On the consistency of the SIFT METHOD (Scale Invariant Feature Transformation)

Guoshen Yu, Jean-Michel Morel

http://www.ipol.im/pub/algo/my_affine_sift/

Figure 1: Various snapshots of a “Mural”

An image matching method should be:

1. invariant to illuminance changes;
2. independent of the viewpoint, and therefore covariant by a subgroup of the projective group;
3. insensitive to the noise inherent to any image acquisition device;
4. robust to partial occlusions, and therefore local enough;
5. robust to scaling.

PLAN : A mathematical analysis of Lowe's *Scale-Invariant Feature Transform* (SIFT) method

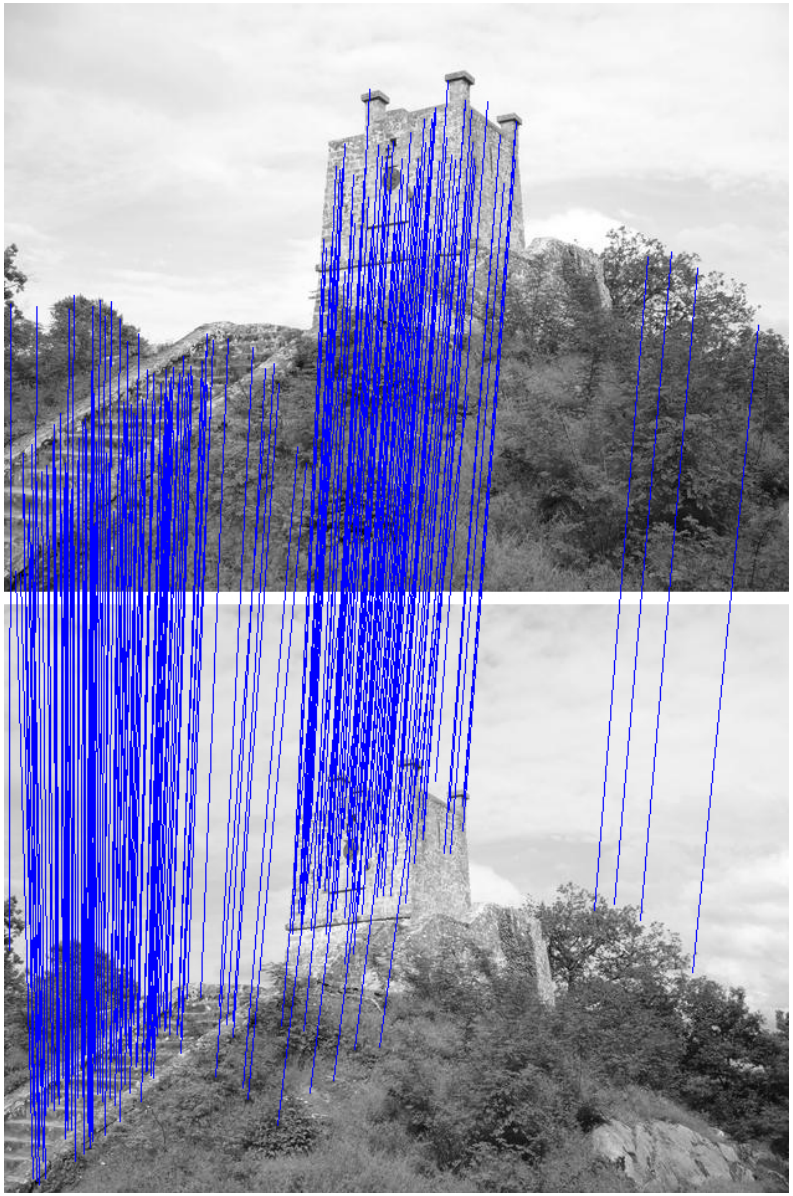
- Striking examples
- Scale space and Gaussian convolution
- Comprehensive description of the SIFT local invariant encoding method
- Proof that : SIFT is indeed similarity invariant, exactly if the images have the same blur, otherwise approximately

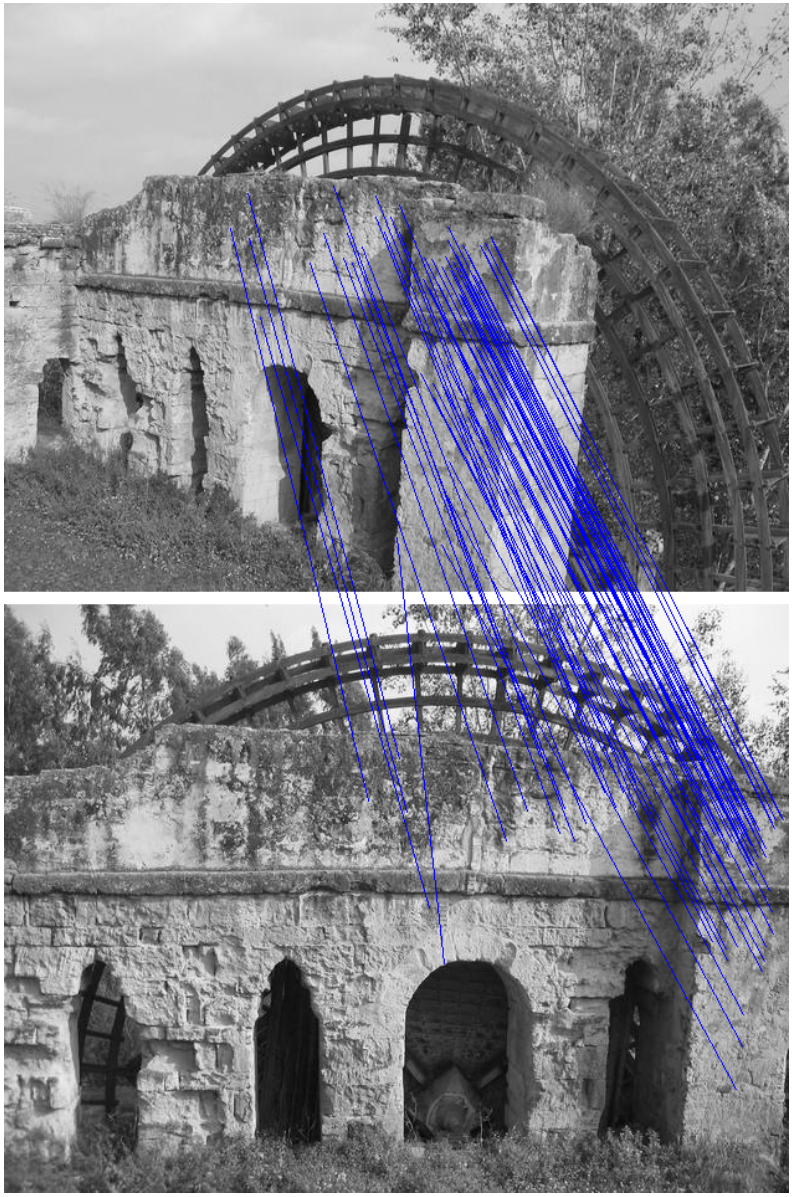
D. G. Lowe, Object recognition from local scale-invariant features International Conference on Computer Vision, 2, 1150–1157, 1999

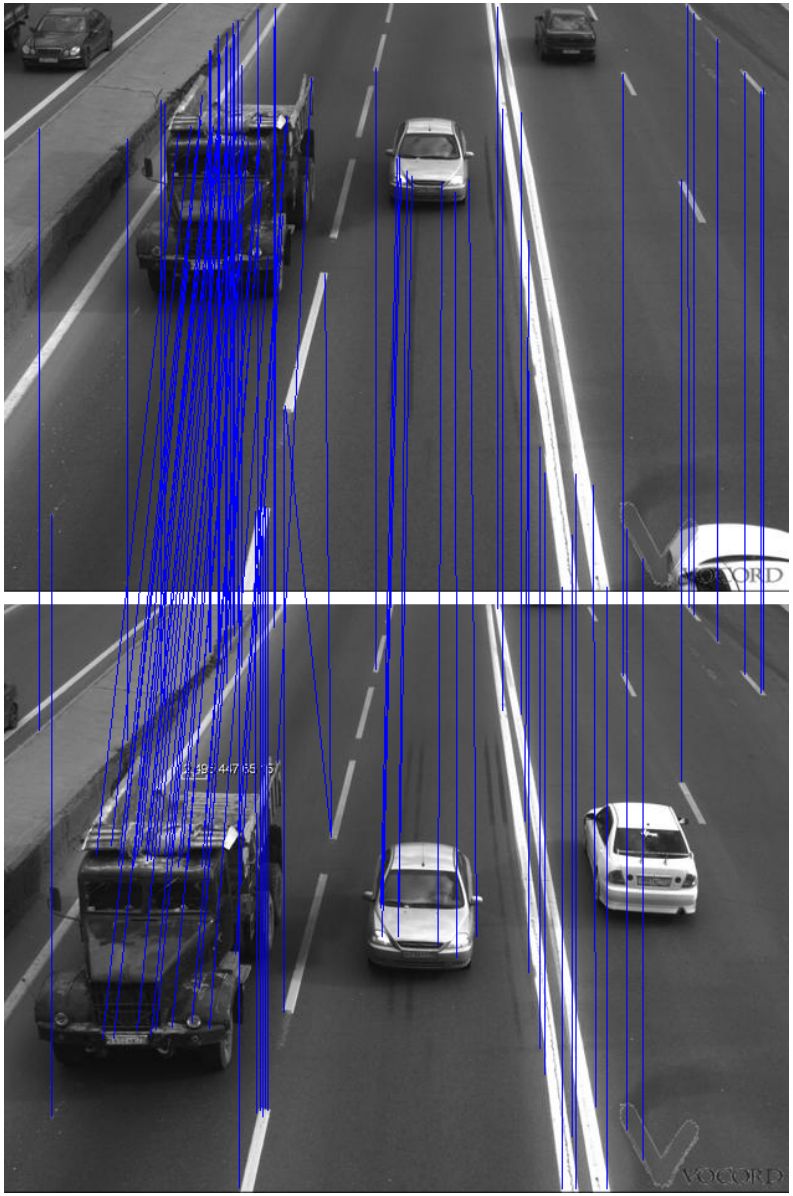
D. G. Lowe, Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision, 60(2):91-110, 2004

