On the consistency of the SIFT Method

Jean-Michel Morel Guoshen Yu

August 11, 2008

Abstract

This note is devoted to the mathematical arguments proving that Lowe's *Scale-Invariant Feature Transform* (SIFT [23]), a very successful image matching method, is indeed similarity invariant. The mathematical proof is given under the assumption that the gaussian smoothing performed by SIFT gives aliasing free sampling. The validity of this main assumption is confirmed by a rigorous experimental procedure. 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 may then be established by matching the descriptors of both images. Many variations exist on the computation of interest points, following the pioneering work of Harris and Stephens [14]. The Harris-Laplace and Hessian-Laplace region detectors [25, 28] are invariant to rotation and scale changes. Some moment-based region detectors [22, 2] including Harris-Affine and Hessian-Affine region detectors [26, 28], an edge-based region detector [41], an intensity- based region detector [41], an entropy-based region detector [16], and two independently developed level line-based region detectors MSER ("maximally stable extremal region") [24] and LLD ("level line descriptor") [33, 34, 5] are designed to be invariant to affine transformations. These two methods stem from the Monasse image registration method [31] 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 [30]. However, as pointed out in [23], no known detector 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 actually an under-sampling, which means that the images differ by a blur.

In his milestone paper [23], Lowe has addressed this central problem and has proposed the so called scale-invariant feature transform (SIFT) descriptor, that is 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). Based on the scale-space theory [21], 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 [17] and gradient location-orientation histogram (GLOH) [29], that claim to have better robustness and distinctiveness with scaled-down complexity have been developed ever since [9, 20]. Demonstrated to be superior to other descriptors [27, 29], SIFT has been popularly applied for scene recognition [7, 32, 39, 43, 12, 40] and detection [10, 35], robot localization [3, 36, 15], image registration [46], image retrieval [13], motion tracking [42, 18], 3D modeling and reconstruction [38, 44], building panoramas [1, 4], or photo management [45, 19, 6].

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. The method turned out to be also robust to large enough changes in view point angle, which explains its success. 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 $\mathbf{u}_0$ are smoothed at several scales to obtain $\mathbf{w}(\sigma, x, y) =: (G_\sigma * \mathbf{u}_0)(x, y)$, where

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

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 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 space Laplacian 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 $\sqrt{\sigma_i^2 + \mathbf{c}^2}$.*

The constant $\mathbf{c} \simeq 0.8$ is the tentative standard deviation of the image blur. The resulting samples of the digital patch at scale σ_i are encoded by their 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 short paper is to give the mathematical arguments proving that the method indeed is scale invariant, and that its main assumption, that images are well-sampled under gaussian blur, is right. Thus, this note is not intended to propose a new variant or extension of the SIFT method; on the contratry it is intended to demonstrate that no other method will ever improve more than marginally the SIFT scale invariance (see Figs 1 and 6 for striking examples). To the best of our knowledge, and in spite of the more than thousand papers quoting and using SIFT, the analysis presented here does not seem to have been done previously.

Plan. A simple formalism (Sect. 2) is introduced to obtain a condensed description of the SIFT shape encoding method. Using this formalism Sect. 4 proves mathematically that the SIFT method indeed computes translation, rotation and scale invariants. This proof is correct under the main assumption that image blur can be assumed to be gaussian, and that images with a gaussian blur larger than 0.6 (SIFT takes 0.8) are approximately (but accurately) well-sampled and can therefore be interpolated. Sect. 3 gives a procedure and checks the validity of this crucial gaussian blur assumption.

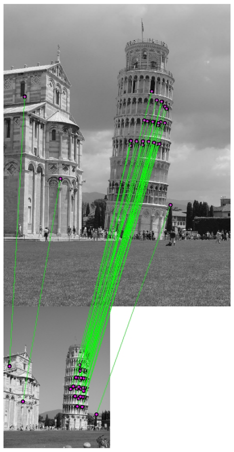

Figure 1: A result of the SIFT method, using an outliers elimination method [37]. Pairs of matching points are connected by segments.

2 Image operators formalizing SIFT

All continuous *image* operators including the sampling will be written in bold capital letters $\mathbf{A}$, $\mathbf{B}$ and their composition as a mere juxtaposition $\mathbf{AB}$. For any affine map A of the plane consider the affine transform of $\mathbf{u}$ defined by $\mathbf{Au}(\mathbf{x}) =: \mathbf{u}(A\mathbf{x})$. For instance $\mathbf{H}_\lambda\mathbf{u}(\mathbf{x}) =: \mathbf{u}(\lambda\mathbf{x})$ denotes an expansion of $\mathbf{u}$ by a factor λ^{-1}. In the same way if $\mathbf{R}$ is a rotation, $\mathbf{Ru} =: \mathbf{u} \circ R$ is the image rotation by R^{-1}.

