Chapter 4

From continuous to digital images, and back

Consider a continuous and bounded image $u(\mathbf{x})$ defined for every $\mathbf{x} = (x, y) \in \mathbb{R}^2$. All continuous *image* operators including the sampling will be written in capital letters A, B and their composition as a mere juxtaposition AB. For any affine map A of the plane consider the affine transform of a continuous image u defined by $Au(\mathbf{x}) =: u(A\mathbf{x})$. 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, $Ru =: u \circ R$ is the image rotation by R^{-1}.

Sampling and interpolation

Digital images, only defined for $(n_1, n_2) \in \boldsymbol{Z}^2$, will be denoted by $\mathbf{u}(n_1, n_2)$. The δ-sampled image $\mathbf{u} = \mathbf{S}_\delta u$ is defined on $\boldsymbol{Z}^2$ by

$$\mathbf{u}(n_1, n_2) = (\mathbf{S}_\delta)u(n_1, n_2) =: u(n_1\delta, n_2\delta); \tag{4.1}$$

Conversely, the Shannon interpolate of a digital image $\mathbf{u}$ is defined as follows [99]. Let $\mathbf{u}$ be a digital image, defined on $\boldsymbol{Z}^2$ and such that $\sum_{n\in\boldsymbol{Z}^2} |\mathbf{u}(n)|^2 < \infty$ and $\sum_{n\in\boldsymbol{Z}^2} |\mathbf{u}(n)| < \infty$. (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 $I\mathbf{u}$ of $\mathbf{u}$ the only $L^2(\mathbb{R}^2)$ function u having $\mathbf{u}$ as samples and with spectrum support contained in $(-\pi, \pi)^2$. $u = I\mathbf{u}$ is defined by the Shannon-Whittaker formula

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

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

$$I\mathbf{S}_1 u = u. \tag{4.2}$$

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

4.0.1 The practical Shannon interpolation: zero-padding

Of course, the Shannon interpolate is unpractical in that it assumes the knowledge of infinitely many samples. In practice image samples and image interpolation will be performed on rectangle. For a sake of simplicity we describe here what happens on a square. Let $a > 0$ and consider a function u from $[0, a]^2$ to $\mathbb{R}$ such that $u(x + a, y + a) = u(x, y)$. Fix an integer N, and consider the N^2 samples of u, $\mathbf{u}_{k,l} = (\mathbf{S}u)\left(\frac{ka}{N}, \frac{la}{N}\right)$ on $[0, a]^2$.

Definition 4.1. *The discrete Fourier transform (DFT) of the N^2 samples $u = (u_{k,l})_{k,l=0,1,...,N-1}$ is the double sequence of discrete Fourier coefficients for $m, n \in \{-\frac{N}{2}, ..., \frac{N}{2} - 1\}$ defined by*

$$DFT(u)_{m,n} = \tilde{u}_{m,n} = \frac{1}{N^2} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} u_{k,l} \omega_N^{-mk} \omega_N^{-nl}, \tag{4.3}$$

where $\omega_N =: e^{\frac{2i\pi}{N}}$ is the first N-root of 1.

Proposition 4.2. *Consider the trigonometric polynomial*

$$I\mathbf{u}(x, y) = \sum_{m,n=-\frac{N}{2}}^{\frac{N}{2}-1} \tilde{u}_{m,n} \exp(\frac{2i\pi mx}{a}) \exp(\frac{2i\pi ny}{a}). \tag{4.4}$$

Then its coefficients $\tilde{u}_{m,n}$ are the only complex numbers such that for every $k, l \in \{0, ..., N–1\}$, $I\mathbf{u}\left(\frac{ka}{N}, \frac{la}{N}\right) = \mathbf{u}\left(\frac{ka}{N}, \frac{la}{N}\right)$. In consequence, the discrete inverse transform of the DFT $\mathbf{u} \to \tilde{u}$ is nothing but the calculation of the value of the polynomial at the samples $(\frac{ka}{N}, \frac{la}{N})$, $0 \leq k, l \leq N-1$. In other terms, setting $\tilde{u} = DFT(\mathbf{u})$, the inverse transform DFT^{-1} is given by

$$\mathbf{u}(k, l) = \sum_{-\frac{N}{2}}^{\frac{N}{2}-1} \sum_{-\frac{N}{2}}^{\frac{N}{2}-1} \tilde{\mathbf{u}}_{m,n} \omega_N^{km+ln},$$

for every $k, l = 0, 1, \ldots N - 1$.

