the DoG scale-space is computed as a simple difference between two successive scales of the Gaussian scale-space so that $\kappa = 2^{1/n_{\text{spo}}}$. In the present work, we have modified this definition by unlinking the parameters κ and n_{spo}. This will allow us to better analyze the implications of the mathematical definition of the DoG operator (given by the κ-value) and the algorithmic implementation (given by the sampling parameter n_{spo}).

The Gaussian convolution implementation The architecture of the Gaussian scale-space requires for the Gaussian convolution to be implemented so it satisfies the semi-group property (1). In SIFT, the Gaussian convolution is implemented as a discrete convolution with a sampled truncated Gaussian kernel. Such an implementation satisfies the semi-group property for the SIFT default parameters ($n_{\text{spo}} = 3$), but it fails for larger values of n_{spo}, as the level of blur to be added approaches zero.

To illustrate and quantify how the discrete Gaussian convolution fails to satisfy the semi-group property, we carried out the following experiment. A sampled Gaussian function of standard deviation $c = 1.0$ was filtered $N = 10$ times using a discrete Gaussian filter of standard deviation σ. If the Gaussian semi-group property were valid, then, applying N times a Gaussian filter of parameter σ should produce the same result as filtering only once with a Gaussian function of parameters $\sqrt{N}\sigma$. We fitted a Gaussian function to the filtered image by least squares. We compared the estimated standard deviation to the theoretical expected value $\sqrt{\sigma_{\text{in}}^2 + N\sigma^2}$ (Fig. 1a). For low values of σ (i.e., $\sigma < 0.7$), the estimated blur deviates from the theoretical value $\sqrt{N}\sigma$ indicating that the method fails to satisfy the semi-group property. This is a direct consequence of image aliasing produced by excessive undersampling of the Gaussian kernel [11,29]

To avoid this undesired phenomenon in our experiments that will consider strong scale oversampling, we replaced the discrete convolution by a Fourier-domain-based convolution using the discrete cosine transform (DCT). This can be interpreted as the continuous convolution between the DCT interpolation the discrete input image and the Gaussian kernel. The implementation details along with a comparison of the Fourier-based convolution with the discrete convolution

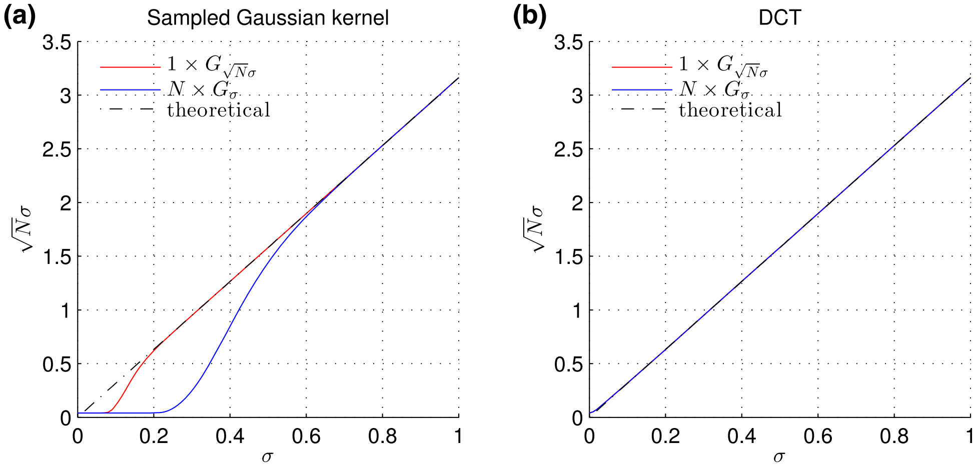

Fig. 1 Analysis of the Gaussian convolution implementation through the semi-group property. An image having a Gaussian blob of standard deviation $c = 1.1$ was filtered by (i) a Gaussian convolution of parameter $\sqrt{N}\sigma$, and (ii) by applying $N = 10$ iterations of a Gaussian convolution of parameter σ for different values of σ. Then the blur levels of the filtered images were estimated and compared to the theoretical expected value. **a** Discrete convolution with sampled Gaussian kernel. For low values of σ, the estimated blur deviates from the theoretical value $\sqrt{N}\sigma$. This is due to image aliasing when sampling the Gaussian kernel. **b** The DCT convolution fully satisfies the semi-group property

and Lindeberg's *discrete scale-space* smoothing method [16] can be found in [29].

Figure 1b shows that the Fourier-based convolution satisfies the semi-group property even for low values of σ.

2.3 Building an Ideal SIFT for Parameter Exploration

Since our goal was to explore extrema detection, we implemented an ideal SIFT where not only the convolution is exact, but also the extrema filters were turned off. The implementation of SIFT used in the present work differs from the original one on two aspects (besides the replacement of the discrete convolution by the Fourier-based one). First, SIFT proposes two filters to discard unreliable keypoints. The first one eliminates poorly contrasted extrema (those with low DoG value) and the second one discards extrema laying on edges (using a threshold on the local Hessian spectrum). These filters were deactivated to gain a full control of all detected extrema and to isolate the impact of each of them in terms of keypoints stability. This choice will be *a posteriori* justified, as we demonstrate in Sect. 5.3 that the DoG contrast threshold is inefficient.

Secondly, we decided to implement the DoG operator in such a way that the same mathematical definition is kept (i.e., using the same κ-value) regardless of the scale sampling rate n_{spo}. SIFT approximates the normalized Laplacian $\sigma^2\Delta$ by the difference of Gaussian operator. Different DoG definitions lead to different extrema. Consider for instance an image with a Gaussian blob of standard deviation σ_{blob} as input. The normalized Laplacian will have an extremum at the center of the Gaussian blob, and scale $\sigma_{\text{detect}} = \sigma_{\text{blob}}$. On the other hand, the DoG scale-space of parameter κ yields an extremum at scale $\sigma_{\text{detect}} = {}^{\sigma_{\text{blob}}}/\sqrt{\kappa}$. Consequently, the range of scales simulated in the scale-space is affected by the parameter κ.

For the requirements of the present work, and to investigate thoroughly how the operator definition affects extrema extraction, the considered DoG scale-space implementation allows us to set κ and n_{spo} independently.

Implementation details The input image was oversampled by a factor $1/\delta_{\min}$ to reach the $\delta_{\min}$ sampling rate. This was done by using a cubic B-spline interpolation of order 3. From this interpolated image all images in the scale-space were computed using a combination of DCT Gaussian convolution and subsampling. For each scale σ simulated in an octave, the algorithm computes two images, the first one corresponding to scale σ and the second one corresponding to scale $\kappa\sigma$ (both being directly computed from the input image). Although we lost the benefit of a low computational cost, this gave us flexibility and allowed us to investigate the influence of the operator definition regardless of the scale-space sampling rate.

3 The Theoretical Scale Invariance

In this section, we give the correct proof that SIFT is scale invariant and stress the fact that this proof also indicates that knowing exactly the initial camera blur is crucial for the method's consistency.

3.1 The Camera Model

In the SIFT framework, the camera point spread function is modeled by a Gaussian kernel G_c and all digital images are frontal snapshots of an ideal planar object described by the infinite resolution image u_0. In the underlying SIFT invariance model, the camera is allowed to rotate around its optical axis, to take some distance, or to translate while keeping the same optical axis direction. All digital images can therefore be expressed as

$$\mathbf{u} =: \mathbf{S}_1 G_c H_\lambda \mathcal{T}_\mathbf{y} R_\theta u_0, \tag{3}$$

where $\mathbf{S}_1$ denotes the sampling operator, H_λ a homothety, defined by $H_\lambda u(\mathbf{x}) := u(\lambda \mathbf{x})$), $\mathcal{T}_\mathbf{y}$ a translation such that $\mathcal{T}_\mathbf{y} u(\mathbf{x}) := u(\mathbf{x} - \mathbf{y})$ and with R_θ denoting both the rotation matrix and the corresponding rotation $R_\theta u(\mathbf{x}) := u(R_\theta \mathbf{x})$.

3.2 The SIFT Method is Theoretically Invariant to Zoom O uts

It is not difficult to prove that SIFT is consistent with the camera model. Nevertheless, the proof in [24] is inexact, as pointed out in [33]. Let $\mathbf{u}_\lambda$ and $\mathbf{u}_\mu$ denote two digital snapshots of the scene u_0. More precisely,

$$\mathbf{u}_\lambda = \mathbf{S}_1 G_c H_\lambda u_0 \quad \text{and} \quad \mathbf{u}_\mu = \mathbf{S}_1 G_c H_\mu u_0. \tag{4}$$

Assuming that the images are well sampled, namely that $\mathbf{S}_1$ is invertible by Shannon interpolation, and taking advantage of the semi-group property (1), the respective scale-spaces are

$$v_\lambda(\sigma, \mathbf{x}) = G_{\sqrt{\sigma^2 - c^2}} \mathbf{I}_1 \mathbf{S}_1 G_c H_\lambda u_0(\mathbf{x}) = G_\sigma H_\lambda u_0(\mathbf{x}) \tag{5}$$

$$v_\mu(\sigma, \mathbf{x}) = G_\sigma H_\mu u_0(\mathbf{x}), \tag{6}$$

where $\mathbf{I}_1$ denotes the Shannon interpolation operator. These formulae imply that both scale-spaces only differ by a reparameterization. Indeed, if v_0 denotes the Gaussian scale-space of the infinite resolution image u_0 (i.e., $v_0(\sigma, \mathbf{x}) = G_\sigma u_0(\mathbf{x})$) we have

$$v_\lambda(\sigma, \mathbf{x}) = H_\lambda(G_{\lambda\sigma} u_0(\mathbf{x})) = v_0(\lambda\sigma, \lambda\mathbf{x}), \tag{7}$$

$$v_\mu(\sigma, \mathbf{x}) = v_0(\mu\sigma, \mu\mathbf{x}), \tag{8}$$

thanks to a commutation relation between homothety and convolution.[1]

By a similar argument, the two respective DoG functions are related to the DoG function w_0 derived from u_0. For a ratio $\kappa > 1$ we have

$$\begin{aligned} w_\lambda(\sigma, \mathbf{x}) &= v_\lambda(\kappa\sigma, \mathbf{x}) - v_\lambda(\sigma, \mathbf{x}) && (9)\\ &= v_0(\kappa\lambda\sigma, \lambda\mathbf{x}) - v_0(\lambda\sigma, \lambda\mathbf{x}) && (10)\\ &= w_0(\lambda\sigma, \lambda\mathbf{x}) && (11) \end{aligned}$$

and similarly $w_\mu(\sigma, \mathbf{x}) = w_0(\mu\sigma, \mu\mathbf{x})$.