Sampling and interpolation

Let us denote by $\mathbf{u}(\mathbf{x})$ a continuous and bounded image defined for every $\mathbf{x} = (x, y) \in \mathbb{R}^2$, and by u a digital image, only defined for $(n_1, n_2) \in \mathbb{Z}^2$. The δ-sampled image $u = \mathbf{S}_\delta\mathbf{u}$ is defined on $\mathbb{Z}^2$ by

$$\mathbf{S}_\delta\mathbf{u}(n_1, n_2) = \mathbf{u}(n_1\delta, n_2\delta); \tag{1}$$

Conversely, the Shannon interpolate of a digital image is defined as follows [11]. Let u be a digital image, defined on $\mathbb{Z}^2$ and such that $\sum_{n\in\mathbb{Z}^2} |u(n)|^2 < \infty$ and $\sum_{n\in\mathbb{Z}^2} |u(n)| < \infty$. (Of course, these conditions are automatically satisfied if the digital has a finite number of non-zero samples, which is the case here.) We call Shannon interpolate Iu of u the only $L^2(\mathbb{R}^2)$ function having u as samples and with spectrum support contained in $(-\pi, \pi)^2$. Iu is defined by the Shannon-Whittaker formula

$$Iu(x, y) =: \sum_{(n_1,n_2)\in\mathbb{Z}^2} u(n_1, n_2)\mathrm{sinc}(x - n_1)\mathrm{sinc}(y - n_2),$$

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

$$I\mathbf{S}_1\mathbf{u} = \mathbf{u}. \tag{2}$$

In that case we simply say that $\mathbf{u}$ is *band-limited.* We shall also say that a digital image $u = \mathbf{S}_1\mathbf{u}$ is *well-sampled* if it was obtained from a band-limited image $\mathbf{u}$.

The Gaussian semigroup

$\mathbf{G}$ denotes the convolution operator on $\mathbb{R}^2$ with the gauss kernel $\mathbf{G}_\sigma(x_1, x_2) = \frac{1}{2\pi(\mathbf{c}\sigma)^2} e^{-\frac{x_1^2+x_2^2}{2(\mathbf{c}\sigma)^2}}$, namely $\mathbf{G}\mathbf{u}(x, y) =: (\mathbf{G} * \mathbf{u})(x, y)$. $\mathbf{G}_\sigma$ satisfies the semigroup property

$$\mathbf{G}_\sigma \mathbf{G}_\beta = \mathbf{G}_{\sqrt{\sigma^2+\beta^2}}. \tag{3}$$

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

$$\mathbf{G}_\sigma \mathbf{H}_\gamma \mathbf{u} = \mathbf{H}_\gamma \mathbf{G}_{\sigma\gamma} \mathbf{u}. \tag{4}$$

Using the above notation, the next paragraph formalizes the SIFT method.

Formalized SIFT scale invariant features transform

The SIFT method is easily formalized in the continuous setting, while in practice images are always digital. The main assumption of the SIFT method being that all blurs can be assumed gaussian, it will be crucial to prove that gaussian blur gives in practice well-sampled images.

1. **Geometry:** there is an underlying infinite resolution bounded planar image $\mathbf{u}_0(\mathbf{x})$ that has undergone a similarity $\mathbf{A}\mathbf{u}_0$ (modeling a rotation, translation, and homothety) before sampling.

2. **Sampling and blur:** the camera blur is assimilated to a Gaussian with standard deviation $\mathbf{c}$. The typical value of $\mathbf{c}$ will be fixed thereafter. In Lowe's paper, $\mathbf{c}$ belongs to $[0.5, 0.8]$. The initial digital image is therefore $u = \mathbf{S}_1\mathbf{G}_\mathbf{c}\mathbf{A}\mathbf{u}_0$;

3. **Sampled scale space:** at all scales $\sigma > 0$, the SIFT method computes a good sampling of $\mathbf{u}(\sigma, \cdot) = \mathbf{G}_\sigma\mathbf{G}_\mathbf{c}\mathbf{A}\mathbf{u}_0$ and "key points" $(\sigma, \mathbf{x})$, namely scale and space extrema of $\Delta\mathbf{u}(\sigma, \cdot)$;

4. **Covariant resampling:** the blurred $\mathbf{u}(\sigma, \cdot)$ image is sampled around each key point at a rate proportional to $\sqrt{\mathbf{c}^2 + \sigma^2}$. The directions of the sampling axes are fixed by a dominant direction of $\nabla\mathbf{u}(\sigma, \cdot)$ in a σ-neighborhood of the key point. This yields rotation, translation and scale invariant samples in which the 4 parameters of $\mathbf{A}$ have been eliminated (see Fig. 3);

5. **Illumination invariance:** the final SIFT descriptors keep only 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 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

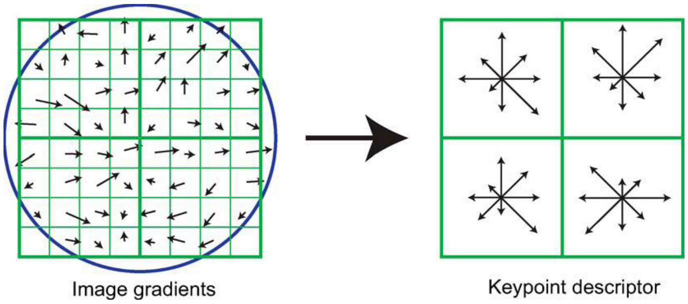

Figure 2: Each key-point is associated *a square image patch whose size is proportional to the scale and whose side direction is given by the assigned direction.* Example of a 2 × 2 descriptor array of orientation histograms (right) computed from an 8×8 set of samples (left). The orientation histograms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