Exercise 4.1. Recall that $\omega_N = \exp\left(\frac{2i\pi}{N}\right)$, N-th root of 1. Show that $\sum_{k=0}^{N-1} \omega_N^k = 0$, and that $\sum_{k=0}^{N-1} \omega_N^{kl} = 0$ for $l \neq 0$ modulo N and finally that for every k_0, $\sum_{k=k_0}^{k_0+N-1} \omega_N^{kl} = 0$ for all $l \neq 0$ modulo N. Using these relation show the above proposition, namely that $DFT(DFT^{-1}) = Id$. ■

In conclusion, the interpolation and sampling operators we shall consider both in theory and practice are the usual sampling $\mathbf{S}$, implicitly restricted to a square. The inverse interpolation operator I is defined by (4.4), and Proposition 4.2 tells us that $\mathbf{S}I\mathbf{u} = \mathbf{u}$. The next statement gives the converse statement.

Proposition 4.3. *If $u(x, y)$ is a a-periodic band-limited function, then it is a trigonometric polynomial. If its highest degree is $\frac{N}{2} - 1$, with N even, then its coefficients $\tilde{u}_{m,n}$ are obtained by DFT from the samples $\mathbf{u}(k, l) = u(\frac{ka}{N}, \frac{la}{N})$. In consequence for such functions we have $I\mathbf{S}u = u$.*

Exercise 4.2. Give a detailed proof of Proposition 4.3. It is a direct consequence of Proposition 4.2. ■

In the rest of this chapter and of the book, we shall always take the functional setting of Proposition 4.3.

Zoom in by zero-padding

Let $(u_{k,l})$ be a digital image and define its zoomed version $(v_{i,j})_{i,j=0,\ ...,\ 2N-1}$ as the inverse discrete Fourier transform of $\tilde{v}_{i,j}$ defined for $i,j=-N,\ ...,\ N-1$ by

$$\tilde{v}_{m,n}=\tilde{u}_{m,n} \text{ if } -\frac{N}{2}\leq m,n\leq \frac{N}{2}-1,\ \ \tilde{v}_{m,n}=0 \text{ otherwise.} \tag{4.5}$$

Proposition 4.4. *The image v whose Discrete Fourier Transform is given by (4.5) satisfies $v_{2k,2l}=u_{k,l}$, for $k,l=0,...,N-1$.*

Proof. Here is the proof in dimension 1:

$$v_{2k}=\sum_{-N}^{N-1}\tilde{v}_n\omega_{2N}^{2nk}=\sum_{-\frac{N}{2}}^{\frac{N}{2}-1}\tilde{u}_n\omega_N^{nk}=u_k.$$

Indeed, $\omega_{2N}^{2nk}=\omega_N^{nk}$. □

Exercise 4.3. Prove Proposition (4.4) in two dimensions. ■

4.1 The Gaussian semigroup

For a sake of simple notation, G_σ denotes the convolution operator on $\mathbb{R}^2$ with the gauss kernel $G_\sigma(x_1,x_2)=\frac{1}{2\pi(\sigma)^2}e^{-\frac{x_1^2+x_2^2}{2(\sigma)^2}}$, namely $G_\sigma\mathbf{u}(x,y)=:(G_\sigma * \mathbf{u})(x,y)$. Notice that the paramaterization of the gaussian is *not* the same as the parameterization used for the heat equation. To make a difference in notation, we use Greek letters for the new parameter. It is easily checked that G_σ satisfies the semigroup property

$$G_\sigma G_\beta = G_{\sqrt{\sigma^2+\beta^2}}. \tag{4.6}$$

Exercise 4.4. Prove (4.6). ■

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

$$G_\sigma H_\gamma u = H_\gamma G_{\sigma\gamma} u. \tag{4.7}$$

Exercise 4.5. Prove (4.7). ■

Discrete Gaussians

Many algorithms in computer vision and image processing assume that all blurs can be assumed gaussian. Thus, it will be crucial to prove that gaussian blur gives *in practice* well-sampled images. Thus, in all that follows, we are dealing with initial digital images obtained by sampling a continuous image with gaussian blur, $u=\mathbf{S}_1G_\mathbf{c}u_0$;

Another question we need to deal with is how to perform a gaussian convolution on a discrete (digital) image. This is valid if and only if a discrete convolution can give an account of the underlying continuous one.

Definition 4.5. *The* **discrete gaussian convolution** *applied to a digital image* $\mathbf{u}$ *is defined as a digital operator by*

$$\mathbf{G}_\delta \mathbf{u} =: \mathbf{S}_1 G_\delta I \mathbf{u}. \tag{4.8}$$

Proposition 4.6. *This definition maintains the gaussian semi-group property,*

$$\mathbf{G}_\delta \mathbf{G}_\beta = \mathbf{G}_{\sqrt{\delta^2+\beta^2}}. \tag{4.9}$$

Proof Indeed, using twice (4.8) and once (4.6) and (4.2),

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

□

The SIFT method that we will study in detail uses repeatedly the semi-group formula and a 2-sub-sampling of images with a gaussian blur larger than 1.6. These SIFT sampling manoeuvres are valid if and only if the *empirical proposition* below is true.