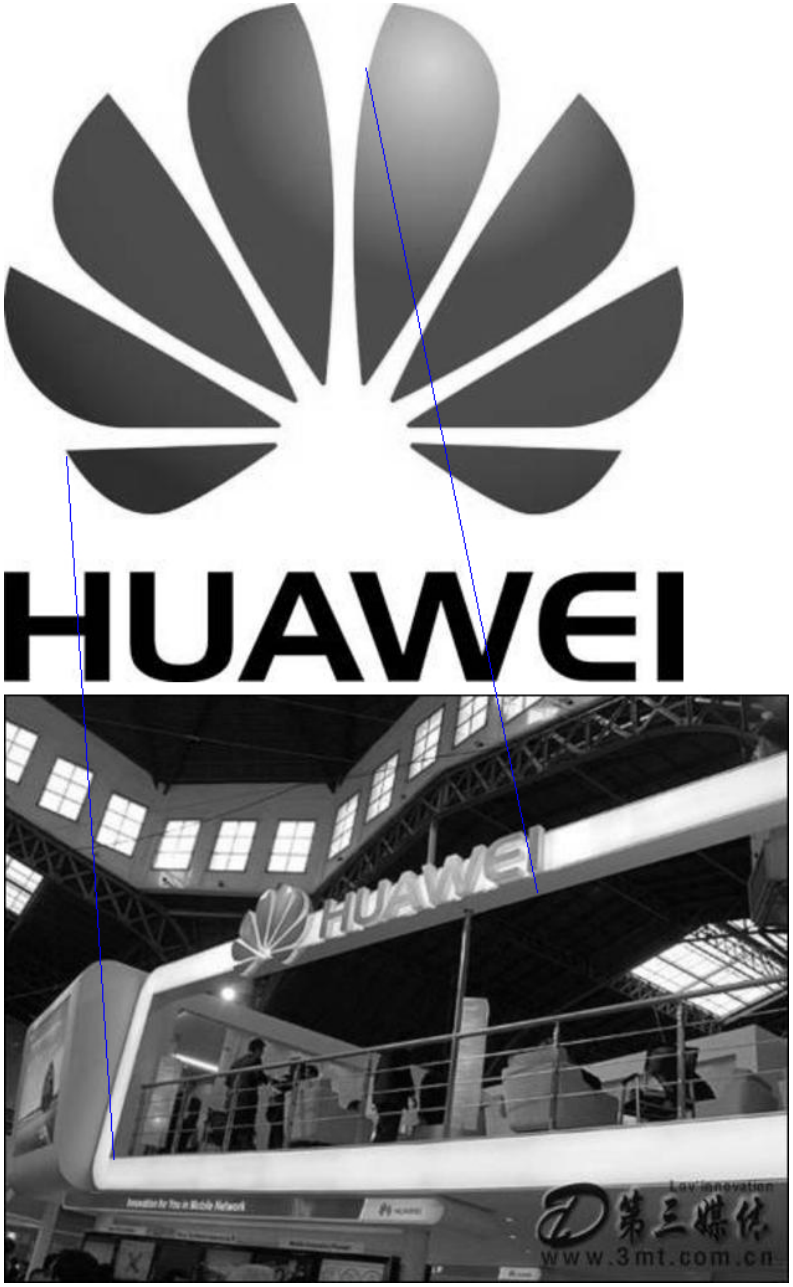
Guoshen Yu, J.M. Morel: Is SIFT scale invariant? Imaging Inverse Problems, 2010