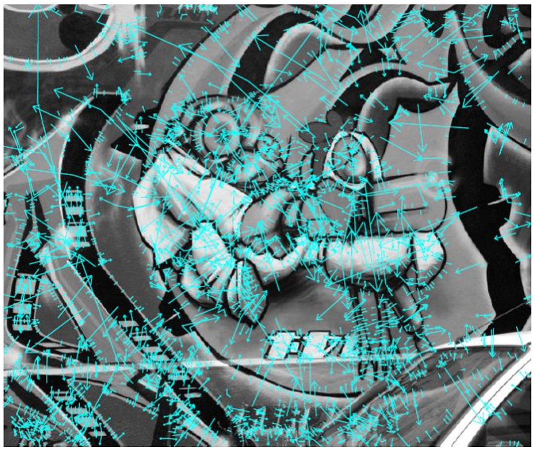

Figure 3: SIFT key points. The arrow starting point, length and the orientation signify respectively the key point position, scale, and dominant orientation. These features are covariant to any image similarity.

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.)

Two more features, however, must be commented upon. Lowe assumes that the initial image has a $\mathbf{c} = 0.5$ gaussian blur. (We call $\mathbf{c}$ gaussian blur a convolution with a gaussian with standard deviation $\mathbf{c}$). This implies a slight under-sampling that is compensated by a complementary gaussian blur applied to the image, that puts the actual initial blur to 0.8. In accordance with this choice, a 2-sub-sampling in the SIFT scale-space computations is always preceded by a $2 \times 0.8 = 1.6$ gaussian blur.

Of course, the gaussian convolution cannot be applied to the continuous image but only to the samples. This is valid if and only if a discrete convolution can give an account of the underlying continuous one, that is, if the image is well-sampled.

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

$$G_\delta u =: S_1 \mathbf{G}_\delta I u. \tag{5}$$

This definition maintains the gaussian semi-group used repeatedly in SIFT,

$$G_\delta G_\beta = G_{\sqrt{\delta^2+\beta^2}}. \tag{6}$$

Indeed, using twice (5) and once (3) and (2),

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

The SIFT method uses repeatedly this formula and a 2-sub-sampling of images with gaussian blur larger than 1.6. To summarize, the SIFT sampling manoeuvres are valid if and only if:

Proposition 1. *For every σ larger than 0.8 and every continuous and bounded image $\mathbf{u}_0$, the gaussian blurred image $\mathbf{G}_\sigma \mathbf{u}_0$ is well sampled, namely $I\mathbf{S}_1 \mathbf{G}_\sigma \mathbf{u}_0 = \mathbf{G}_\sigma \mathbf{u}_0$.*

This proposition is not a mathematical statement, but it will be checked experimentally in the next section, where we shall see that in fact a 0.6 blur is enough.

3 The right gaussian blur to achieve well-sampling

Images need to be blurred before they are sampled. In principle gaussian blur cannot lead to a good sampling because it is not *stricto sensu* band limited. Therefore the Shannon-Whittaker formula does not apply. However, in practice it does. The aim here is to define a procedure that checks that a gaussian blur works and to fix the minimal variance of the blur ensuring well-sampling (up to a minor mean square and visual error).

One must 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 case of gaussian blur, because of the semi-group formula (3), the relation between the absolute and relative blur is

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

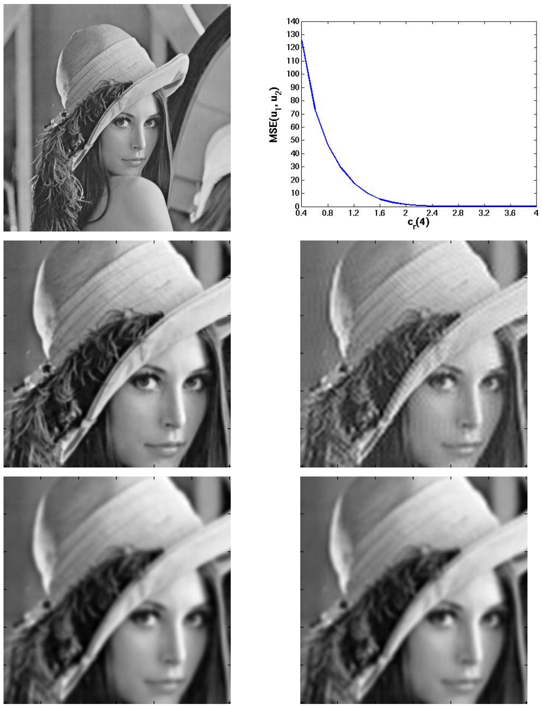

Figure 4: Top left: $\mathbf{u}$. Top right: MSE(u_1, u_2) vs $\mathbf{c}_r(4)$. Middle (from left to right): u_1 and u_2 with $\mathbf{c}_r(4) = 1.2$. MSE$(\mathbf{u}_1, \mathbf{u}_2)$=17.5. Bottom (from left to right): u_1 and u_2 with $\mathbf{c}_r(4) = 2.4$. MSE(u_1, u_2)=0.33.