Consider an extremum point $(\sigma_0, \mathbf{x}_0)$ of the DoG scale-space w_0. Then if $\sigma_0 \geq \max(\lambda c, \mu c)$, this extremum corresponds to extrema $(\sigma_1, \mathbf{x}_1)$ and $(\sigma_2, \mathbf{x}_2)$ in w_λ and w_μ, respectively, satisfying $\sigma_0 = \lambda\sigma_1 = \mu\sigma_2$. This equivalence of extrema between the two scale-space guarantees that the SIFT descriptors are identical.

Note that this same relation links the two normalized Laplacians applied on v_λ and v_μ, denoted, respectively, nL_λ and nL_μ, both related to the normalized Laplacian of v_0 denoted nL_0. We have

$$\begin{aligned} nL_\lambda(\sigma, \mathbf{x}) &= \sigma^2 \Delta v_\lambda(\sigma, \mathbf{x}) && (12)\\ &= (\lambda\sigma)^2 \Delta v_0(\lambda\sigma, \lambda\mathbf{x}) && (13)\\ &= nL_0(\lambda\sigma, \lambda\mathbf{x}) && (14)\\ nL_\mu(\sigma, \mathbf{x}) &= nL_0(\mu\sigma, \mu\mathbf{x}) && (15) \end{aligned}$$

Therefore, considering extrema of the normalized Laplacian as keypoints will also lead to SIFT descriptors that are identical.

3.3 Knowing the Camera Blur is Crucial for Scale Invariance

The knowledge of the camera blur is crucial to ensure the theoretical invariance to zoom-outs [33]. Indeed, DoG scale-spaces computed with a wrong camera blur have in general unrelated extrema. Starting again from the two digital snapshots $\mathbf{u}_\lambda$ and $\mathbf{u}_\mu$, but assuming a wrong blur c' instead of the correct blur c, the respective Gaussian scale-spaces are

$$\begin{aligned} v_\lambda(\sigma, \mathbf{x}) &= G_{\sqrt{\sigma^2 - c'^2}} \mathbf{I}_1 \mathbf{S}_1 G_c H_\lambda u_0(\mathbf{x}) && (16)\\ &= G_{\sqrt{\sigma^2 - c'^2 + c^2}} H_\lambda u_0(\mathbf{x}) && (17)\\ &= v_0\left(\lambda\sqrt{\sigma^2 - c'^2 + c^2}, \lambda\mathbf{x}\right) && (18) \end{aligned}$$

and

$$v_\mu(\sigma, \mathbf{x}) = v_0\left(\mu\sqrt{\sigma^2 - c'^2 + c^2}, \mu\mathbf{x}\right). \tag{19}$$

We see that, because of the wrong blur assumption, the scale-space function v_0 is shrunken or dilated along scale. The corresponding DoG scale-spaces are

$$\begin{aligned} w_\lambda(\sigma, \mathbf{x}) = v_0&\left(\lambda\sqrt{\kappa^2\sigma^2 - c'^2 + c^2}, \lambda\mathbf{x}\right)\\ &- v_0\left(\lambda\sqrt{\sigma^2 - c'^2 + c^2}, \lambda\mathbf{x}\right), \end{aligned} \tag{20}$$

[1] Indeed, thanks to a change of variable, we have

$$\begin{aligned} G_c H_\lambda u_0(\mathbf{x}) &= \int_{\mathbb{R}^2} G_c(\mathbf{x}') u_0(\lambda\mathbf{x} - \lambda\mathbf{x}') d\mathbf{x}'\\ &= \int_{\mathbb{R}^2} G_{\lambda c}(\mathbf{x}'') u_0(\lambda\mathbf{x} - \mathbf{x}'') d\mathbf{x}'' = H_\lambda G_{\lambda c} u_0(\mathbf{x}). \end{aligned}$$

$$w_\mu(\sigma, \mathbf{x}) = v_0\left(\mu\sqrt{\kappa^2\sigma^2 - c'^2 + c^2}, \mu\mathbf{x}\right) - v_0\left(\mu\sqrt{\sigma^2 - c'^2 + c^2}, \mu\mathbf{x}\right). \tag{21}$$

None of these are linear reparameterizations of the DoG function w_0 anymore. They yield therefore unrelated extrema. Such bias is maximal with detections at finer scales and with large zoom factors.

4 Simulating the Digital Camera

Controlling the image formation process permits us to measure how invariant SIFT is in different scenarios. Such a control was achieved by simulating images that are consistent with the SIFT camera model. Images at different zoom levels were simulated from a large reference real digital image u_{ref} through Gaussian convolution and subsampling. To simulate a camera having a Gaussian blur level c, a Gaussian convolution of standard deviation cS, with $S > 10$ was first applied to the reference image. The convolved image was then subsampled by a factor S. Assuming that the reference image has an intrinsic Gaussian blur level of $c_{\text{ref}} \ll cS$, the resulting Gaussian blur level is $\sqrt{c^2 + (c_{\text{ref}}/S)^2} \approx c$. We estimated the blur level introduced by a digital reflex camera by fitting a Gaussian function to the estimated camera point spread function (following [9]). The obtained Gaussian blur levels varied from $c = 0.35$–0.95, depending on the aperture of the lens (blur level increases with aperture size). Different zoomed-out and translated versions were simulated by adjusting the scale parameter S and by translating the sampling grid. Thanks to the large subsampling factor, the generated images are noiseless. In addition, the images were stored with 32-bit precision to mitigate quantization effects. Figure 2 shows some examples of simulated images used in the experiments.

It might be objected that our simulations are highly unrealistic as the images to be compared by SIFT in a real scenario are not perfectly sampled or noiseless. Nevertheless, with an ever growing image resolution, more and more images will be compared by SIFT in large octaves, and therefore after a large subsampling, so that these properties can become realistic in practice. Furthermore, even if applying SIFT to the originals and regardless of initial noise and blur, the images at large scales also become anyway perfect so that the accuracy and repeatability issues under such favorable conditions are relevant.

5 Empirical Analysis of the Digital Scale-Space Sampling

The SIFT method aims at locating accurately the extrema of the DoG scale-space. Ideally, one would like to detect and locate all extrema from the underlying continuous DoG scale-space. However, in practice, we do not have access to the continuous scale-space but to its discrete counterpart. In theory, as $\delta_{\min} \to 0$ and $n_{\text{spo}} \to \infty$ the discrete scale-space better approximates the continuous scale-space therefore allowing to extract reliably all continuous extrema. This section investigates what happens when the sampling rates increase and how sampling affects the successive steps of the rudimentary procedure for detecting 3D scale-space discrete extrema, namely the extraction of discrete extrema, their quadratic interpolation and their filtering based on their DoG response.

To focus on the influence of the scale-space sampling, the study was carried out in the most favorable conditions: noiseless and aliasing-free input images ($c = 1.1$ and $S = 10$). In all experiments, we set $\kappa = 2^{1/3}$ to separate the mathematical definition of the DoG analysis operator from the scale-space discretization.

5.1 Number of Detections

To evaluate how the scale-space sampling rates affects the number of detections we generated different scale-space discretizations by varying the parameters $(\delta_{\min}, n_{\text{spo}})$, and extracted the 3D discrete extrema for each one of them.

Figure 3a shows the number of detected extrema for the different scale-space samplings. At first sight, it seems that some digital scale-space samplings produce many more keypoints than the SIFT default sampling ($\delta_{\min} = 1/2$, $n_{\text{spo}} = 3$). However, this increase in detections happens for discretizations that are significantly unbalanced in space and in scale. By unbalance we mean that the scale and the space dimensions are sampled with very different sampling rates.

Boundary effect To do a fair comparison of the different *discrete* detected extrema when changing the scale-space

Fig. 2 Examples of simulated images consistent with SIFT's image camera model. The respective blur levels are $c = 0.5$, $c = 1.0$, and $c = 0.6$

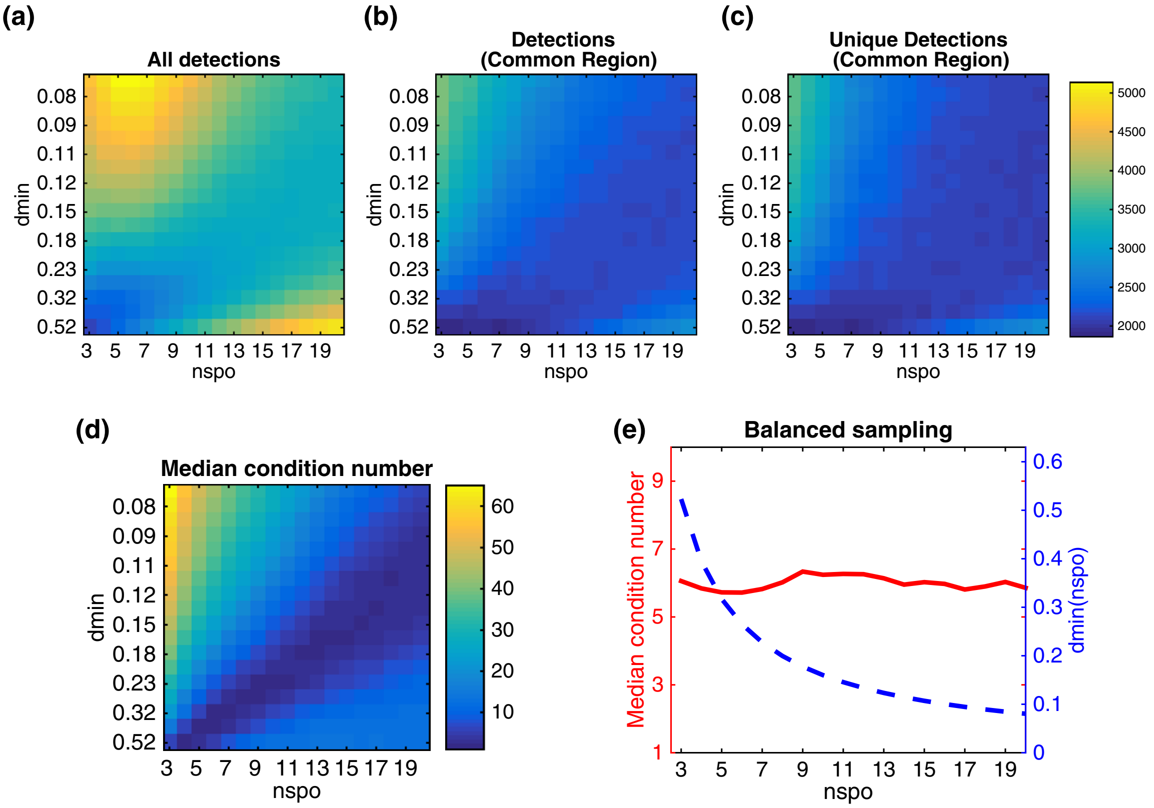

Fig. 3 Influence of the scale-space sampling rate ($n_{\text{spo}}, \delta_{\min}$) on the number of detected DoG extrema. **a** Number of 3D DoG discrete extrema. Unbalanced discretizations can produce twice as many detections as the default scale-space sampling used in SIFT ($n_{\text{spo}} = 3$, $\delta_{\min} = 1/2$). This gap is reduced after compensating for a boundary effect by discarding 3D discrete extrema with detected scale below $\sigma_{\min}2^{1/3}$ (**b**), and after removing duplicate detections (**c**). Unbalanced discretizations may lead to inaccurate local models for the extrema refinement proposed in SIFT. **d** Median of the condition numbers of DoG 3D Hessians used for extrema interpolations. Unbalanced sampling grids (shown in the top-right or bottom-left parts of this graph) produce extrema with significantly poor Hessian condition number. This leads to unstable extrema interpolations. **e** Balanced sampling rates (those satisfying (23), shown in the dotted *blue line*) lead to extrema having well-conditioned Hessian matrices (*red line*) (Color figure online)

