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

Jean-Michel Morel, Ives Rey Otero, Guoshen Yu

Gestalt psychophysical invariance laws proposed by Wertheimer, Attneave and Kanizsa: a good 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.

Figure 1: Various snapshots of a “Mural”

Detectors in competition and their invariance

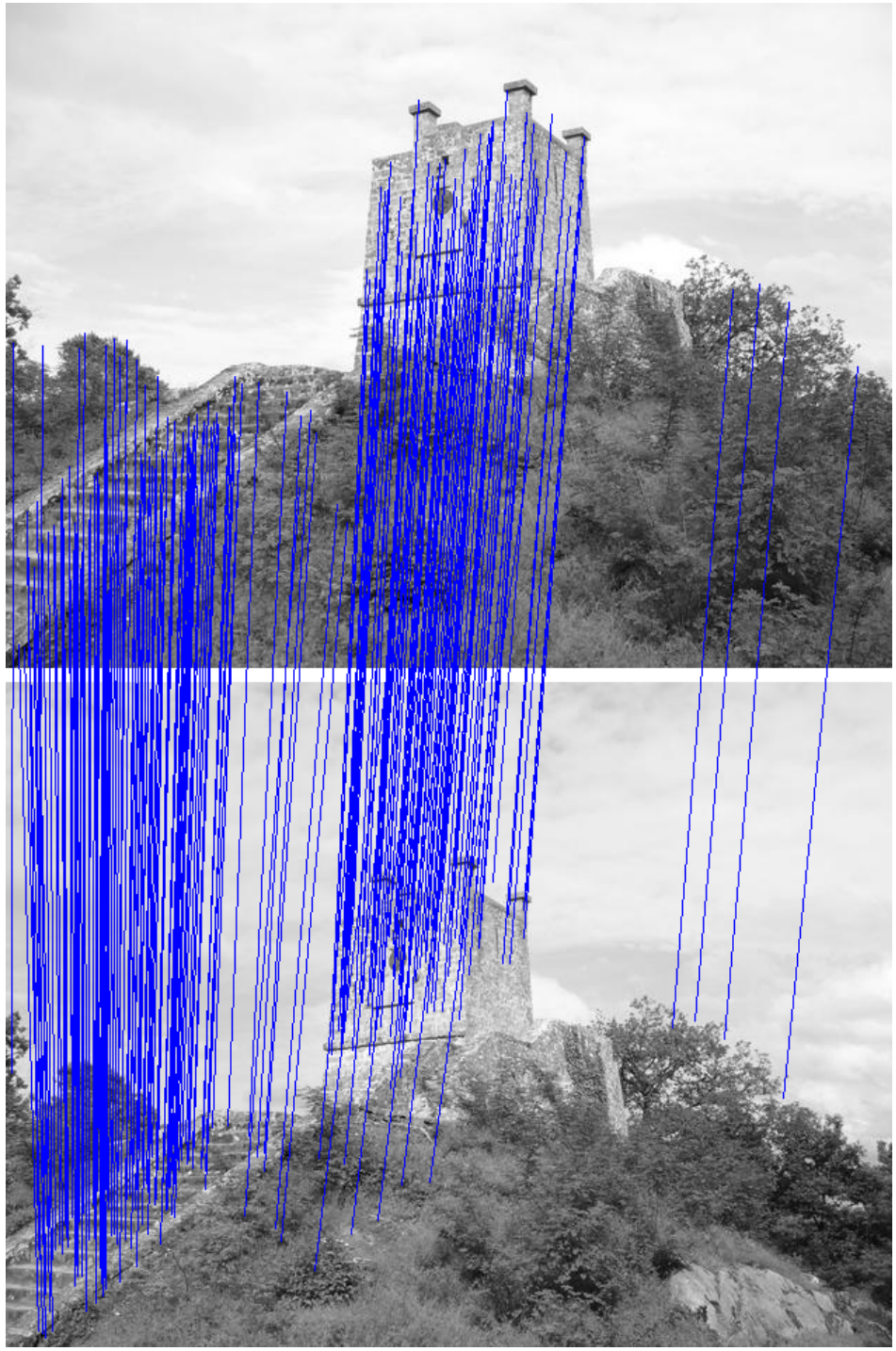
- Harris point detector (1988) : rotation, translation
- Harris-Laplace and Hessian-Laplace region detectors (Mikolajczyk et al. 20001, 2004) invariant to rotation and scale changes, extended to some affine invariance
- edge-based region detector Tuytelaars 1999, 2004, Entropy-based region detector (Kadir 2004)
- level line-based region detectors: MSER ("maximally stable extremal region") (Matas 2002) special affine invariant
- LLD ("level line descriptor") (Musé 2003) special affine invariant
- MSER, better performance than other affine invariant detectors
- Some are special affine invariant (MSER, LLD), some are scale invariant but imperfectly affine invariant
- the surprise: although in principle only scale invariant, SIFT has good affine invariance

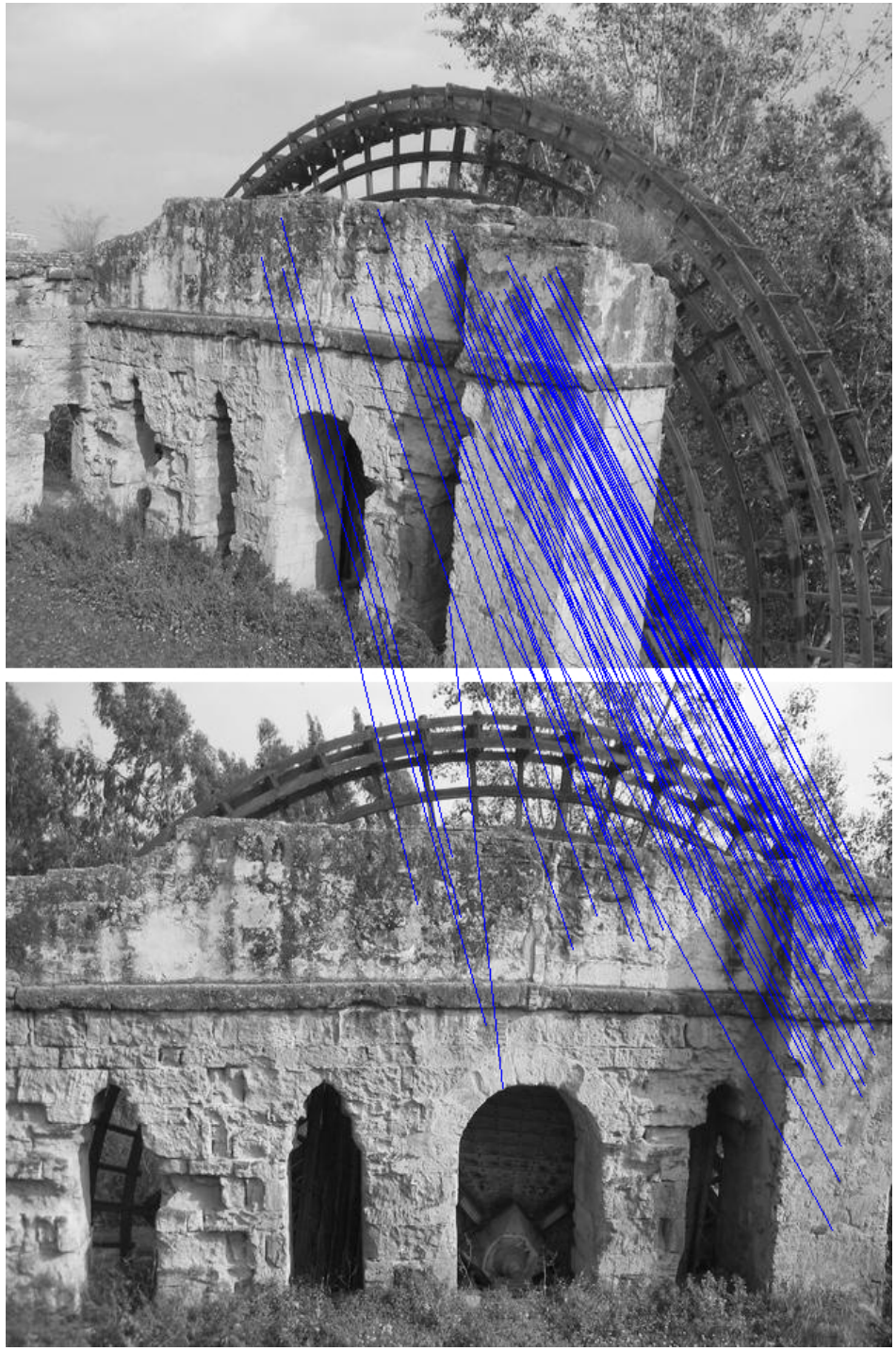
- D. G. Lowe, Object recognition from local scale-invariant features International Conference on Computer Vision, 2, 1150–1157, 1999
- D. G. Lowe, Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision, 60(2):91-110, 2004
- J. Matas, O. Chum, M.Urban, T. Pajdla, Robust Wide Baseline Stereo from Maximally Stable Extremal Regions, BMVC, 2002
- F. Cao, J.L. Lisani, J.M. Morel, P. Musé, and F. Sur, *A theory of shape identification*, LNM Springer 2008, forthcoming
- Julien Rabin, Julie Delon, Yann Gousseau, Mise en correspondance de descripteurs géométriques locaux par méthode a contrario, preprint, ENST, GRETSI 2007
- Guoshen Yu, J.M. Morel: Is SIFT scale invariant? submitted to IPI (Imaging Inverse Problems, 2010)

PLAN : A mathematical analysis of the most popular shape descriptor: Lowe's *Scale-Invariant Feature Transform* (SIFT) method