which yields

$$\mathbf{c}_r(t) = \mathbf{c}_a\sqrt{t^2 - 1}. \tag{7}$$

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 compares for several values of $\mathbf{c}_r(t)$ the digital images

$$u_1 =: G_{\mathbf{c}_r(t)}u = \mathbf{S}_1\mathbf{G}_{\mathbf{c}_r(t)}Iu \quad \text{and} \quad u_2 =: (\mathbf{S}_{1/t}I)\mathbf{S}_t G_{\mathbf{c}_r(t)}u = (\mathbf{S}_{1/t}I)\mathbf{S}_t\mathbf{G}_{\mathbf{c}_r(t)}Iu,$$

where u is an initial digital image that is well-sampled, $\mathbf{S}_t$ is a t sub-sampling operator, $\mathbf{S}_{\frac{1}{t}}$ a t over-sampling operator, and I a Shannon-Whitakker interpolation operator. The discrete convolution by a gaussian is defined in (5). Since t is an integer, the t sub-sampling is trivial. The Shannon over-sampling $\mathbf{S}_{1/t}I$ with an integer zoom factor t is obtained by the classic zero-padding method. This method is exactly Shannon interpolation if the initial image is both band-limited and periodic [11].

If the anti-aliasing filter size $\mathbf{c}_r(t)$ is too small, u_1 and u_2 can be very different. The right value of $\mathbf{c}_r(t)$ should be the smallest value permitting $u_1 \approx u_2$. Fig. 4 shows u_1 and u_2 with $t = 4$ and plots their mean square error $\text{MSE}(u_1, u_2)$. An anti-aliasing filter with $\mathbf{c}_r(4) = 1.2$ is clearly not broad enough: u_2 presents strong ringing artifacts. The ringing artifact is instead hardly noticeable with $\mathbf{c}_r(4) = 2.4$. The value $\mathbf{c}_r(4) \simeq 2.4$ is a good visual candidate, and this choice is confirmed by the curve showing that $\text{MSE}(u_1, u_2)$ decays rapidly until $\mathbf{c}_r(4)$ gets close to 2.4, and is stable and small thereafter. By (7), this value of $\mathbf{c}_r$ yields $\mathbf{c}_a = 0.62$. 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 u is equal to $\mathbf{c}_a$.

In a second experiment, $\mathbf{c}_a$ has been evaluated directly by using a binary image u_0 that does not contain any blur. As illustrated in Fig. 5, u_0 is obtained by binarizing Lena (Fig. 4) the threshold being the median value. Since u_0 is now blur-free, we can compare for several values of $\mathbf{c}_a$ and for $t = 4$, which is large enough, the digital images

$$u_1 =: G_{t\mathbf{c}_a}u = \mathbf{S}_1\mathbf{G}_{t\mathbf{c}_a}Iu \quad \text{and} \quad u_2 =: (\mathbf{S}_{1/t}I)\mathbf{S}_t G_{t\mathbf{c}_a}u = (\mathbf{S}_{1/t}I)\mathbf{S}_t\mathbf{G}_{t\mathbf{c}_a}Iu,$$

As shown in Fig. 5, $\mathbf{c}_a = 0.6$ is the smallest value ensuring no visual ringing in u_2. Under this value, for example for $\mathbf{c}_a = 0.3$, clear ringing artifacts are present in u_2. That $\mathbf{c}_a = 0.6$ is the correct value is confirmed by the $\text{MSE}(u_1, u_2)$ curve showing that the mean square error decays rapidly until $\mathbf{c}_a$ goes down to 0.6, 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.

4 Scale and SIFT: consistency of the method

We denote by $\mathcal{T}$ an arbitrary image translation, by $\mathbf{R}$ an arbitrary image rotation, by $\mathbf{H}$ an arbitrary image homothety, and by $\mathbf{G}$ 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.) $\mathbf{R}\mathcal{T} = \mathcal{T}'\mathbf{R}$, meaning that given $\mathbf{R}$ and $\mathcal{T}$ there is $\mathcal{T}'$ such that the former relation occurs. The next lemma is straightforward.