sampling rates, we have to consider that depending on the scale-space sampling, some extrema close to the finer scale boundary are not detected. Indeed, due to the scale discretization there are no detected keypoints with scale below $\sigma_{\min}2^{1/2n_{\text{spo}}}$. To compensate for this dead range, which is a function of n_{spo}, we restricted the analysis to a common scale range independent of n_{spo}. This was achieved by discarding all extrema with scale below $\sigma_{\min}2^{1/3}$. To avoid issues due to the coarse scale discretization, we used the keypoint scale obtained after refinement (2). Figure 3b shows, for all scale-space tested configurations, the number of detections in the common scale region. The number of detected extrema lying in the common region is much more similar for all the scale-space samplings.

Duplicate detections We will say that detections $(\sigma_0, \mathbf{x}_0)$ and $(\sigma_1, \mathbf{x}_1)$ are the same, if

$$||\mathbf{x}_0 - \mathbf{x}_1||_\infty \leq \epsilon \quad \text{and} \quad R^{-1} \leq \sigma_1/\sigma_0 \leq R, \tag{22}$$

where ϵ and R are the spatial tolerance and scale relative tolerance values respective.

Clearly, there is a compromise between saying that two detections are not the same and allowing some displacement due to numerical errors. Currently, we are not tackling the problem of precision but the problem of not mixing two different detections. With that aim, it seems reasonable that the tolerance values are set in order to avoid that one detection be mistaken for another. We opted to set tolerance values to $\epsilon = 1.0$ and $R = 2^{1/2}$ independently of the scale-space sampling.

Let $\mathcal{D}$ be the set of detected DoG extrema. We call duplicates of $(\mathbf{x}_0, \sigma_0) \in \mathcal{D}$ the subset of detected extrema $\mathrm{D}(\mathbf{x}_0, \sigma_0) \subset \mathcal{D}$ that satisfy (22). Given the set of all detected keypoints $\mathcal{D}$, we say that $\mathcal{U}$ is a representative set of unique detections if

$$\mathcal{U} = \arg\min |U| \quad \text{s.t.} \quad U \subset \mathcal{D} \text{ and } \cup_{(\mathbf{x},\sigma)\in U} D(\mathbf{x}, \sigma) = \mathcal{D},$$

where the number of keypoints in the set U is denoted by $|U|$. Figure 3c shows the number of unique detections in the common scale region. The number of unique detections is similar to the number of detections (Fig. 3b). This indicates that in general duplicate detections are negligible.

Balancing the scale and space DoG sampling The SIFT algorithm proposes to refine the position of a discrete extremum using a quadratic interpolation. Having an unbalance sampling in scale and space may lead to an unreliable interpolation due to the very different discretization. As we presented in Sect. 2, the refinement of a keypoint is done by solving a linear system [from (2)]. The sensitiveness to numerical errors can be measured by the linear system's condition number (i.e., the condition number of the Hessian at the extrema to be refined). Figure 3d shows the median of the condition number for the sets of detected extrema associated with different scale-space samplings. It shows that using a balanced sampling rate improves the overall stability of the extrema interpolation.

By balanced sampling we mean that the distance separating adjacent samples in the scale dimension is similar to the distance separating adjacent samples in space. For a DoG scale-space with parameter κ, the distance between the first two simulated scales is

$$\Delta\sigma = \kappa\sigma_{\min}(2^{1/n_{\text{spo}}} - 1).$$

Thus, to equally sample the Gaussian kernel

$$G(\mathbf{x}, \sigma) = \frac{1}{2\pi\sigma^2} e^{-||\mathbf{x}||^2/2\sigma^2}$$

in scale and space, the spatial inter-pixel distance should be

$$\delta_{\min} = \sqrt{2}\Delta\sigma = \sqrt{2}\kappa\sigma_{\min}(2^{1/n_{\text{spo}}} - 1). \tag{23}$$

This relation between both sampling rates is plotted in Fig. 3e along with the median condition numbers on this set of balanced sampling rates. The condition number is mostly constant for balanced samplings.

5.2 Stability of DoG Extrema to Scale-Space Sampling

To evaluate if all 3D discrete extrema are equally stable to an increase of the DoG sampling rate, we simulated a set of increasingly dense balanced scale-spaces. We set the minimal scale-space blur level to $\sigma_{\min} = 1.1$. We simulated increasingly dense scale-space samplings $(n_{\text{spo}}, \delta_{\min})_i$, for $i = 1, \dots, n$ with $n_{\text{spo}} = 3, \dots, 19$ and the balanced spatial sampling rate $\delta_{\min} := \delta_{\min}(n_{\text{spo}})$ given by (23) ($i = 1$ being the coarsest one and $i = n$ the finest one). Figure 4a shows that the number of detections is approximately constant for different balanced sampling rates.

Let $\mathcal{D}_i$ for $i = 1 \dots, n$ be the sets of detected 3D extrema for the discretizations described above. Given a detected extremum $(\mathbf{x}_0, \sigma_0) \in \mathcal{D}_i$, we say the extremum is detected in $\mathcal{D}_j$ if there exists $(\mathbf{x}, \sigma) \in \mathcal{D}_j$ such that they are the same detection according to the precision conditions (22). We say that a detected extremum $(\mathbf{x}_0, \sigma_0) \in \mathcal{D}_i$ is new if it was not detected in $\mathcal{D}_{i-1}$. Given the sampling i, the rate of new extrema is computed as the proportion of new detected keypoints among the total number of detections. In the same way, we define the rate of lost extrema as the proportion of those present in the (coarser) sampling i and not present in the (finer) sampling $i+1$. Figure 4b shows the rate of new and lost detections as a function of the sampling rate. The new detection rate decreases with the sampling rate and stabilizes to a minimal rate of 10 % of the total number of detections for $n_{\text{spo}} \geq 14$. The same observations apply to the rate of lost extrema.

This surprising result means that despite sampling the scale-space very finely, 3D discrete extrema keep appearing and disappearing when changing the sampling.

We additionally considered pairs of translated, scaled, and rotated images and evaluated the stability of the extracted discrete extrema for various sampling grids to understand the influence of sampling. This led to the same conclusions, namely that a dense and balanced sampling of the scale-space leads to more stable discrete extrema and that even finely sampled scale-spaces are not enough to achieve perfect stability.

To illustrate how discrete extrema appear and disappear as scale-space sampling rates change, we decided to investigate the stability of each single detected extremum. The set of all unique detected extrema was formed by gathering the extrema detected on all the simulated scale-spaces $\mathcal{D}_{\text{all}} = \cup_{i=1,\dots,n}\mathcal{D}_i$ and then by extracting a unique set of detections $\mathcal{U}_{\text{all}}$. For each detected extremum $(\mathbf{x}, \sigma) \in \mathcal{U}_{\text{all}}$, we checked for its presence in each of the $\mathcal{D}_i$ detection sets. This was done by using the same definition as in (22). The results are summarized in the *occurrence* matrix shown in Fig. 4c. Each simulated discretization is indexed by the n_{spo} value. Each entry in this matrix indicates if a keypoint in $\mathcal{U}_{\text{all}}$ (column) was found in the scale-space with a given discretization $i = 1, \dots, n$ (where i is the row index in the matrix).

We define the *stability* of a unique keypoint as the proportion of discretizations where it is detected. Figure 4d shows the normalized *occurence* matrix, where each entry in the occurence matrix is multiplied by the *stability* value (therefore each column has the same color). Also, keypoints (columns) were reorganized from less to more stable (left to right).

The normalized occurrence matrix confirms that a majority of the keypoints are stable as they appear on at least 80 % of the discretizations, and that some keypoints tend to appear and disappear repeatedly as sampling rates increase. It also