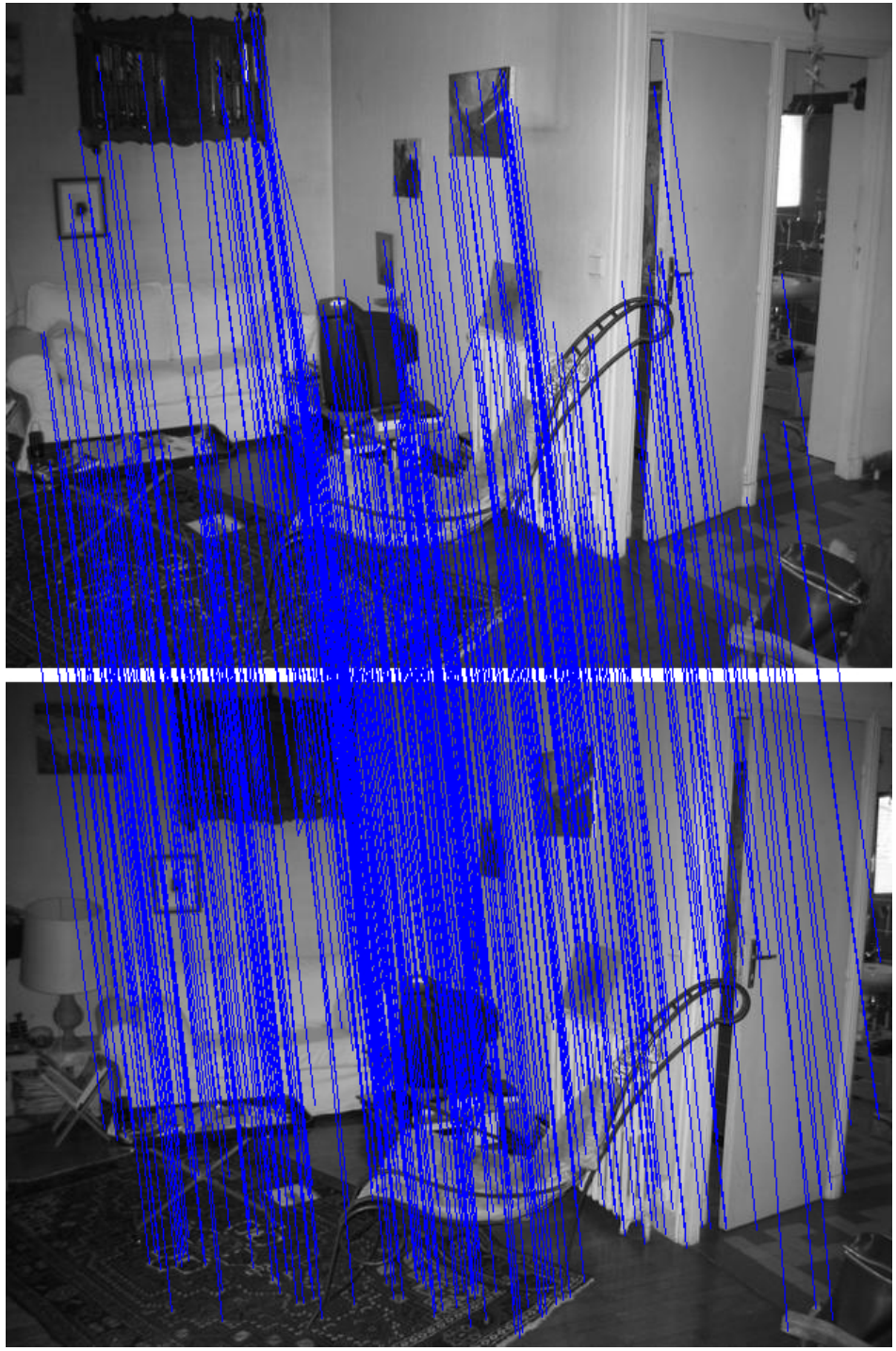
- Many examples;
- Comprehensive description of the SIFT shape encoding method
- Proof that : SIFT method achieves perfect performance only with, zoomed, rotated, translated versions of two images
- if not exactly, SIFT is factually more and more scale-invariant when the scale grows in the image scale-space