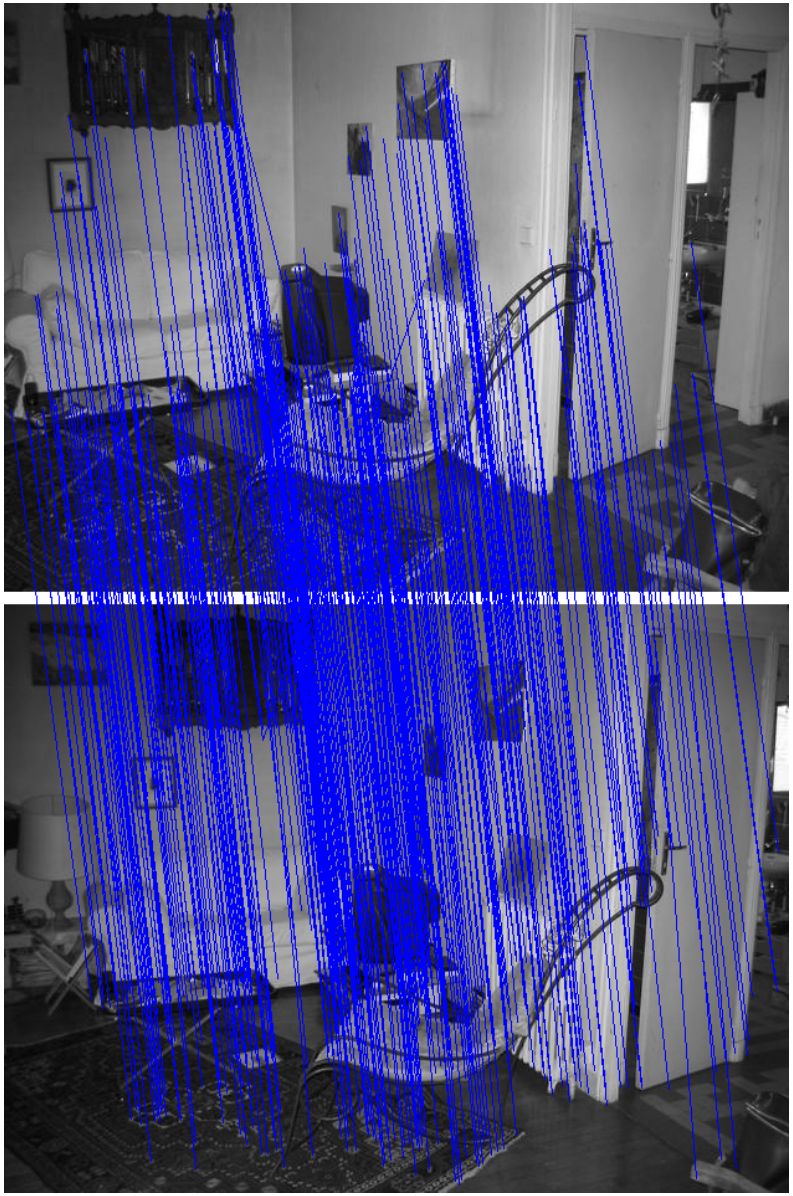
STRIKING EXAMPLES

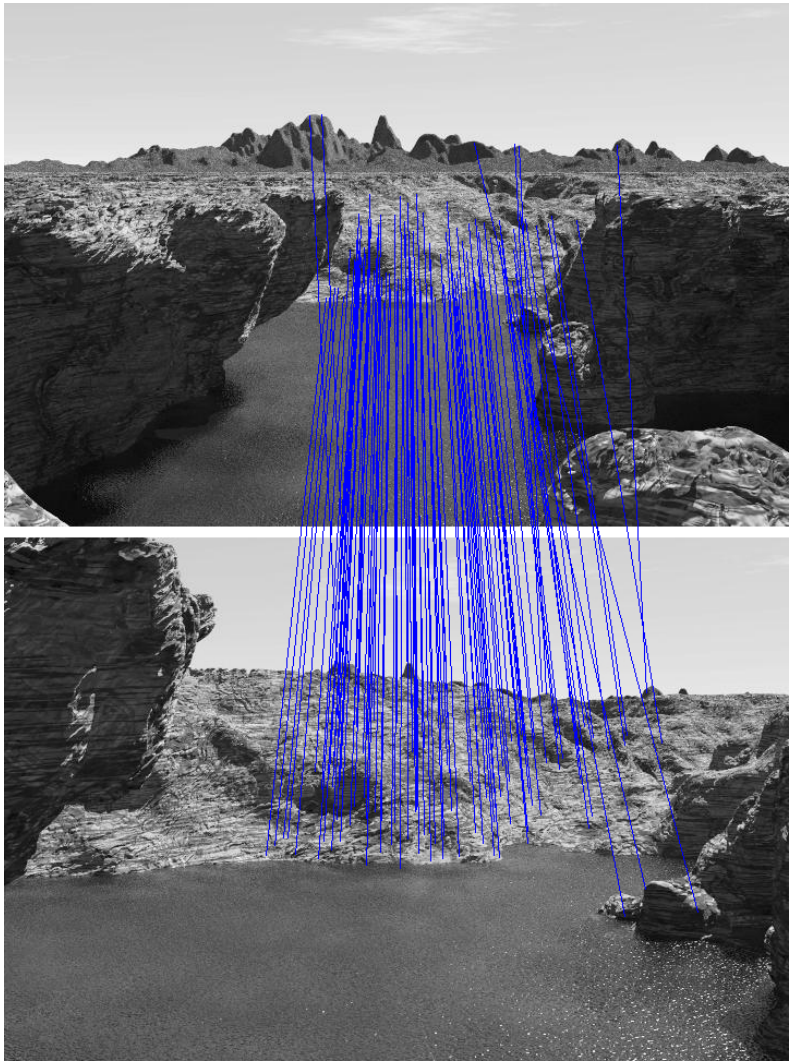


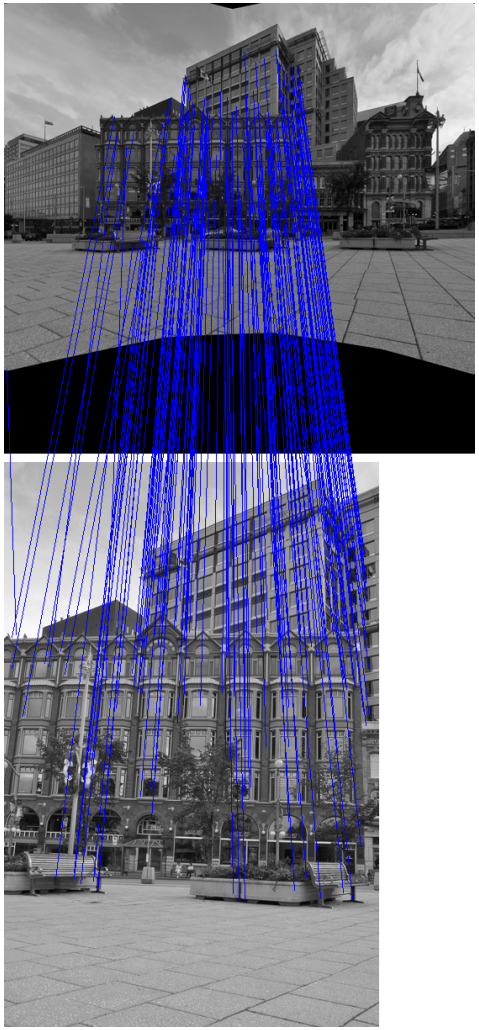


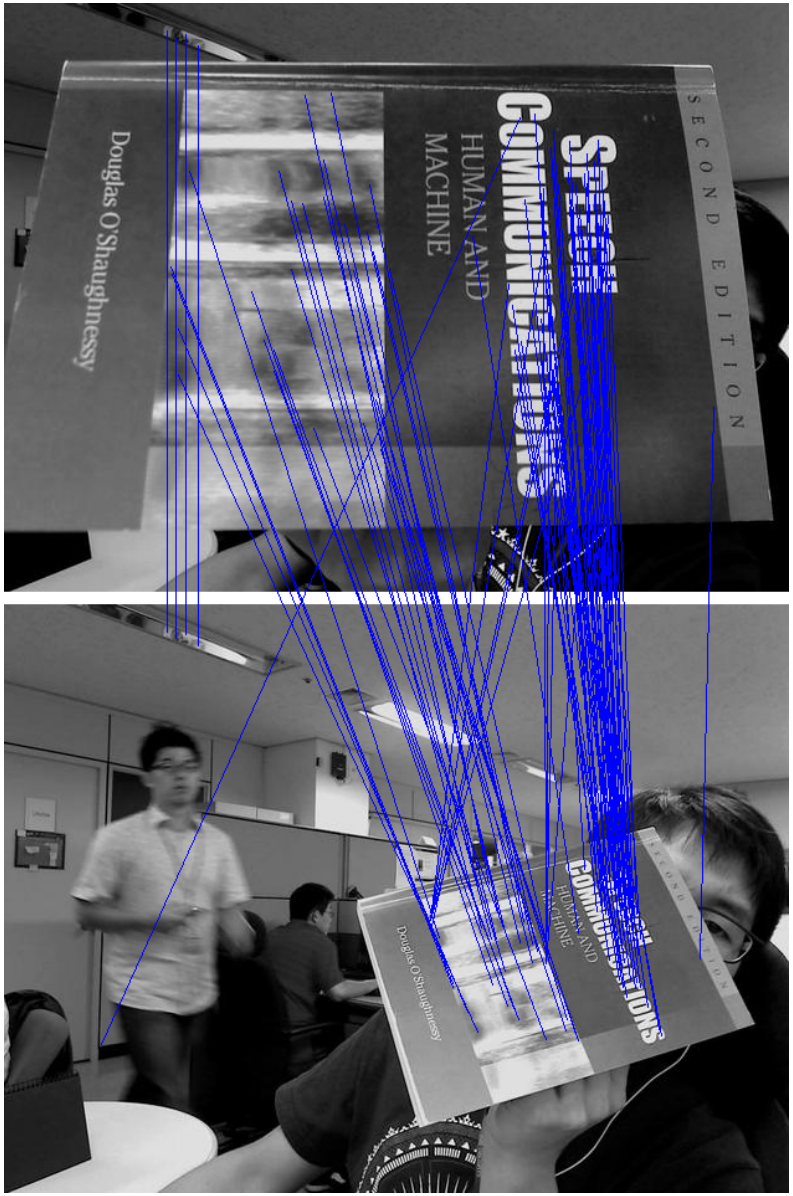


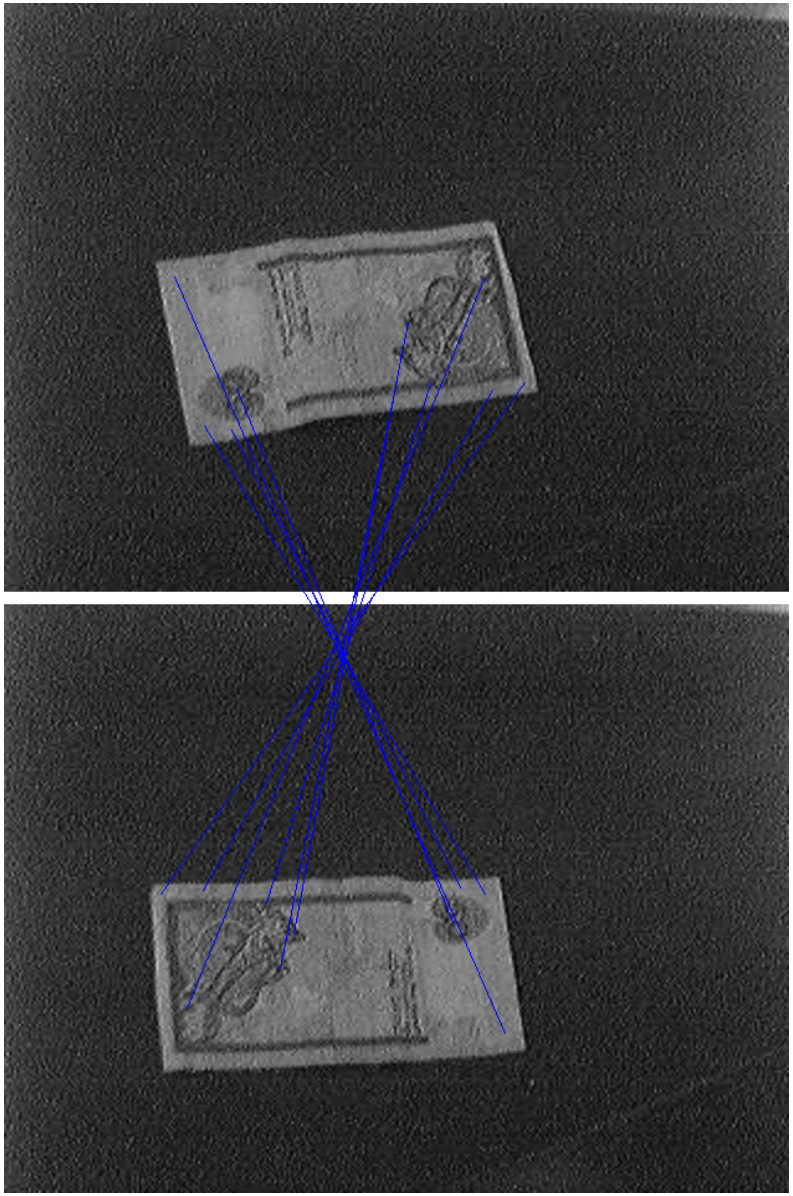

HUAWEI
HUAWEI
www.3mt.com.cn

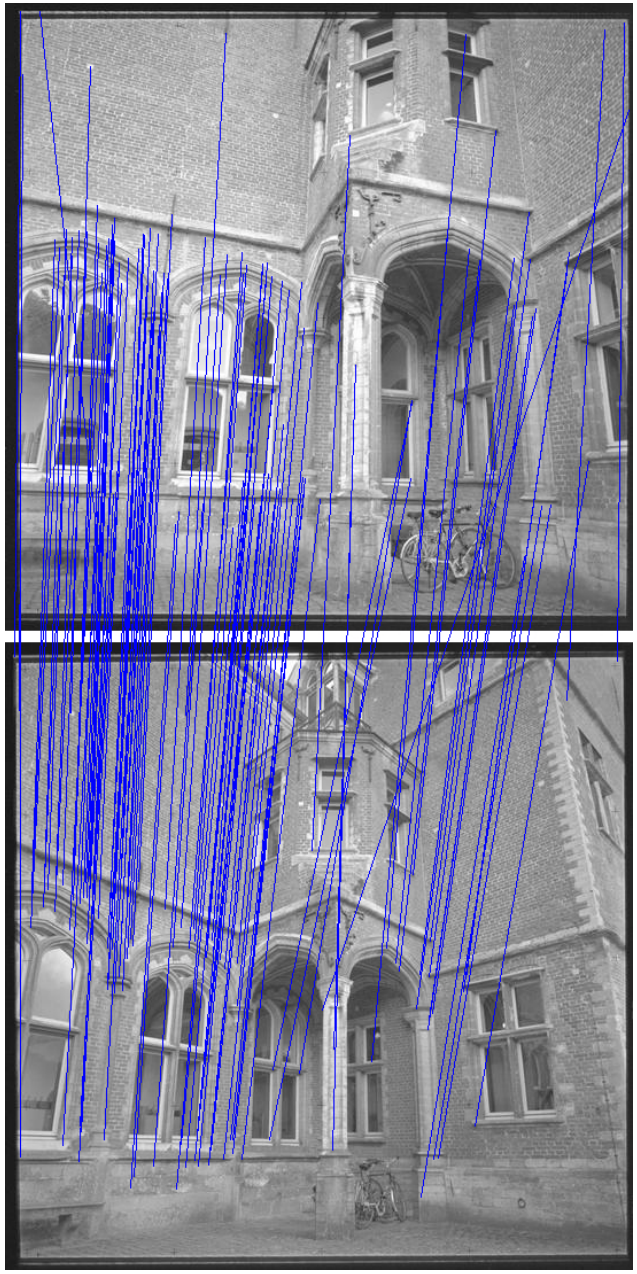


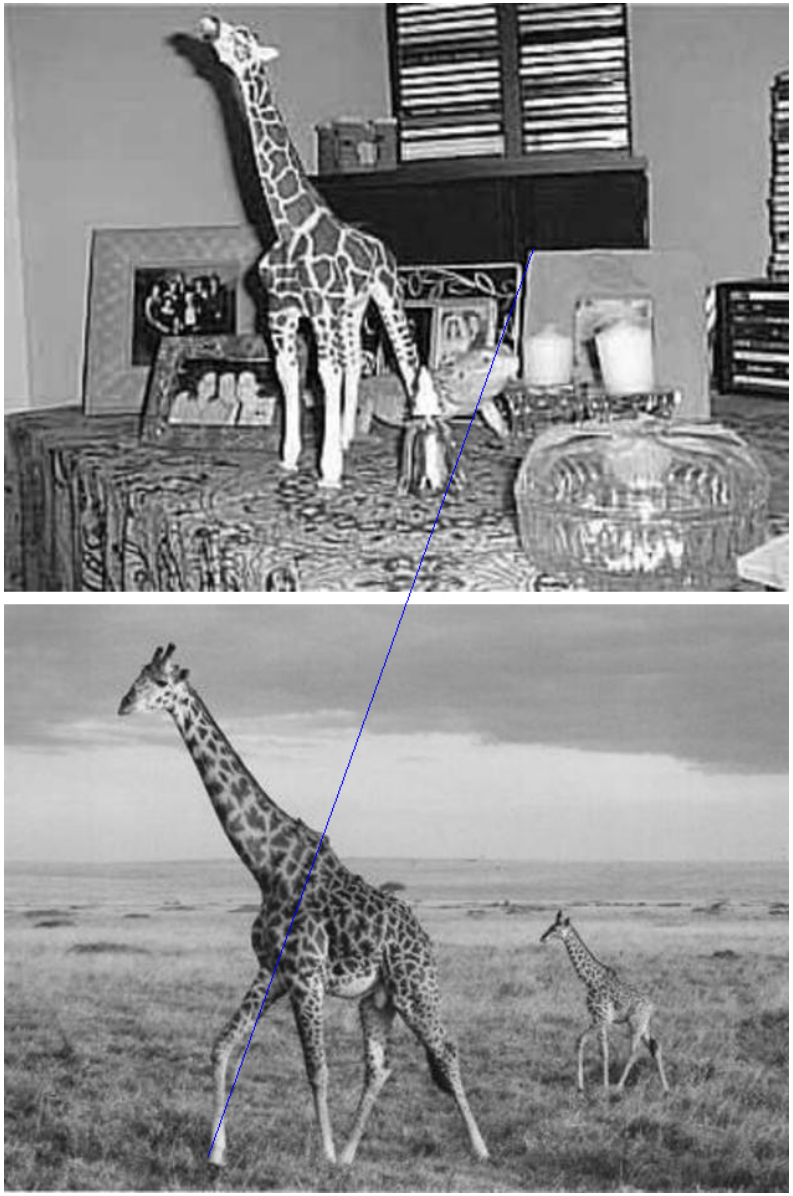




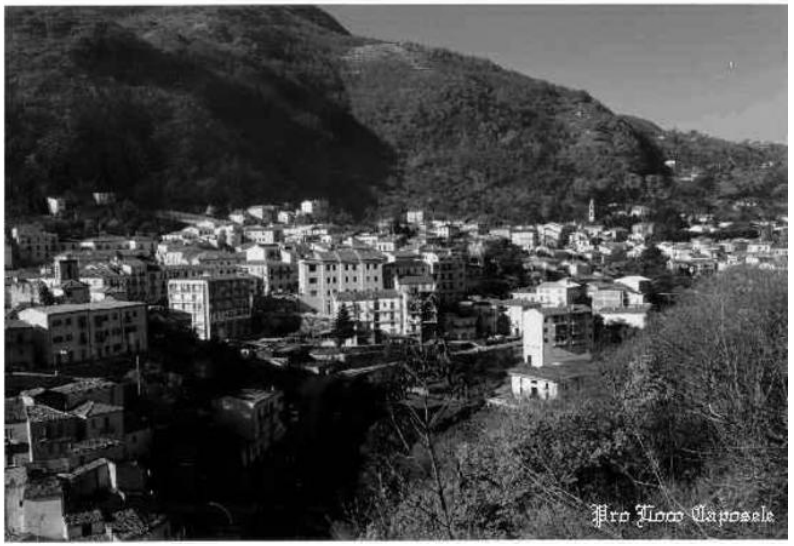

SECOND EDITION
SPEECH COMMUNICATIONS
HUMAN AND MACHINE
Douglas O'Shaughnessy

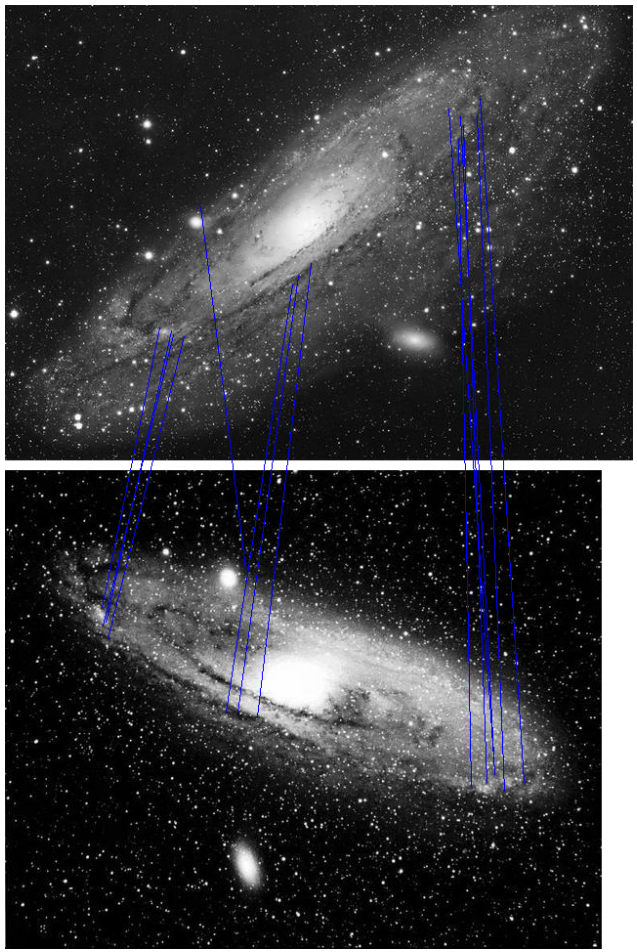


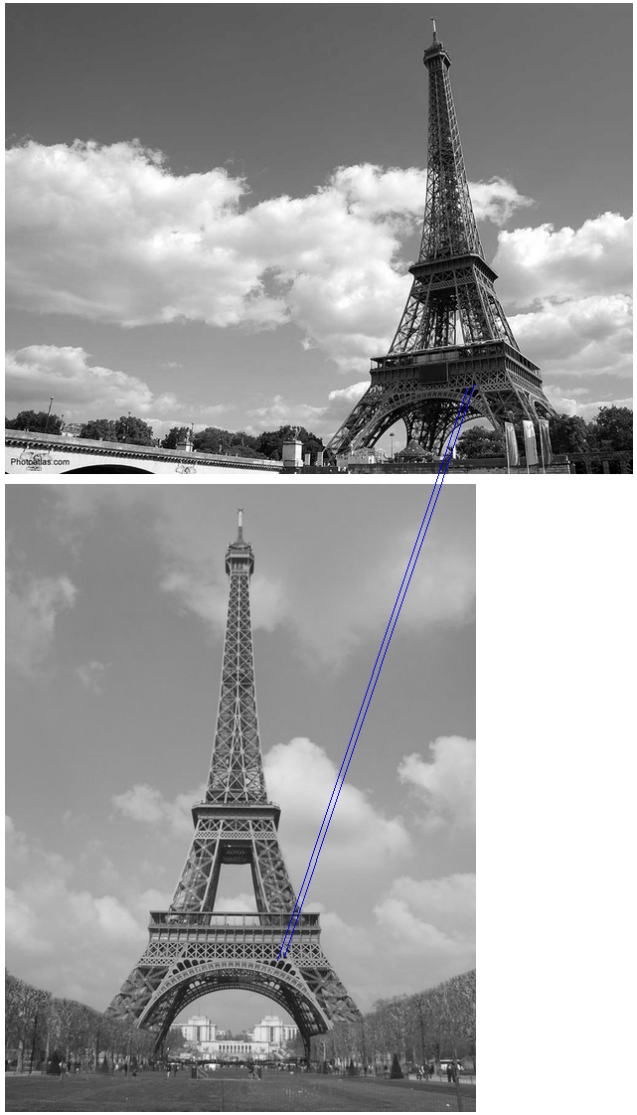




Panorama (Caposele)