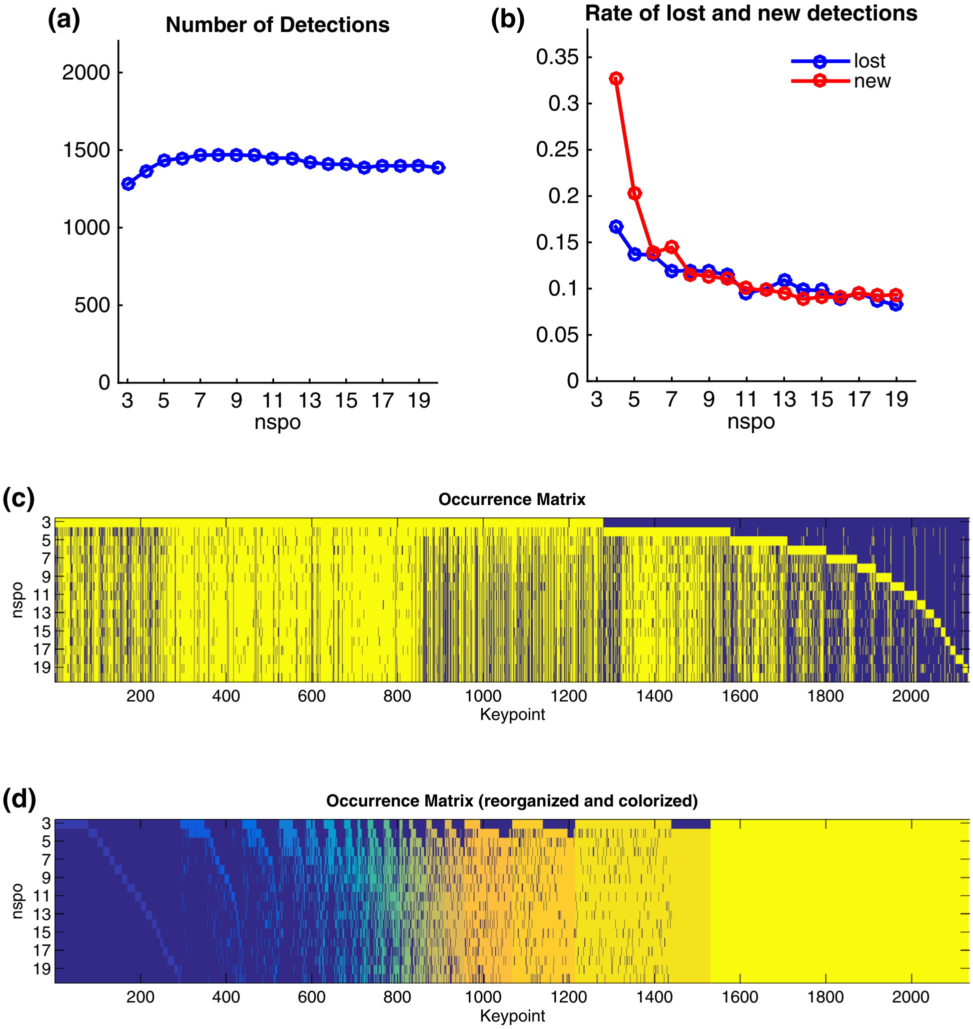

Fig. 4 Influence of sampling density on keypoint stability. A set of increasingly dense and balanced scale-spaces is computed. The scale-space samplings are indexed by the n_{spo} value, and $\delta_{\min}$ is given by (23). **a** The number of detections is roughly constant for different sampling rates. **b** The rates of lost extrema (detected in the current sampling but not in the immediately finer sampling) and of new extrema (detected in the current sampling but not in the immediately coarser sampling) decrease with the sampling rate n_{spo} and stabilize around 10 % of the total number of detections. **c** The occurrence matrix. Each row in this matrix corresponds to one of the simulated samplings (n_{spo}), while each column indicates if a keypoint was detected in that particular sampling. **d** For better visualization, the columns are colored and reorganized in increasing order of stability (*yellow* always detected, *blue* detected only once). Almost 20 % of the detections appear for all scale-space sampling rates (Color figure online)

shows that the proportion of unstable keypoints (e.g., those appearing less than 20 % of the times) is low overall but is significantly larger for coarse discretizations than in denser ones.

5.3 Can Unstable (Intermittent) Detections Be Detected?

To increase its overall detection stability, SIFT discards non-contrasted extrema based on their absolute DoG value. However, many other features, computed from the values of the extremum and its neighbors, could be used as well. The DoG value, the Laplacian of the DoG, the DoG Hessian condition number, and the minimal absolute value of the difference between the extremum and its adjacent samples are some of them.

To find out if any of these simple features is good at predicting if a discrete extremum is stable (to different sampling rates), we proceeded as follows. Given the set of unique detections $\mathcal{U}_{\text{all}}$ computed by gathering all detections from the different scale-spaces with different sampling rates, we considered two subsets of unique keypoints: one subset of *stable* unique extrema (with occurrence rate above 80%) and one subset of *unstable* unique extrema (occurrence rate below 20%). Figure 5a–d shows the proportion of extrema in both stable/unstable sets, respectively, that have a feature value below a certain threshold. The considered features are (a) the DoG value, (b) the Laplacian of the DoG, (c) the DoG Hessian condition number, and (d) the minimal absolute value of the difference between the extremum and its adjacent samples.

This figure demonstrates that none of these features manages to faithfully separate the stable from the unstable ones. This is confirmed by the ROC curve shown in Fig. 5e (see figure caption for details). Noticeably, the keypoint feature giving the lowest discrimination performance is the DoG value used by SIFT.

5.4 Visualizing Unstable (Intermittent) Detections

In an attempt to understand why the rudimentary detection and filtering procedures fail to avoid spurious detections,

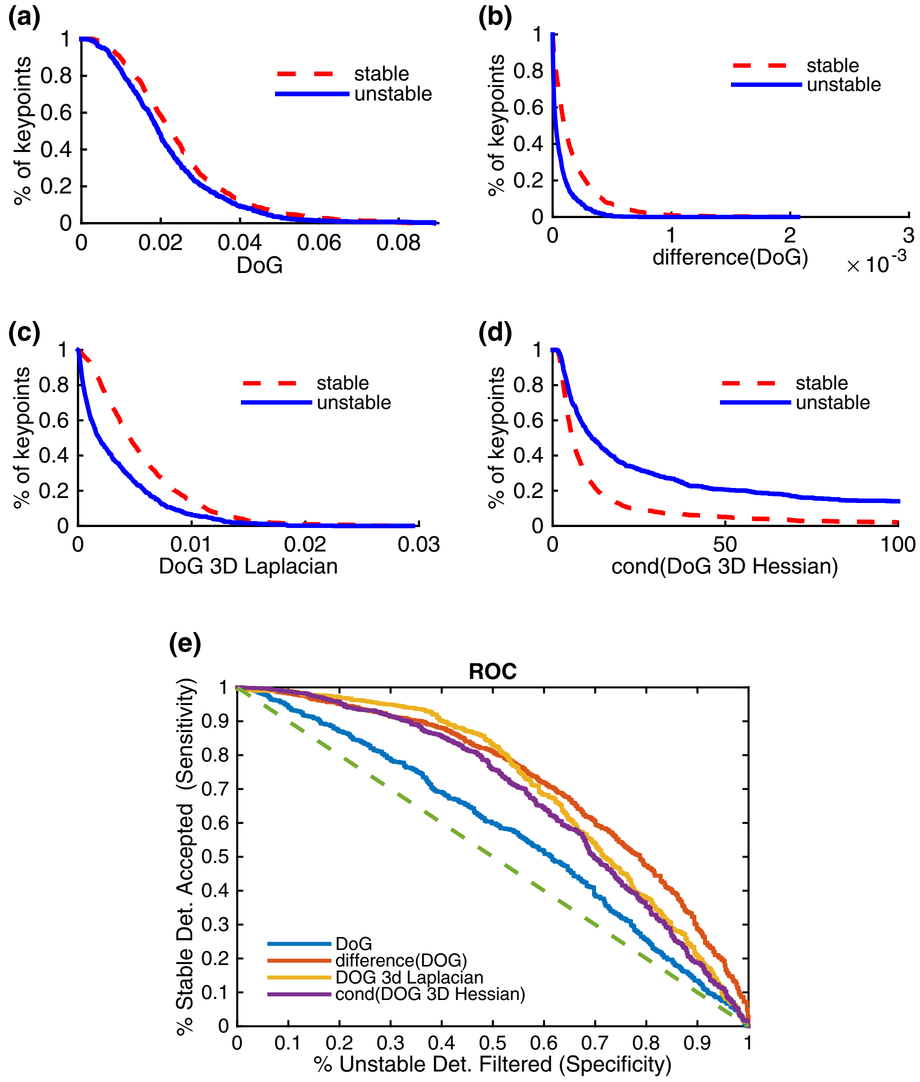

Fig. 5 Attempts at filtering keypoints that are unstable to changes in the scale-space sampling. Increasing thresholds are applied, respectively, to the set of stable and unstable detections. The considered features are **a** the extremum DoG value, **b** the difference of extremum DoG value and the adjacent samples in the scale-space, **c** the DoG 3D Laplacian value at the extremum, and **d** the condition number of the DoG 3D Hessian at the extremum. None of the tested features separates convincingly the unstable from the stable detections. This is confirmed by the ROC curves, illustrating the performance of each feature, shown in **e**. A point in a ROC *curve* indicates the proportion of non-filtered stable keypoints (good detections—sensitivity) as a function of the filtered unstable ones (good removals—specificity) for a particular threshold value. A perfect feature should produce a ROC that is always equal to 1. According to this experiment, the worst feature for eliminating keypoints unstable to changes in the scale-space sampling is the DoG value.

we examined visually some of the detected scale-space local structures. Figure 6 shows the DoG iso-surface computed around several stable and unstable keypoints from a very dense scale-space. Some detections are associated to isotropic shapes while others stem from elongated structures. There is no obvious link between how isotropic a structure is and its overall stability. As shown in the figure, some elongated structures produce stable detections. It seems therefore that a local analysis of the scale-space structure is not sufficient to characterize unstable detections.

5.5 The Influence of Extrema Interpolation on Stability, Precision and Invariance

The refinement of the discrete extrema position proposed in SIFT has two main purposes. First, it allows to locate the extrema to subpixel accuracy thanks to a local continuous model of the DoG scale-space. But this refinement procedure also detects and discards unstable discrete extrema.

In this section, we analyze the impact of the refinement procedure. To that aim, we considered an input image and a series of transformations simulating small displacements of the camera. Although the analysis was restricted for a sake of simplicity to the case of translations and scale changes, it could be easily generalized to more complex image transformations such as perspective projections.

We examined the influence of the two main parameters in the refinement procedure (see Sect. 2.1): the maximal number of allowed interpolations N_{interp}, and the maximum offset M_{offset} authorized for the extremum at each refinement iteration.