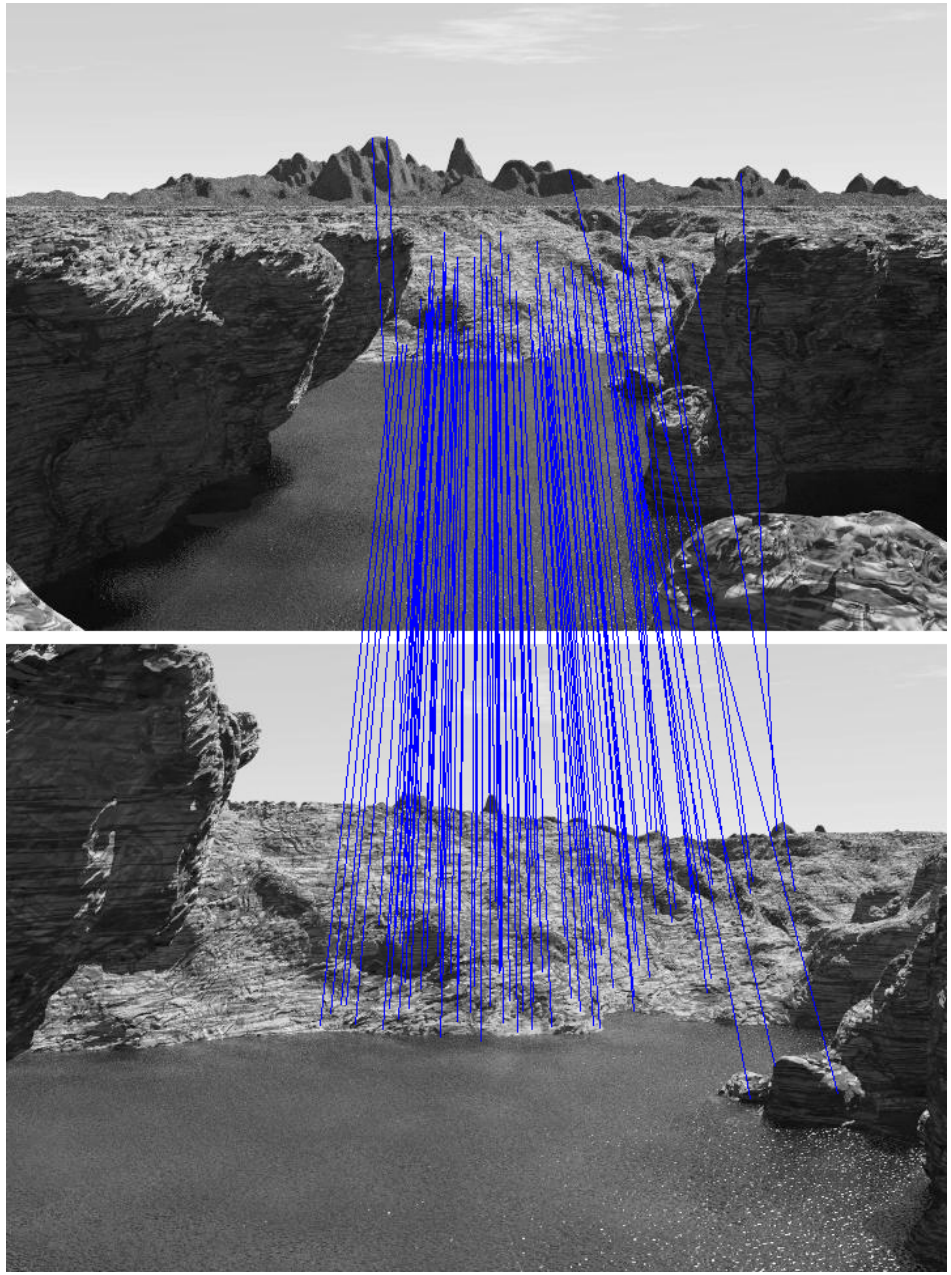
STRIKING EXAMPLES

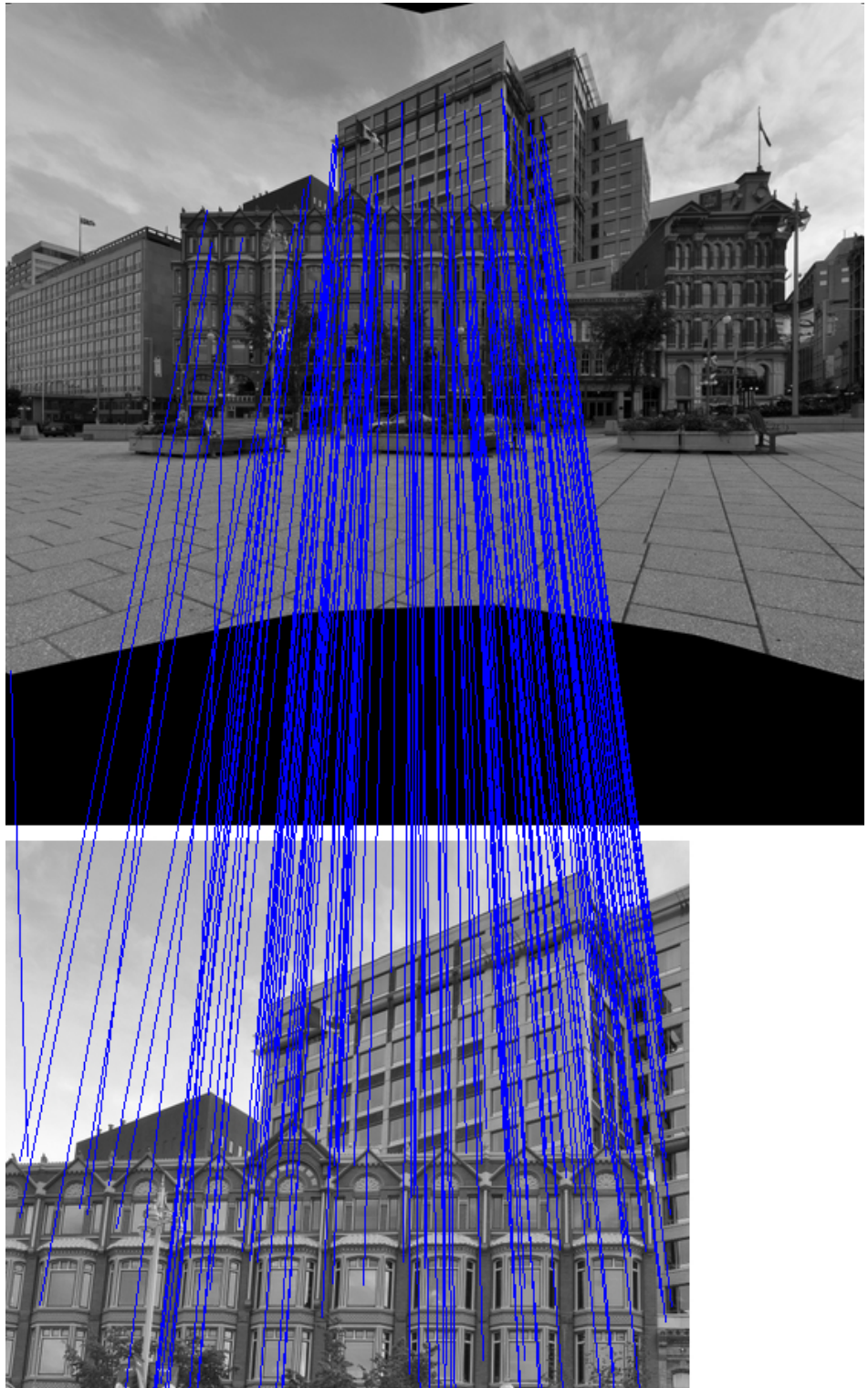




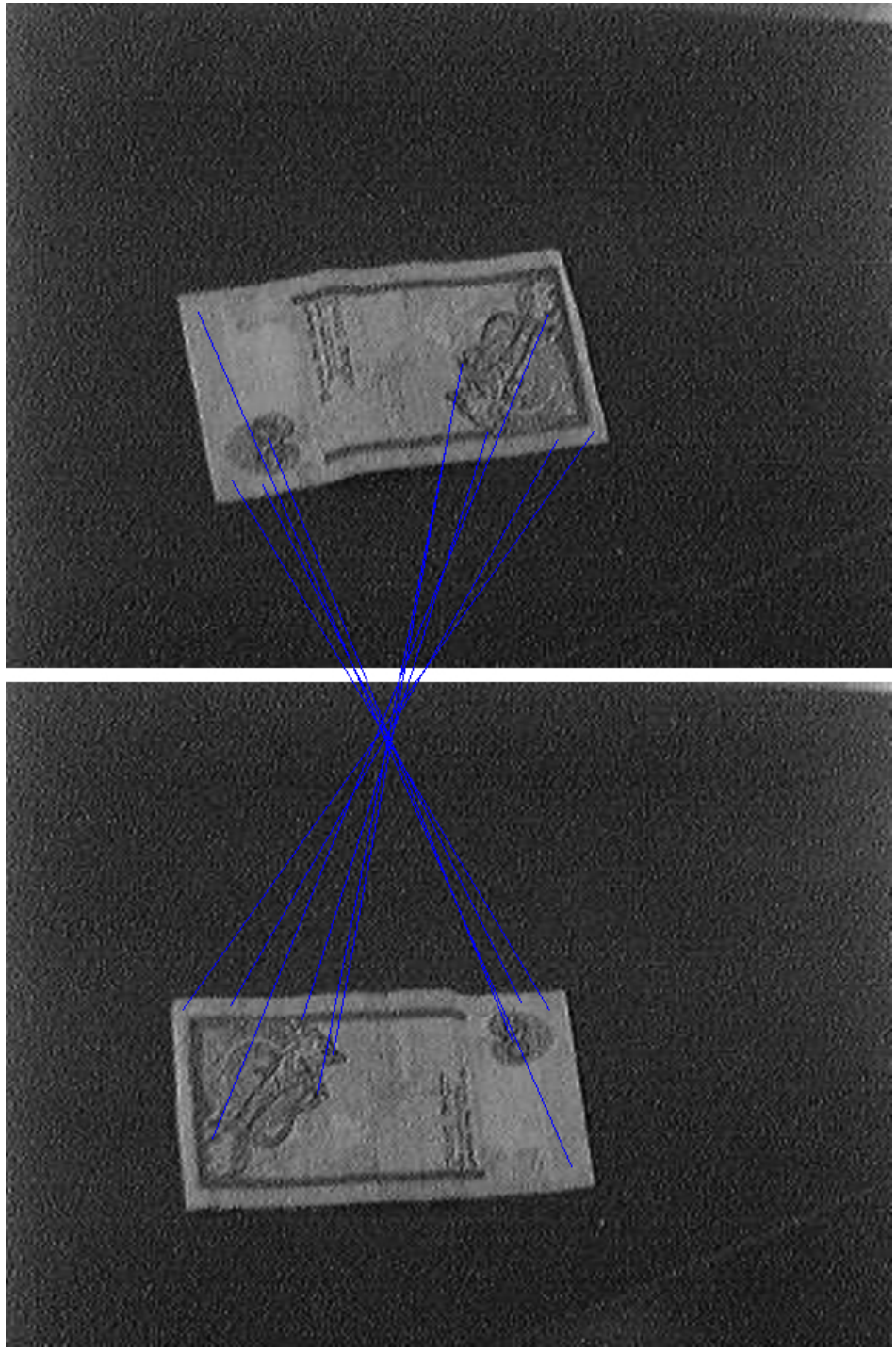

HUAWEI
Love innovation
第三媒体
www.3mt.com.cn

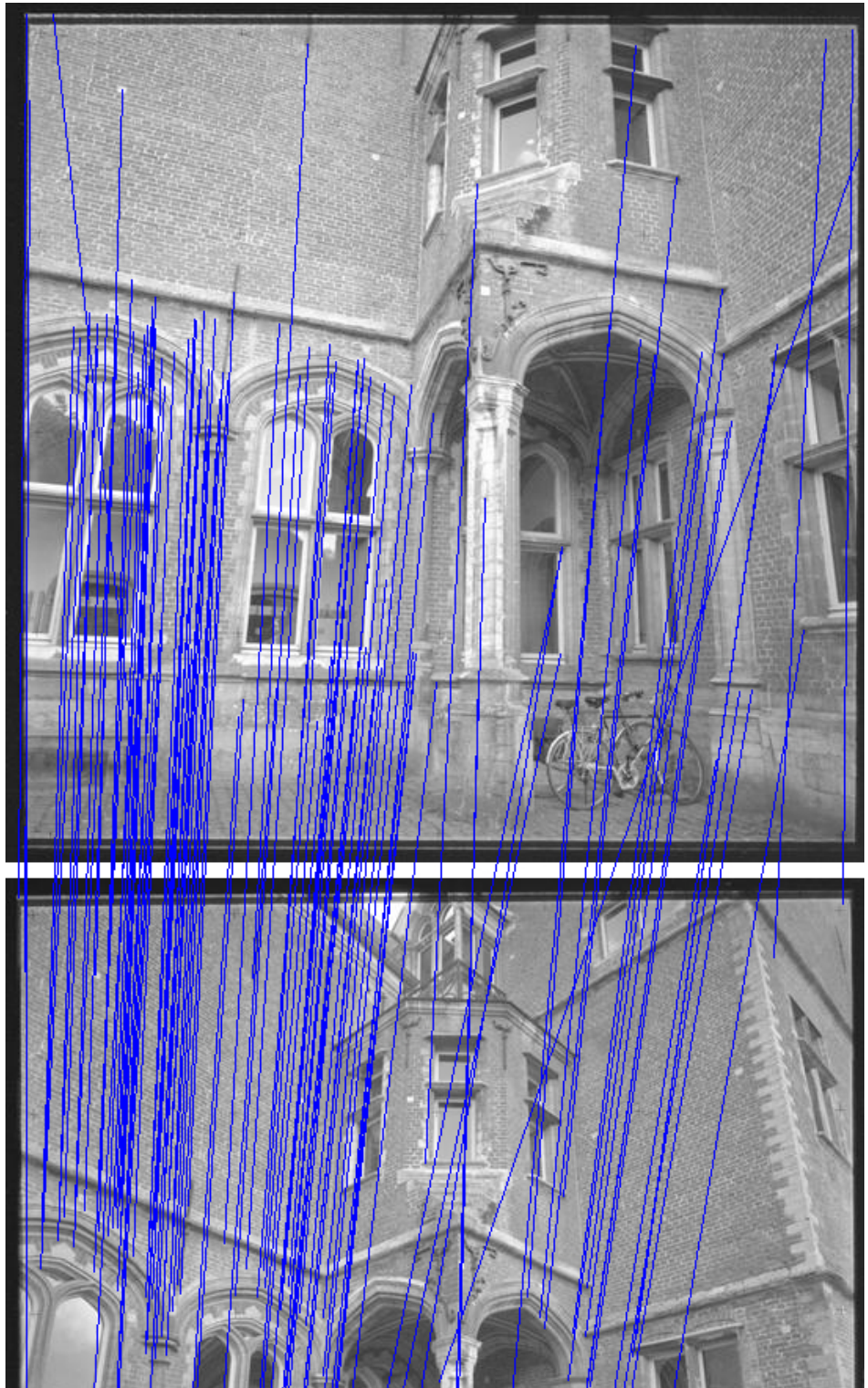


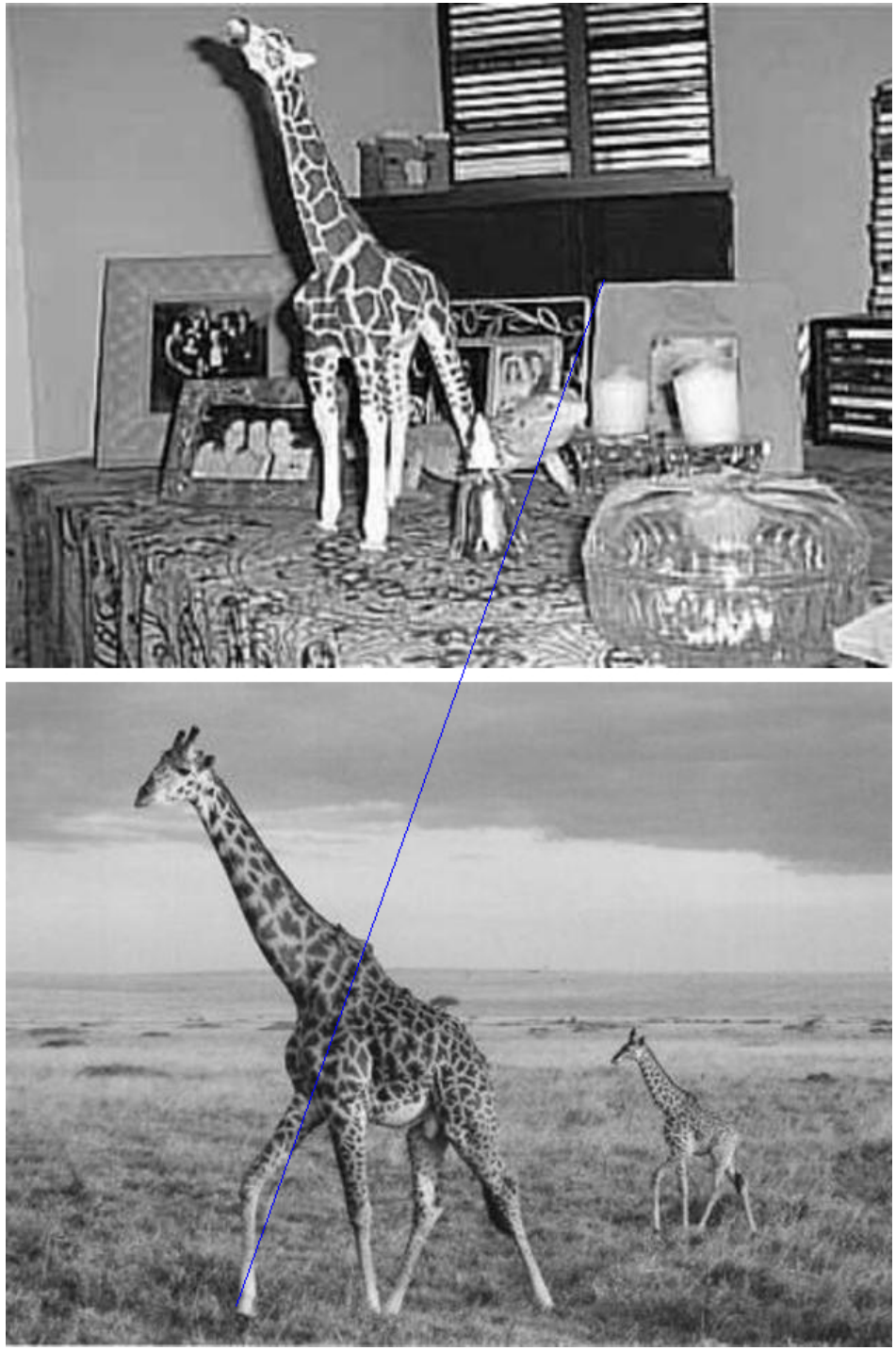




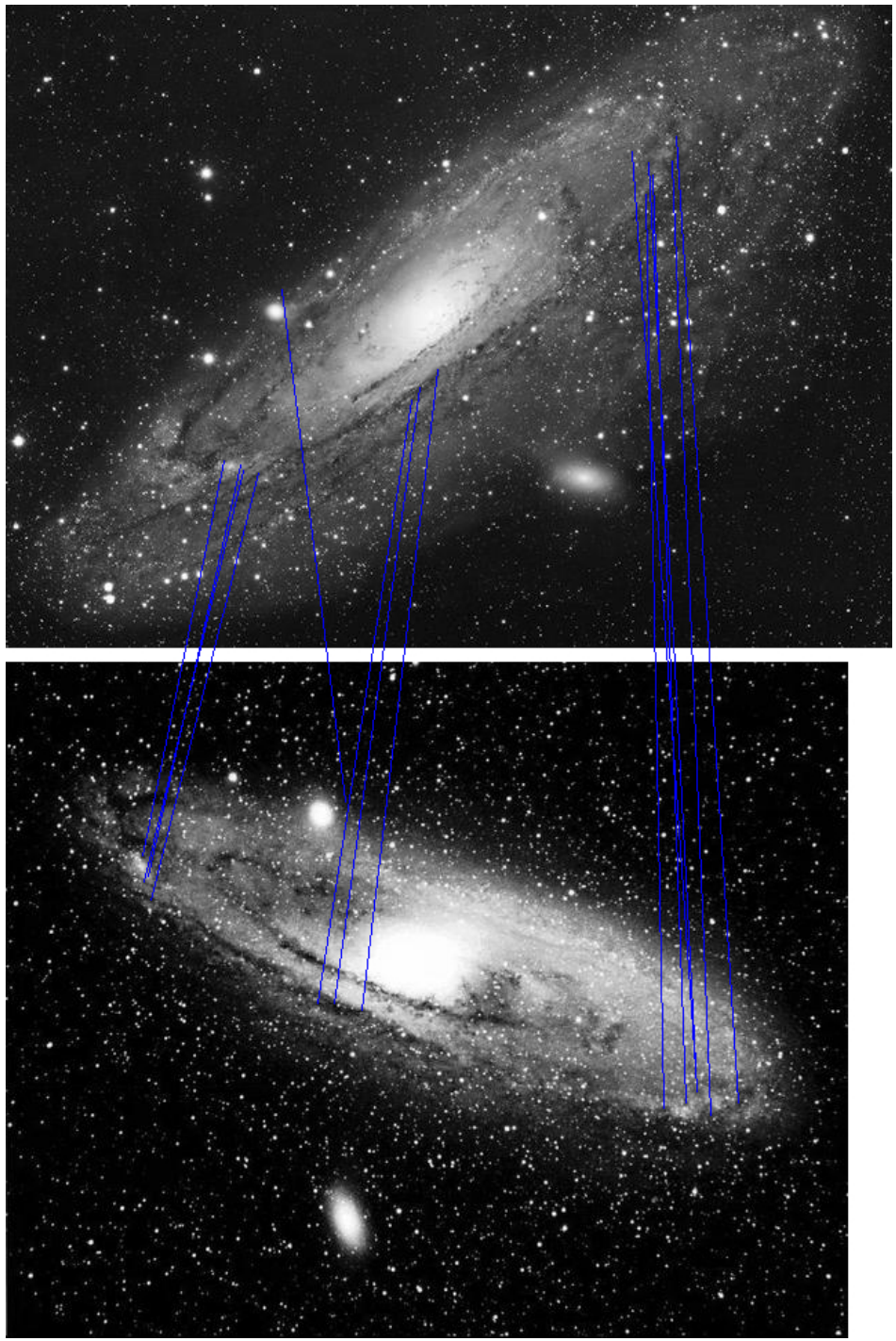
SECOND EDITION