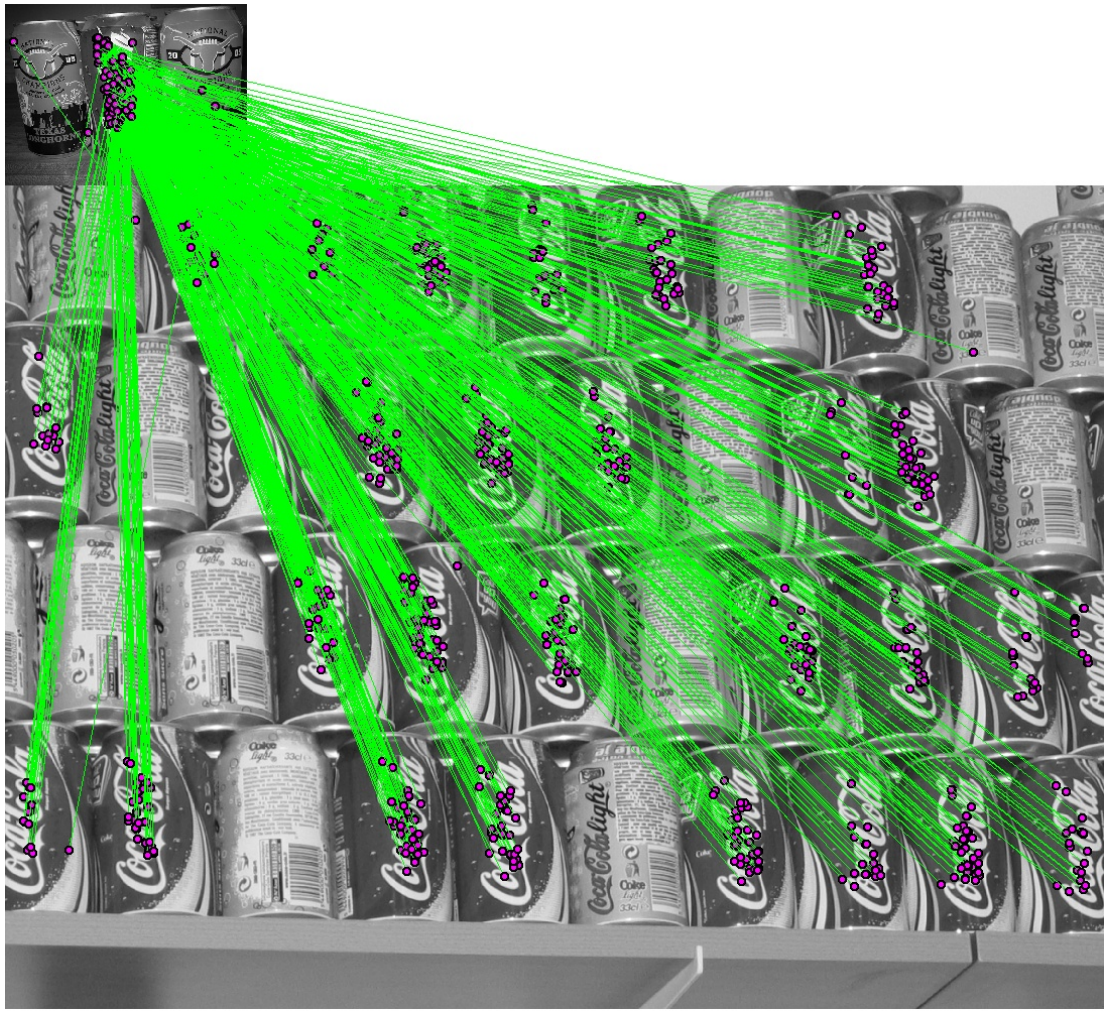


Figure 2: Coke cans: Multiple matches (Rabin, Gousseau, Delon, method which eliminates false alarms)

The "infinite resolution" image u_0 is convolved with a positive integrable kernel g, **which models camera blur**. This means that the observed image is

$$u(\mathbf{x}) = g * u_0(\mathbf{x}) = \int_{\mathbb{R}^N} g(\mathbf{x}-\mathbf{y})u_0(\mathbf{y})\,d\mathbf{y} = \int_{\mathbb{R}^N} g(\mathbf{y})u(\mathbf{x}-\mathbf{y})\,d\mathbf{y}.$$

By the central limit theorem, the most classic model for g is a Gaussian.