Our performance measure was the *stability*, measured by considering the number of keypoints that appear in at least a certain percentage of the simulated image transformations. A perfectly stable keypoint would be one that appears in all the simulated images, while a perfectly unstable keypoint would be one that only appears in one of the images. We also measured the *precision* by computing the average standard

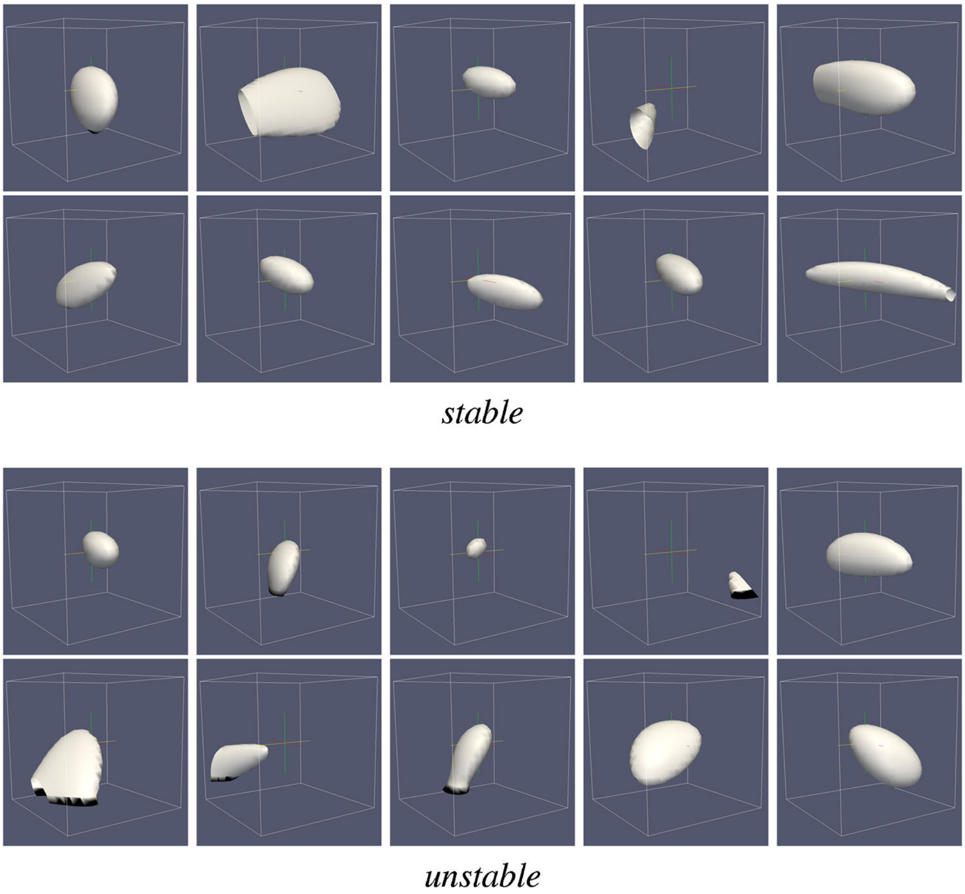

Fig. 6 Illustration of the DoG scale-space around detected keypoints. DoG Iso-surfaces are computed from a dense scale-space. We observe a variety of configurations from isotropic shapes to elongated structures. Furthermore, there seems to be no obvious connection between the local structures and the keypoint's stability level

deviation of the location of the stable keypoints, where keypoints were considered stable if they appeared in at least 50% of the simulated transformations.

Figure 7a,b shows the percentage of unique keypoints that appear in at least a given percentage of the translations for different values of M_{offset}. Each figure corresponds to a given sampling rate ($n_{\text{spo}} = 3$ and 15) and a given maximal number of interpolations ($N_{\text{interp}} = 1, 2, \infty$). Ideally, one would like to have a large proportion of stable detections, which would correspond to a flat curve. The percentage of detections for the SIFT sampling rate ($n_{\text{spo}} = 3$) decreases quickly when considering only the more stable ones, present in a large percentage of the simulated transformations. On the other hand, $n_{\text{spo}} = 15$ leads to flatter curves, which implies more stable detections, and demonstrates that increasing the scale-space sampling improves stability. The refinement of the extrema helps discard the unstable ones.

The fact that the results with $N_{\text{interp}} = 2$ and $N_{\text{interp}} = \infty$ are identical (second and third row of Fig. 7), implies that there is no extra benefit in allowing more than two iterations. The present analysis indicates that allowing a maximum of two interpolations ($N_{\text{interp}} = 2$) in combination with a maximum displacement of $M_{\text{offset}} = 0.6$ produce on average keypoints that are more stable. This conclusion is independent of the considered n_{spo}. Therefore, for the remainder of the article, we consider the refinement step with these two values.

Increasing the scale-space sampling rate in conjunction with extrema interpolation has a tremendous impact on the detection precision. Figure 7 shows for both, discrete and interpolated detections, the mean of the precision of stable keypoints (appearing in at least 50 % of translations) as a function of the scale-space sampling rate.

We repeated the same experiment but different camera zoom-outs were simulated. The results are very similar to the pure camera translation case (see Fig. 8). In general, sampling the scale-space finer than what is proposed in SIFT (e.g., $n_{\text{spo}} > 3$) allows to better localize the DoG extrema. In addition, the local refinement of the extrema position increases the extrema precision. We repeated the experiments with different rotations and reached the same conclusions.

5.6 Influence of κ

The DoG scale-space is formed by computing the difference of Gaussians operator at scales $\kappa\sigma$ and σ. To analyze the influence of the DoG parameter κ, we computed the extrema of different DoG scale-spaces produced with $\kappa = 2^{1/30}, 2^{1/29}, \ldots, 2^{1/2}$. In order to minimize sampling related

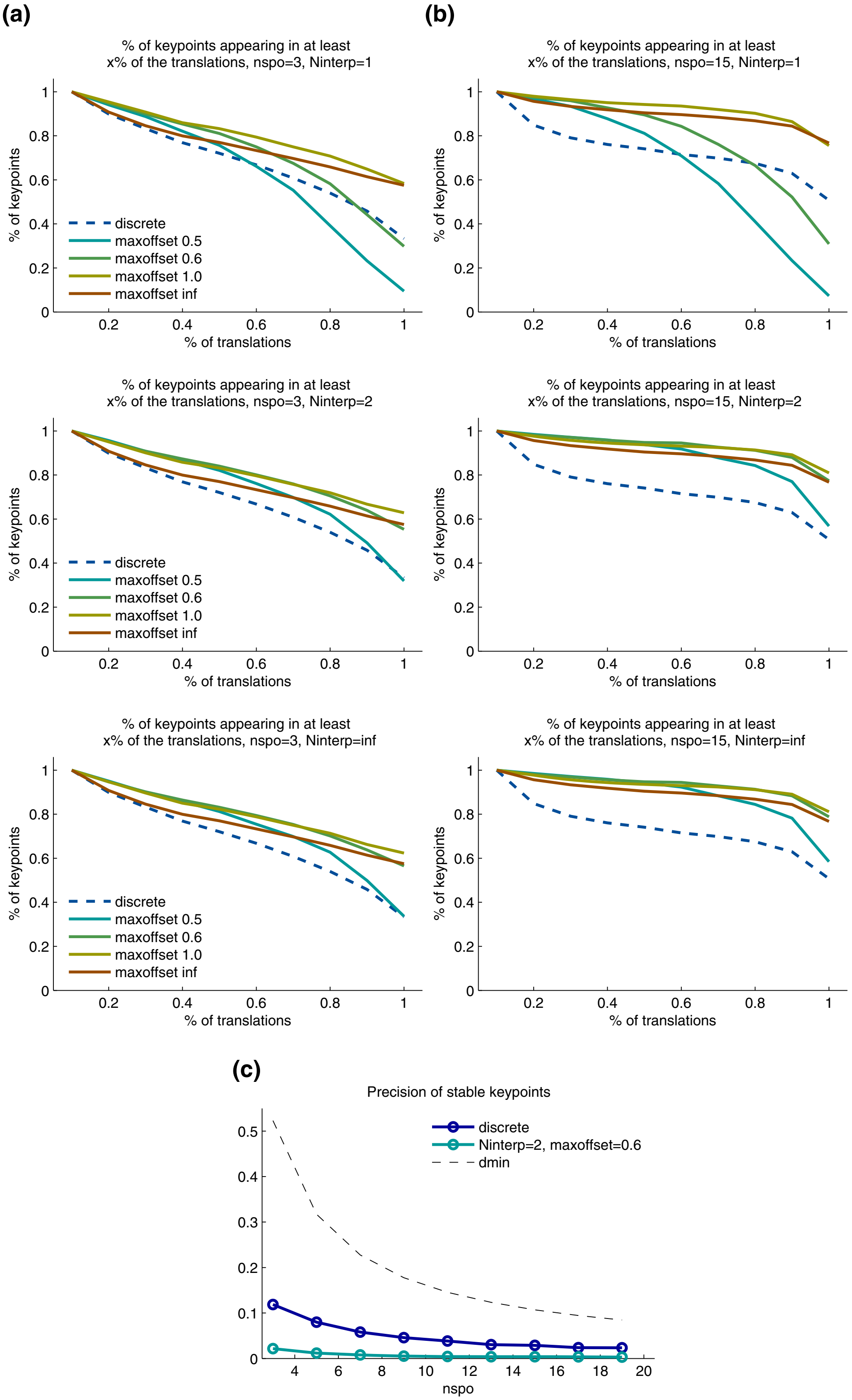

Fig. 7 Influence of extrema refinement parameters M_{offset} and N_{interp} on the detection stability/precision. A set of translated images was simulated and the keypoints extracted. Each *curve* shows the percentage of unique keypoints appearing in at least a certain percentage of the simulated image translations for different values of $M_{\text{offset}} = 0.5, 0.6, 1.0, \infty$. The plots in the first, second, and third row were generated considering a maximum number of interpolations $N_{\text{interp}} = 1, 2,$ and ∞, respectively. The left block of plots **(a)** was generated by sampling the scale-space with $n_{\text{spo}} = 3$ (and the corresponding $\delta_{\min}$), while the right block **(b)** was generated using $n_{\text{spo}} = 15$. Allowing two iterations ($N_{\text{interp}} = 2$) and a maximal offset of $M_{\text{offset}} = 0.6$ gives the best performance in terms of stability of detected keypoints. Allowing for more interpolations, attempts did not increase the performance, as can be seen by comparing the third row to the second row. **c** shows the influence of the extrema refinement on the precision of the stable set of keypoints (appearing in at least 50 % of the simulated images). In this pure translation scenario, it appears that the precision of the detected extrema significantly increases when using extrema interpolation and when sampling finely the scale-space (e.g., $n_{\text{spo}} > 3$)

instability, the scale-spaces were sampled at $n_{\text{spo}} = 15$ and the respective $\delta_{\min}$.

The number of detected extrema is more or less constant for different values of κ (Fig. 9 a). Depending on the κ value, the same structure is detected at a different scale. As pointed out in Sect. 2.3, a Gaussian blob of standard deviation σ produces an extrema of the DoG at scale $\sigma/\sqrt{\kappa}$. Thus, we have normalized the detections scale by $\sigma_{\text{normalized}} = \sigma\sqrt{\kappa}$.