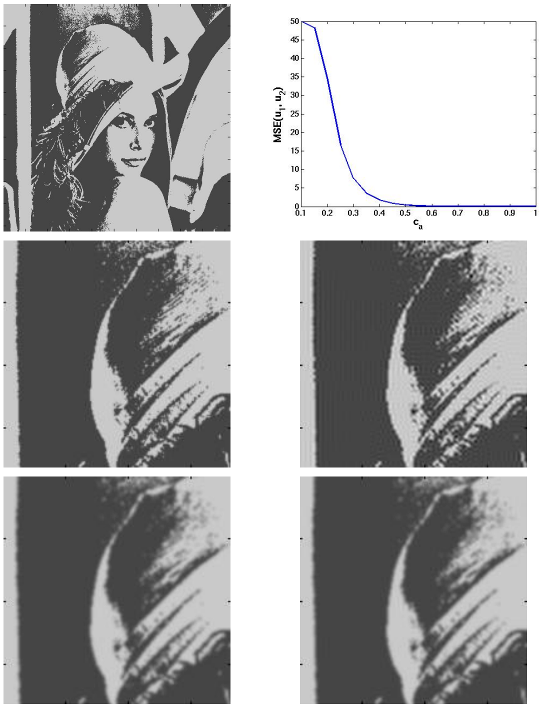

Figure 5: Top left: $\mathbf{u}$. Top right: MSE(u_1, u_2) vs $\mathbf{c}_a$. Middle (from left to right): u_1 and u_2 with $\mathbf{c}_a = 0.3$. MSE(u_1, u_2)=7.46. Bottom (from left to right): u_1 and u_2 with $\mathbf{c}_a = 0.6$. MSE(u_1, u_2)=0.09.

Lemma 1. *All of the aforementioned operators weakly commute. In addition,* $\mathbf{R}$ *and* $\mathbf{G}$ *commute strongly.*

In this section, in conformity with the SIFT model of Sect. 2, the digital image is a frontal view of an infinite resolution ideal image $\mathbf{u}_0$. In that case, $\mathbf{A} = \mathbf{H}\mathcal{T}\mathbf{R}$ is the composition of a homthety $\mathbf{H}$, a translation $\mathcal{T}$ and a rotation $\mathbf{R}$. Thus the digital image is $u = \mathbf{S}_1\mathbf{G}_\delta\mathbf{H}\mathcal{T}\mathbf{R}\mathbf{u}_0$, for some $\mathbf{H}$, $\mathcal{T}$, $\mathbf{R}$ as above. Assuming that the image is not aliased boils down, by the experimental results of Sect. 3, to assuming $\delta \geq 0.6$. (Notice that Lowe always takes $\delta = 0.8$, which is more conservative.)

Lemma 2. *For any rotation* $\mathbf{R}$ *and any translation* $\mathcal{T}$*, the SIFT descriptors of* $\mathbf{S}_1\mathbf{G}_\delta\mathbf{H}\mathcal{T}\mathbf{R}\mathbf{u}_0$ *are identical to those of* $\mathbf{S}_1\mathbf{G}_\delta\mathbf{H}\mathbf{u}_0$*.*

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* u *and* v *be two digital images that are frontal snapshots of the same continuous flat image* $\mathbf{u}_0$*,* $u = \mathbf{S}_1\mathbf{G}_\beta\mathbf{H}_\lambda\mathbf{u}_0$ *and* $v =: \mathbf{S}_1\mathbf{G}_\delta\mathbf{H}_\mu\mathbf{u}_0$*, taken at different distances, with different gaussian blurs and possibly different sampling rates. Let* $\mathbf{w}(\sigma, \mathbf{x}) = (\mathbf{G}_\sigma\mathbf{u})(\mathbf{x})$ *denote the scale space of* $\mathbf{u}$*. Then the scale spaces of* u *and* v *are*

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

If $(s_0, \mathbf{x}_0)$ *is a key point of* $\mathbf{w}$ *satisfying* $s_0 \geq \max(\lambda\beta, \mu\delta)$*, then it corresponds to a key point of* $\mathbf{u}$ *at the scale* σ_1 *such that* $\lambda\sqrt{\sigma_1^2 + \beta^2} = s_0$*, whose SIFT descriptor is sampled with mesh* $\sqrt{\sigma_1 + \mathbf{c}^2}$*. In the same way* $(s_0, \mathbf{x}_0)$ *corresponds to a key point of* $\mathbf{v}$ *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 (2)

$$\mathbf{u} =: I\mathbf{S}_1\mathbf{G}_\beta\mathbf{H}_\lambda\mathbf{u}_0 = \mathbf{G}_\beta\mathbf{H}_\lambda\mathbf{u}_0 \text{ and } \mathbf{v} =: I\mathbf{S}_1\mathbf{G}_\delta\mathbf{H}_\mu\mathbf{u}_0 = \mathbf{G}_\delta\mathbf{H}_\mu\mathbf{u}_0.$$

Computing the scale-space of these images amounts to convolve these images for every $\sigma > 0$ with $\mathbf{G}_\sigma$, which yields, using the commutation relation (4) and the semigroup property (3):

$$\mathbf{u}(\sigma, \cdot) = \mathbf{G}_\sigma\mathbf{G}_\beta\mathbf{H}_\lambda\mathbf{u}_0 = \mathbf{G}_{\sqrt{\sigma^2+\beta^2}}\mathbf{H}_\lambda\mathbf{u}_0 = \mathbf{H}_\lambda\mathbf{G}_{\lambda\sqrt{\sigma^2+\beta^2}}\mathbf{u}_0.$$

