Is the "Scale Invariant Feature Transform" (SIFT) really Scale Invariant?

Course prepared by: Jean-Michel Morel, Guoshen Yu and Ives Rey Otero

October 24, 2010

Abstract

This note is devoted to a mathematical exploration of whether Lowe's *Scale-Invariant Feature Transform* (SIFT) [21], a very successful image matching method, is similarity invariant as claimed. It is proved that the method is scale invariant only if the initial image blurs is exactly guessed. Yet, even a large error on the initial blur is quickly attenuated by this multiscale method, when the scale of analysis increases. In consequence, its scale invariance is almost perfect. This result explains why SIFT outperforms all other image feature extraction methods when it comes to scale invariance.

1 Introduction

Image comparison is a fundamental step in many computer vision and image processing applications. A typical image matching method first detects points of interest, then selects a region around each point, and finally associates with each region a descriptor. Correspondences between two images can then be established by matching the descriptors of both images. The images under comparison may have been taken under arbitrary viewpoints. Therefore the invariance to the viewpoint is crucial in image comparison. Many variations exist on the computation of invariant interest points, following the pioneering work of Harris and Stephens [13]. The Harris-Laplace and Hessian-Laplace region detectors [23, 26] are invariant to rotation and scale changes. Some moment-based region detectors [20, 2] including the Harris-Affine and Hessian-Affine region detectors [24, 26], an edge-based region detector [42], an intensity- based region detector [42], an entropy-based region detector [14], and two independently developed level line-based region detectors MSER ("maximally stable extremal region") [22] and LLD ("level line descriptor") [33, 34] are designed to be invariant to affine transformations. These two methods stem from the Monasse image registration method [29] that used well contrasted extremal regions to register images. MSER is the most efficient one and has shown better performance than other affine invariant detectors [28]. However, as pointed out in [21], none of these detectors is actually fully affine invariant: All of them start with initial feature scales and locations selected in a non-affine invariant manner. The difficulty comes from the scale change from an image to another: This change of scale is in fact an under-sampling, which means that the images differ by a blur.

In his milestone paper [21], Lowe has addressed this central problem and has proposed the so called scale-invariant feature transform (SIFT) descriptor, that is claimed to be invariant to image

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

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

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

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

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

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

The constant $\mathbf{c} \approx 0.5$ is the tentative standard deviation of the initial image blur. The resulting samples of the digital patch at scale σ_i are encoded by their 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. 10).

The goal of this course is to give the mathematical formalism to examine whether the method indeed is scale invariant, and to discuss whether its main assumption, that images are well-sampled under Gaussian blur, does not entail significant errors. We shall not propose a new variant or an extension of the SIFT method; on the contrary we intend to demonstrate that no other method will ever improve more than marginally the SIFT scale invariance (see Figs. 5 and 12 for striking examples). To the best of our knowledge, and in spite of the more than ten thousand papers quoting and using SIFT, the analysis presented here does not seem to have been done previously.

The course is organized as follows. In Section 5, a simple formalism is introduced to obtain a

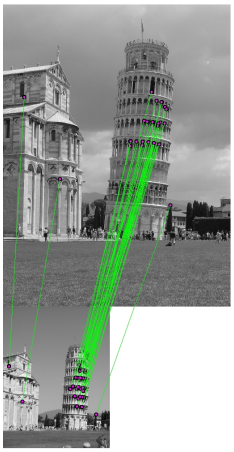

Figure 1: A result of the SIFT method, using an outliers elimination method [37]. Pairs of matching points are connected by segments.

condensed description of the SIFT shape encoding method. Using this formalism Section 6 proves mathematically that the SIFT method indeed computes translation, rotation and scale invariants.

2 The heat equation

Linear image filtering is mainly done by convolving an image u with a positive integrable kernel g. This means that the smoothed image is given by the function $g * u$ defined as

$$g * u(\mathbf{x}) = \int_{\mathbb{R}^N} g(\mathbf{x}-\mathbf{y})u(\mathbf{y})\,d\mathbf{y} = \int_{\mathbb{R}^N} g(\mathbf{y})u(\mathbf{x}-\mathbf{y})\,d\mathbf{y}.$$

Proposition 1 (The Gaussian and the heat equation)**.** *For all $t > 0$, the function $\mathbf{x} \mapsto G_t(\mathbf{x}) = (1/(4\pi t)^{N/2})\mathrm{e}^{-|\mathbf{x}|^2/4t}$ satisfies the semigroup property $G_t * G_s = G_{t+s}$ and the heat equation*

$$\frac{\partial G_t}{\partial t} - \Delta G_t = 0.$$

Theorem 1 (Existence and uniqueness of solutions of the heat equation)**.** *Assume that u_0 is a uniformly continuous and bounded function and define for $t > 0$ and $\mathbf{x} \in \mathbb{R}^N$, $u(t, \mathbf{x}) = (G_t * u_0)(\mathbf{x})$, and $u(0, \mathbf{x}) = u_0(\mathbf{x})$. Then*

(i) *u is C^∞, uniformly continuous and bounded on $(0, +\infty) \times \mathbb{R}^N$;*

(ii) *$u(t, \mathbf{x})$ tends uniformly to $u_0(\mathbf{x})$ as $t \to 0$;*

(v) *$u(t, \mathbf{x})$ satisfies the heat equation with initial value u_0;*

$$\frac{\partial u}{\partial t} = \Delta u \quad \textit{and} \quad u(0, \mathbf{x}) = u_0(\mathbf{x}); \tag{1}$$