Proposition 4.7. *For every* σ *larger than 0.8 and every continuous and bounded image* u_0*, the gaussian blurred image* $G_\sigma u_0$ *is well sampled, namely* $I\mathbf{S}_1 G_\sigma u_0 = G_\sigma u_0$*.*

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

4.2 The right gaussian blur for 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 in this section 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 t factor. In the case of gaussian blur, because of the semi-group formula (4.6), the relation between the absolute and relative blur is

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

which yields

$$\mathbf{c}_r(t) = \mathbf{c}_a \sqrt{t^2 - 1}. \tag{4.10}$$

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

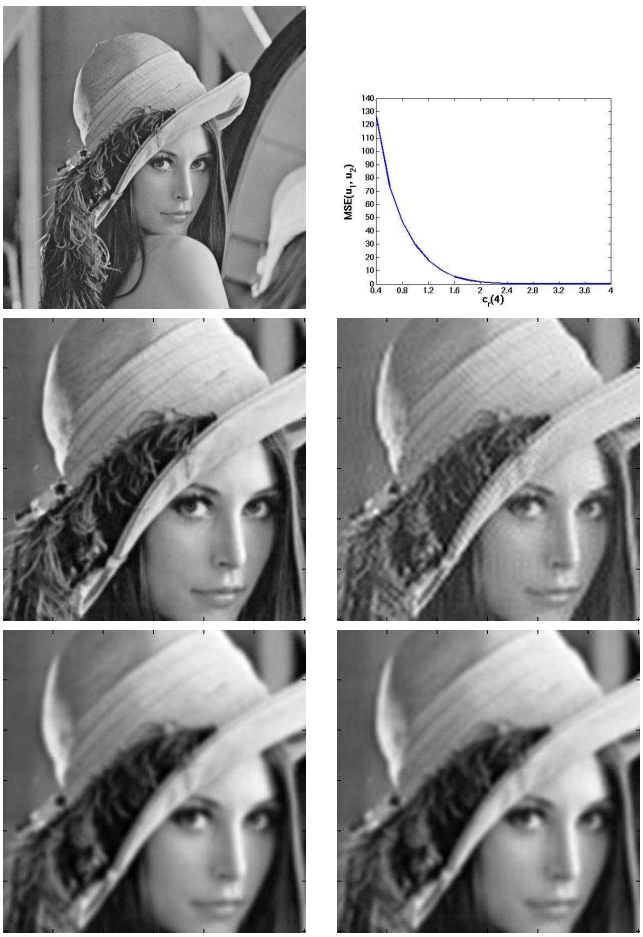

Figure 4.1: Top left: $\mathbf{u}$. Top right: MSE$(\mathbf{u}_1, \mathbf{u}_2)$ vs $\mathbf{c}_r(4)$. Middle (from left to right): $\mathbf{u}_1$ and $\mathbf{u}_2$ with $\mathbf{c}_r(4) = 1.2$. MSE$(\mathbf{u}_1, \mathbf{u}_2)$=17.5. Bottom (from left to right): $\mathbf{u}_1$ and $\mathbf{u}_2$ with $\mathbf{c}_r(4) = 2.4$. MSE$(\mathbf{u}_1, \mathbf{u}_2)$=0.33. Digital images are always displayed by coloring each square pixel with its central sample value.

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

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

where $\mathbf{u}$ is an initial digital image that is (intuitively) 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 (4.8). Since t is an integer, the t sub-sampling is trivial.

If the anti-aliasing filter size $\mathbf{c}_r(t)$ is too small, $\mathbf{u}_1$ and $\mathbf{u}_2$ can be very different. The right value of $\mathbf{c}_r(t)$ should be the smallest value permitting $\mathbf{u}_1 \approx \mathbf{u}_2$. Fig. 4.1 shows $\mathbf{u}_1$ and $\mathbf{u}_2$ with $t = 4$ and plots their mean square error $\mathrm{MSE}(\mathbf{u}_1, \mathbf{u}_2)$. An anti-aliasing filter with $\mathbf{c}_r(4) = 1.2$ is clearly not broad enough: $\mathbf{u}_2$ 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 $\mathrm{MSE}(\mathbf{u}_1, \mathbf{u}_2)$ decays rapidly until $\mathbf{c}_r(4)$ gets close to 2.4, and is stable and small thereafter. By (4.10), 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 $\mathbf{u}$ is equal to $\mathbf{c}_a$.