SPEECH
COMMUNICATIONS
HUMAN AND
MACHINE
Douglas O'Shaughnessy

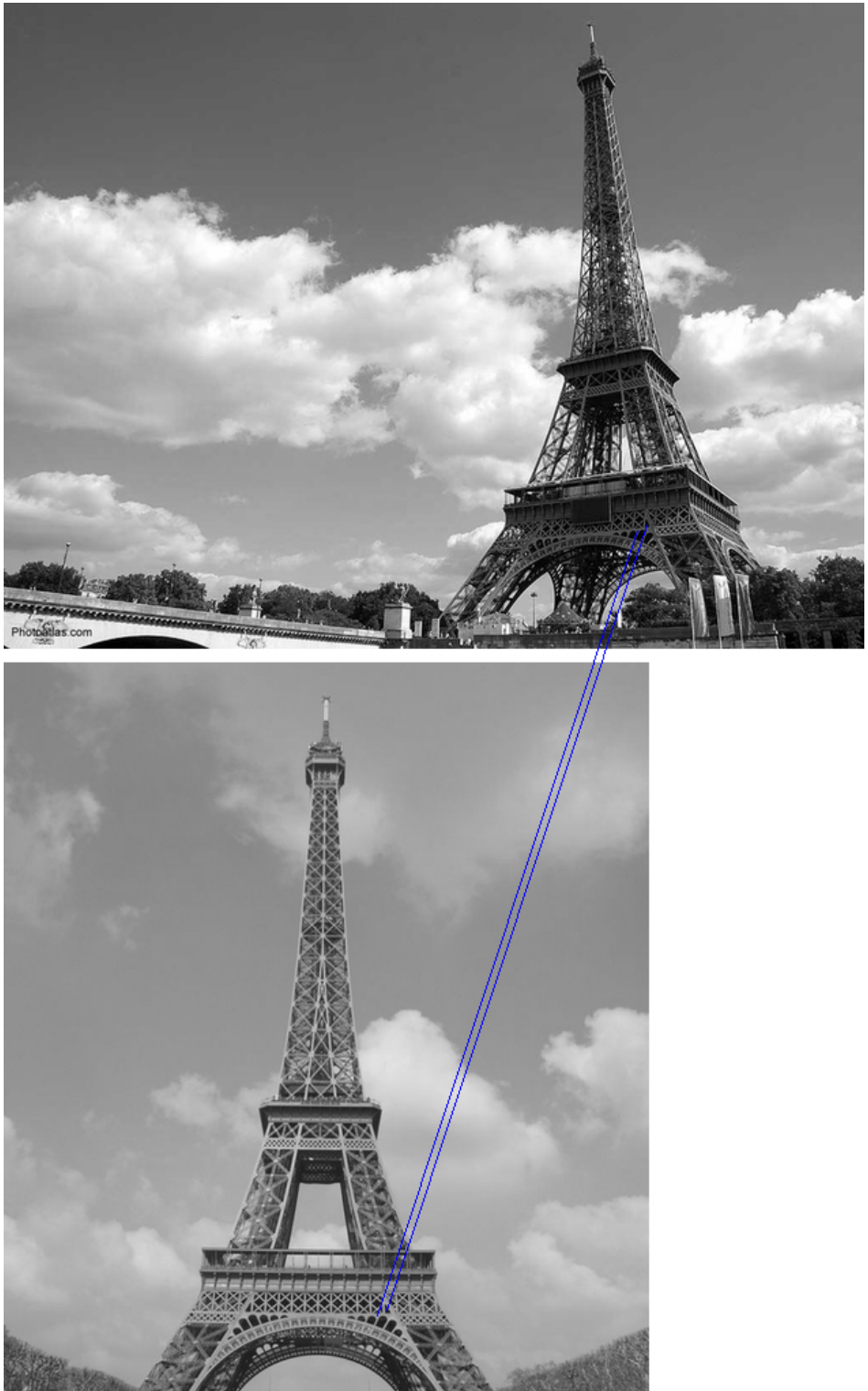






Panorama (Caposele)




Photoatlas.com

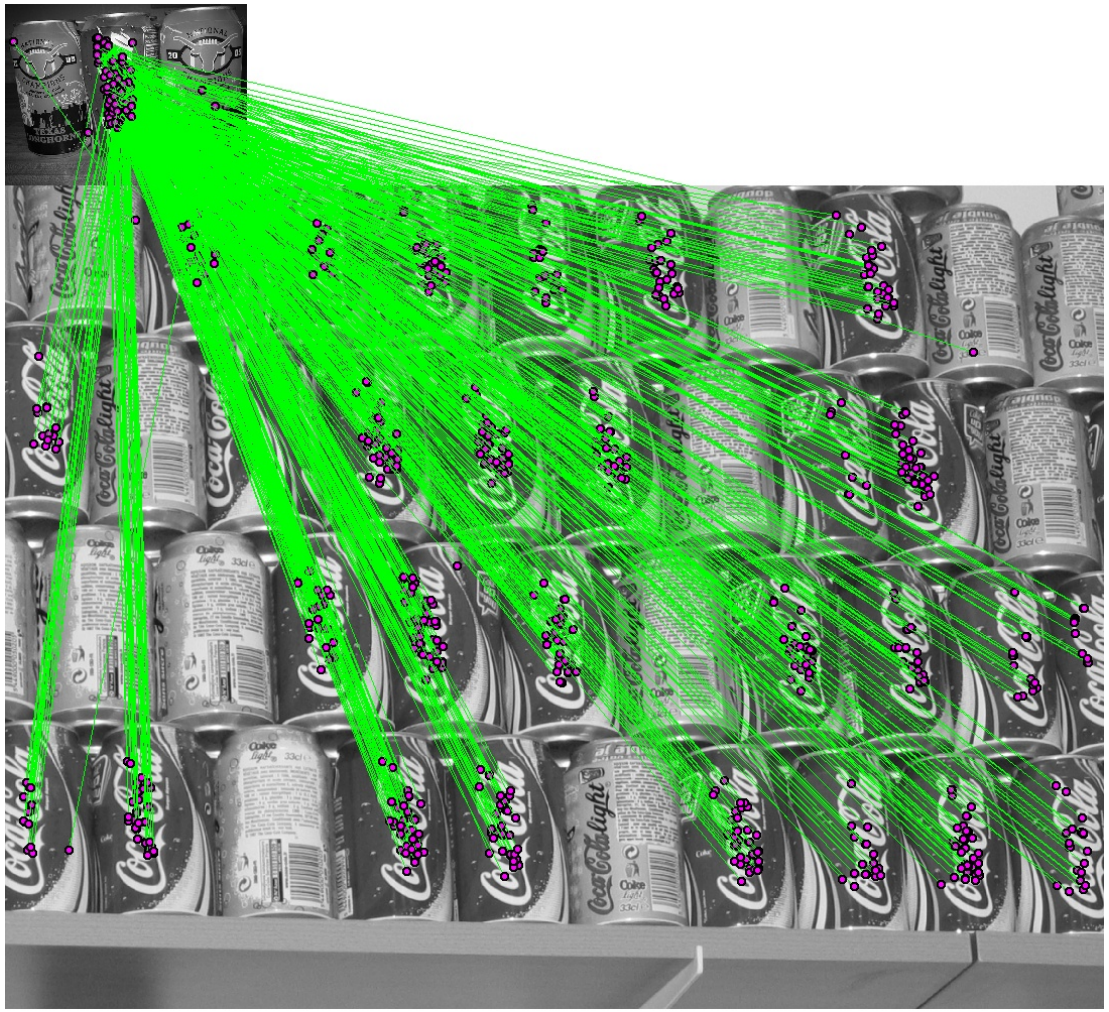


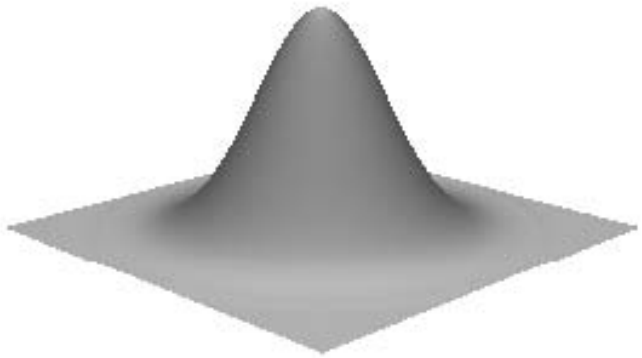
Figure 2: Coke cans: Multi scale matches (Rabin, Gousseau, Delon, method which eliminates false alarms)