- $G_\sigma(x_1, x_2) = \frac{1}{2\pi(\sigma)^2} e^{-\frac{x_1^2+x_2^2}{2(\sigma)^2}}$, G_σ satisfies the heat equation

$$\frac{\partial G_\sigma}{\partial \sigma} = \sigma \Delta \mathbf{G}_\sigma,$$

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

- Notation: $G_\sigma u(\mathbf{x}) =: (G_\sigma * u)(\mathbf{x})$.

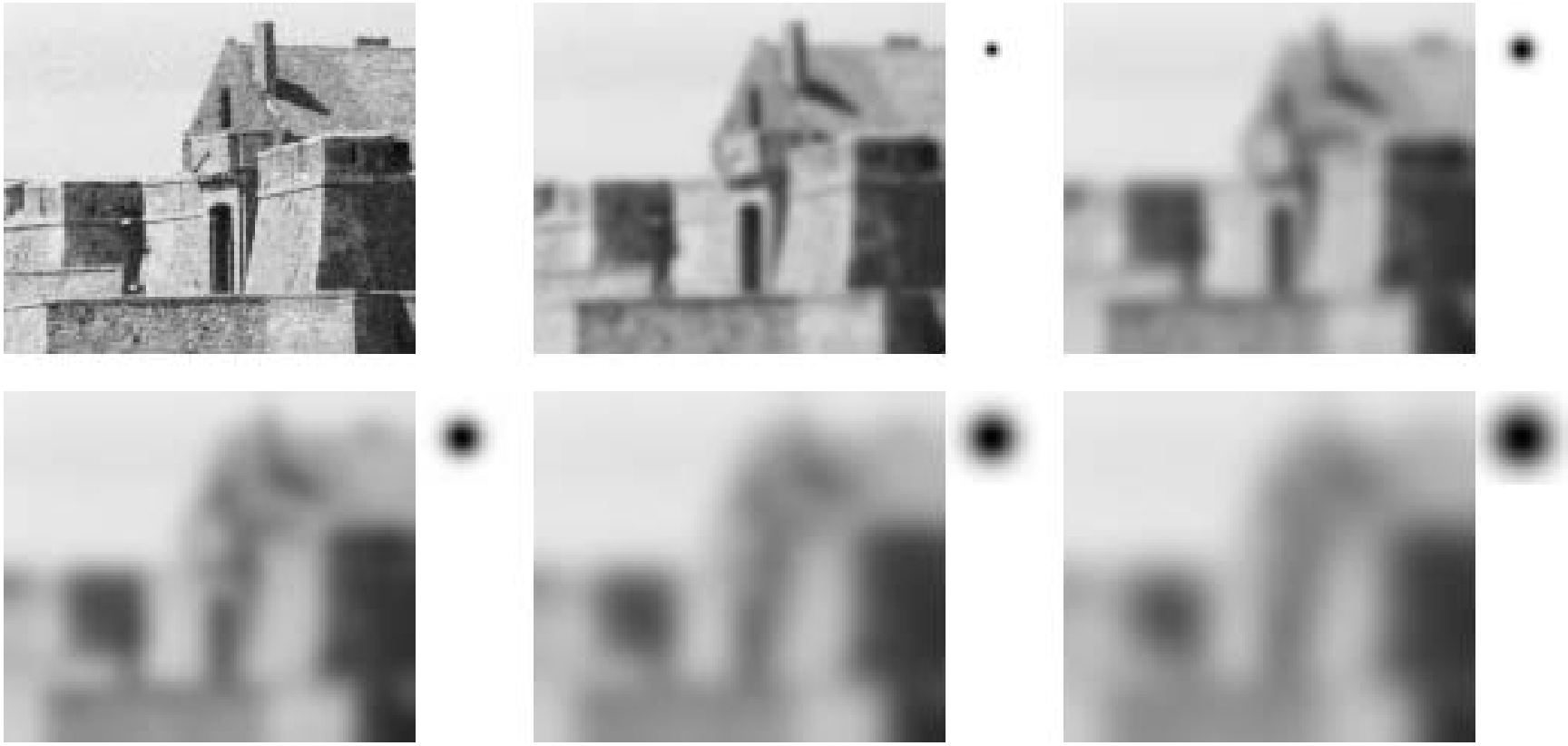
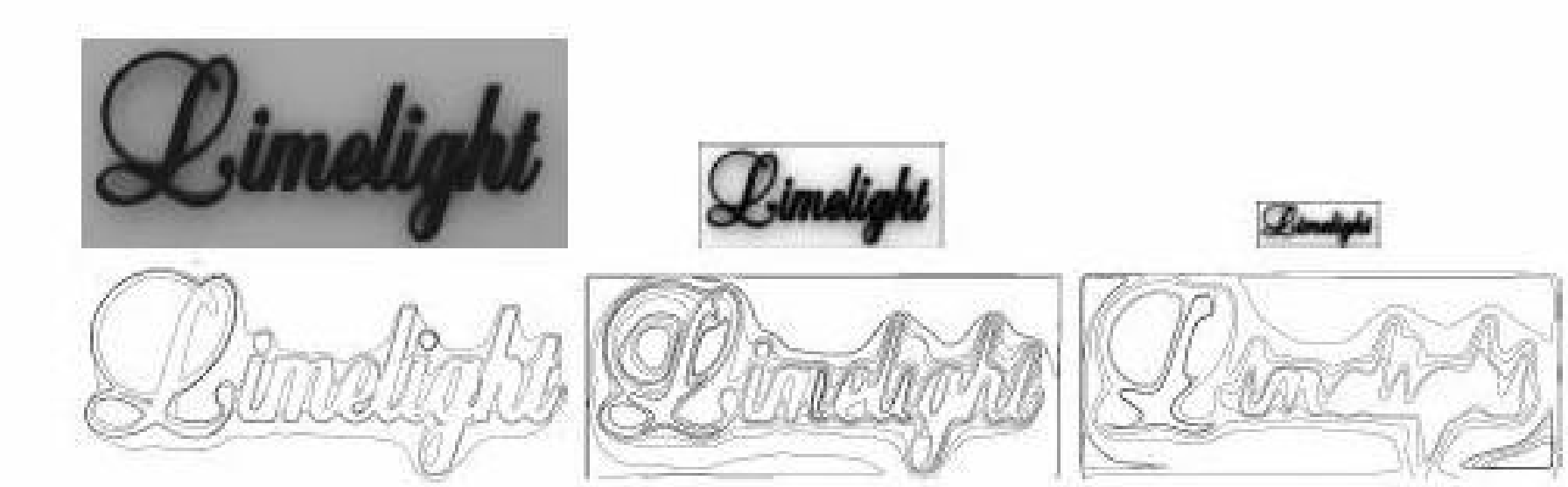


Figure 3: Convolution with Gaussian kernels (heat equation).

Distance means blur and subsampling!



How to compute exactly the convolution of a digital image with a Gaussian.

The rectangular image u is given by $M \times N$ samples noted

$$(u_{k,l})_{0 \leq k \leq M-1;\ 0 \leq l \leq N-1}$$

We define the *Discrete Fourier Transform* (DFT) of the sample matrix $(u_{k,l})_{k,l}$ in $\mathbf{C}^{MN}$ by

$$\tilde{u}_{m,n} = \frac{1}{MN} \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} u_{k,l} e^{-\frac{2i\pi mk}{M}} e^{-\frac{2i\pi nl}{N}}$$

for all $m = -\lfloor \frac{M}{2} \rfloor, ..., -\lfloor \frac{M}{2} \rfloor + M - 1$ and $n = -\lfloor \frac{N}{2} \rfloor, ..., -\lfloor \frac{N}{2} \rfloor + N - 1$.

A digital image is usually given by its $M \times N$ samples on a rectangle. These samples are (implicitly) extended by N-periodicity into a periodic image on $\mathbb{R}^2$, and it is assumed that the resulting underlying image is band-limited with spectrum in $[-\pi, \pi]^2$.

Given a digital image $\mathbf{u}(\mathbf{k})$ *with* $\mathbf{k} = (k, l) \in \{0, \ldots, N-1\}^2$ *its band-limited* N*-periodic interpolate* $u(x, y)$ *is nothing but the trigonometric polynomial*

$$u(x,y) := (\mathbf{I}_d \mathbf{u})(x,y) = \sum_{(m,n) \in \lfloor -M/2, M/2-1 \rfloor \times \lfloor -N/2, N/2-1 \rfloor} \tilde{u}_{m,n} e^{2i\pi(\frac{m}{M}x + \frac{n}{N}y)},$$

where $\tilde{u}_{m,n}$ with $\mathbf{b} = (m, n)$ are the discrete Fourier transform (DFT) coefficients of the $M \times N$ samples $\mathbf{u}(k, l)$.

Let f in $L^1(\mathbb{R}^2)$. Its *Fourier Transform* is

$$\hat{f}(\xi,\eta) = \int_{(x,y)\in\mathbb{R}^2} f(x,y)e^{-i(x\xi+y\eta)}dxdy$$

Normalized 2D Gaussian function of standard deviation σ :

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

Its Fourier transform for all $(\xi,\mu)\in\mathbb{R}^2$ is

$$\hat{G}_\sigma(\xi,\mu) = e^{-\sigma^2\frac{\xi^2+\mu^2}{2}}$$

The main point is that

$$(G_\sigma * e^{2i\pi(\frac{m}{M}x+\frac{n}{N}y)})(x_0,y_0) = \int_{\mathbb{R}^2} G_\sigma(x,y)e^{2i\pi(\frac{m}{M}(x_0-x)+\frac{n}{N}(y_0-y))}dxdy =$$

$$= e^{2i\pi(\frac{m}{M}x_0+\frac{n}{N}y_0)}\hat{G}_\sigma\left(\frac{2\pi m}{M},\frac{2\pi n}{N}\right).$$

Convolving the Gaussian kernel with the image DFT interpolate is equivalent to weighting the digital image's DFT coefficients. Formally if $v(x, y, \sigma) = (G_\sigma * u)(x, y)$,

$$u(x,y) = \sum_{m=-\lfloor \frac{M}{2} \rfloor}^{\left(-\lfloor \frac{M}{2} \rfloor + M - 1\right)} \sum_{n=-\lfloor \frac{N}{2} \rfloor}^{\left(-\lfloor \frac{N}{2} \rfloor + N - 1\right)} \tilde{u}_{m,n} e^{2i\pi\left(\frac{mx}{M} + \frac{ny}{N}\right)}$$

$$v(x,y,\sigma) = \sum_{m=-\lfloor \frac{M}{2} \rfloor}^{\left(-\lfloor \frac{M}{2} \rfloor + M - 1\right)} \sum_{n=-\lfloor \frac{N}{2} \rfloor}^{\left(-\lfloor \frac{N}{2} \rfloor + N - 1\right)} \left[\hat{G}_\sigma(\frac{2\pi m}{M}, \frac{2\pi n}{N})\tilde{u}_{m,n}\right] e^{2i\pi\left(\frac{mx}{M} + \frac{ny}{N}\right)}$$

Each weighting factor $\hat{G}_\sigma(\frac{2\pi m}{M}, \frac{2\pi n}{N})$ is the continuous Fourier transform of the Gaussian function evaluated on the frequency $(\frac{2\pi m}{M}, \frac{2\pi n}{N})$,

$$\hat{G}_\sigma(\frac{2\pi m}{M}, \frac{2\pi n}{N}) = e^{-\frac{\sigma^2}{2}\left(\left(\frac{2\pi m}{M}\right)^2 + \left(\frac{2\pi n}{N}\right)^2\right)}$$

.