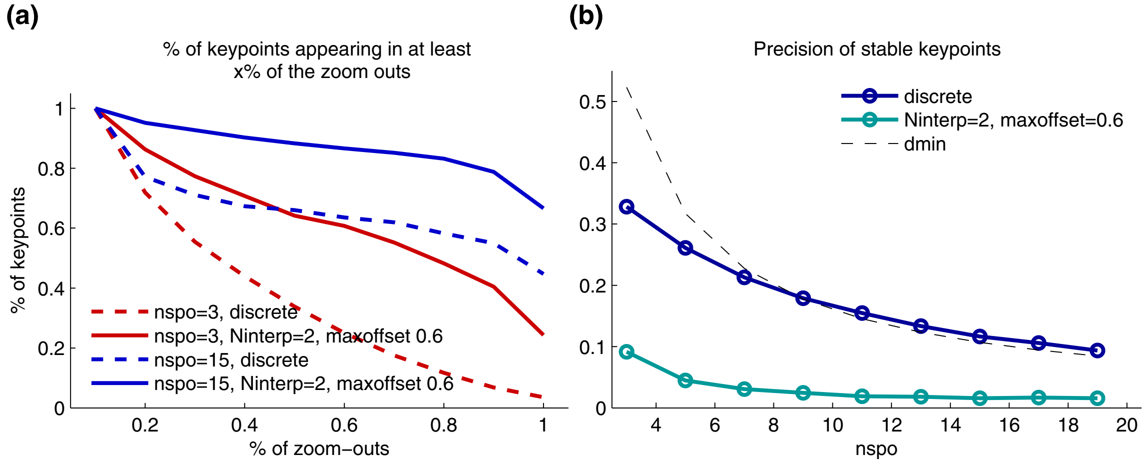

Fig. 8 Influence of scale-space sampling and extrema refinement on the invariance to zoom-outs. A set of zoomed-out images was simulated, scale-space were computed and the keypoints extracted and those which were detected outside the commonly covered scale range were discarded. **a** The percentage of unique detections appearing in at least a certain percentage of the simulated images for different scale-space sampling and refinements. The best performance is obtained by significantly oversampling the scale-space, with $n_{\text{spo}} = 15$, and by refining the extrema with the local interpolation. In this case, most of the detected keypoints are present in all the simulated images. On the other hand, the original SIFT sampling $n_{\text{spo}} = 3$ leads to low stability even with the extrema refinement step. **b** Mean precision of stable keypoints location (appearing in at least 50 % of the zoom-outs) plotted as a function of the sampling rate n_{spo}. The local refinement of the extrema position significantly increases the precision of the extrema detection. Also, using a finer grid than the one proposed in SIFT (e.g., $n_{\text{spo}} > 3$) allows to better localize the extrema

To compare the keypoints detected with different κ values, we also restricted the analysis to those lying on the common scale range, that is, $\sigma_{\min}\sqrt{2^{1/2}} \leq \sigma \leq 2\sigma_{\min}\sqrt{2^{1/30}}$.

We proceeded similarly as before by gathering all the detections from the different DoG scale-spaces and computed a set of unique detections. Then, we proceeded to create the occurrence matrix. The occurrence matrix in Fig. 9b shows that the different κ's lead for the most part to identical detections. Almost half the keypoints are detected in every DoG scale-space and a large percentage of the keypoints is detected in most simulated scale-spaces.

6 Impact of Deviations from the Perfect Camera Model

In order to achieve perfect invariance, SIFT formally requires that the image is acquired in perfect conditions. This means that the input image should be noiseless, well sampled (according to the Nyquist–Shannon sampling theorem) and with an *a priori* known level of Gaussian blur c. These ideal conditions justify the construction of the image scale-space. In this section, we evaluate what happens when there are deviations from these ideal requirements.

6.1 Image Aliasing

Let us assume that the input image was generated with a camera having a Gaussian point spread function of standard deviation c. If c is low (i.e., $c \leq 0.7$) the acquired image will be subject to aliasing artifacts. We shall assume first that this camera blur c is known beforehand, so that the SIFT method can be applied consistently.

To evaluate the SIFT performance in this aliasing situation, we simulated random translations of the digital camera. Then, we computed the extrema of the DoG scale-spaces generated with each translated image and compared the extrema. All scale-space consisted of one octave computed with $n_{\text{spo}} = 15$, $\sigma_{\min} = 1.1$ and the interpolation parameters were set to $N_{\text{interp}} = 2$ and $M_{\text{offset}} = 0.6$.

Figure 10a shows the average number of keypoints detected as a function of the camera blur c. The number of detections is independent of the camera blur. Indeed, a sharper shot does not increase the number of keypoints.

In Fig. 10b we show the percentage of unique keypoints that appear in at least a certain percentage of the translated images. Keypoints detected from well-sampled images (e.g., $c > 0.6$) are stable to translation (the curves are almost flat) while those from severely undersampled images ($c \approx 0.3$) are very sensitive to the position of the sampling grid, as expected.

6.2 Unknown Input Image Blur Level

A more realistic scenario is the case where the level of blur of the input image c is unknown. SIFT requires this value to create the scale-space starting at a known level of image blur $\sigma_{\min}$. A wrong assumption of the input camera blur affects

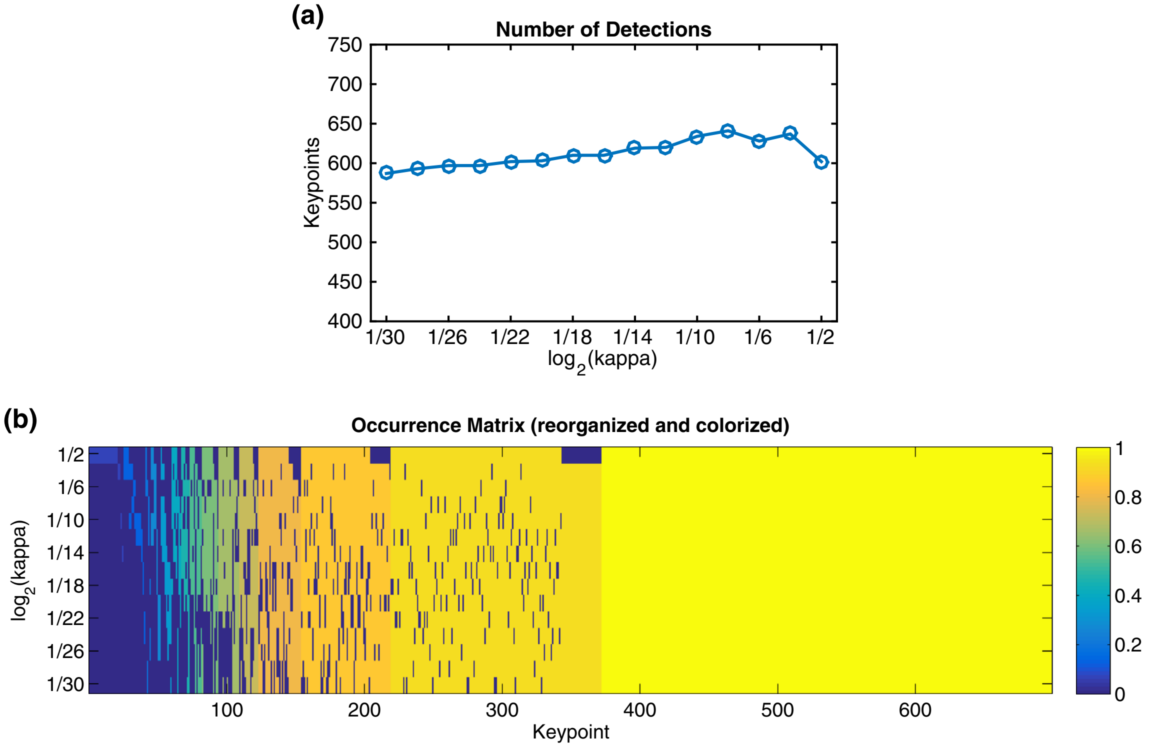

Fig. 9 Influence of the DoG parameter κ. The number of detected keypoints is roughly constant for different values of κ (**a**). The occurrence matrix for the set of unique normalized keypoints detected in the different DoG scale-spaces (**b**). A large majority of the keypoints are detected in most simulated scale-spaces when changing the value of κ

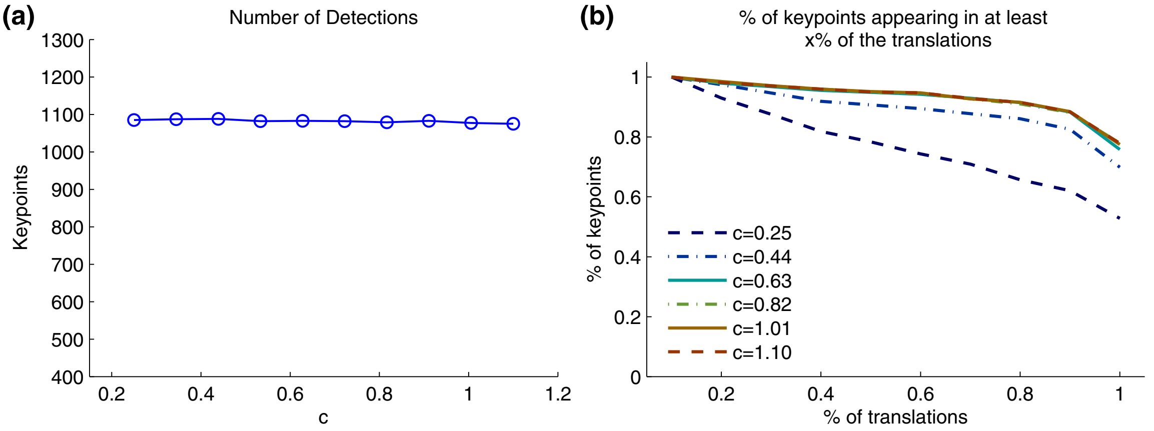

Fig. 10 Impact of image aliasing. For various camera blurs, $0.25 \leq c \leq 1.1$, a set of translated images were simulated and the DoG keypoints extracted ($n_{\text{spo}} = 15$, $\sigma_{\text{min}} = 1.1$). Aliasing does not affect the number of detections (**a**). In **b** we show the percentage of unique keypoints appearing in at least a certain percentage of the simulated translations. Detections are less stable for severely aliased images ($c = 0.25$), while for $c > 0.6$, the impact of aliasing is negligible

the range of simulated scales simulated in the Gaussian scale-space.

To demonstrate to what extent the wrong knowledge of the input camera blur produces unrelated keypoints, we compared the keypoints extracted assuming an image blur of $c = 0.7$ from a set of images having actual random blur c_{real} uniformly picked from $[c - \Delta c, c + \Delta c]$.