In a second experiment, $\mathbf{c}_a$ has been evaluated directly by using a binary image $\mathbf{u}_0$ that does not contain any blur. As illustrated in Fig. 4.2, $\mathbf{u}_0$ is obtained by binarizing the digital test image Lena (Fig. 4.1), the threshold being the median value of Lena. Since $\mathbf{u}_0$ 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

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

As shown in Fig. 4.2, $\mathbf{c}_a = 0.6$ is the smallest value ensuring no visual ringing in $\mathbf{u}_2$. Under this value, for example for $\mathbf{c}_a = 0.3$, clear ringing artifacts are present in $\mathbf{u}_2$. That $\mathbf{c}_a = 0.6$ is the correct value is confirmed by the $\mathrm{MSE}(\mathbf{u}_1, \mathbf{u}_2)$ 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.

4.2.1 Discrete sampling

If $\mathbf{u}$ is digital and therefore only defined on $\boldsymbol{Z}^2$ and if δ is an integer, then one can define any sub- or over-sampling operations on $\mathbf{u}$. But this requires interpolating $\mathbf{u}$ first.

Definition 4.8. *Thus we define a digital re-sampling operator by*

$$\mathcal{S}_\delta \mathbf{u} =: \mathbf{S}_\delta I \mathbf{u}. \tag{4.11}$$

$\mathcal{S}_\delta$ is a discrete filter. If $\delta < 1$ $\mathcal{S}_\delta$ is an over-sampling, and it is invertible. If $\delta > 1$ it is an sub-sampling, and may be not invertible.

Exercise 4.6. Show that if $\delta < 1$, then $\mathcal{S}_{\delta^{-1}} \mathcal{S}_\delta = Id$. What can happen if $\delta > 1$? ■

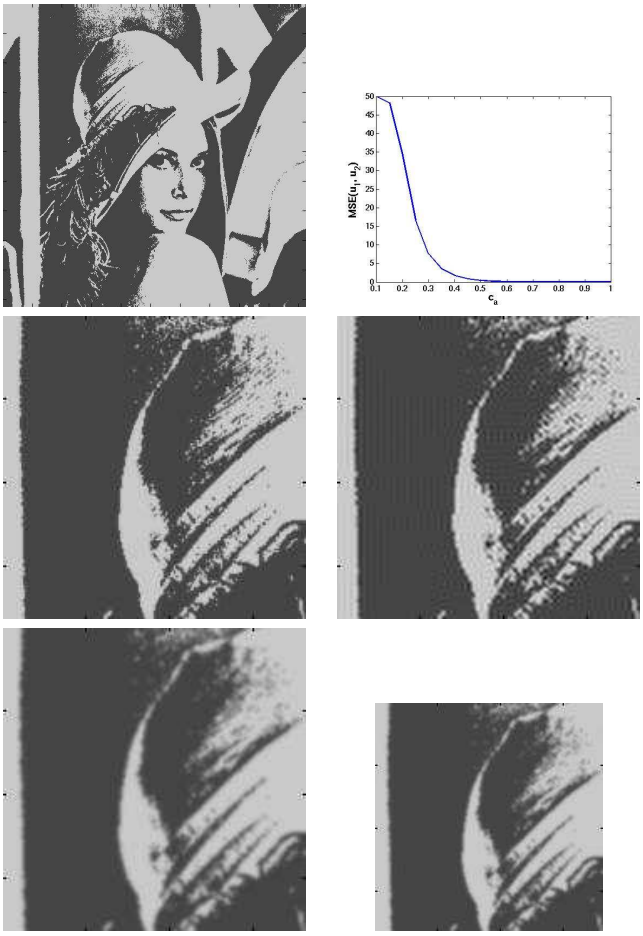

Figure 4.2: Top left: $\mathbf{u}$. Top right: MSE($\mathbf{u}_1, \mathbf{u}_2$) vs $\mathbf{c}_a$. Middle (from left to right): $\mathbf{u}_1$ and $\mathbf{u}_2$ with $\mathbf{c}_a = 0.3$. MSE($\mathbf{u}_1, \mathbf{u}_2$)=7.46. Bottom (from left to right): $\mathbf{u}_1$ and $\mathbf{u}_2$ with $\mathbf{c}_a = 0.6$. MSE($\mathbf{u}_1, \mathbf{u}_2$)=0.09.

Over-sampling can be interpreted as *a zoom in*. A zoom in is not the same as a *blow up*. Blow up is a photographic term involving the use of a system of lenses increasing the image resolution : it permits to see more details of the observed object. Zoom in is instead is a digital term. Being just an interpolation, it adds no detail to the image. The next proposition confirms that it is just an image enlargement.