Summary of algorithm computing the convolution:

- Input image:

$$U = [u_{k,l}]_{0 \leq k \leq M-1;\ 0 \leq l \leq N-1};$$

- $DFT(U) = [\tilde{u}_{m,n}]$, $-\lfloor \frac{M}{2} \rfloor \leq m \leq -\lfloor \frac{M}{2} \rfloor + M - 1$; $-\lfloor \frac{N}{2} \rfloor \leq n \leq -\lfloor \frac{N}{2} \rfloor + N - 1$;

- Pointwise multiplication in the Fourier domain:

$$DFT(V) = \hat{G}.DFT(U) = \left[\hat{u}_{m,n}.e^{-\frac{\sigma^2}{2}\left(\left(\frac{2\pi m}{M}\right)^2 + \left(\frac{2\pi n}{N}\right)^2\right)} \right];$$

- Inverse Discrete Fourier Transform to get the $M \times N$ real samples of the output smoothed image:

$$V = DFT^{-1}(DFT(V)).$$

Try it: `http://dev.ipol.im/~reyotero/ipol_demo/rorm_the_heat_equation3/` `http://dev.ipol.im/~reyotero/ipol_demo/rorm_the_heat_equation2/` (user: demo, password: demo)

Figure 4: Barbara

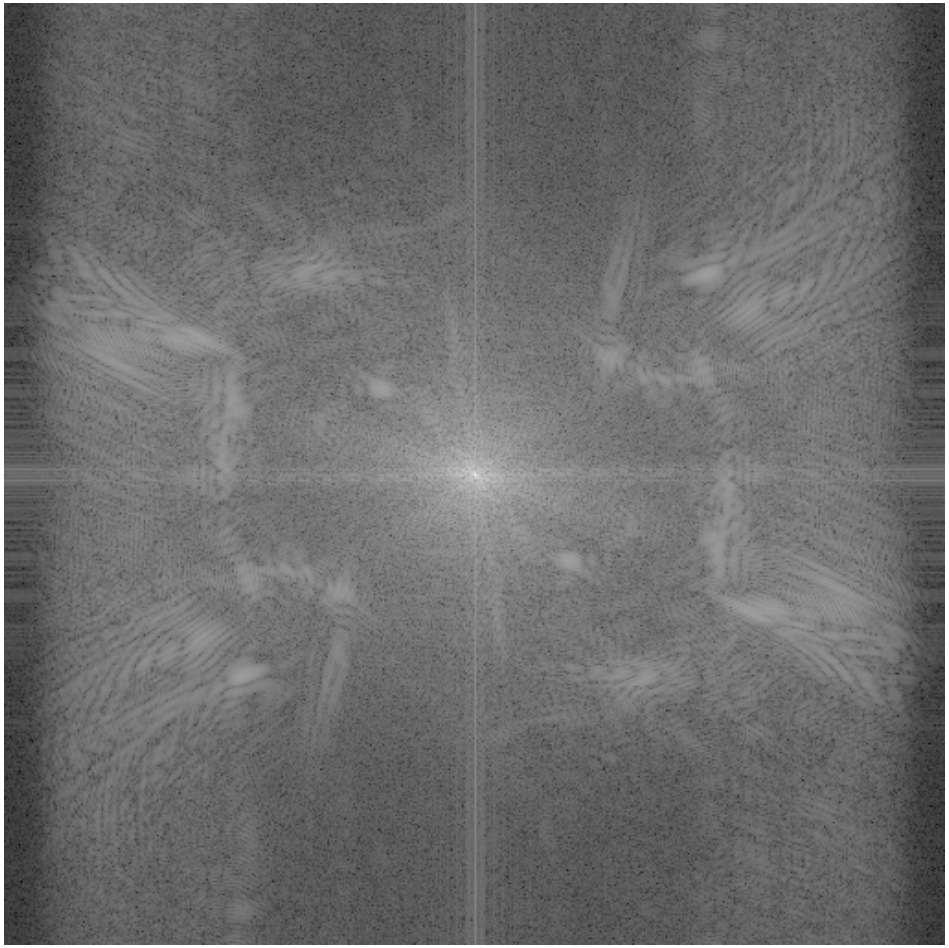


Figure 5: Barbara, DFT

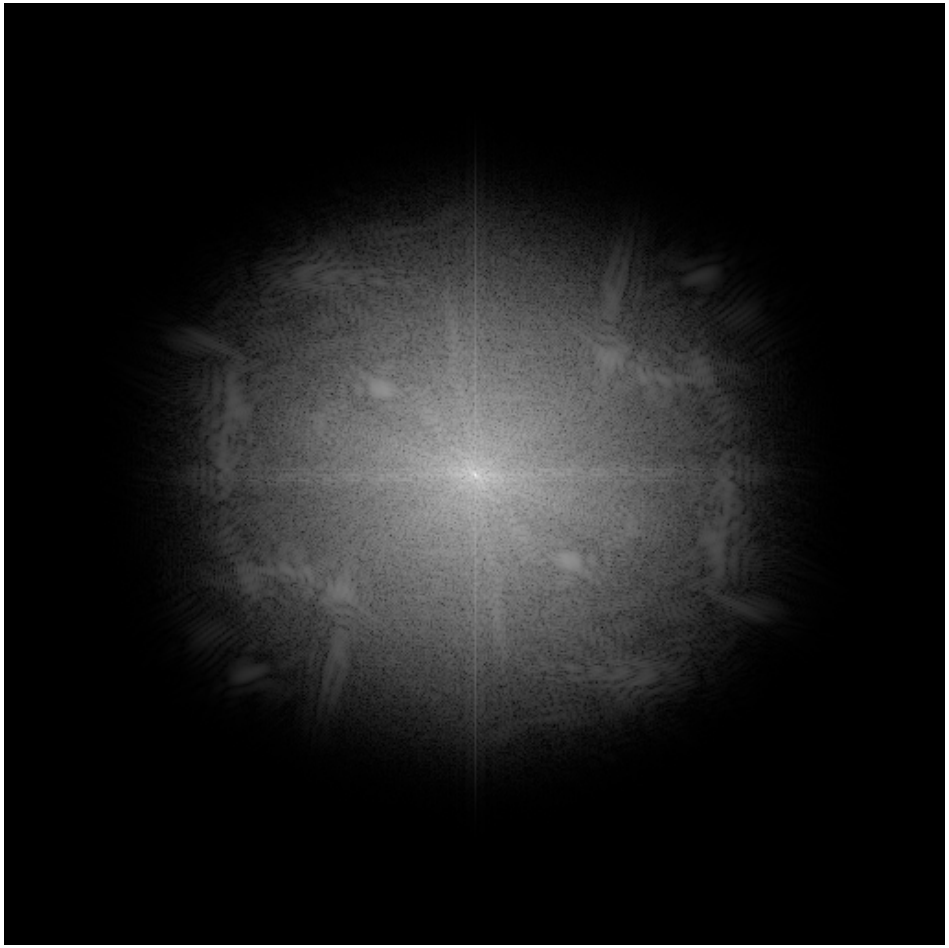


Figure 6: Barbara, filtered DFT, $\sigma = 2$

Figure 7: Barbara, filtered by Gaussian

From continuous to digital and conversely

- $u(\mathbf{x})$: a continuous and bounded image defined for every $\mathbf{x} = (x, y) \in \mathbb{R}^2$.
- $\mathbf{u}$: a digital image, only defined for $(n_1, n_2) \in \mathbb{Z}^2$.
- $\mathbf{S}_1$: the 1-sampling operator. Let u be a continuous image on $\mathbb{R}^2$. The associated sampled digital image $\mathbf{S}_1 u$ is defined on $\mathbb{Z}^2$ by

$$\mathbf{S}_1 u(n_1, n_2) = u(n_1, n_2);$$

Better for theoretical results: from a digital image back to a continuous image by Shannon interpolation

- $\mathbf{u}(n_1, n_2)$ digital image, $\sum_{(n_1,n_2)\in\mathbb{Z}^2} |\mathbf{u}(n_1, n_2)|^2 < \infty$.
- Shannon interpolate of $\mathbf{u}$: the $L^2(\mathbb{R}^2)$ function $u = I\mathbf{u}$ having $\mathbf{u}(n_1, n_2)$ as samples and with spectrum supported in $(-\pi, \pi)^2$.
- Shannon-Whittaker :

$$I\mathbf{u}(x, y) =: \sum_{(n_1,n_2)\in\mathbb{Z}^2} \mathbf{u}(n_1, n_2) \frac{\sin \pi(x - n_1)}{\pi(x - n_1)} \frac{\sin \pi(y - n_2)}{\pi(y - n_2)}.$$

- $\mathbf{S}_1 I\mathbf{u} = \mathbf{u}$. Conversely, if u is L^2 and band-limited in $(-\pi, \pi)^2$, then $I\mathbf{S}_1 u = u$.
- If $\mathbf{c} \geq 0.8$, $\mathbf{G_c} * u_0$ is experimentally "well-sampled" and therefore $I\mathbf{S}_1 G_{\mathbf{c}} u_0 = G_{\mathbf{c}} u_0$.

SIFT assumptions and condensed description of the method

1. the initial digital image is $\mathbf{S}_1 G_{\mathbf{c}} A u_0$, A is any similarity (homothety, translation, rotation, 4 parameters);
2. at all scales $\sigma > 0$, the SIFT method computes "good samplings" of the "scale space" $u(\sigma, \cdot) = G_\sigma G_{\mathbf{c}} A u_0$;
3. key points $(\sigma, \mathbf{x})$ are scale and space extrema of $\Delta u(\sigma, \cdot)$;
4. directions of the sampling axes are fixed by a dominant direction of $\nabla u(\sigma, \cdot)$ in a neighborhood of the key point proportional to $\sqrt{\sigma^2 + \mathbf{c}^2}$;
5. the image $u(\sigma, \cdot)$ blurred at scale σ is sampled around each key point at a pace proportional to $\sqrt{\sigma^2 + \mathbf{c}^2}$ (initial blur + added blur);
6. this yields rotation, translation and zoom invariant samples;
7. the final SIFT descriptor keeps only orientations of the gradient to gain invariance w.r. light conditions.