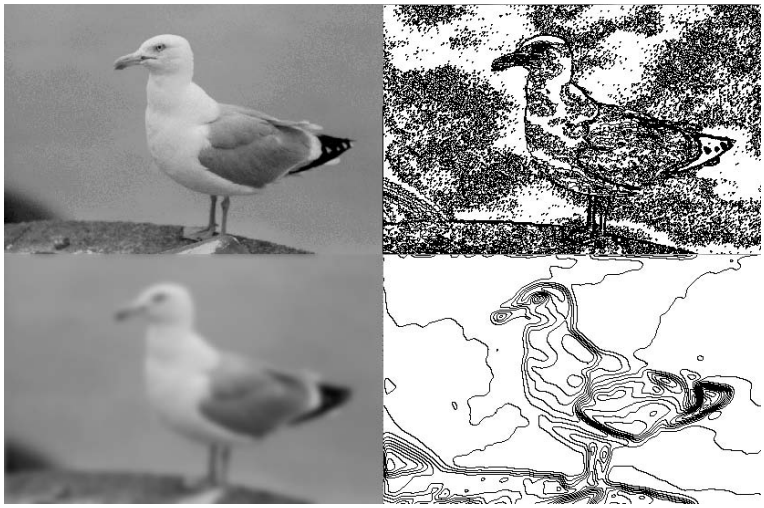

Figure 2: Level lines and the heat equation. Top, left to right: original 410×270 grey level image; level lines of original image for levels at multiples of 12. Bottom, left to right: original image smoothed by the heat equation (convolution with the Gaussian). The standard deviation of the Gaussian is 4, which means that its spatial range is comparable to a disk of radius 4. The image gets blurred by the convolution, which averages grey level values and removes all sharp edges. This can be appreciated on the right, where we have displayed all level lines for levels at multiples of 12. Note how some level lines on the boundaries of the image have split into parallel level lines that have drifted away from each other. The image has become smooth, but it is losing its structure.

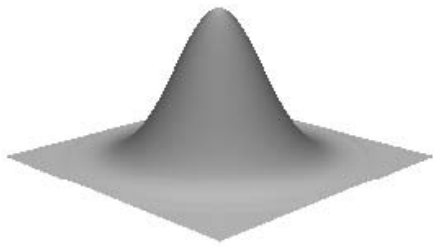

Figure 3: The Gaussian in two dimensions.

(vi) *More specifically,*

$$\sup_{\mathbf{x}\in\mathbb{R}^N,\, t\geq 0} |u(t,\mathbf{x})| \leq \sup_{\mathbf{x}} \|u_0(\mathbf{x})\|. \tag{2}$$

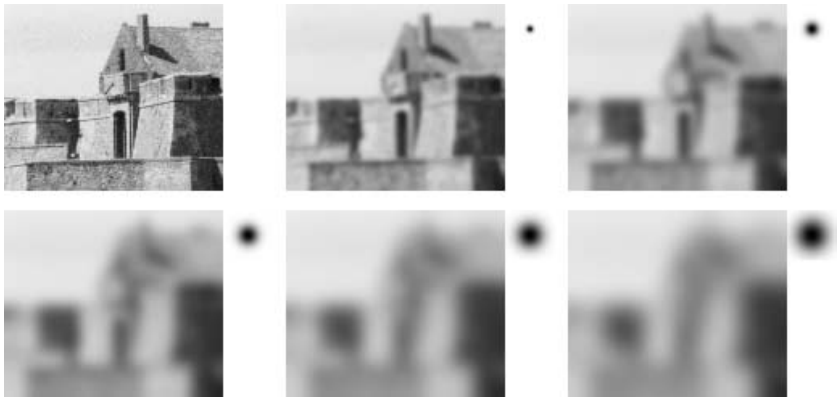

Figure 4: Convolution with Gaussian kernels (heat equation). Displayed from top-left to bottom-right are the original image and the results of convolutions with Gaussians of increasing variance. A grey level representation of the convolution kernel is put on the right of each convolved image to give an idea of the size of the involved neighborhood.

3 A Short Guide to SIFT Encoding

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.

In the SIFT method, stable points of interest are supposed to lie at extrema of the Laplacian of the image in the image scale-space representation. The scale-space representation introduces a smoothing parameter σ. Images u_0 are smoothed at several scales to obtain $w(\sigma, x, y) =: (G_\sigma * u_0)(x, y)$, where we use the parameterization of the gaussian by its standard deviation σ,

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

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 $\mathbf{c} \simeq 0.5$ 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. 10).

Figs 5 and 12 show striking examples of the method scale invariance. Lowe claims that 1) his descriptors are invariant with respect to translation, scale and rotation, and that 2) they provide a

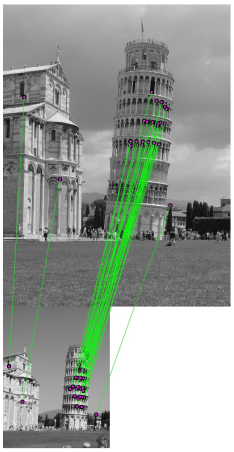

Figure 5: A result of the SIFT method, using an outliers elimination method [37]. Pairs of matching points are connected by segments.

robust matching across a substantial range of affine distortions, change in 3D viewpoint, addition of noise, and change in illumination. In addition, being local, they are robust to occlusion. Thus they match all requirements for shape recognition algorithms except one: they are not really affine invariant but only robust to moderate affine distortions.

The SIFT encoding algorithm consists of four steps: detection of scale-space extrema (Sect. 3.1), accurate localization of key points (Sect. 3.2), and descriptor construction (Sect. 3.3).

3.1 Scale-Space Extrema

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. We recall that the scale-space representation introduces a smoothing parameter σ, the scale, and convolves the image with Gaussian functions of increasing standard deviation σ. By a classical approximation inspired from psychophysics, the Laplacian of the Gaussian is replaced by a Difference of Gaussians at different scales (DOG). Extrema of the Laplacian are then replaced by extrema of DOG functions: $\mathbb{D}(\sigma, x, y) = w(k\sigma, x, y) - w(\sigma, x, y)$, where k is a constant multiplicative factor. Indeed, it is easy to show that $\mathbb{D}(\sigma, x, y)$ is an approximation of the Laplacian:

$$\mathbb{D}(\sigma, x, y) \approx (k-1)\sigma^2(\Delta G_\sigma * u_0)(x, y).$$

In the terms of David Lowe:

> *The factor* $(k-1)$ *in the equation is constant over all scales and therefore does not influence extrema location. The approximation error will go to zero as k goes to 1, but in practice we have found that the approximation has almost no impact on the stability of extrema detection or localization for even significant differences in scale, such as* $k = \sqrt{2}$.