Proposition 4.9. *For every* $\gamma \leq 1$,

$$I\mathcal{S}_\gamma \mathbf{u} = H_\gamma I\mathbf{u}. \tag{4.12}$$

Proof. $I\mathbf{u}$ is well sampled, with spectrum in $[0, 2\pi]^2$. Thus since $\gamma < 1$, $I\mathcal{S}_\gamma \mathbf{u}$ is over-sampled: it has spectrum in $[0, 2\pi\gamma]^2$. Thus $I\mathcal{S}_\gamma \mathbf{u}$ is band limited, as $H_\gamma I\mathbf{u}$. Since

$$I\mathcal{S}_\gamma \mathbf{u}(n_1, n_2) = \mathcal{S}_\gamma \mathbf{u}(n_1, n_2) = I\mathbf{u}(n_1\gamma, n_2\gamma) \text{ and } H_\gamma I\mathbf{u}(n_1, n_2) = I\mathbf{u}(n_1\gamma, n_2\gamma),$$

both functions have the same $\mathbb{Z}^2$ samples and therefore coincide. □

Corollary 4.10. *If* $\gamma \leq 1$, *then*

$$\mathcal{S}_\beta \mathcal{S}_\gamma = \mathcal{S}_{\beta\gamma}. \tag{4.13}$$

Proof. using once (4.12) and twice (4.11),

$$\mathcal{S}_\beta \mathcal{S}_\gamma \mathbf{u} = \mathbf{S}_\beta I\mathcal{S}_\gamma \mathbf{u} = \mathbf{S}_\beta \mathbf{H}_\gamma I\mathbf{u} = \mathbf{S}_{\beta\gamma} I\mathbf{u} = \mathcal{S}_{\beta\gamma}\mathbf{u}.$$

□

Proposition 4.11. *A discrete commutation formula : Assume* $\mathbf{u}$ *is a digital image. Then for* $\gamma < 1$,

$$\mathbf{G}_\beta \mathcal{S}_\gamma \mathbf{u} = \mathcal{S}_\gamma \mathbf{G}_{\beta\gamma}\mathbf{u}. \tag{4.14}$$

Proof.

$$\mathbf{G}_\beta \mathcal{S}_\gamma \mathbf{u} \overset{(4.8)}{=} \mathbf{S}_1(G_\beta I\mathcal{S}_\gamma \mathbf{u}) \overset{(4.12)}{=} \mathbf{S}_1(G_\beta H_\gamma I\mathbf{u})$$

$$\overset{(4.7)}{=} \mathbf{S}_1(H_\gamma(G_{\beta\gamma} I\mathbf{u})) \overset{(4.1)}{=} \mathbf{S}_\gamma(G_{\beta\gamma} I\mathbf{u}) \overset{(4.2)}{=} \mathbf{S}_\gamma I\mathbf{S}_1(G_{\beta\gamma} I\mathbf{u}) \overset{(4.8,4.11)}{=} \mathcal{S}_\gamma \mathbf{G}_{\beta\gamma}\mathbf{u}.$$

Notice that we use $I\mathbf{S}_1 u = u$ with $u = G_{\beta\gamma} I\mathbf{u}$. Indeed, this last function is well sampled, because Iu is. □

Chapter 5

The SIFT Method

This chapter is devoted to Lowe's *Scale-Invariant Feature Transform* (SIFT [166]), a very efficient image comparison method. 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. This method uses as fundamental tool the heat equation or, in other terms, the linear scale space. The heat equation is used to simulate all zooms out of both images that have to be compared. Indeed, these images may contain similar objects taken at different distances. But at least two of the simulated zoomed in images should contain these objects at the same apparent distance. This is the principal ingredient of the SIFT method, but other invariance requirements must be addressed as well.

Sect. 5.2 gives a detailed description of the SIFT shape encoding method. Sect. 5.4 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. Chapter. 4.2 checked the validity of this crucial gaussian blur assumption.

5.1 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.

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,

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

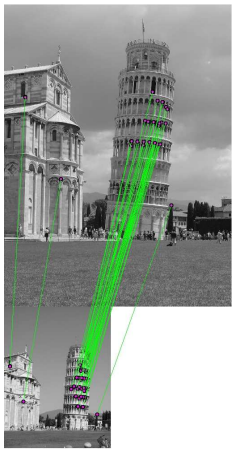

Figure 5.1: A result of the SIFT method, using an outliers elimination method [219]. Pairs of matching points are connected by segments.

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. 5.5).

Figs 5.1 and 5.6 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 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.

5.2 A Short Guide to SIFT Encoding

The SIFT encoding algorithm consists of four steps: detection of scale-space extrema (Sect. 5.2.1), accurate localization of key points (Sect. 5.2.2), and descriptor construction (Sect. 5.2.3).