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

$$\mathbf{v}(\sigma, \cdot) = \mathbf{H}_\mu\mathbf{G}_{\mu\sqrt{\sigma^2+\delta^2}}\mathbf{u}_0.$$

Let us set $w(s, \mathbf{x}) =: \mathbf{G}_s\mathbf{u}_0$. Then the scale spaces compared by SIFT are

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

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) $\mathbf{u}(\sigma, \mathbf{x})$ and σ_2 for (the Laplacian of) $\mathbf{v}(\sigma, \mathbf{x})$ 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 $\mathbf{x} \to \mathbf{u}(\sigma, \mathbf{x})$. 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 $\mathbf{u}(\sigma_1, \cdot)$ 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 also proportional to $\sqrt{\sigma_1^2 + \mathbf{c}^2}$ in $\mathbf{u}(\sigma_1, \mathbf{x})$, and to $\sqrt{\sigma_2^2 + \mathbf{c}^2}$ in $\mathbf{u}(\sigma_2, \mathbf{x})$. □

Theorem 1. *Let u and v be two digital images that are frontal snapshots of the same continuous flat image $\mathbf{u}_0$, $u = \mathbf{S}_1\mathbf{G}_\beta\mathbf{H}_\lambda\mathcal{T}\mathbf{R}\mathbf{u}_0$ and $v =: \mathbf{S}_1\mathbf{G}_\delta\mathbf{H}_\mu\mathbf{u}_0$, taken at different distances, with different gaussian blurs and possibly different sampling rates, and up to a camera translation and rotation around its optical axe. Without loss of generality, assume $\lambda \leq \mu$. Then if the blurs are identical ($\beta = \delta = \mathbf{c}$), all SIFT descriptors of u are identical to SIFT descriptors of v. If $\beta \neq \delta$ (or $\beta = \delta \neq \mathbf{c}$), the SIFT descriptors of u and v become (quickly) similar when their scales grow, namely as soon as $\frac{\sigma_1}{\max(\mathbf{c},\beta)} \gg 1$ and $\frac{\sigma_2}{\max(\mathbf{c},\delta)} \gg 1$.*

Proof. By the result of Lemma 2, we can neglect the effect of translations and rotations. Therefore assume w.l.o.g. that the images under comparison are as in Lemma 3. Assume a key point $(s_0, \mathbf{x}_0)$ of $\mathbf{w}$ has scale $s_0 \geq \max(\lambda\beta, \mu\delta)$. This key point has a sampling rate proportional to s_0. There is a corresponding key point $(\sigma_1, \frac{\mathbf{x}_0}{\lambda})$ for $\mathbf{u}$ with sampling rate $\sqrt{\sigma_2^2 + \mathbf{c}^2}$ and a corresponding key point $(\sigma_2, \frac{\mathbf{x}_0}{\mu})$ with sampling rate $\sqrt{\sigma_2^2 + \mathbf{c}^2}$ 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 $\lambda\sqrt{\sigma_1^2 + \mathbf{c}^2} = \mu\sqrt{\sigma_2^2 + \mathbf{c}^2}$. Now, we have $\lambda\sqrt{\sigma_1^2 + \beta^2} = \mu\sqrt{\sigma_2^2 + \delta^2}$, which implies $\lambda\sqrt{\sigma_1^2 + \mathbf{c}^2} = \mu\sqrt{\sigma_2^2 + \mathbf{c}^2}$ if and only if

$$\lambda^2\beta^2 - \mu^2\delta^2 = (\lambda^2 - \mu^2)\mathbf{c}^2. \tag{9}$$

Since λ and μ correspond to camera distances to the observed object $\mathbf{u}_0$, they are pretty arbitrary. Thus in general the only way to get (9) is to have $\beta = \delta = \mathbf{c}$, which means that the blurs of both images have been guessed correctly. In any case, $\beta = \delta$ does imply that the SIFT descriptors of both images are identical.

The second statement is straighforward: 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 $\lambda\sqrt{\sigma_1^2 + \mathbf{c}^2} \simeq \mu\sqrt{\sigma_2^2 + \mathbf{c}^2}$. □

The almost perfect scale invariance of SIFT stated in Theorem 1 is illustrated by the striking example of Fig. 6. The 28 SIFT key points of a very small image u are compared to the 86 key points obtained by zooming in u by a 32 factor: The resulting digital image is $v = \mathbf{S}_{\frac{1}{32}}Iu$, again obtained by zero-padding. For better observability, both images are displayed with the same size by enlarging the pixels of u. Almost each key point (22 out of 28) of u finds its counterpart in v. 22 matches are detected between the descriptors as shown on the right. If we trust Theorem 1, all descriptors of u should have been retrieved in v. This does not fully happen for two reasons. First, the SIFT method thresholds (not taken into account in the theorem) eliminate many potential key points. Second, the zero-padding interpolation giving v is imperfect near the image boundaries.

