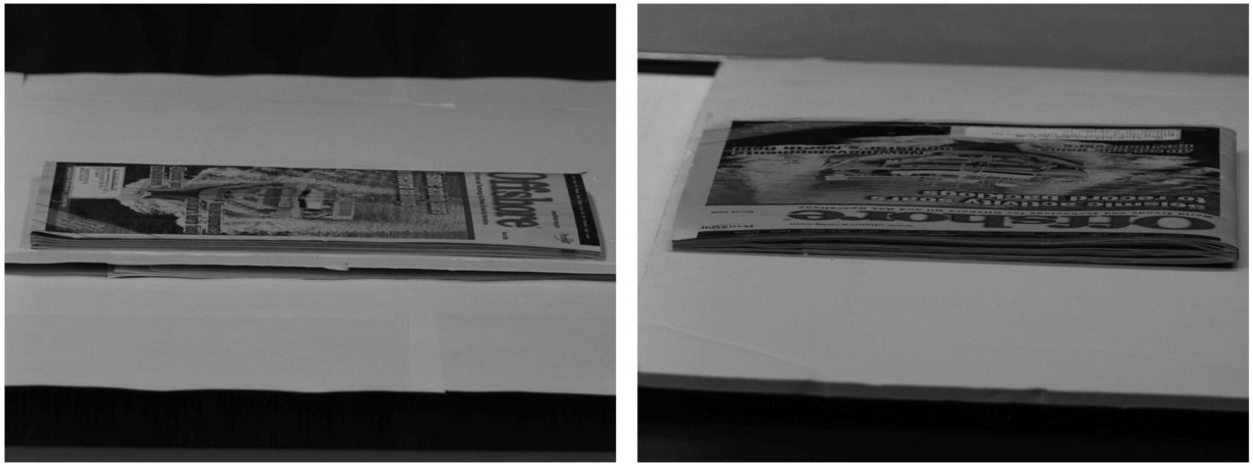
The heat equation for ever

(The SIFT method and its extensions)

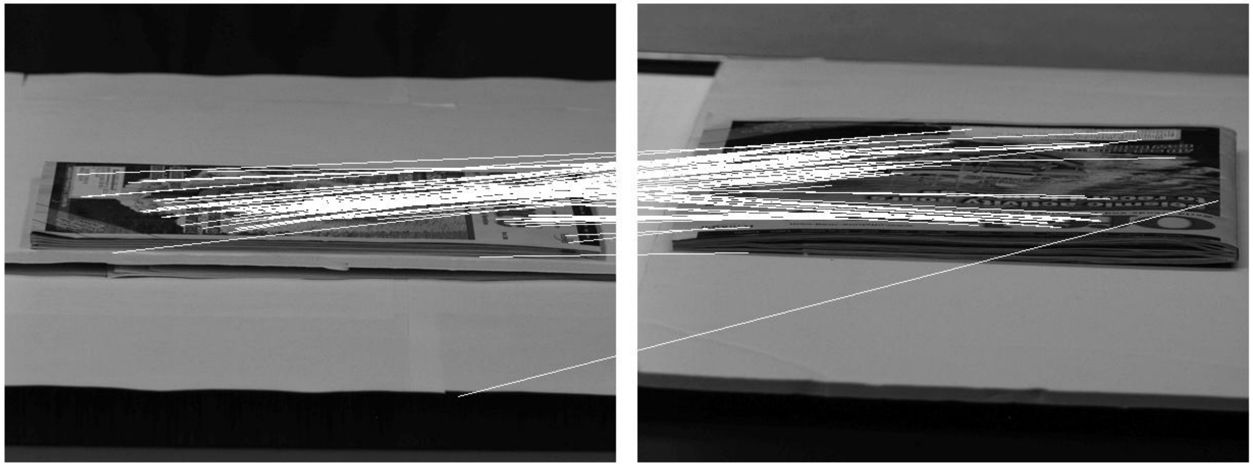
Jean-Michel Morel
ENS Cachan, France
Guoshen Yu,
Ecole Polytechnique, France

D. G. Lowe, *Object recognition from local scale-invariant features,* IJCV 2, 1999



The new state of art: It is by now possible to recognize a solid object in a digital image, no matter what the angle and the distance, up to limits that only depend on resolution.