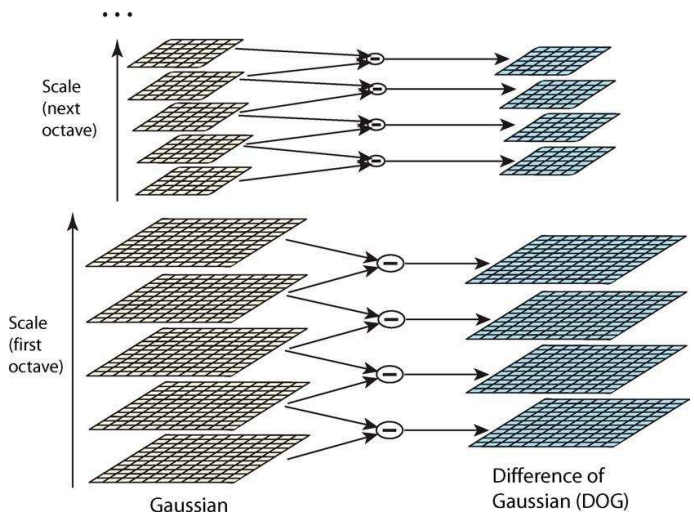

Figure 5.2: Gaussian pyramid for key points extraction (from [166])

5.2.1 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 [178], 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 $k = \sqrt{2}$.*

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*

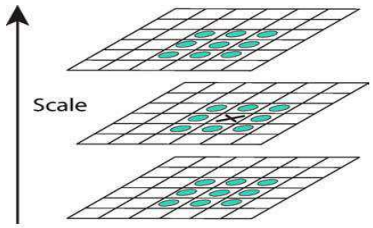

Figure 5.3: Neighborhood for the location of key points (from [166]). Local extrema are detected by comparing each sample point in $\mathbb{D}$ with its eight neighbors at scale σ and its nine neighbors in the scales above and below.

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. 5.2 and 5.3.

Exercise 5.1. Show that the gaussian G_σ parameterized by its standard deviation σ satisfies as stated by Lowe the time-dependent heat equation $\frac{\partial G}{\partial \sigma} = \sigma \Delta G$. ■

5.2.2 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:

$$\mathbf{y} = -\left(D^2\mathbb{D}(\mathbf{x}_0)\right)^{-1}(D\mathbb{D}(\mathbf{x}_0)), \tag{5.1}$$

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

Exercise 5.2. Check that (5.1) 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:

$$\frac{\mathrm{Tr}(H)^2}{\mathrm{Det}(H)} < \frac{(r+1)^2}{r}. \tag{5.2}$$

This rules out standard edge points and puts points of interest at locations which are strong enough extrema, or saddle points.

Exercise 5.3. Explain why (5.2) 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? ■

5.2.3 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. 5.5). 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 5.4: 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.

5.2.4 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

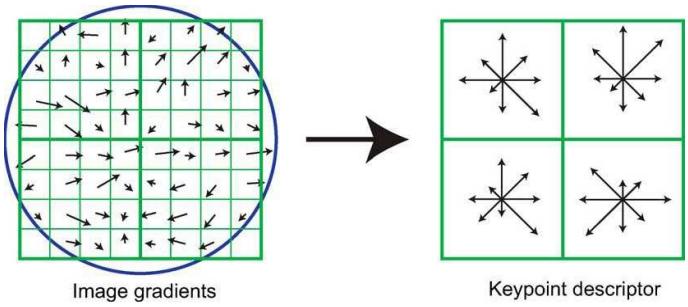

Figure 5.5: Example of a 2×2 descriptor array of orientation histograms (right) computed from an 8×8 set of samples (left). The orientation histograms are quantized into 8 directions and the length of each arrow corresponds to the magnitude of the histogram entry. (From [166])

> 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.

5.3 Image acquisition model underlying SIFT

5.3.1 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 $\boldsymbol{Z}^2$. 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 $u = \mathbf{S}_1 G_\delta A u_0$, where $\delta \geq \mathbf{c}$, $\mathbf{c} \simeq 0.6$ ensures well-sampling (see Chapter 4.2), and A is a similarity with positive determinant.

Definition 5.1. *We model all digital images obtained from a given ideal planar object whose frontal infinite resolution image is u_0 as*

$$\mathbf{u}_0 =: \mathbf{S}_1 G_\delta A u_0 \tag{5.3}$$

where $\delta \geq \mathbf{c}$ and A is any affine map.

So the possibility of aliasing (under-sampling, $\delta < \mathbf{c}$ is discarded). 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. We refer to section 4.2.1 for this passage from continuous to discrete operations in the well-sampled world.