Figure 11 shows the number of unique keypoints that appear in at least a certain percentage of the simulated images. This was evaluated for different ranges of uncertainty (i.e., $\Delta c = 0.05 - 0.4$). The larger the range of uncertainty Δc, the more unrelated the extrema are (the curve decreases very fast, indicating the presence of many unique keypoints appearing in only a few of the simulated images). Figure 11b

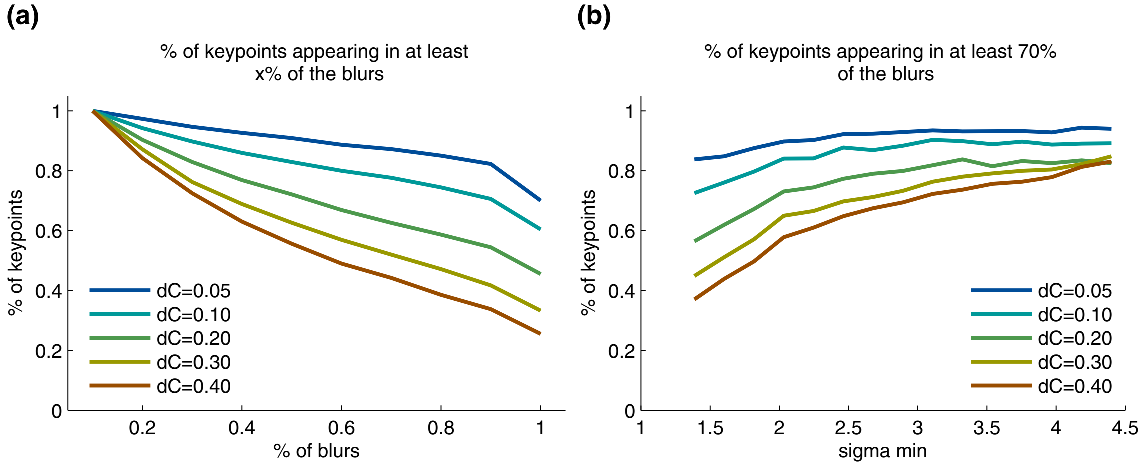

Fig. 11 The impact of a wrong assumption on the camera blur. Comparison of the keypoints extracted assuming $c = 0.7$ when the real camera blur was picked randomly in $[c - \Delta c, c + \Delta c]$. **a** The percentage of unique keypoints that appear in at least a certain percentage of the simulated images is plotted for different levels of uncertainty on camera blur ($\Delta c = 0.05 - 0.4$). **b** Influence of scale on stability to wrong blur assumption. For keypoints detected at scales ranging from σ_{min} and $2\sigma_{\text{min}}$, the proportion of unique keypoints that appear in at least 70 % of the simulated images is shown as a function of scale σ_{min}. The impact of a wrong blur level assumption decreases as we consider detections at larger scale (i.e., large σ_{min})

explores the influence of detection scale on stability to wrong blur assumption. The percentage of unique keypoints appearing in at least 70 % of the simulated images is shown as a function of scale. The influence of a wrong assumption decreases with detection scales.

6.3 Image Noise

The digital image acquisition is always affected by noise that undermines the performance of SIFT. To evaluate the impact of image noise we simulated different image acquisition, by adding random white Gaussian noise to the input image. Then, we proceeded to compute the keypoints that are detected in a certain percentage of the simulated images.

Figure 12 shows results when considering set of input images with increasing level of noise.

Specifically, Fig. 12a shows the percentage of unique keypoints that appear in at least a certain percentage of the simulated images.

It demonstrates the strong impact of noise level on keypoint stability. Such impact however is mitigated for detections at larger scales. In a Gaussian scale-space, the level of noise decreases as the scale increases. In fact, the noise standard deviation observed in a given octave is half the one observed in the previous octave. This is confirmed in Fig. 12d, which shows, for keypoints detected in a range of scale $[\sigma_{\text{min}}, 2\sigma_{\text{min}}]$, the proportion of unique keypoints that appear in at least 70 % of the simulated noisy image as a function of scale σ_{min}.

7 Matching Scenario

To illustrate some of the findings of the present analysis, we examined the following matching scenario. Three pairs of snapshots are considered: one pair of slightly translated snapshots, one pair of snapshots taken with two different zoom factors, and a pair of rotated and zoomed-out snapshots (Snapshots in Fig. 13). Keypoints extracted using the default sampling parameters ($n_{\text{spo}} = 3$, $\delta_{\text{min}} = 1/2$) and the default interpolation setup ($M_{\text{offset}} = 0.6$, $N_{\text{interp}} = 5$) were compared to the keypoints extracted from a densely sampled scale-space ($n_{\text{spo}} = 10$, $\delta_{\text{min}} = 0.081$) with interpolation parameters $M_{\text{offset}} = 0.6$ and $N_{\text{interp}} = 2$ (the rest of the algorithm was left unchanged). To prevent border effects biasing this comparison towards the densely sampled variants (see Sect. 5.1), detections with scale $\sigma \leq 1.0$ were filtered out. For all the SIFT feature vectors in one image we found the most similar feature vector on the other image (in terms of the Euclidean distance). If the distance to the most similar one was less than 60% of the distance to the second nearest feature vector, then the pair of detections was considered as a match (as proposed in [20]).

For each pair, the numbers of keypoints along with the number of matches are summarized in Table 1. In general, slightly more keypoints were detected in the balanced, densely sampled scale-spaces. Also, the dense setup led to proportionally more matches than the default sampling setting, thus confirming the stability gain.

While this simple experiment hints at how to improve the overall performance of the SIFT method, it also demonstrates that the influence of the numerical implementation should not

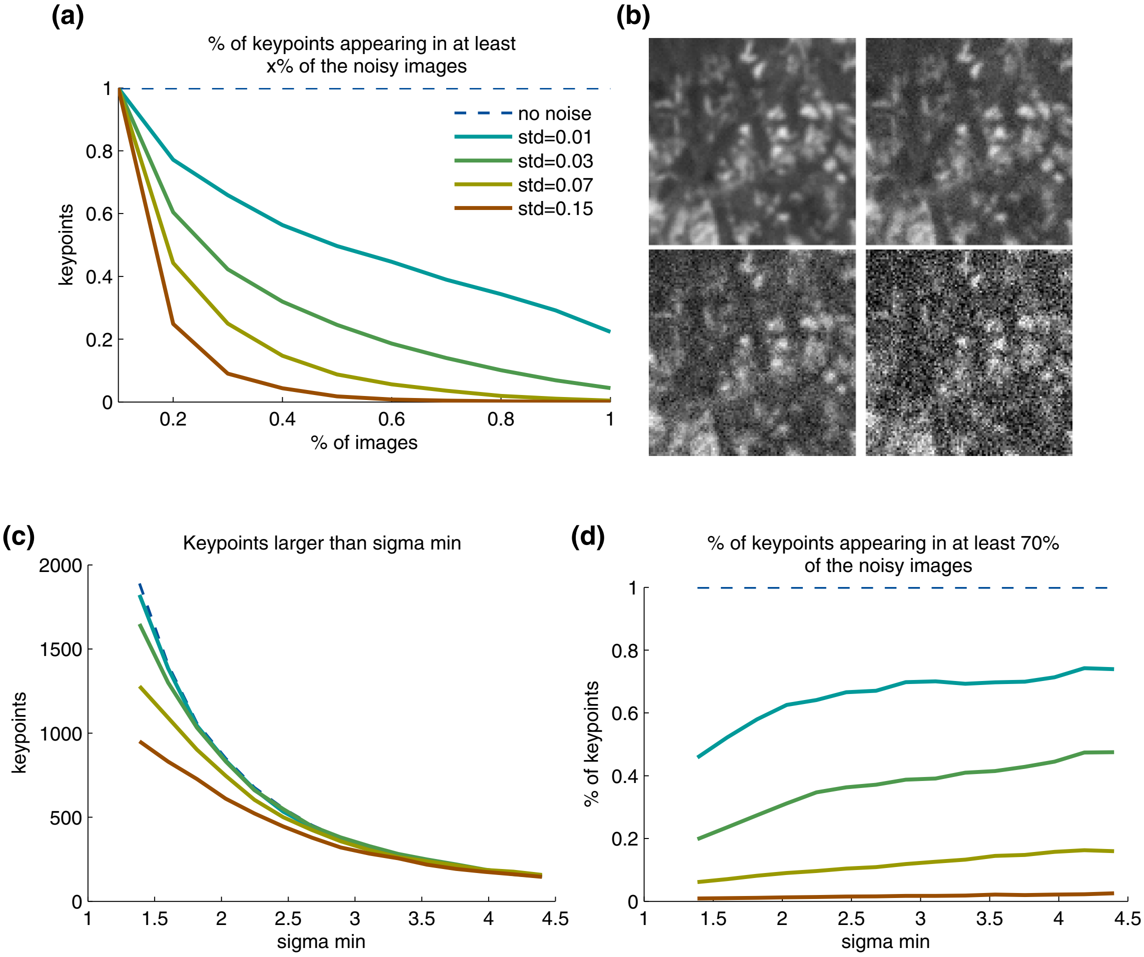

Fig. 12 Impact of image noise. **a** The proportion of unique keypoints that appear in at least a certain proportion of the simulated images is plotted for different levels of image noise. Noise has a significant impact on the DoG extrema detection. **b** Crops of the input images simulated with $c = 0.8$ and added Gaussian white noise of standard deviation $\sigma_{\text{noise}} = 0.01, 0.03, 0.07,$ and 0.15. **c** Number of keypoints detected at a scale larger than $\sigma_{\min}$ as a function of $\sigma_{\min}$. The number of detections decreases as the level of noise increases. **d** Influence of scale on stability to noise. For keypoints detected at scales ranging from $\sigma_{\min}$ to $2\sigma_{\min}$, the proportion of unique keypoints that appear in at least 70 % of the simulated images is shown a function of scale $\sigma_{\min}$. Unsurprisingly we observe that, for a given level of noise, the stability in the second octave is comparable to the stability achieved in the first octave with half the level of noise

Fig. 13 Matching scenario. Snapshots considered to constitute pairs of translated, zoomed-out, and rotated images. Keypoints were extracted using the default scale-space sampling setup as well as a densely sampled scale-space ($n_{\text{spo}} = 10$, $\delta_{\min} = 0.081$)

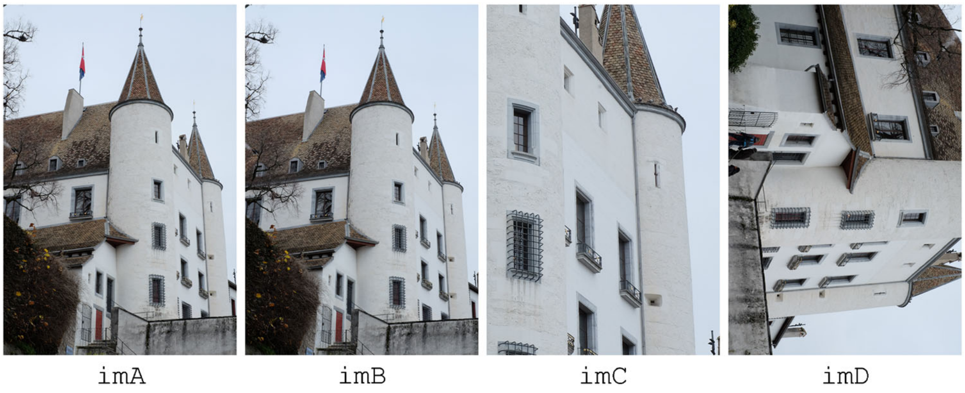

Table 1 Matching scenario: (a) Number of keypoints, respectively, detected with the standard and the dense scale-space discretization for each of the four snapshots. In general, slightly more keypoints were detected in the densely sampled scale-space. Keypoints detected in the reference image `imA` were matched to the keypoints in the translated `imB`, the zoomed-out `imC` and the rotated `imD` (Fig. 13). (b) Number of matches, the densely sampled scale-space leads to proportionally more matches demonstrating that the corresponding keypoints are more stable overall

(a) Number of keypoints	`imA`	`imB`	`imC`	`imD`
Standard	6487	6426	6423	6576
Dense	7393	7408	7973	7354
(b) Number of matches	To `imB`	To `imC`	To `imD`	
Standard	2950	430	2295	
Dense	4360	585	4076	

be overlooked. In particular, the sampling of the multiscale representations must be taken into account while comparing the SIFT method to any of its numerous variants to reliably identify the root causes of eventual improvements.