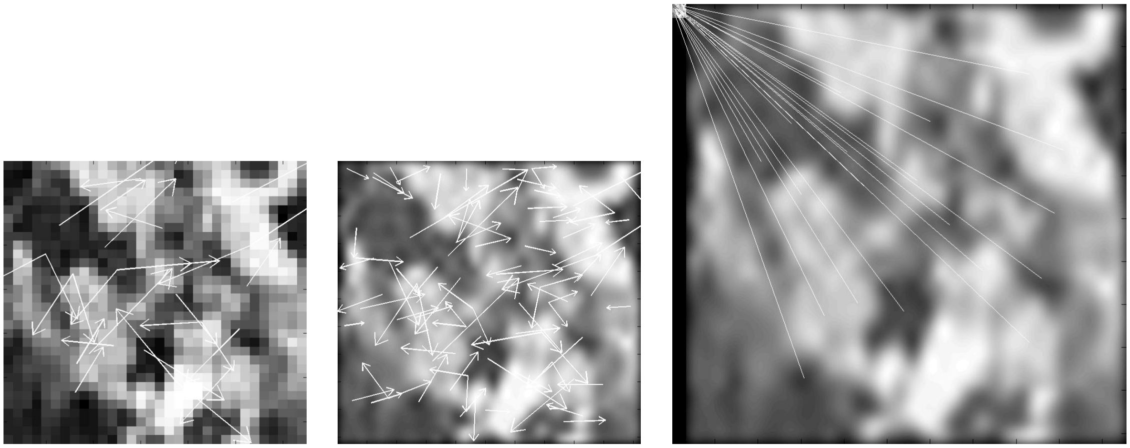

Figure 6: Scale invariance of SIFT, an illustration of Theorem 1. Left: a very small digital image u with its 28 key points. For the conventions to represent key points and matches, see the comments in Fig. 3. Middle: this image is over sampled by a 32 factor to $\mathbf{S}_{\frac{1}{32}}Iu$. It has 86 key points. Right: 22 matches found between $\mathbf{u}$ and $\mathbf{H}_{\frac{1}{32}}\mathbf{u}$.

By the second part of Theorem 1 the reliability of the SIFT matching increases with scale. This fact is illustrated in Fig. 7. Starting from a high resolution image $\mathbf{u}_0$, two images u and v are obtained by simulated zoom out, $u = \mathbf{S}_1\mathbf{G}_\beta\mathbf{H}_\lambda\mathbf{u}_0 = \mathbf{S}_\lambda\mathbf{G}_{\lambda\beta}\mathbf{u}_0$ and $v = \mathbf{S}_\mu\mathbf{G}_{\mu\delta}\mathbf{u}_0$, with $\lambda = 2$, $\mu = 4$, $\beta = \delta = 0.6$. Pairs of SIFT descriptors of u and 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. 7. That $\sigma_1/\sigma_2 \approx 2$ 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, when the scale is small ($\sigma_2 < 1$), σ_1/σ_2 is very different from 2 and $\mathbf{d}$ is large.

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 real invariance of each method when all thresholds have been eliminated. This real invariance has been proved here for SIFT. 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