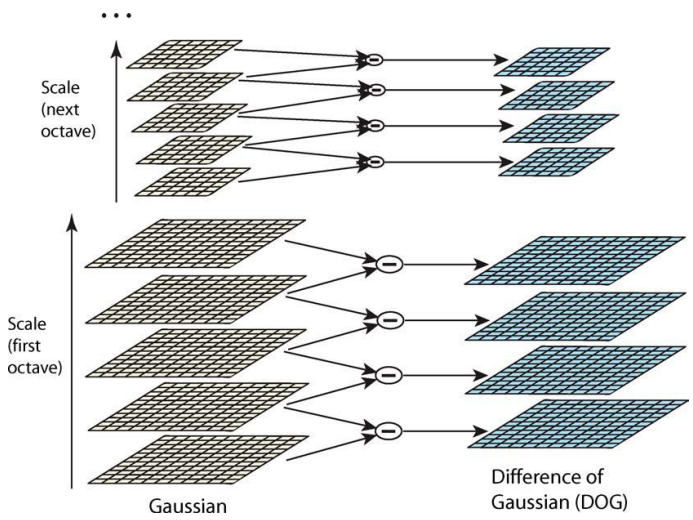

Figure 6: Gaussian pyramid for key points extraction (from [21])

To be more specific, quoting Lowe again:

$$\mathbb{D}(\sigma, x, y) =: (G(k\sigma, x, y) - G(\sigma, x, y)) * u_0(x, y) = w(k\sigma, x, y) - w(\sigma, x, y)$$

The relationship between D and $\sigma^2\Delta G$ can be understood from the heat diffusion equation (parameterized in terms of σ rather than the more usual $t = \sigma^2$):

$$\frac{\partial G}{\partial \sigma} = \sigma \Delta G.$$

From this, we see that ΔG can be computed from the finite difference approximation to $\partial G/\partial\sigma$, using the difference of nearby scales at $k\sigma$ and σ:

$$\sigma\Delta G = \frac{\partial G}{\partial \sigma} \approx \frac{G(k\sigma, x, y) - G(\sigma, x, y)}{k\sigma - \sigma}$$

and therefore,

$$G(k\sigma, x, y) - G(\sigma, x, y) \approx (k-1)\sigma^2\Delta G.$$

This shows that when the difference-of-Gaussian function has scales differing by a constant factor it already incorporates the σ^2 scale normalization required for the scale-invariant Laplacian.

This leads to an efficient computation of local extrema of $\mathbb{D}$ by exploring neighborhoods through a Gaussian pyramid ; see Figs. 6 and 7. The gaussian G_σ parameterized by its standard deviation σ satisfies as stated by Lowe the time-dependent heat equation $\frac{\partial G}{\partial \sigma} = \sigma\Delta G$.

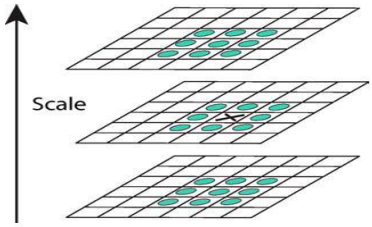

Figure 7: Neighborhood for the location of key points (from [21]). Local extrema are detected by comparing each sample point in $\mathbb{D}$ with its eight neighbors at scale σ and its nine neighbors in the scales above and below.

3.2 Accurate Key Point Detection

In order to achieve sub-pixel accuracy, the interest point position is slightly corrected thanks to a quadratic interpolation. Let us call $\mathbf{x}_0 =: (\sigma_0, x_0, y_0)$ the current detected point in scale space, which is known up to the (rough) sampling accuracy in space and scale. Notice that all points $\mathbf{x} = (\sigma, x, y)$ here are scale-space coordinates. Let us call $\mathbf{x}_1 = \mathbf{x}_0 + \mathbf{y}$ the real extremum of the DOG function. The Taylor expansion of $\mathbb{D}$ yields

$$\mathbb{D}(\mathbf{x}_0 + \mathbf{y}) = \mathbb{D}(\mathbf{x}_0) + (D\mathbb{D})(\mathbf{x}_0) \cdot \mathbf{y} + \frac{1}{2}\left(D^2\mathbb{D}\right)(\mathbf{x}_0)(\mathbf{y}, \mathbf{y}) + o(\|\mathbf{y}\|^2),$$

where $\mathbb{D}$ and its derivatives are evaluated at an interest point and $\mathbf{y}$ denotes an offset from this point. Since interest points are extrema of $\mathbb{D}$ in scale space, setting the derivative to zero gives:

$$\mathbf{y} = -\left(D^2\mathbb{D}(\mathbf{x}_0)\right)^{-1}(D\mathbb{D}(\mathbf{x}_0)), \tag{3}$$

which is the sub-pixel correction for a more accurate position of the key point of interest.

Clearly, (3) is a point where the gradient of $\mathbb{D}$ vanishes.

Since points with low contrast are sensitive to noise, and since points that are poorly localized along an edge are not reliable, a filtering step is called for. Low contrast points are handled through a simple thresholding step. Edge points are swept out following the Harris and Stephen's interest points paradigm. Let H be the following Hessian matrix:

$$H = \begin{pmatrix} \mathbb{D}_{xx} & \mathbb{D}_{xy} \\ \mathbb{D}_{xy} & \mathbb{D}_{yy} \end{pmatrix}.$$

The reliability test is simply to assess whether the ratio between the larger eigenvalue and the smaller one is below a threshold r. This amounts to check:

$$\frac{\mathrm{Tr}(H)^2}{\mathrm{Det}(H)} < \frac{(r+1)^2}{r}. \tag{4}$$

This rules out standard edge points and puts points of interest at locations which are strong enough extrema, or saddle points. It is easily checked that (4) is equivalent imposing that the ratio between the smaller eigenvalue and the larger eigenvalue of H is smaller than r. These eigenvalues are assumed to have the same sign. Why?