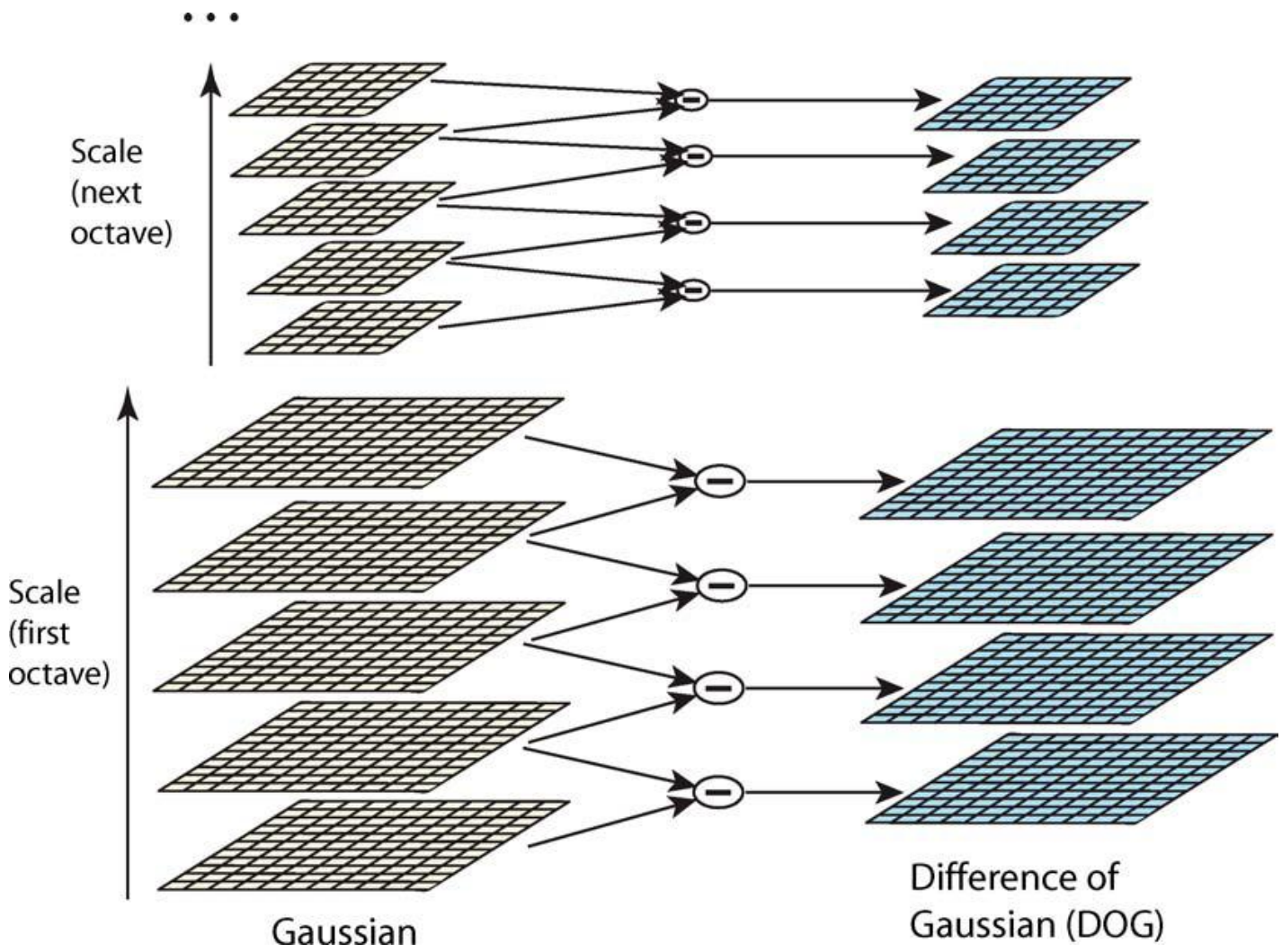


Figure 8: (Digital) Gaussian pyramid for key points extraction (from Lowe). The subsampling is allowed because of the sufficient blur

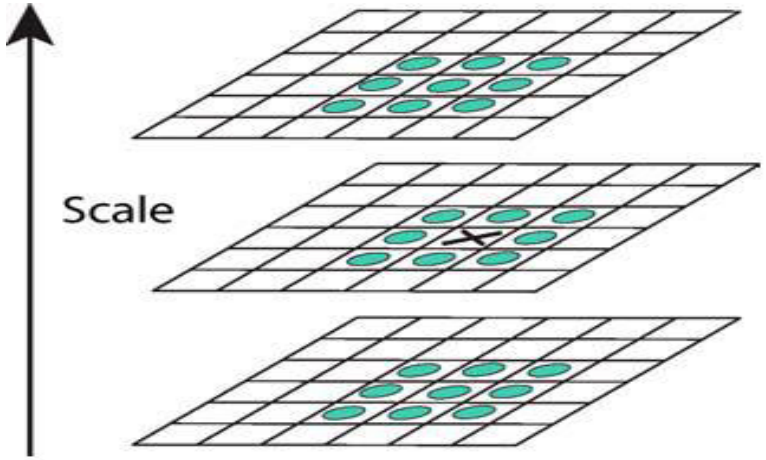


Figure 9: Neighborhood for the location of key points (from Lowe). Local extrema are detected by comparing each sample point in $\mathbb{D}$ with its eight neighbors at scale σ and its nine neighbors in the scales above and below

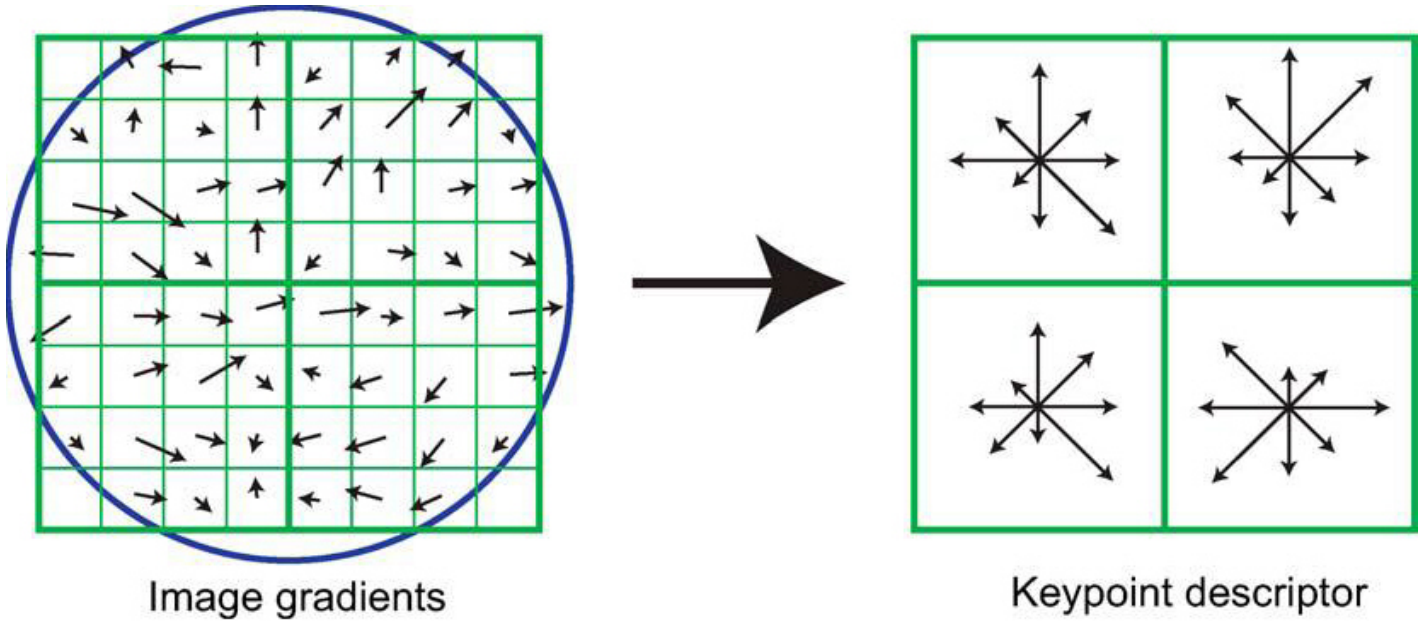


Figure 10: Each key-point is associated *a square image patch whose size is proportional to the scale and whose side direction is given by the assigned direction*. Example of a 2×2 descriptor array of orientation histograms (right) computed from an 8×8 set of samples (left). The orientation histograms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

Figure 11: SIFT key points (scale and orientation)

1 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. 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 the SIFT model 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.

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

PROOF: Using the weak commutation of translations and rotations with all other operators : 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.