Linear image filtering is mainly done by convolving an image u with a positive integrable kernel g, **which simulates a camera blur**. 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 (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;*

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

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

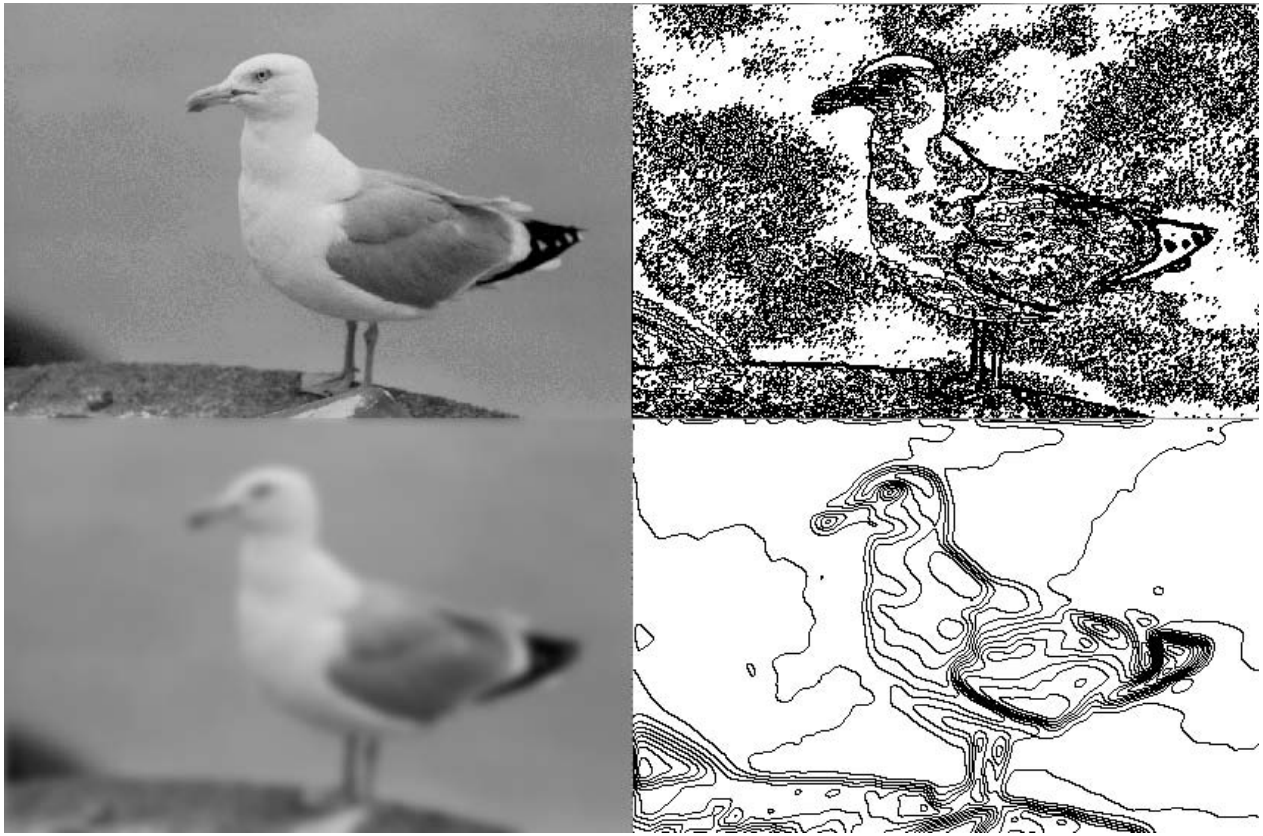


Figure 3: Level lines and the heat equation. Top: original; level lines for levels multiple of 12. Bottom: image smoothed by the heat equation (convolution with the Gaussian, standard deviation 4.)

The Gaussian and the scale space

- $\mathbf{G}_\sigma(x_1, x_2) = \frac{1}{2\pi(\mathbf{c}\sigma)^2} e^{-\frac{x_1^2+x_2^2}{2(\mathbf{c}\sigma)^2}}$, $\mathbf{G}_\sigma$ satisfies the heat equation

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

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

- $\mathbf{G}$: convolution operator on $\mathbb{R}^2$,

$$\mathbf{Gu}(\mathbf{x}) =: (\mathbf{G} * \mathbf{u})(\mathbf{x}) = \int_{\mathbb{R}^2} \mathbf{G}(\mathbf{y})\mathbf{u}(\mathbf{x}-\mathbf{y})d\mathbf{y}.$$

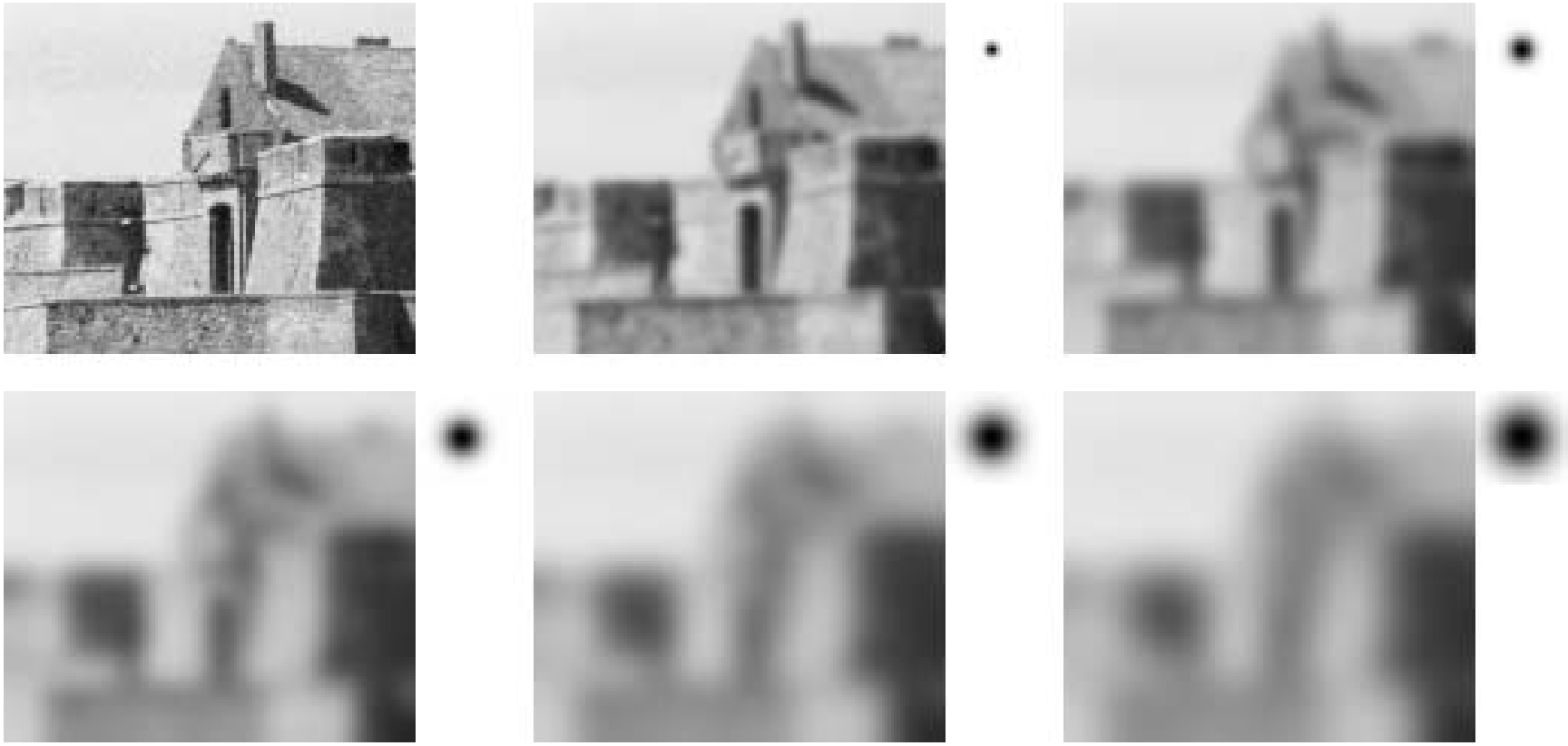
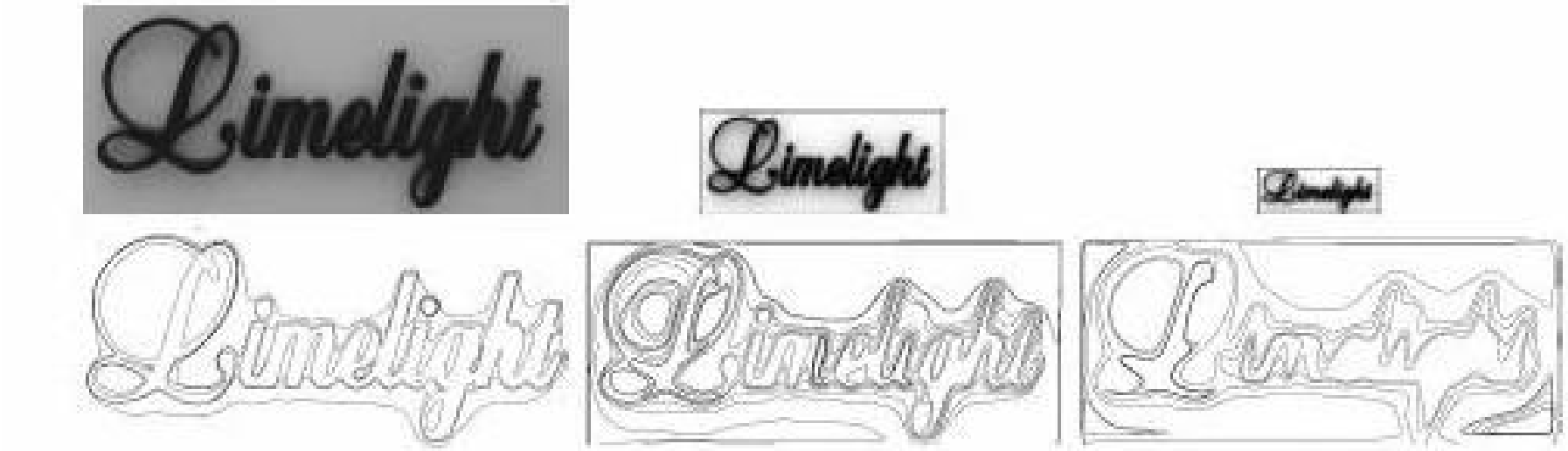


Figure 4: Convolution with Gaussian kernels (heat equation).

Distance means blur!