Figure 8: 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.

3.3 Construction of the SIFT descriptor

In order to extract rotation-invariant patches, an orientation must be assigned to each key point. Lowe proposes to estimate a semi-local average orientation for each key point. From each sample image L_σ, gradient magnitude and orientation is pre-computed using a 2×2 scheme. An orientation histogram is assigned to each key point by accumulating gradient orientations weighted by 1) the corresponding gradient magnitude and by 2) a Gaussian factor depending on the distance to the considered key point and on the scale. The precision of this histogram is 10 degrees. Peaks simply correspond to dominant directions of local gradients. Key points are created for each peak with similar magnitude, and the assigned orientation is refined by local quadratic interpolation of the histogram values.

Once a scale and an orientation are assigned to each key point, 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*. The next step is to extract from this patch *robust* information. Gradient samples are accumulated into orientation histograms summarizing the contents over 4×4 subregions surrounding the key point of interest. Each of the 16 subregions corresponds to a 8-orientations bins histogram, leading to a 128 element feature for each key point (see Fig. 10). Two modifications are made in order to reduce the effects of illumination changes: histogram values are thresholded to reduce importance of large gradients (in order to deal with a strong illumination change such as camera saturation), and feature vectors are normalized to unit length (making them invariant to affine changes in illumination).

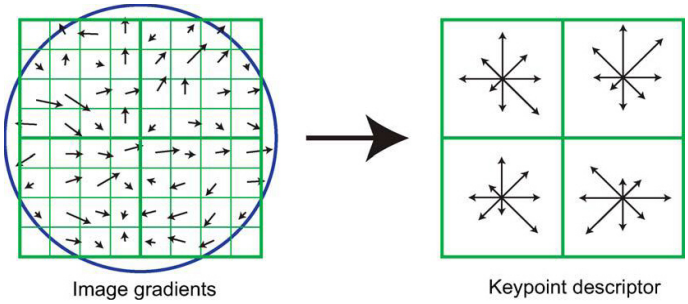

Figure 9: Example of a 2×2 descriptor array of orientation histograms (right) computed from an 8×8 set of samples (left). The orientation histograms are quantized into 8 directions and the length of each arrow corresponds to the magnitude of the histogram entry. (From [21])

3.4 Final matching

The outcome is for each image, a few hundreds or thousands SIFT descriptors associated with as many key points. The descriptors of any image can be compared to the descriptors of any other image, or belonging to a database of descriptors built up from many images. The only remaining question is to decide when two descriptors match, or not. In the terms of Lowe again:

> The best candidate match for each keypoint is found by identifying its nearest neighbor in the database of keypoints from training images. The nearest neighbor is defined as the keypoint with minimum Euclidean distance for the invariant descriptor vector. However, many features from an image will not have any correct match in the training database because they arise from background clutter or were not detected in the training images. Therefore, it would be useful to have a way to discard features that do not have any good match to the database. A global threshold on distance to the closest feature does not perform well, as some descriptors are much more discriminative than others. A more effective measure is obtained by comparing the distance of the closest neighbor to that of the second-closest neighbor. (...) This measure performs well because correct matches need to have the closest neighbor significantly closer than the closest incorrect match to achieve reliable matching. For false matches, there will likely be a number of other false matches within similar distances due to the high dimensionality of the feature space. We can think of the second-closest match as providing an estimate of the density of false matches within this portion of the feature space and at the same time identifying specific instances of feature ambiguity. (...) For our object recognition implementation, we reject all matches in which the distance ratio is greater than 0.8, which eliminates 90% of the false matches while discarding less than 5% of the correct matches.

4 Image acquisition model underlying SIFT

4.1 The camera model

We always work on the camera CCD plane, whose mesh unit is taken to be 1. We shall always assume that the camera pixels are indexed by $\mathbf{Z}^2$. The image sampling operator is therefore always $\mathbf{S}_1$. Our second assumption is that the digital initial image is well-sampled and obtained by a gaussian kernel. Thus, the digital image is $\mathbf{u} = \mathbf{S}_1 G_\delta A u_0$, where $\delta \geq \mathbf{c}$, and A is a similarity with positive determinant. (Lowe's original paper assumes $\mathbf{c} \simeq 0.5$, which amounts to assume a slight under-sampling of the original image).

Definition 1. *We model all digital frontal images obtained from a given ideal planar object whose frontal infinite resolution image is u_0 as*

$$\mathbf{u}_0 =: \mathbf{S}_1 G_\delta A u_0 \tag{5}$$

where A is a $A = RH_\lambda \mathcal{T}$ is the composition of a translation and of a similarity.

Taking into account the way the digital image is blurred and sampled in the SIFT method, we can now list the SIFT assumptions and formalize the method itself. The description is by far simpler if we do it without mixing in sampling issues. We need not mix them in, since the fact that images are well-sampled at all stages permits equivalently to describe all operations with the continuous images directly, and to deduce afterwards the discrete operators on samples.

4.2 Condensed description of the SIFT method

1. There is an underlying infinite resolution bounded planar image u_0;
2. The initial digital image is $\mathbf{S}_1 G_\delta A u_0$ where $\delta \geq 0.5$, and $A = RH_\lambda \mathcal{T}$ is the composition of a rotation, a zoom, and a translation;
3. the SIFT method computes a sufficient scale-space sampling of $u(\sigma, \mathbf{x}) = (G_\sigma G_\delta A u_0)(\mathbf{x})$, and deduces by the Newton method the accurate location or key points defined as extrema in scale-space of the spatial image Laplacian, $\Delta u(\sigma; \mathbf{x})$;
4. The blurred $u(\sigma, \cdot)$ image is then re-sampled around each characteristic point with sampling mesh $\sqrt{\sigma^2 + \mathbf{c}^2}$;
5. the directions of the sampling axes are fixed by a dominant direction of the gradient of $u(\sigma, \cdot)$ in a neighborhood, with size proportional to $\sqrt{\sigma^2 + \mathbf{c}^2}$ around the characteristic point;
6. the rest of the operations in the SIFT method is a contrast invariant encoding of the samples around each characteristic point. It is not needed for the discussion to follow.