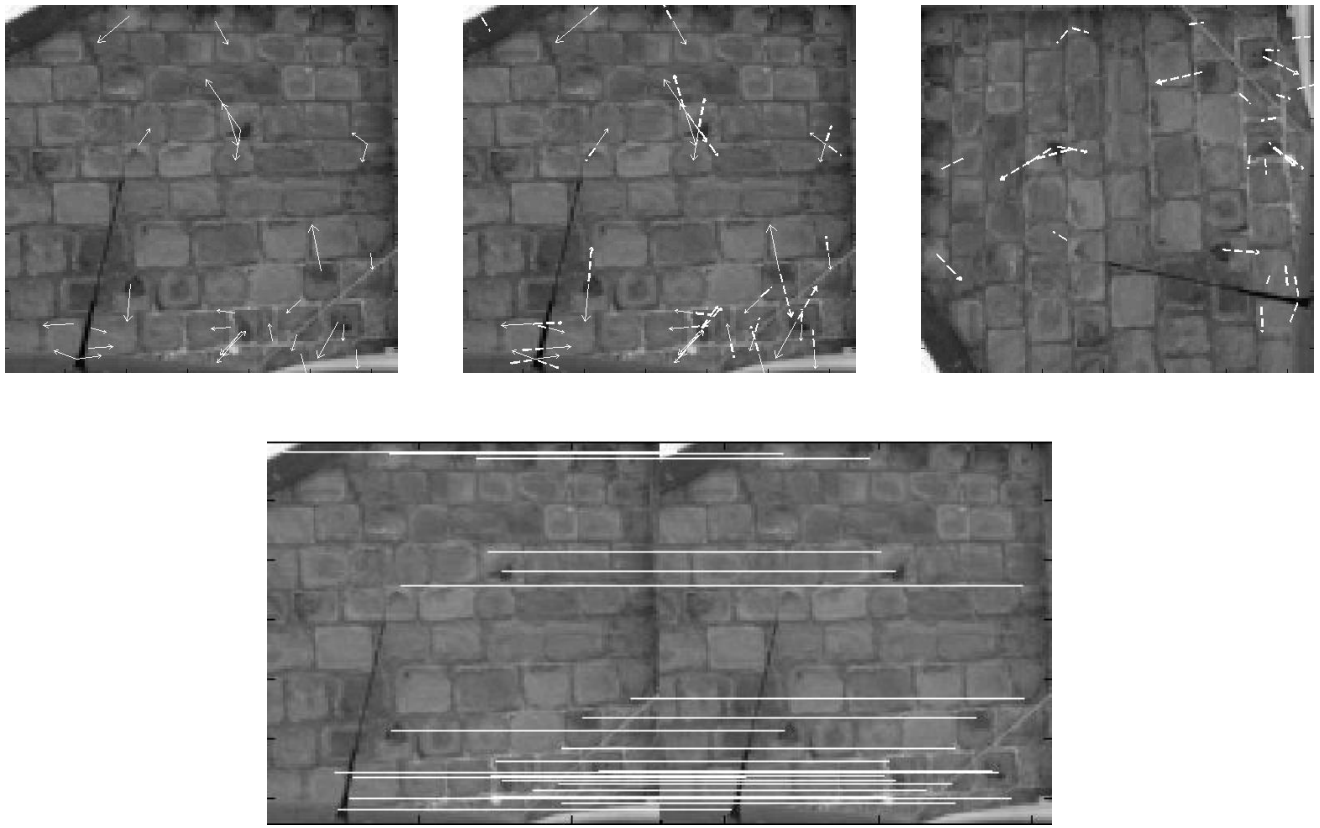


Figure 12: Rotation invariance of SIFT. Top left and right: $\mathbf{u}$ and $\mathbf{R}_{\frac{\pi}{2}}\mathbf{u}$ superposed with their 31 keypoints. Top middle: descriptors of $\mathbf{R}_{\frac{\pi}{2}}\mathbf{u}$ are projected on $\mathbf{u}$ and their orientations are inverted for better observability. Bottom: 31 matches between $\mathbf{u}$ and $\mathbf{R}_{\frac{\pi}{2}}\mathbf{u}$ ($\mathbf{R}_{\frac{\pi}{2}}\mathbf{u}$ are rotated by 90° for better preservability).

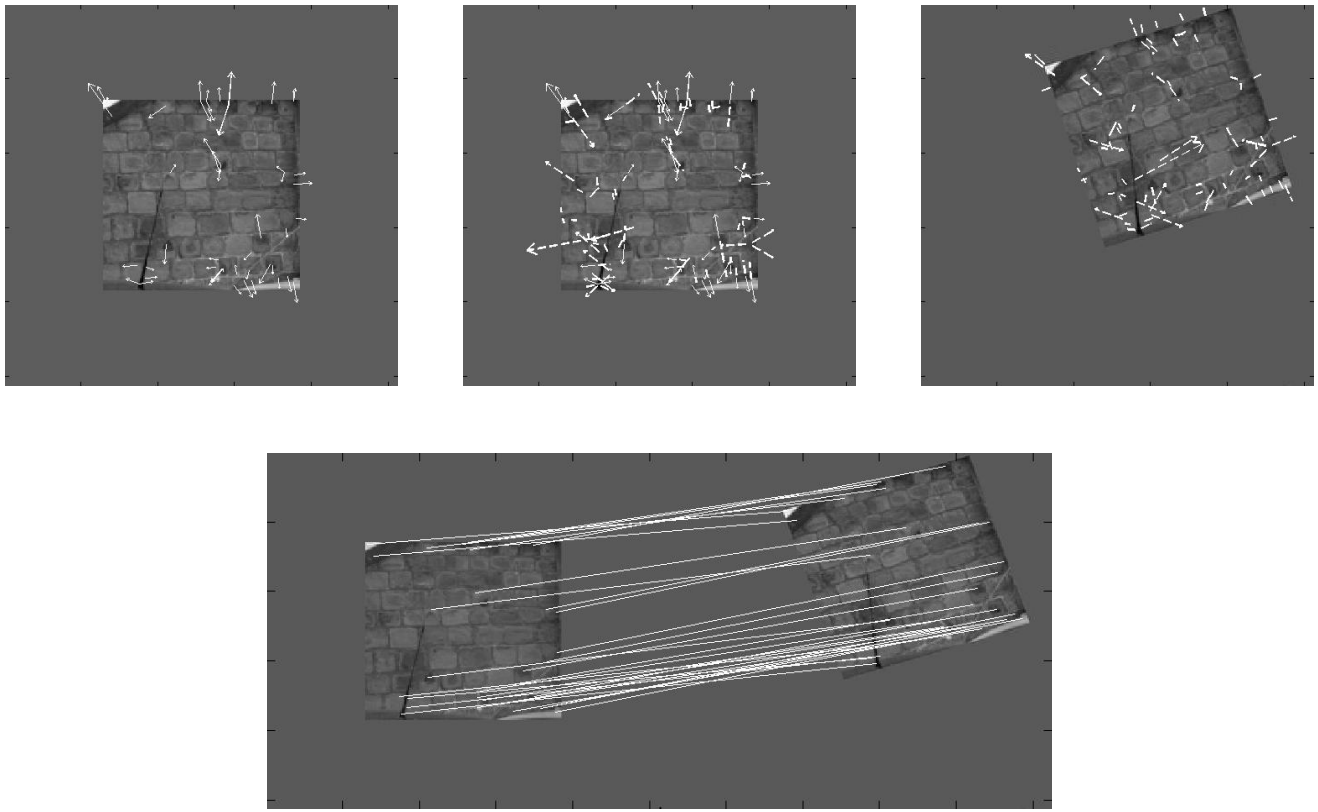


Figure 13: Rotation invariance of SIFT. Top left: $\mathbf{u}$ superposed with its 52 keypoints. Top right: $\mathbf{R}_{\frac{\pi}{10}}\mathbf{u}$ (obtained with Shannon interpolation) superposed with its 73 keypoints. Top middle: descriptors of $\mathbf{R}_{\frac{\pi}{2}}\mathbf{u}$ are projected on $\mathbf{u}$ and their orientations are inverted for better observability. Bottom: 37 matches between $\mathbf{u}$ and $\mathbf{R}_{\frac{\pi}{2}}\mathbf{u}$.

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*

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

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 ideal initial image blur assumed by SIFT. 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: Computing the scale-space for both images u and v amounts to convolve them for every $\sigma > 0$ with G_σ.

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

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

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

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

For any extremal point $(s_0, \mathbf{x}_0)$ of the Laplacian of w, if $s_0 \geq \max(\lambda\beta, \mu\delta)$, an extremal point occurs at scales σ_1 for $u(\sigma, \mathbf{x})$ and σ_2 for $v(\sigma, \mathbf{x})$ satisfying

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

The SIFT descriptor sampling rate around the key point $(\sigma_1, \mathbf{x}_1)$ 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 $u(\sigma_2, \mathbf{x})$.

Theorem 1 *Let* $\mathbf{u}$ *and* $\mathbf{v}$ *be two 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: We can neglect the effect of translations and rotations. Consider a key point $(s_0, \mathbf{x}_0)$ of w with scale $s_0 \geq \max(\lambda\beta, \mu\delta)$. 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 $\mathbf{v}$. The corresponding sampling rates for $w(s_0, \mathbf{x})$, 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. 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

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

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

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

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 (53) 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 $\lambda\sqrt{\sigma_1^2+\mathbf{c}^2}\approx\mu\sqrt{\sigma_2^2+\mathbf{c}^2}$.

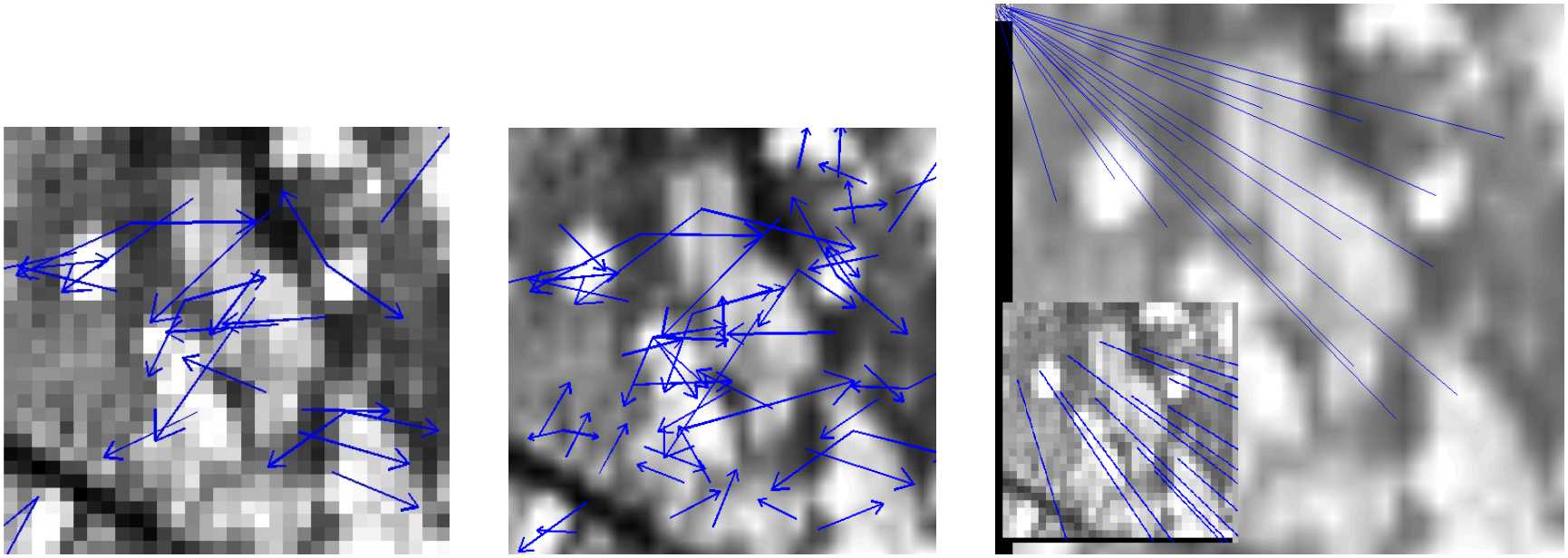


Figure 14: Scale invariance of SIFT, an illustration of Theorem 1. Left: a very small digital image $\mathbf{u}$ with its 25 key points. Middle: this image is over sampled by a 32 factor to $\mathbf{S}_{\frac{1}{32}}\mathbf{I}_d\mathbf{u}$. It has 60 key points. Right: 18 matches found between $\mathbf{u}$ 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.