In this pair: A very large transition tilt (extreme angle); $\simeq 36$. The transition tilt will be defined later.



120 correct matches (not all shown), 4 outliers. Each match is indicated by a white segment

Recognition in spite of a very large view point change. The matches were obtained by the Affine SIFT method (A-SIFT), a variant of the SIFT method. Both methods will be explained.

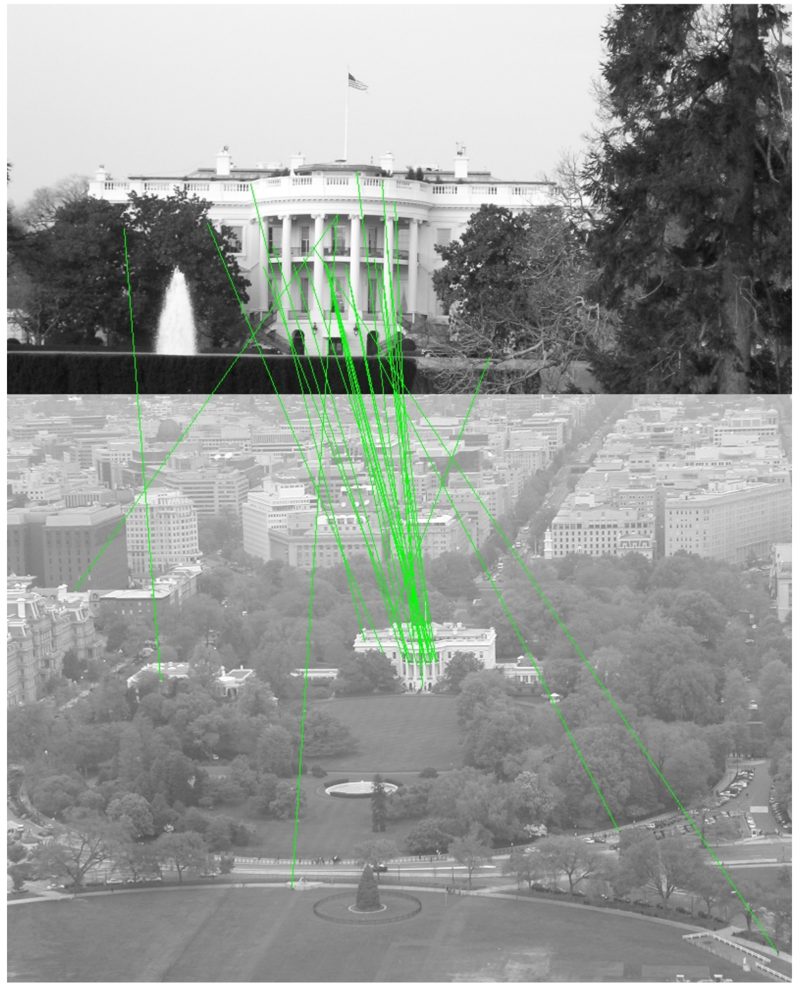


Figure 1: Recognition with extreme scale difference. 26 matches, 6 outliers. Exp. : Rabin, Gousseau, Delon, SIFT method