5 Image operators formalizing SIFT

The analysis of the scale invariance is much easier on the continuous images whose samples form the digital image. The Shannon-Whittaker interpolation permits a perfect reconstruction of a

continuous image from a discrete one, when the continuous image has been well-sampled [41]. Under the assumption that a Gaussian filtering can deliver well-sampled images up to a negligible error, this section gives a formalized description of the SIFT procedure.

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

5.1 Sampling and interpolation

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

$$(\mathbf{S}_\delta u)(k, l) := u(k\delta, l\delta); \tag{6}$$

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

$$\mathbf{Iu}(x, y) := \sum_{(k,l)\in\mathbf{Z}^2} \mathbf{u}(k, l)\text{sinc}(x - k)\text{sinc}(y - l), \tag{7}$$

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

$$\mathbf{IS}_1 u = u. \tag{8}$$

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

5.2 The Gaussian filtering

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

$$G_\sigma G_\beta = G_{\sqrt{\sigma^2+\beta^2}}. \tag{9}$$

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

$$G_\sigma H_\gamma u = H_\gamma G_{\sigma\gamma} u. \tag{10}$$

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

$$\mathbf{G}_\delta \mathbf{u} =: \mathbf{S}_1 G_\delta \mathbf{I}\mathbf{u}. \tag{11}$$

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

$$\mathbf{G}_\delta \mathbf{G}_\beta = \mathbf{G}_{\sqrt{\delta^2+\beta^2}}. \tag{12}$$

Indeed, using twice (11) and once (9) and (8),

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

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

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

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

5.3 Formalized Scale Invariant Features Transform

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

1. **Geometry:** there is an underlying infinite resolution planar image $u_0(\mathbf{x})$ that has undergone a similarity Au_0 (modeling the composition of a rotation, a translation, and a homothety) before sampling.

2. **Sampling and blur:** the camera blur is assumed to be a Gaussian with standard deviation $\mathbf{c}$. The initial digital image is therefore $\mathbf{u} = \mathbf{S}_1 G_\mathbf{c} A u_0$;

3. **Sampled scale space:** the SIFT method computes enough samples of the scale space function $u(\sigma, \cdot) = G_\sigma G_\mathbf{c} A u_0$ to detect accurately "key points" $(\sigma, \mathbf{x})$, defined as scale and space local extrema of $\Delta u(\sigma, \cdot)$.

4. **Covariant resampling:** a 32×32 grid centered at the key point is used to sample $u(\sigma, \cdot)$ around each key point $(\sigma, \mathbf{x})$. The grid mesh is proportional to $\sqrt{\sigma^2 + \mathbf{c}^2}$. The directions of the sampling grid axes are fixed by a dominant direction of $\nabla u(\sigma, \cdot)$ in a neighborhood of the key point, whose size is also proportional to the key point scale σ. This yields for each key point a rotation, translation and scale invariant square sampled subimage samples in which the four parameters of A have been eliminated (see Fig. 11);

5. **Illumination invariance:** the final SIFT descriptors keep mainly the orientation of the samples gradient, to gain invariance with respect to light conditions.

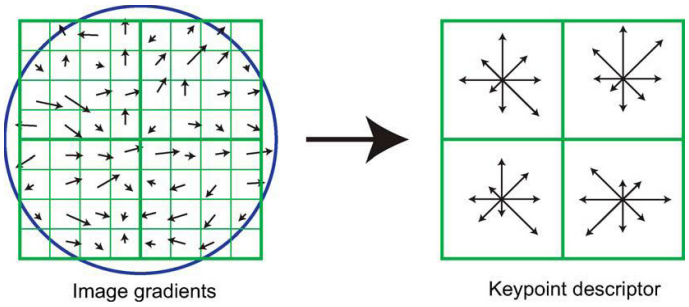

Figure 10: Each key-point is associated *a square image patch whose size is proportional to the scale and whose side direction is given by the assigned direction.* Example of a 2×2 descriptor array of orientation histograms (right) computed from an 8×8 set of samples (left). The orientation histograms are quantized into 8 directions and the length of each arrow corresponds to the magnitude of the histogram entry.

Figure 11: SIFT key points. The arrow starting point, length and the orientation signify respectively the key point position, scale, and dominant orientation. These features are claimed to be covariant to any image similarity.

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 be close enough to 1 to ensure a good key point localization. (Typically, straight edge points have only one large Hessian eigenvalue, are poorly localized, and are therefore ruled out by this second threshold, which is scale invariant.)

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

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

6 Scale and SIFT: consistency of the method

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

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

In this Section, in conformity with the SIFT model of Section 5, the digital image is a frontal view of an infinite resolution ideal image u_0. In that case, $A = H\mathcal{T}R$ is the composition of a rotation R, a translation $\mathcal{T}$ and a homothety H. Thus the digital image is $\mathbf{u} = \mathbf{S}_1 G_\delta H\mathcal{T}Ru_0$, for some H, $\mathcal{T}$, R. The next Lemma shows that SIFT is rotation- and translation-invariant.