From continuous to digital and conversely

- $\mathbf{u}(\mathbf{x})$: a continuous and bounded image defined for every $\mathbf{x} = (x, y) \in \mathbb{R}^2$.
- u: a digital image, only defined for $(n_1, n_2) \in \mathbb{Z}^2$.
- $\mathbf{S}_\delta$: the sampling operator at rate $\delta > 0$. Let $\mathbf{u}$ be a continuous image on $\mathbb{R}^2$. The associated sampled digital image $\mathbf{S}_\delta \mathbf{u}$ is defined on $\mathbb{Z}^2$ by

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

From a digital image back to a continuous image by Shannon interpolation

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

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

- $\mathbf{S}_1 Iu = u$. Conversely, if $\mathbf{u}$ is L^2 and band-limited in $(-\pi, \pi)^2$, then $I\mathbf{S}_1\mathbf{u} = \mathbf{u}$.
- If $\mathbf{c} \geq 0.8$, $\mathbf{G}_1\mathbf{u}_0$ is experimentally "well-sampled" and therefore $I\mathbf{S}_1\mathbf{G}_1\mathbf{u}_0 = \mathbf{G}_1\mathbf{u}_0$.

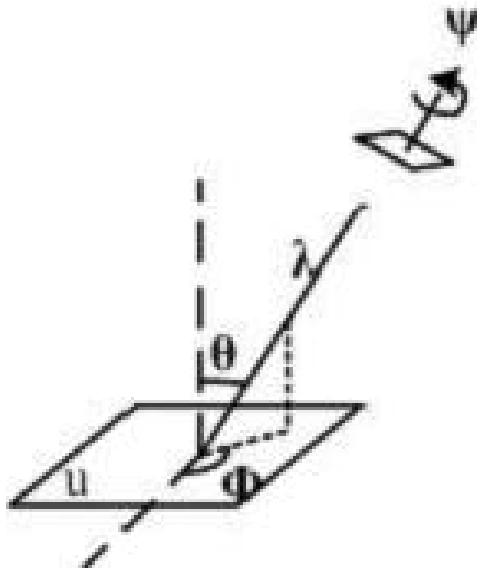
Definition We model all digital images obtained from a given ideal planar object whose frontal infinite resolution image is $\mathbf{u}_0$ by

$$u =: \mathbf{S}_1 \mathbf{G}_\delta \mathbf{A} \mathbf{u}_0$$

where $\mathbf{A}$ is any affine map *(six parameters)!*.

$$\mathbf{A} = \lambda R_\psi T_{\tan\theta} R_\phi,$$

Rotation, tilt, rotation again, zoom. Model correct if λ is large.



SIFT assumptions and condensed description of the method

1. the initial digital image is $\mathbf{S}_1\mathbf{G_c}\mathbf{A}\mathbf{u}_0$, $\mathbf{A}$ is any affine map with positive eigenvalues;
2. at all scales $\sigma > 0$, the SIFT method computes good samplings of $\mathbf{u}(\sigma, \cdot) = \mathbf{G}_\sigma\mathbf{G_c}\mathbf{A}\mathbf{u}_0$;
3. key points $(\sigma, \mathbf{x})$ are extrema of $\Delta\mathbf{u}(\sigma, \cdot)$ in the scale space;
4. the blurred $\mathbf{u}(\sigma, \cdot)$ image is sampled around each key point at a pace proportional to $\sqrt{\sigma^2 + \mathbf{c}^2}$;
5. directions of the sampling axes are fixed by a dominant direction of $\nabla\mathbf{u}(\sigma, \cdot)$ in a neighborhood of the key point proportional to $\sqrt{\sigma^2 + \mathbf{c}^2}$;
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.

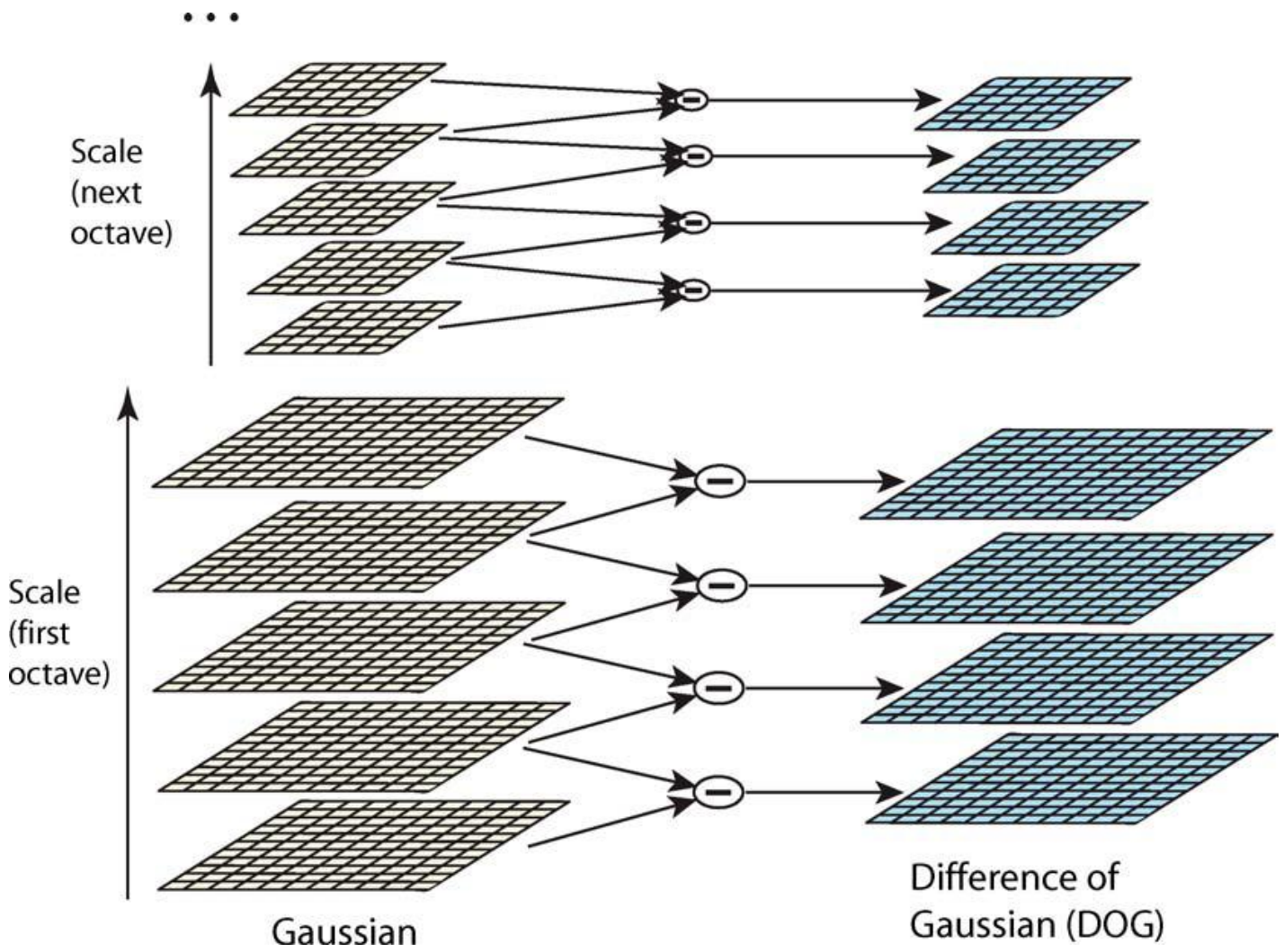


Figure 5: Gaussian pyramid for key points extraction (from Lowe) DOG functions proportional to Laplacian

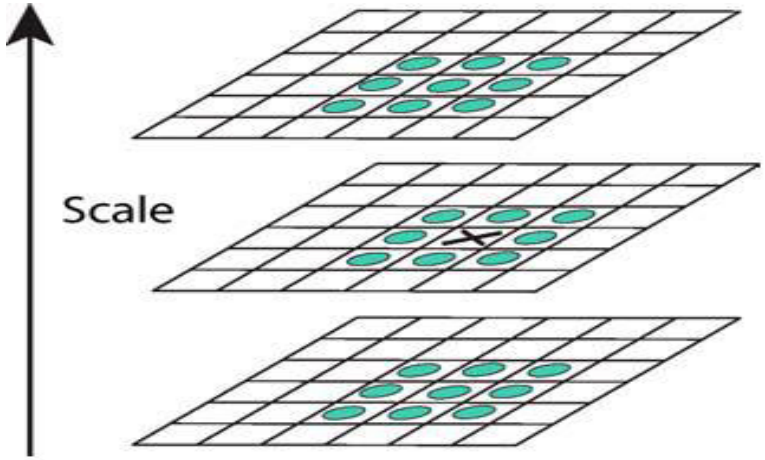


Figure 6: Neighborhood for the location of key points (from Lowe). Local extrema are detected by comparing each sample point in $\mathbb{D}$ with its eight neighbors at scale σ and its nine neighbors in the scales above and below

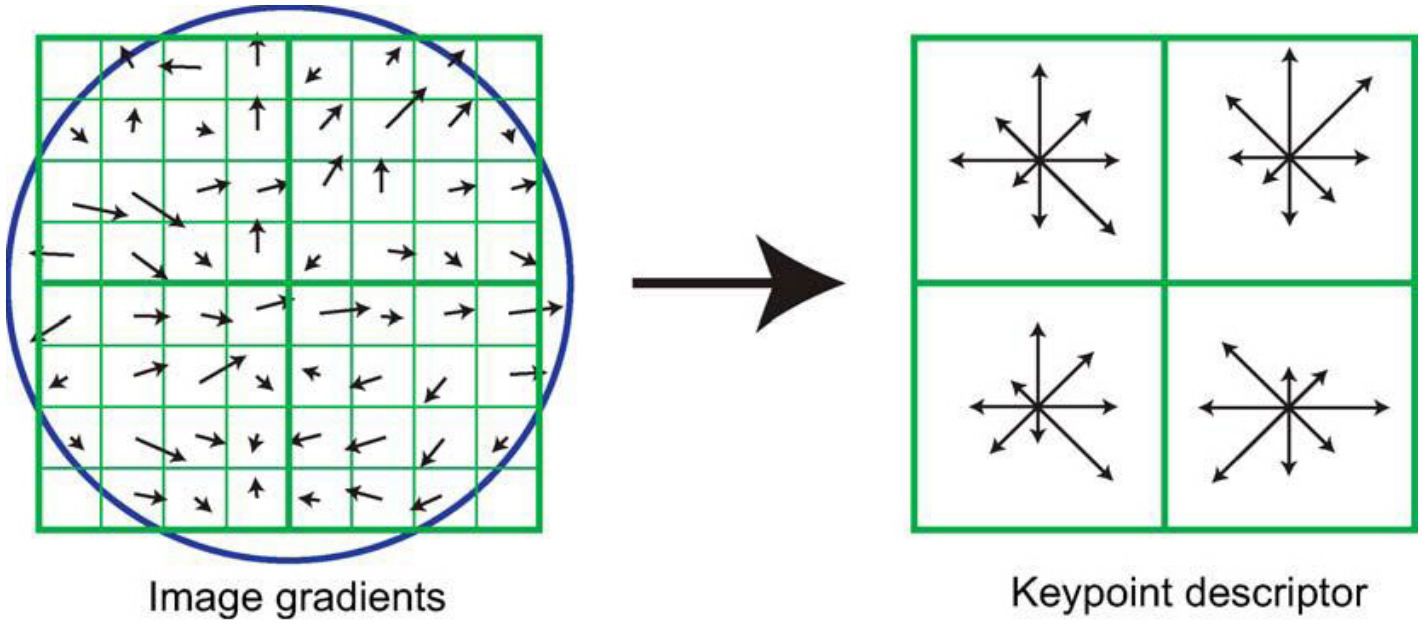


Figure 7: Each key-point is associated *a square image patch whose size is proportional to the scale and whose side direction is given by the assigned direction*. Example of a 2×2 descriptor array of orientation histograms (right) computed from an 8×8 set of samples (left). The orientation histograms are quantized into 8 directions and the length of each arrow corresponds to the magnitude of the histogram entry.