Camera Model

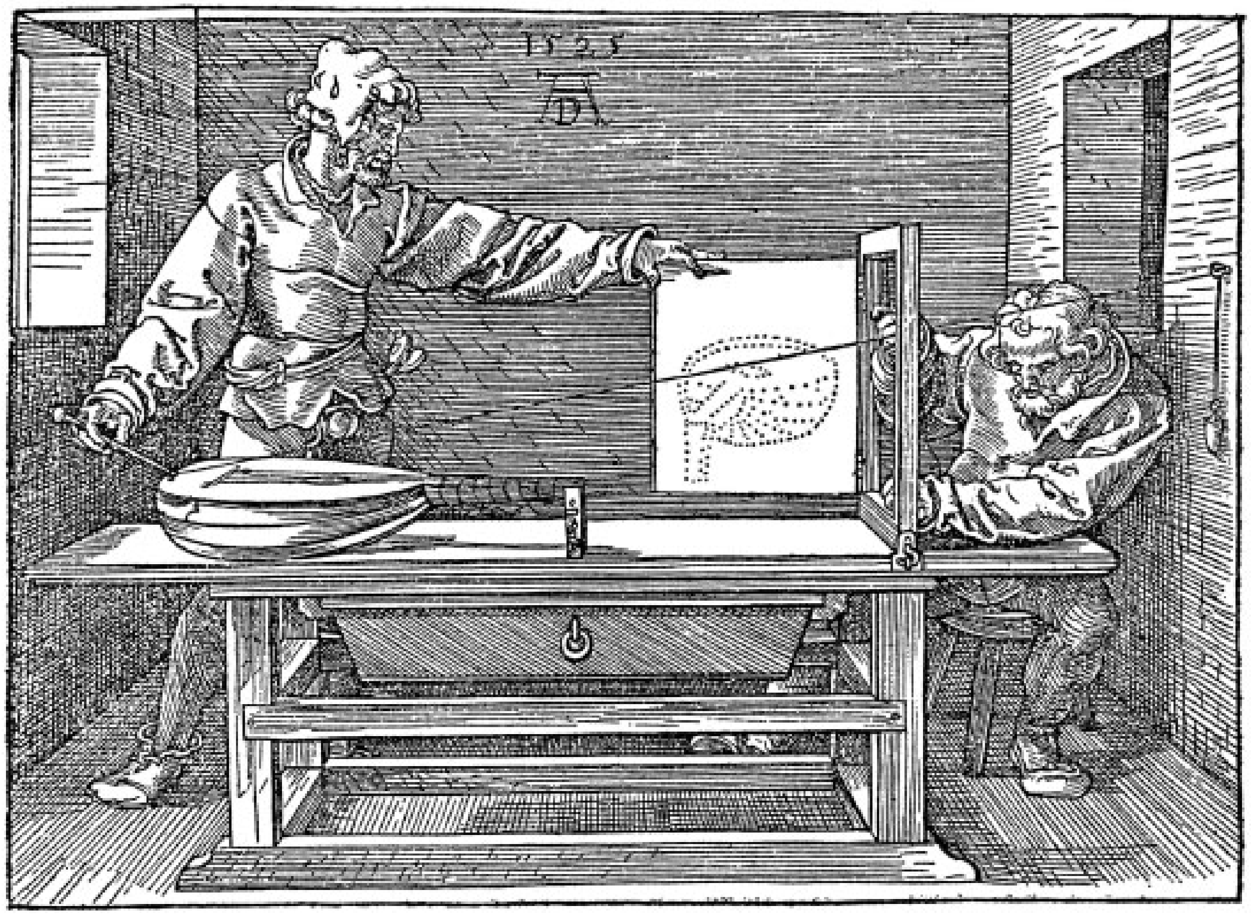
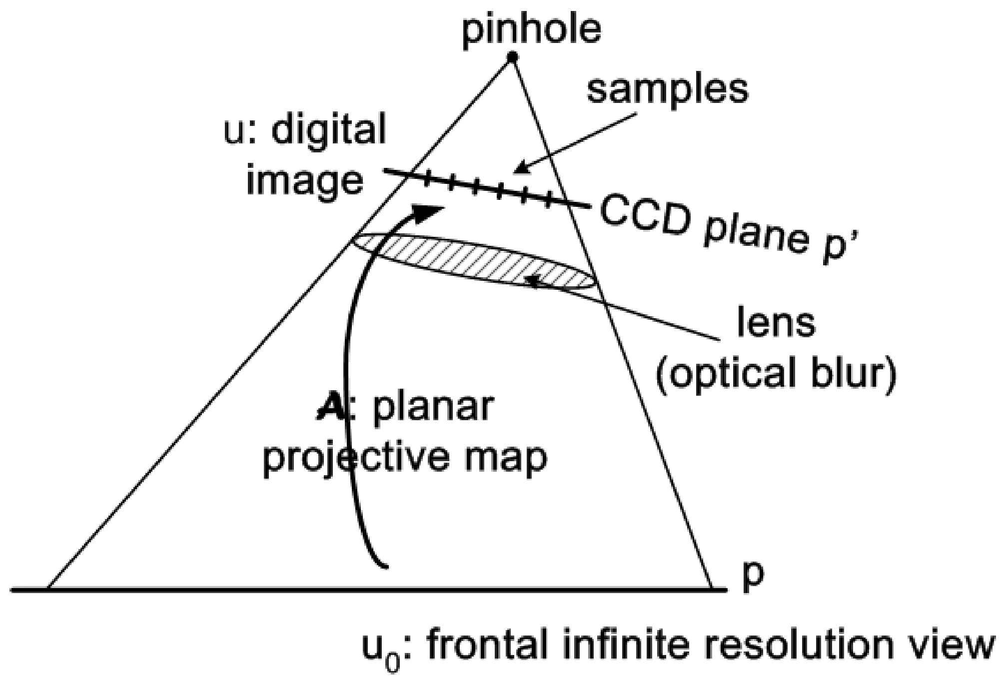