8 Concluding Remarks

We presented a systematic analysis of the main steps involved in the detection of keypoints in the SIFT algorithm. One of the main conclusions is that the original parameter choice in SIFT is not sufficient to ensure a theoretical and practical scale (and even translation) invariance, which was the main claim of the SIFT method. In addition, we showed that the SIFT invariance claim is strongly affected if the assumption on the level of blur in the input image is wrong.

A series of practical conclusions can be drawn from our analysis:

– Increasing the scale-space sampling from $n_{\text{spo}} = 3$ to $n_{\text{spo}} = 15$ (and respectively the space sampling rate δ_{min}) improves the stability of the detected keypoints.
This implies that if a series of image transformations (e.g., translations, zoom-outs) are applied to an image, the keypoints detected in one of them will be detected with high probability in all the others.
This stability property is fundamental for fulfilling the scale invariance claim.

– The extrema refinement improves the precision as well as the stability of the detected keypoints.
We showed that the largest number of stable keypoints is achieved with parameters $M_{\text{offset}} = 0.6$ and $N_{\text{interp}} = 2$ (while SIFT recommends $N_{\text{interp}} = 5$).

– The DoG threshold fails to filter out unstable keypoints, and that the different definitions of the DoG scale-space (parameter κ) lead for the most part to identical detections up to a normalization of the scale.

– Finally, we showed how the presence of aliasing and specially of noise in the acquired image deteriorate detections stability.
The effects of aliasing and noise can be mitigated by considering in priority the keypoints detected at larger scales.

One could interpret the large number of variants focusing on speed rather than stability and precision as the demonstration that the SIFT method is "good enough" in terms of stability and localization. In the light of the results of this paper, we interpret that the large number of existent variants, as well as the lack of a general consensus on which method is the best, is an illustration of the fact that very little is actually known with regard to the actual phenomena affecting keypoint stability in scale-space. Because of its ubiquity, the SIFT method deserved to be thoroughly analyzed. The submitted analysis demonstrates that implementation details such as the sampling rates have an important impact on the method's performance. On the one hand, this contributions has pointed out different practical conclusions on the SIFT method to make it more stable and precise. On the other hand, it sets the ground for rigorous and interpretable evaluation of keypoint detectors.

References

1. Agrawal, M., Konolige, K., Blas, M.: CenSurE: center surround extremas for realtime feature detection and matching. In: ECCV, pp. 102–115. Springer, Heidelberg (2008)
2. Ancuti, C., Bekaert, P.: SIFT-CCH: increasing the SIFT distinctness by color co-occurrence histograms. In: ISPA. 5th International Symposium on IEEE, pp. 130–135. (2007)
3. Bay, H., Tuytelaars, T., van Gool, L.: SURF: Speeded up robust features. In: ECCV (2006)
4. Brown, M., Lowe, D.: Automatic panoramic image stitching using invariant features. IJCV **74**(1), 59–73 (2007)
5. Brown, M., Szeliski, R., Winder, S.: Multi-image matching using multi-scale oriented patches. In: CVPR (2005)
6. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: BRIEF: Binary robust independent elementary features. In: ECCV. pp. 778–792. Springer, Heidelberg (2010)
7. Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X., Gao, W.: WLD: a robust local image descriptor. PAMI **32**(9), 1705–1720 (2010)
8. Cordes, K., Muller, O., Rosenhahn, B., Ostermann, J.: HALF-SIFT: High-accurate Localized Features for SIFT. In: CVPR Workshops. pp. 31–38 (2009)
9. Delbracio, M., Musé, P., Almansa, A.: Non-parametric sub-pixel local point spread function estimation. IPOL (2012)
10. Dickscheid, T., Schindler, F., Förstner, W.: Coding images with local features. IJCV **94**(2), 154–174 (2011)

11. Florack, L.: A spatio-frequency trade-off scale for scale-space filtering. IEEE Trans. Pattern Anal. Mach. Intell. **22**(9), 1050–1055 (2000)
12. Förstner, W., Dickscheid, T., Schindler, F.: Detecting interpretable and accurate scale-invariant keypoints. In: ICCV (2009)
13. Grabner, M., Grabner, H., Bischof, H.: Fast approximated SIFT. In: ACCV, pp. 918–927. Springer, Heidelberg (2006)
14. Ke, Y., Sukthankar, R.: PCA-SIFT: a more distinctive representation for local image descriptors. In: CVPR (2004)
15. Leutenegger, S., Chli, M., Siegwart, R.: BRISK: binary robust invariant scalable keypoints. In: ICCV (2011)
16. Lindeberg, T.: Scale-Space Theory in Computer Vision. Springer, New York (1993)
17. Liu, C., Yuen, J., Torralba, A., Sivic, J., Freeman, W.: SIFT Flow: Dense correspondence across different scenes. In: ECCV, pp. 28–42. Springer, Heidelberg (2008)
18. Loncomilla, P., Ruiz-del Solar, J.: Improving sift-based object recognition for robot applications. In: Image Analysis and Processing ICIAP 2005, vol. 3617, pp. 1084–1092. Springer, Berlin (2005). http://dx.doi.org/10.1007/11553595_133
19. Lowe, D.: Object recognition from local scale-invariant features. In: ICCV (1999)
20. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **60**, 91–110 (2004)
21. Mainali, P., Lafruit, G., Yang, Q., Geelen, B., Van Gool, L., Lauwereins, R.: SIFER: scale-invariant feature detector with error resilience. IJCV **104**(2), 172–197 (2013)
22. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. PAMI **27**(10), 1615–1630 (2005)
23. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. IJCV **65**(1–2), 43–72 (2005)
24. Morel, J.M., Yu, G.: Is SIFT scale invariant? Inverse Probl. Imaging **5**(1), 115–136 (2011)
25. Moreno, P., Bernardino, A., Santos-Victor, J.: Improving the SIFT descriptor with smooth derivative filters. Pattern Recognit. Lett. **30**(1), 18–26 (2009)
26. Pele, O., Werman, M.: A linear time histogram metric for improved SIFT matching. In: ECCV. pp. 495–508. Springer, Heidelberg (2008)
27. Rabin, J., Delon, J., Gousseau, Y.: A statistical approach to the matching of local features. SIAM J. Imaging Sci. **2**(3), 931–958 (2009)
28. Rey-Otero, I., Delbracio, M.: Anatomy of the SIFT method. Image Process. Line **4**, 370–396 (2014)
29. Rey-Otero, I., Delbracio, M.: Computing an exact gaussian scale-space. Image Process. Line **6**, 8–26 (2016)
30. Rey-Otero, I., Morel, J.M., Delbracio, M.: An analysis of scale-space sampling in SIFT. In: Image processing (ICIP), 2014 IEEE international conference. pp. 4847–4851. (2014)
31. Riggi, F., Toews, M., Arbel, T.: Fundamental matrix estimation via TIP-transfer of invariant parameters. In: ICPR (2006)
32. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to SIFT or SURF. In: ICCV (2011)
33. Sadek, R.: Some problems on temporally consistent video editing and object recognition. Ph.D. thesis, Universitat Pompeu Fabra (2012)
34. Sadek, R., Constantinopoulos, C., Meinhardt, E., Ballester, C., Caselles, V.: On affine invariant descriptors related to SIFT. SIAM **5**(2), 652–687 (2012)
35. Strecha, C., von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: CVPR (2008)
36. Tola, E., Lepetit, V., Fua, P.: A fast local descriptor for dense matching. In: CVPR (2008)
37. Tola, E., Lepetit, V., Fua, P.: DAISY: an efficient dense descriptor applied to wide-baseline stereo. PAMI **32**(5), 815–830 (2010)
38. Tuytelaars, T., Mikolajczyk, K.: Local invariant feature detectors: a survey. Found. Trends Comput. Graph. Vis. **3**(3), 177–280 (2008)
39. Van De Sande, K., Gevers, T., Snoek, C.: Evaluating color descriptors for object and scene recognition. PAMI **32**(9), 1582–1596 (2010)
40. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. In: Proceedings of the 18th ACM international conference on multimed (2010)
41. Weickert, J., Ishikawa, S., Imiya, A.: Linear scale-space has first been proposed in Japan. J. Math. Imaging Vis. **10**(3), 237–252 (1999)
42. Winder, S., Brown, M.: Learning local image descriptors. In: CVPR (2007)
43. Winder, S., Hua, G., Brown, M.: Picking the best DAISY. In: CVPR (2009)
44. Zeisl, B., Georgel, P.F., Schweiger, F., Steinbach, E.G., Navab, N., Munich, G.: Estimation of location uncertainty for scale invariant features points. In: BMVC. pp. 1–12 (2009)

Ives Rey-Otero received his M.Sc. and Ph.D. degrees in applied mathematics from ENS-Cachan, France, in 2010 and 2015, respectively. His research interests include image processing and computer vision.

Jean-Michel Morel received his Ph.D. degree in applied mathematics from University Pierre et Marie Curie, Paris, France in 1980. He started his career in 1979 as an assistant professor in Marseille Luminy and then moved in 1984 to University Paris-Dauphine where he was promoted as a professor in 1992. He is a professor of Applied Mathematics at the ENS-Cachan since 1997. His research is focused on the mathematical analysis of image processing. He received in 2013 the Grand Prix INRIA delivered by the French Academy of Science.

Mauricio Delbracio received his graduate degree from the Universidad de la República, Uruguay, in electrical engineering in 2006, and his M.Sc. and Ph.D. degrees in applied mathematics from ENS-Cachan, France, in 2009 and 2013, respectively. He currently has a postdoctoral position at the Department of Electrical and Computer Engineering, Duke University. His research interests include image and signal processing, computer graphics, photography, and computational photography.