Figure 8: 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 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.

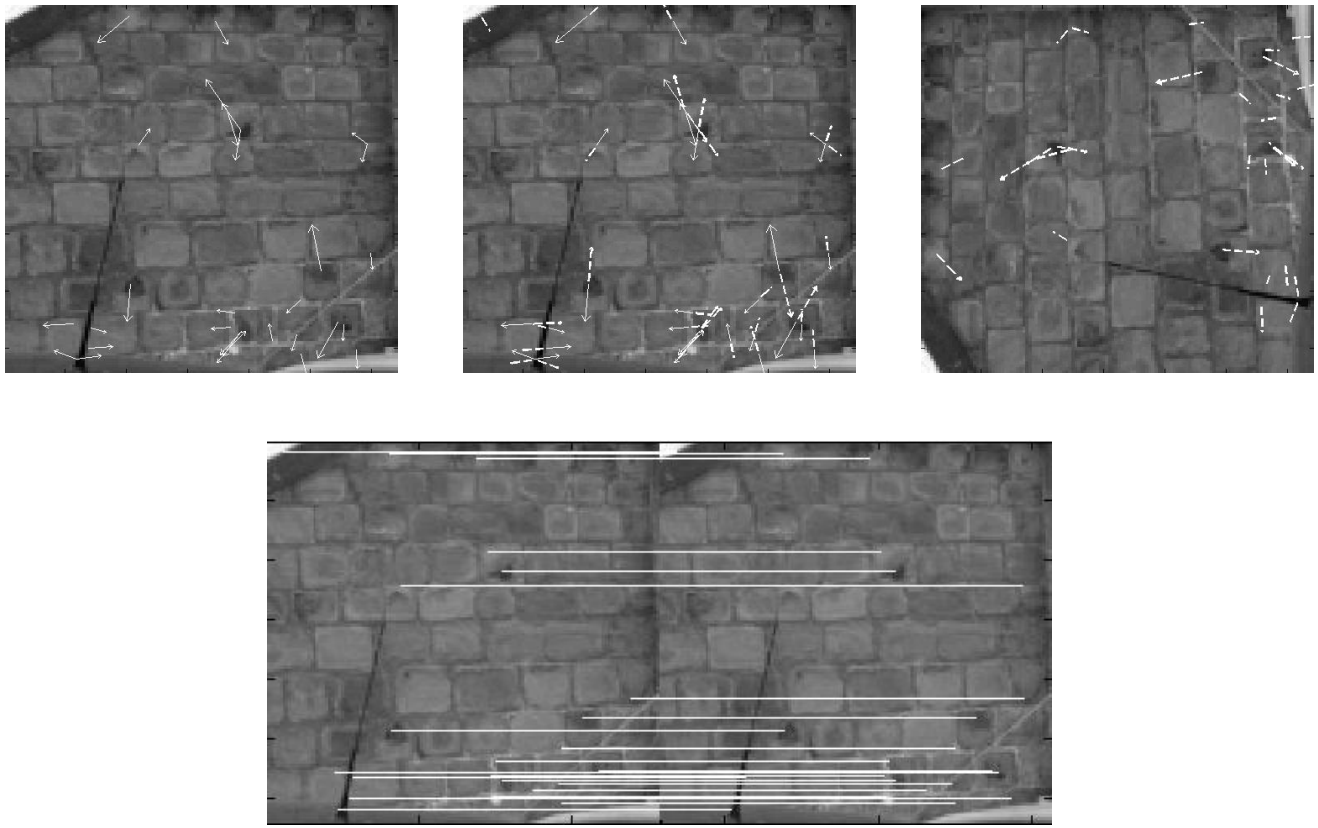


Figure 9: Rotation invariance of SIFT. Top left and right: $\mathbf{u}$ and $\mathbf{R}_{\frac{\pi}{2}}\mathbf{u}$ superposed with their 31 keypoints. Top middle: descriptors of $\mathbf{R}_{\frac{\pi}{2}}\mathbf{u}$ are projected on $\mathbf{u}$ and their orientations are inverted for better observability. 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).

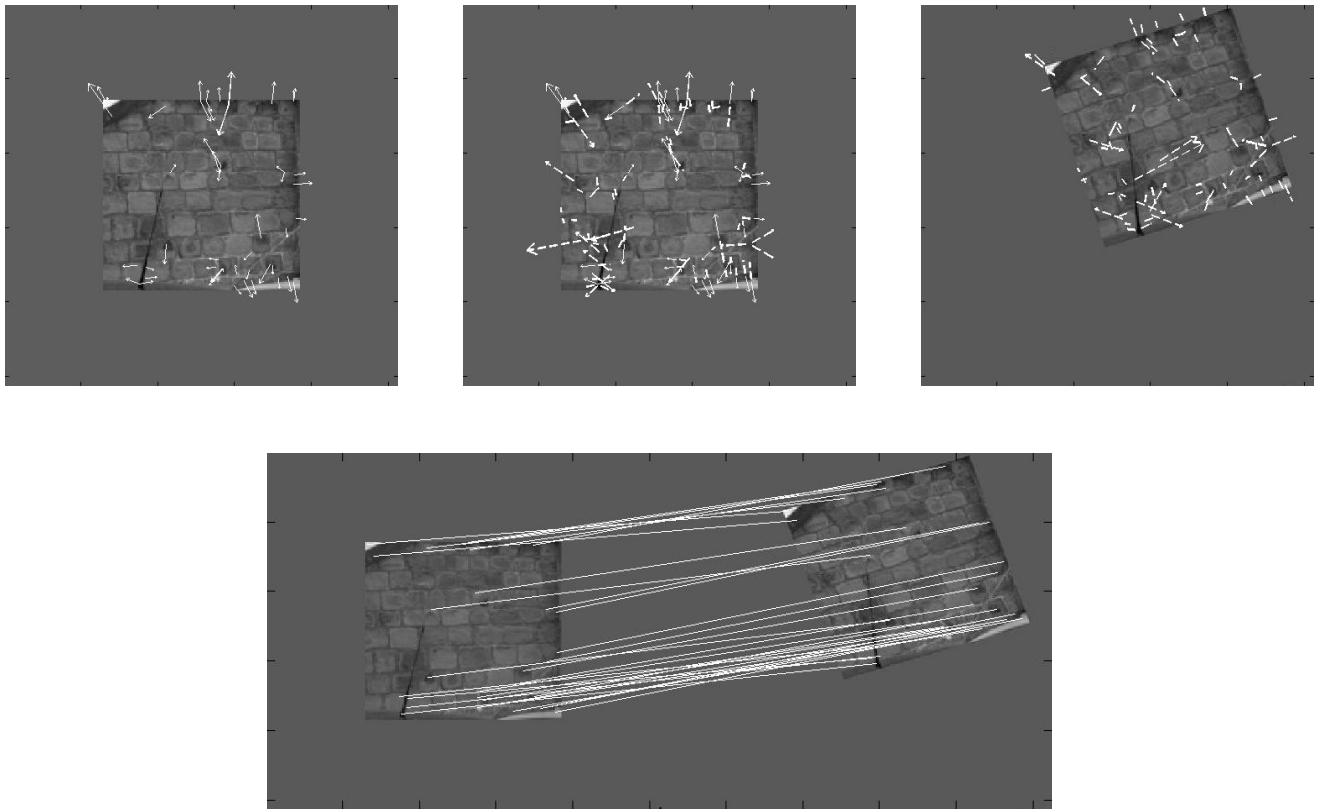


Figure 10: Rotation invariance of SIFT. Top left: $\mathbf{u}$ superposed with its 52 keypoints. Top right: $\mathbf{R}_{\frac{\pi}{10}}\mathbf{u}$ (obtained with Shannon interpolation) superposed with its 73 keypoints. Top middle: descriptors of $\mathbf{R}_{\frac{\pi}{2}}\mathbf{u}$ are projected on $\mathbf{u}$ 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 tentative standard deviation of the initial image blur as described in Section 57. 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 images 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

$$s_0 = \lambda\sqrt{\sigma_1^2+\beta^2} = \mu\sqrt{\sigma_2^2+\delta^2}. \tag{1}$$

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 $u(\sigma_2, \mathbf{x})$.

Theorem 2 *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

$$\lambda\sqrt{\sigma_1^2 + \beta^2} = \mu\sqrt{\sigma_2^2 + \delta^2} \Rightarrow \lambda\sqrt{\sigma_1^2 + \mathbf{c}^2} = \mu\sqrt{\sigma_2^2 + \mathbf{c}^2}. \quad (1)$$