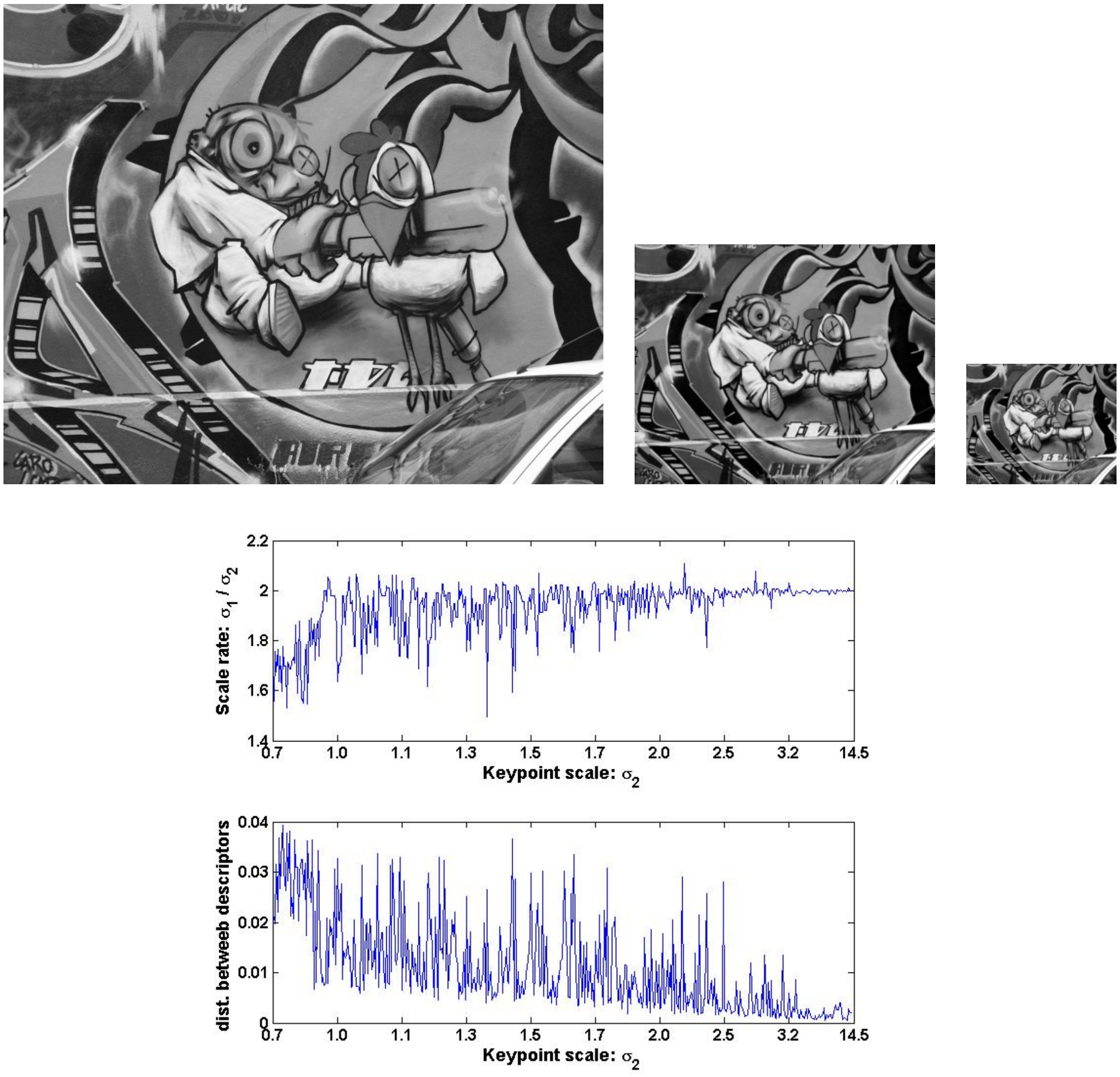

Figure 7: Top (from left to right): $\mathbf{u}_0$, u, v. Middle: Rate of scales σ_1/σ_2 of matched keypoints in u and v against σ_2. Bottom: Distance between matched descriptors of u and v against σ_2.

ensuring robustness, repeatability, and a low false alarm rate. This one has been extensively tested and confirmed in previous benchmark papers (see also the very recent and complete report [8]). 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. IEEE CVPR*, 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 2006*, 2006.

[4] M. Brown and D.G. Lowe. Recognising panoramas. *Proc. ICCV*, 1(2):3, 2003.

[5] F. Cao, J.-L. Lisani, J.-M. Morel, Musé P., and F. Sur. *A Theory of Shape Identification.* Number Vol. 1948 in Lecture Notes in Mathematics. Springer Verlag, 2008.

[6] 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.

[7] 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.

[8] L. Fevrier. A wide-baseline matching library for Zeno. *Technical report*, 2007.

[9] J.J. Foo and R. Sinha. Pruning SIFT for scalable near-duplicate image matching. *Proceedings of the eighteenth conference on Australasian database-Volume 63*, pages 63–71, 2007.

[10] 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.

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

[12] I. Gordon and D.G. Lowe. What and Where: 3D Object Recognition with Accurate Pose. *Lecture Notes in Computer Science*, 4170:67, 2006.

[13] 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.

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

[15] Aniket Murarka Joseph. Building local safety maps for a wheelchair robot using vision and lasers.

[16] T. Kadir, A. Zisserman, and M. Brady. An Affine Invariant Salient Region Detector. In *European Conference on Computer Vision*, pages 228–241, 2004.

[17] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. *Proc. CVPR*, 2:506–513, 2004.

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

[19] 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.

[20] 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.

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

[22] 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.

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

[24] 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.

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

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

[27] 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.

[28] K. Mikolajczyk and C. Schmid. Scale and Affine Invariant Interest Point Detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[29] K. Mikolajczyk and C. Schmid. A Performance Evaluation of Local Descriptors. *IEEE Trans. PAMI*, pages 1615–1630, 2005.

[30] 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.

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

[32] P. Moreels and P. Perona. Common-frame model for object recognition. *Proc. NIPS*, 2004.

[33] P. Musé, F. Sur, F. Cao, and Y. Gousseau. Unsupervised thresholds for shape matching. *Image Processing, 2003. Proceedings. 2003 International Conference on*, 2, 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. CVPR*, pages 2161–2168, 2006.

[37] J. Rabin, Y. Gousseau, and J. Delon. A contrario matching of local descriptors. Technical Report hal-00168285, Ecole Nationale Supérieure des Télécommunications, Paris, France, 2007.

[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, 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] 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.

[42] L. Vacchetti, V. Lepetit, and P. Fua. Stable Real-Time 3D Tracking Using Online and Offline Information. *IEEE Trans PAMI*, pages 1385–1391, 2004.

[43] M. Veloso, F. von Hundelshausen, and PE Rybski. Learning visual object definitions by observing human activities. *Humanoid Robots, 2005 5th IEEE-RAS International Conference on*, pages 148–153, 2005.

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

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

[46] G. Yang, CV Stewart, M. Sofka, and CL Tsai. Alignment of challenging image pairs: Refinement and region growing starting from a single keypoint correspondence. *IEEE Trans. Pattern Anal. Machine Intell*, 2007.