5.3.2 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 $\delta \geq \mathbf{c}$, and $A = RH_\lambda \mathcal{T}$ is the composition of a rotation, a zoom, and a translation;

3. at all scales $\sigma > 0$, the SIFT method computes good samplings of $u(\sigma, \cdot) = G_\sigma G_\delta A u_0$ and deduces by the Newton method key points; The computation of these key points involves the computation of spatial derivatives of $u(\sigma, \cdot)$. Notice that if $u(\sigma, \cdot)$ is well-sampled, then so are its derivatives;

4. the blurred $u(\sigma, \cdot)$ image is then sampled around each characteristic point at a pace proportional to $\sqrt{\sigma^2 + \mathbf{c}^2}$. For a sake of simplicity we shall assume without loss of generality in our discussion that the fixed proportion is 1, so that the sampling is at $\sqrt{\sigma^2 + \mathbf{c}^2}$ rate;

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 σ around the characteristic point.

The rest of the operations in the SIFT method is just a contrast invariant encoding of the samples around each characteristic point. It is not needed for the discussion to follow.

5.4 Scale and SIFT: consistency of the method

In this section, in conformity with the SIFT model of Sect. 5.3.2, the digital image is a frontal view of an infinite resolution ideal image u_0. In that case, $A = H\mathcal{T}R$ is the composition of a homothety H, a translation $\mathcal{T}$ and a rotation R. Thus the digital image is $u = \mathbf{S}_1 G_\delta H \mathcal{T} R u_0$, for some H, $\mathcal{T}$, R as above. Assuming that the image is not aliased boils down, by the experimental results of Sect. 4.2, to assuming $\delta \geq 0.6$. (Lowe always takes $\delta = 0.8$.)

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

Lemma 5.2. *All of the aforementioned operators weakly commute. In addition, R and G commute strongly.*

Exercise 5.4. Check all of the mentioned weak and strong commutations and give their exact formula. There are four kinds of operators: translations, rotations, homotheties, gaussian convolutions. Thus, there are 6 verifications to make. ■

Lemma 5.3. *For any rotation R and any translation $\mathcal{T}$, the SIFT descriptors of $\mathbf{S}_1 G_\delta H \mathcal{T} R u_0$ 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 5.2), 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. □

Exercise 5.5. The above proof uses that gradients and Laplacians are covariant with respect to translations and rotations. Prove that for any C^2 function, $\Delta(Ru) = R(\Delta u)$ and $D(Ru) = RDu$. ■

Lemma 5.4. *Let u and v be two digital images that are frontal snapshots of the same continuous flat image u_0, $u = \mathbf{S}_1 G_\beta H_\lambda u_0$ and $v =: \mathbf{S}_1 G_\delta H_\mu u_0$, taken at different distances, with different gaussian blurs and possibly different sampling rates. Let $w(\sigma, \mathbf{x}) = (G_\sigma u)(\mathbf{x})$ denote the scale space of u. Then the scale spaces of u and v are*

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

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 $\sqrt{\sigma_1 + \mathbf{c}^2}$. 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 (4.2)

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

Computing the scale-space of these images amounts to convolve these images for every $\sigma > 0$ with G_σ, which yields, using the commutation relation (4.7) and the semigroup property (4.6):

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

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

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

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

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

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) $u(\sigma, \mathbf{x})$ and at scale σ_2 for (the Laplacian of) $\mathbf{v}(\sigma, \mathbf{x})$ satisfying

$$s_0 = \lambda\sqrt{\sigma_1^2 + \beta^2} = \mu\sqrt{\sigma_2^2 + \delta^2}. \tag{5.4}$$

□

Theorem 5.5. *Let u and v be two digital images that are frontal snapshots of the same continuous flat image u_0, $u = \mathbf{S}_1 G_\beta H_\lambda \mathcal{T} R u_0$ and $v =: \mathbf{S}_1 G_\delta H_\mu u_0$, taken at different distances, with different gaussian blurs and possibly different sampling rates, and up to a camera translation and rotation around its optical axe. Without loss of generality, assume $\lambda \leq \mu$. Then if the camera blurs are standard ($\beta = \delta = \mathbf{c}$), 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 5.3, we can neglect the effect of translations and rotations. Therefore assume w.l.o.g. that the images under comparison are as in Lemma 5.4. Assume a key point $(s_0, \mathbf{x}_0)$ of 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 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 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 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

$$\lambda^2\beta^2 - \mu^2\delta^2 = (\lambda^2 - \mu^2)\mathbf{c}^2. \tag{5.5}$$

Since λ and μ are proportional to camera distances to the observed object u_0, they are arbitrary and generally different. Thus, the only way to ensure (5.5) is to have $\beta = \delta = \mathbf{c}$, which means that the blurs of both images (or of both cameras) are ideal and gaussian). In any case, $\beta = \delta = \mathbf{c}$ does imply that the SIFT descriptors of both images are identical.