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

$$\lambda^2\beta^2-\mu^2\delta^2=(\lambda^2-\mu^2)\mathbf{c}^2. \tag{1}$$

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

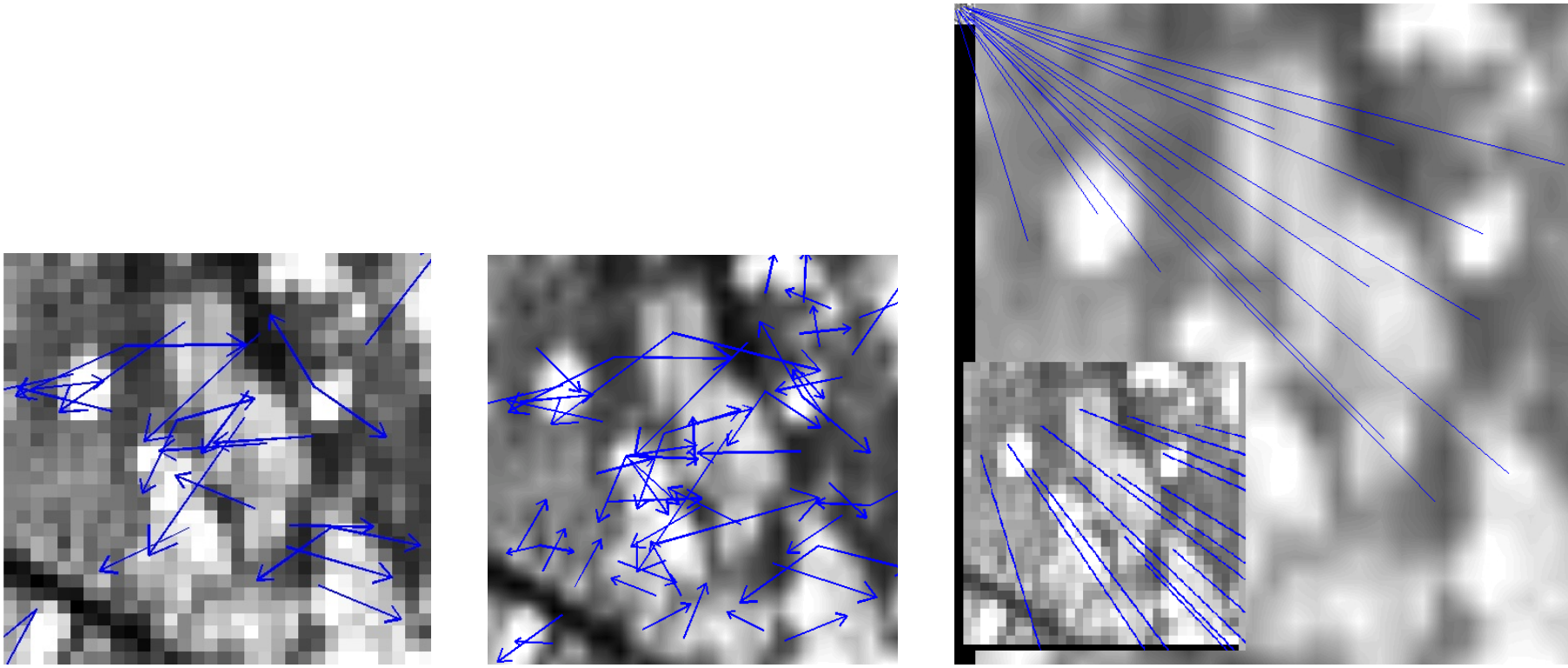
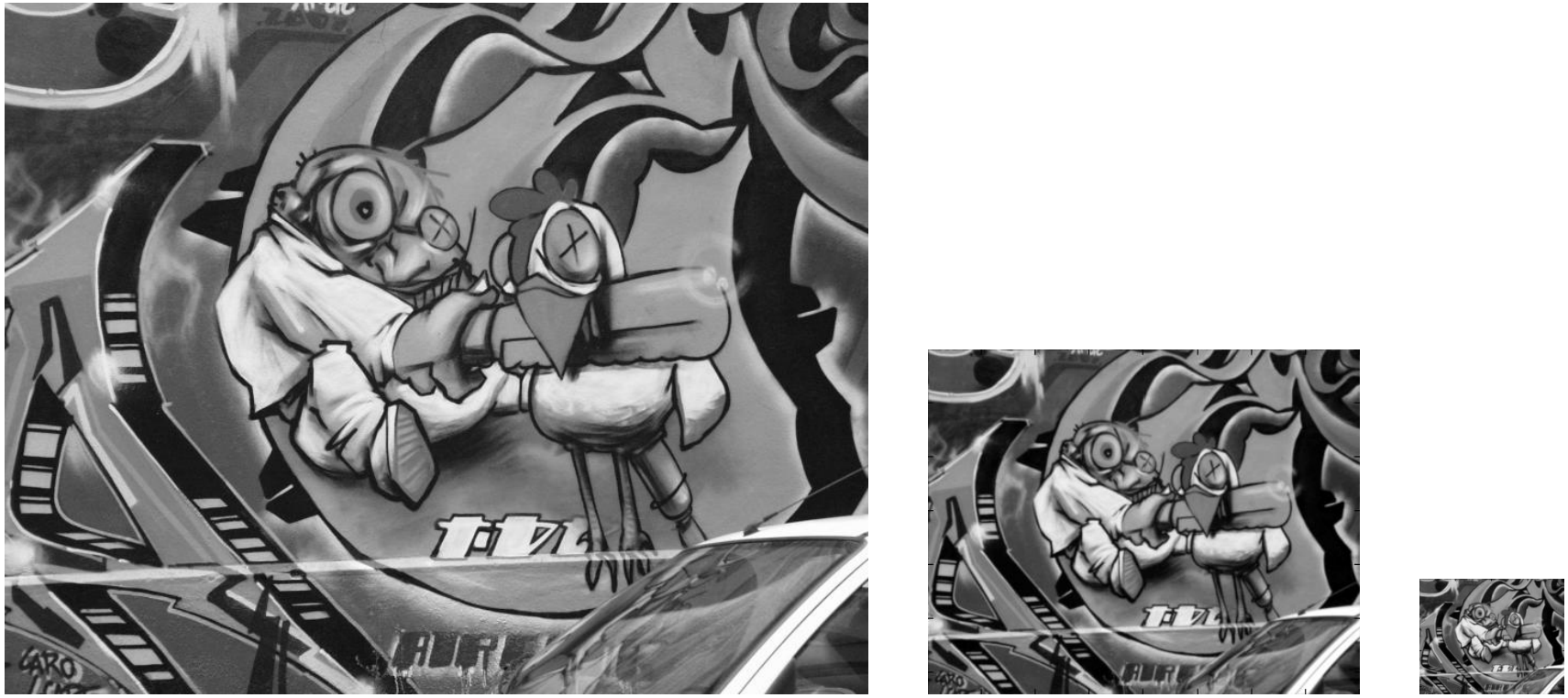


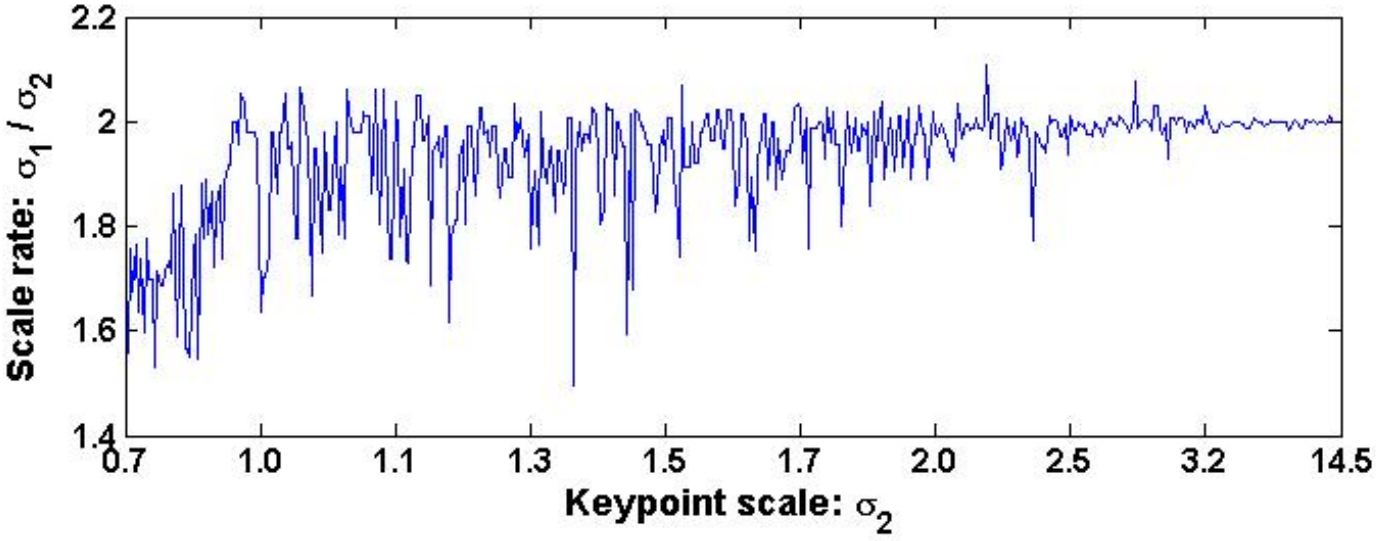
Figure 11: Scale invariance of SIFT, an illustration of Theorem 2. Left: a very small digital image $\mathbf{u}$ with its 25 key points. Middle: this image is over sampled by a 32 factor to $\mathbf{S}_{\frac{1}{32}}\mathbf{I}_d\mathbf{u}$. It has 60 key points. Right: 18 matches found between $\mathbf{u}$ 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.

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 commutation

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

Here the numerical application of the relations in the above proof give: We want (1) 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.




Scale rate: σ_1 / σ_2
2.2
2
1.8
1.6
1.4
0.7
1.0
1.1
1.3
1.5
1.7
2.0
2.5
3.2
14.5
Keypoint scale: σ_2


0.04

A 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.

> 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).

If the tilt is moderate : $u =: \mathbf{G}_1\mathbf{A}\mathbf{u}_0 \simeq \mathbf{A}\mathbf{G}_{\det A}\mathbf{u}_0$. If $\det A$ is not far from 1, *normalization methods can work*: Proposed by (Monasse, Guichard 1997), but more recently known as MSER method (Matas et al.) and LLD : extract connected components of level sets and apply to them an **affine normalization.** *This eliminates the effect of all 6 parameters.* But not scale invariant!

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.

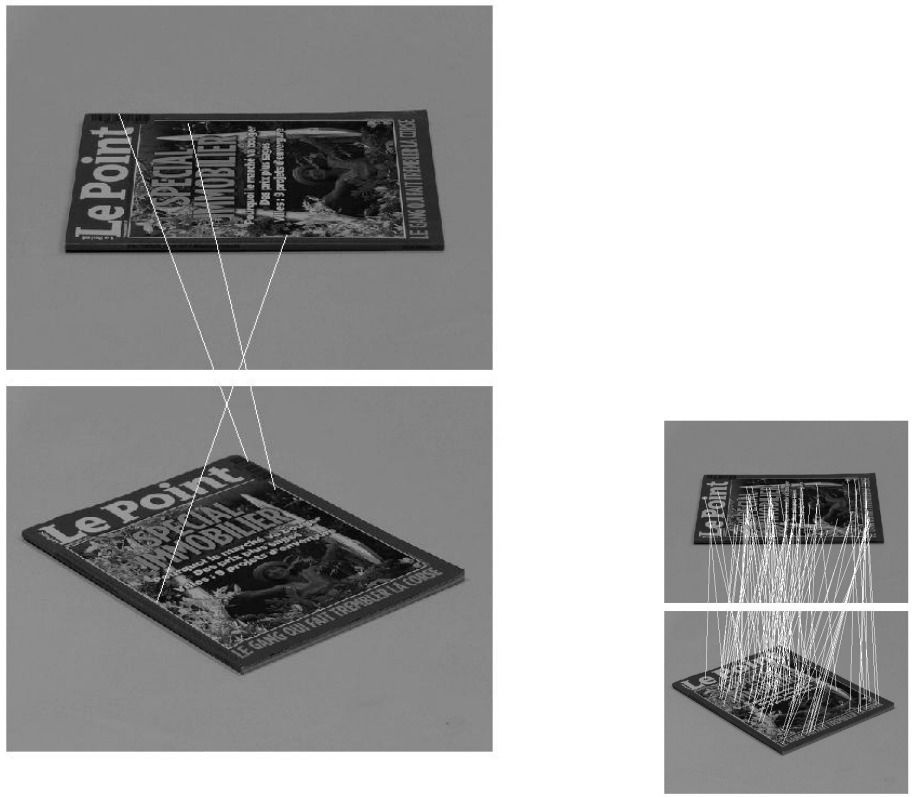


Figure 13: SIFT 3 matches, MSER 87 matches

A second 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 SURF's 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.