Lemma 2. *For any rotation R and any translation $\mathcal{T}$, the SIFT descriptors of $\mathbf{S}_1 G_\delta H\mathcal{T}Ru_0$ are identical to those of $\mathbf{S}_1 G_\delta H u_0$.*

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

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

If $(s_0, \mathbf{x}_0)$ *is a key point of* w *satisfying* $s_0 \geq \max(\lambda\beta, \mu\delta)$*, then it corresponds to a key point of* u *at the scale* σ_1 *such that* $\lambda\sqrt{\sigma_1^2 + \beta^2} = s_0$*, whose SIFT descriptor is sampled with mesh* $\sqrt{\sigma_1^2 + \mathbf{c}^2}$*, where* $\mathbf{c}$ *is the tentative standard deviation of the initial image blur as described in Section 5.3. In the same way* $(s_0, \mathbf{x}_0)$ *corresponds to a key point of* v *at scale* σ_2 *such that* $s_0 = \mu\sqrt{\sigma_2^2 + \delta^2}$*, whose SIFT descriptor is sampled with mesh* $\sqrt{\sigma_2^2 + \mathbf{c}^2}$*.*

Proof. The interpolated initial images are by (8)

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

Computing the scale-space of these images amounts to convolve them for every $\sigma > 0$ with G_σ, which yields, using the Gaussian semigroup property (9) and the commutation relation (10):

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

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

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

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

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

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

$$s_0 = \lambda\sqrt{\sigma_1^2 + \beta^2} = \mu\sqrt{\sigma_2^2 + \delta^2}. \tag{13}$$

We recall that each SIFT descriptor at a key point $(\sigma_1, \mathbf{x}_1)$ is computed from space samples of $\mathbf{x} \to u(\sigma, \mathbf{x})$. The origin of the local grid is $\mathbf{x}_1$, the intrinsic axes are fixed by one of the dominant directions of the gradient of $u(\sigma_1, \cdot)$ around $\mathbf{x}_1$, in a circular neighborhood whose size is proportional to σ_1. The SIFT descriptor sampling rate around the key point is proportional to $\sqrt{\sigma_1^2 + \mathbf{c}^2}$ for $u(\sigma_1, \mathbf{x})$, and to $\sqrt{\sigma_2^2 + \mathbf{c}^2}$ for $u(\sigma_2, \mathbf{x})$, as described in Section 5.3. □

Theorem 2. *Let* $\mathbf{u}$ *and* $\mathbf{v}$ *be two digital images that are frontal snapshots of the same continuous flat image* u_0*,* $\mathbf{u} = \mathbf{S}_1 G_\beta H_\lambda \mathcal{T} R u_0$ *and* $\mathbf{v} := \mathbf{S}_1 G_\delta H_\mu u_0$*, taken at different distances, with different Gaussian blurs and possibly different sampling rates, and up to a camera translation and rotation around its optical axis. Without loss of generality, assume* $\lambda \leq \mu$*. Then if the initial blurs are identical for both images (if* $\beta = \delta = \mathbf{c}$*), then each SIFT descriptor of* $\mathbf{u}$ *is identical to a SIFT descriptor of* $\mathbf{v}$*. If* $\beta \neq \delta$ *(or* $\beta = \delta \neq \mathbf{c}$*), the SIFT descriptors of* $\mathbf{u}$ *and* $\mathbf{v}$ *become (quickly) similar when their scales grow, namely as soon as* $\frac{\sigma_1}{\max(\mathbf{c},\beta)} \gg 1$ *and* $\frac{\sigma_2}{\max(\mathbf{c},\delta)} \gg 1$*, where* σ_1 *and* σ_2 *are respectively the scales of the key points in the two images.*

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

$$\lambda\sqrt{\sigma_1^2 + \beta^2} = \mu\sqrt{\sigma_2^2 + \delta^2} \Rightarrow \lambda\sqrt{\sigma_1^2 + \mathbf{c}^2} = \mu\sqrt{\sigma_2^2 + \mathbf{c}^2}. \tag{14}$$

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

$$\lambda^2\beta^2 - \mu^2\delta^2 = (\lambda^2 - \mu^2)\mathbf{c}^2. \tag{15}$$

Since λ and μ correspond to camera distances to the observed object u_0, their values are arbitrary. Thus in general the only way to get (15) is to have $\beta = \delta = \mathbf{c}$, which means that the blurs of both images have been guessed correctly.

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

The almost perfect scale invariance of SIFT stated in Theorem 2 is illustrated by the striking example of Fig. 12. The 25 SIFT key points of a very small image $\mathbf{u}$ are compared to the 60 key points obtained by zooming in $\mathbf{u}$ by a 32 factor: The resulting digital image is $\mathbf{v} := \mathbf{S}_{\frac{1}{32}}\mathbf{I}_d\mathbf{u}$, again obtained by zero-padding. For better observability, both images are displayed with the same size by enlarging the pixels of $\mathbf{u}$. Almost each key point (18 out of 25) of $\mathbf{u}$ finds its counterpart in $\mathbf{v}$. 18 matches are detected between the descriptors as shown on the right. Let us check how this extreme example is covered by Theorem 2. We compare an initial image $\mathbf{u} = \mathbf{S}_1 G_\delta \mathbf{I}_d \mathbf{u}_0$ (with $\delta = \mathbf{c}$) with its zoomed in version $\mathbf{v} = \mathbf{S}_{\frac{1}{32}} G_\delta \mathbf{I}_d \mathbf{u}_0$. But we have by (10)

$$\mathbf{v} = \mathbf{S}_{\frac{1}{32}} G_\delta \mathbf{I}_d \mathbf{u}_0 = \mathbf{S}_1 H_{\frac{1}{32}} G_\delta \mathbf{I}_d \mathbf{u}_0 = \mathbf{S}_1 G_{32\delta} H_{\frac{1}{32}} \mathbf{I}_d \mathbf{u}_0.$$

Here the numerical application of the relations in the above proof give: We want (14) to hold approximately, where $\mu = 1$, $\lambda = \frac{1}{32}$, $\beta = 32\delta$. Thus we want $\frac{1}{32}\sqrt{\sigma_1^2 + (32\delta)^2} = \sqrt{\sigma_2^2 + \delta^2}$ to imply $\frac{1}{32}\sqrt{\sigma_1^2 + \mathbf{c}^2} \approx \sqrt{\sigma_2^2 + \mathbf{c}^2}$ which means $\sqrt{(\frac{\sigma_1}{32})^2 + \mathbf{c}^2} = \sqrt{\sigma_2^2 + \mathbf{c}^2}$ to imply $\sqrt{(\frac{\sigma_1}{32})^2 + (\frac{\mathbf{c}}{32})^2} \approx \sqrt{\sigma_2^2 + \mathbf{c}^2}$. This is true only if σ_1 is significantly larger than 32, which is true, since σ_1 is the scale of the SIFT descriptors in the image $\mathbf{v}$, which has been zoomed in by a 32 factor.