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 $\lambda\sqrt{\sigma_1^2 + \mathbf{c}^2} \simeq \mu\sqrt{\sigma_2^2 + \mathbf{c}^2}$. □

The almost perfect scale invariance of SIFT stated in Theorem 5.5 is illustrated by the striking example of Fig. 5.6. 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 $v = \mathbf{S}_{\frac{1}{32}} I u$, 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 5.5, 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.

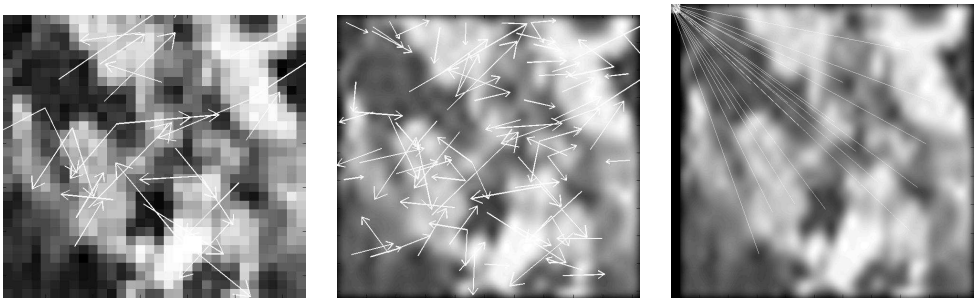

Figure 5.6: Scale invariance of SIFT, an illustration of Theorem 5.5. Left: a very small digital image u with its 28 key points. For the conventions to represent key points and matches, see the comments in Fig. 5.4. Middle: this image is over sampled by a 32 factor to $\mathbf{S}_{\frac{1}{32}} Iu$. It has 86 key points. Right: 22 matches found between u and $\mathbf{S}_{\frac{1}{32}} Iu$.

5.5 Exercises

Exercise 5.6. The aim of the exercise is to explain why the experiment of Fig. 5.6 works, and to illustrate Theorem 5.5. The digital zoom in by a factor λ is nothing but the over-sampling operator $\mathcal{S}_{\frac{1}{\lambda}}$ with sampling step $\frac{1}{\lambda}$, defined in (4.11). Here, $\lambda = 32$. In the experiment an original digital image $\mathbf{u} = \mathbf{S}_1 G_\delta u$ is zoomed into $\mathbf{v} = \mathcal{S}_{\frac{1}{\lambda}} \mathbf{u}$.

1) Using the definition of the discrete zoom and the right commutation relations given in this chapter and in the former one (give their numbers), show that

$$\mathbf{v} = \mathcal{S}_{\frac{1}{\lambda}} G_\delta u = \mathbf{S}_1 G_{\lambda\delta} H_{\frac{1}{\lambda}} u.$$

2) Is $\mathbf{v}$ well-sampled if $\mathbf{u}$ was?

3) By applying carefully Theorem 5.5, assuming that $\delta \simeq \mathbf{c}$, discuss why SIFT manages to match SIFT descriptors of $\mathbf{u}$ and $\mathbf{v}$. ■

5.6 Comments and references

Many variations exist on the computation of interest points, following the pioneering work of Harris and Stephens [115]. The Harris-Laplace and Hessian-Laplace region detectors [186, 189] are invariant to rotation and scale changes. Some moment-based region detectors [160, 29] including Harris-Affine and Hessian-Affine region detectors [187, 189], an edge-based region detector [256], an intensity-based region detector [256], an entropy-based region detector [137], and two independently developed level line-based region detectors MSER ("maximally stable extremal region") [182] and LLD ("level line descriptor") [198, 200, 202] are designed to be invariant to affine transformations. These two methods stem from the Monasse image registration method [193] 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 [191]. However, as pointed out in [166], 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 [166], 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 [159], 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 [139] and gradient location-orientation histogram (GLOH) [190], that claim to have better robustness and distinctiveness with scaled-down complexity have been developed ever since [93, 157]. Demonstrated to be superior to other descriptors [188, 190], SIFT has been popularly applied for scene recognition [86, 196, 230, 261, 105, 240] and detection [94, 206], robot localization [31, 208, 133], image registration [275], image retrieval [114], motion tracking [257, 142], 3D modeling and reconstruction [224, 262], building panoramas [3, 39], or photo management [274, 155, 55].

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 Chapter 4.2. The last criterion is the clever fixing of several thresholds in the SIFT method ensuring robustness, repeatability, and a low false alarm rate. This one has been extensively tested and confirmed in previous benchmark papers (see also the very recent and complete report [87]). We think, however, that the success of SIFT in these benchmarks is primarily due to its full scale invariance.