By the second part of Theorem 2 the reliability of the SIFT matching increases with scale. This fact is illustrated in Fig. 13. Starting from a high resolution image $\mathbf{u}_0$, two images $\mathbf{u}$ and $\mathbf{v}$ are obtained by simulated zoom out, $\mathbf{u} = \mathbf{S}_1 G_\beta H_\lambda \mathbf{I}_d \mathbf{u}_0 = \mathbf{S}_\lambda G_{\lambda\beta} \mathbf{I}_d \mathbf{u}_0$ and $\mathbf{v} = \mathbf{S}_\mu G_{\mu\delta} \mathbf{I}_d \mathbf{u}_0$, with $\lambda = 2$, $\mu = 4$, $\beta = \delta = 0.8$. Pairs of SIFT descriptors of $\mathbf{u}$ and $\mathbf{v}$ in correspondence, established by a SIFT matching, are compared using an Euclidean distance $\mathbf{d}$. The scale rate σ_1/σ_2 as well as the distance d between the matched key points are plotted against σ_2 in Fig. 13. That $\sigma_1/\sigma_2 \approx 2$

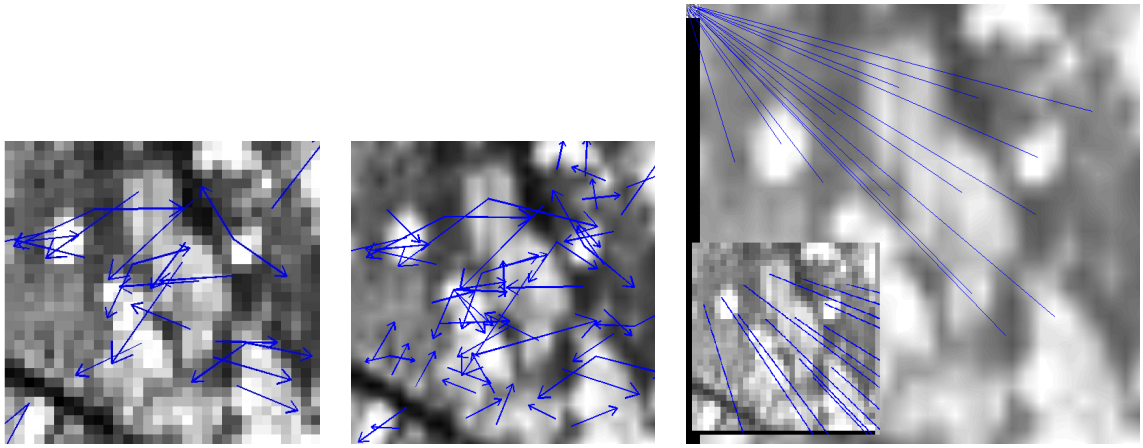

Figure 12: Scale invariance of SIFT, an illustration of Theorem 2. Left: a very small digital image $\mathbf{u}$ with its 25 key points. For the conventions to represent key points and matches, see the comments in Fig. 11. Middle: this image is over sampled by a 32 factor to $\mathbf{S}_{\frac{1}{32}}\mathbf{I}_d\mathbf{u}$. It has 60 key points. Right: 18 matches found between $\mathbf{u}$ and $\mathbf{S}_{\frac{1}{32}}\mathbf{I}_d\mathbf{u}$. A zoom of the small image $\mathbf{u}$ on the up-left corner is shown in the bottom left. It can be observed that all the matches are correct.

for all key points confirms that the SIFT matching process is reliable. As stated by the theorem, the rate σ_1/σ_2 goes to $\mu/\lambda = 2$ when σ_2 increases, and the distance $\mathbf{d}$ goes down. However, as also apparent in the numerical result, when the scale is small ($\sigma_2 < 1$), σ_1/σ_2 is very different from 2 and $\mathbf{d}$ is large.

7 Conclusion

Our overall conclusion is that no substantial improvement of the SIFT method can be ever hoped, as far as translation, rotation and scale invariance are concerned. As pointed out by several benchmarks, the robustness and repeatability of the SIFT descriptors outperforms other methods. However, such benchmarks mix three very different criteria that, in our opinion, should have been discussed separately. The first one is the formal invariance of each method when all thresholds have been eliminated. This formal invariance has been proved here for SIFT when the initial blurs of both images are equal to a known value $\mathbf{c}$, and it has been proved to be approximately true even with images having undergone very different blurs, like in the surprising experiment of fig. 13. The second main criterion is the clever fixing of several thresholds in the SIFT method ensuring robustness, repeatability, and a low false alarm rate. This one has been extensively tested and confirmed in previous benchmark papers (see also the recent and complete report [7]). We think, however, that the success of SIFT in these benchmarks is primarily due to its full scale invariance.

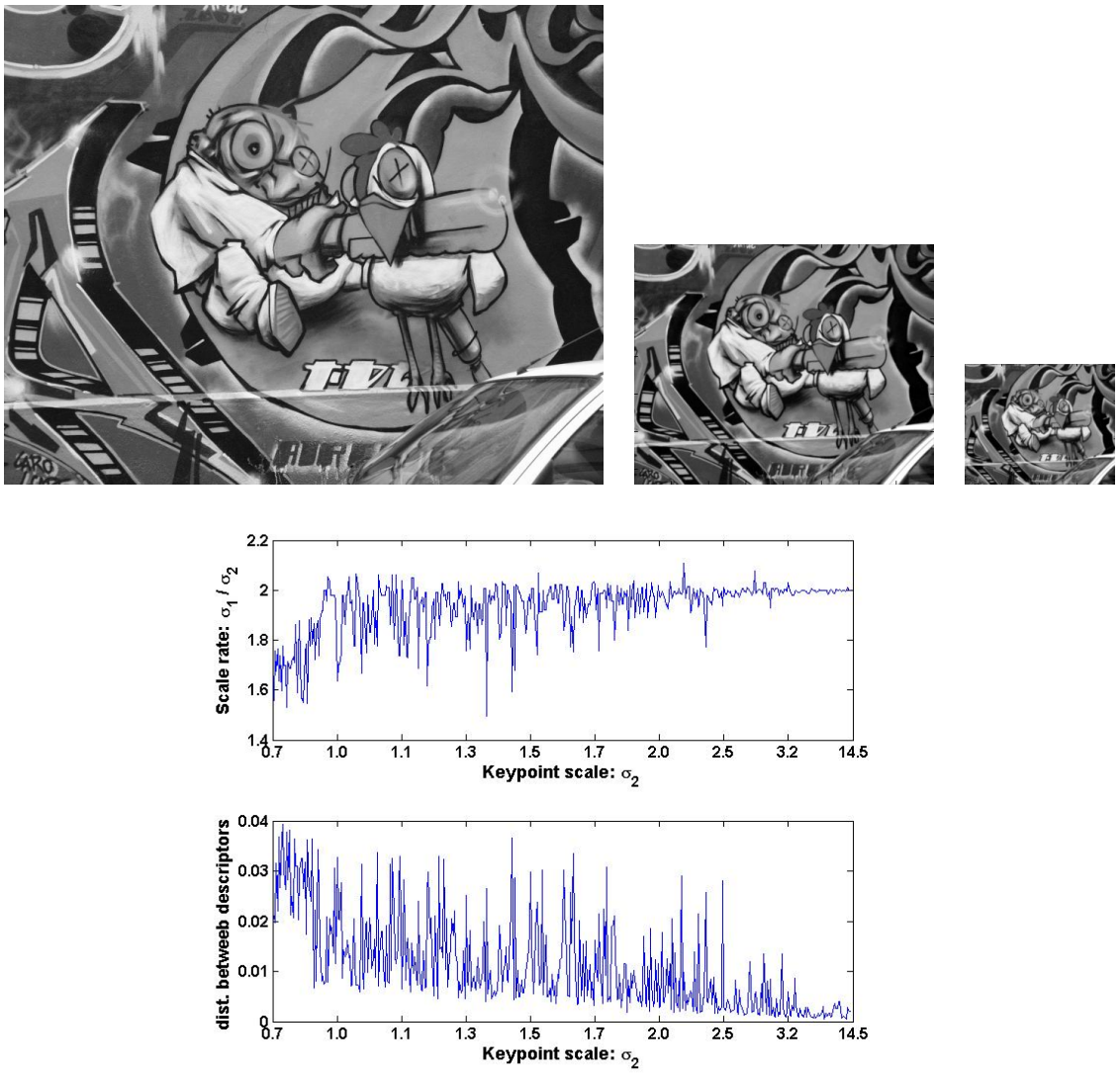

Figure 13: Top (from left to right): $\mathbf{u}_0$, $\mathbf{u}$, $\mathbf{v}$. Middle: Rate of scales σ_1/σ_2 of matched key points in $\mathbf{u}$ and $\mathbf{v}$ against σ_2. Bottom: Distance between matched descriptors of $\mathbf{u}$ and $\mathbf{v}$ against σ_2.

8 Projects

8.1 A famous competitor: the SURF method

Surf means: "Speeded up robust features." It is basically a SIFT method where everything has been indeed sped up and kept to the essentials. It is very used and quoted 1199 times in four years, according to Google scholar. In the terms of the authors:

> (...) we present a novel scale- and rotation-invariant interest point detector and descriptor, coined SURF (Speeded Up Robust Features). It approximates or even outperforms previously proposed schemes with respect to repeatability, distinctiveness, and robustness, yet can be computed and compared much faster. This is achieved by relying on integral images for image convolutions; by building on the strengths of the leading existing detectors and descriptors (in casu, using a Hessian matrix-based measure for the detector, and a distribution-based descriptor); and by simplifying these methods to the essential. This leads to a combination of novel detection, description, and matching steps. The paper presents experimental results on a standard evaluation set, as well as on imagery obtained in the context of a real-life object recognition application. Both show SURFs strong performance.

The goal of the project is first detail completely SURF, to discuss on the mathematical side its invariance properties, compared to SIFT, and possibly to explore and discuss available implementation source codes to compare them to the original paper. This kind of comparison is usually very enlightening.

8.2 A second famous competitor: the MSER method

The famous "Maximally stable extremal region" method extracts from images all contrasted shapes. These shapes receive affine invariant descriptors and are used to compare several snapshots of a scene from different viewpoints. This method is again acclaimed (for example quoted 1028 times in six years, according to Google Scholar). Very fast algorithms have been proposed since then to implement it.

In the terms of the authors:

> A new set of image elements that are put into correspondence, the so called extremal regions, is introduced. Extremal regions possess highly desirable properties: the set is closed under (1) continuous (and thus projective) transformation of image coordinates and (2) monotonic transformation of image intensities. An efficient (near linear complexity) and practically fast detection algorithm (near frame rate) is presented for an affinely invariant stable subset of extremal regions, the maximally stable extremal regions (MSER).
>
> A new robust similarity measure for establishing tentative correspondences is proposed. The robustness ensures that invariants from multiple measurement regions (regions obtained by invariant constructions from extremal regions), some that are significantly

larger (and hence discriminative) than the MSERs, may be used to establish tentative correspondences.

The high utility of MSERs, multiple measurement regions and the robust metric is demonstrated in wide-baseline experiments on image pairs from both indoor and outdoor scenes. Significant change of scale (3.5), illumination conditions, out-of-plane rotation, occlusion, locally anisotropic scale change and 3D translation of the viewpoint are all present in the test problems. Good estimates of epipolar geometry (average distance from corresponding points to the epipolar line below 0.09 of the inter-pixel distance) are obtained.

The goal of the project will be:

- Mathematics: discuss the affine invariance of the method, and its scale invariance (compared to SIFT)
- algorithm: study the implementation and the fast implementations, the source code available, propose a simple and robust version for IPOL.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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