Figure 2: Dürer 1525 "Le Portillon". Illustration of perspective deformation of solid objects

Camera Model



$$u = S_1 \, G_1 \, \mathcal{A} \, u_0$$

u	S_1	G_1	$\mathcal{A}$	u_0
digital image	sampling (grid)	Gaussian kernel (blur)	planar projective map	original infinite resolution surface

Affine simplification

Figure 3: Uccello's miracle (1465): "*Oh che dolce cosa è questa prospettiva!* " Projective transforms are differentiable and therefore locally equivalent to affine transforms. The room is a trapezoid, but it is paved with parallelograms. This means that affine invariance is enough for shape recognition.

Conclusion: the local camera model

All digital images obtained from a locally smooth object whose local frontal view is $\mathbf{u}_0$ satisfy, locally,

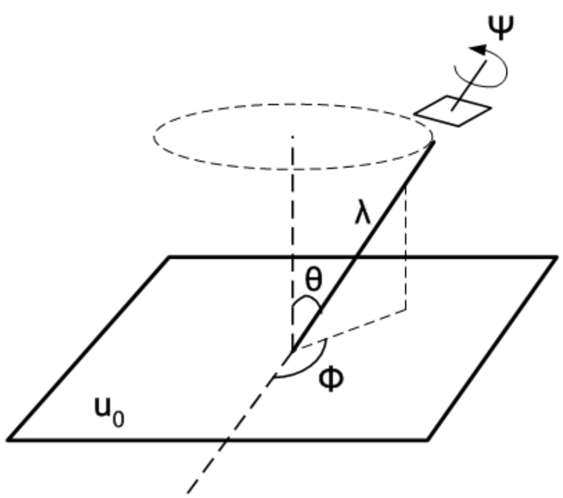
$$u =: \mathbf{S}_1 \mathbf{G}_1 \mathbf{A} \mathbf{u}_0$$

for some planar affine map $\mathbf{A}$ *(six parameters)!*.

Geometric interpretation of the six affine parameters

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$$

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \mathbf{H}_\lambda \mathbf{R}_1(\psi)\mathbf{T}_t\mathbf{R}_2(\phi) = \lambda \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$$

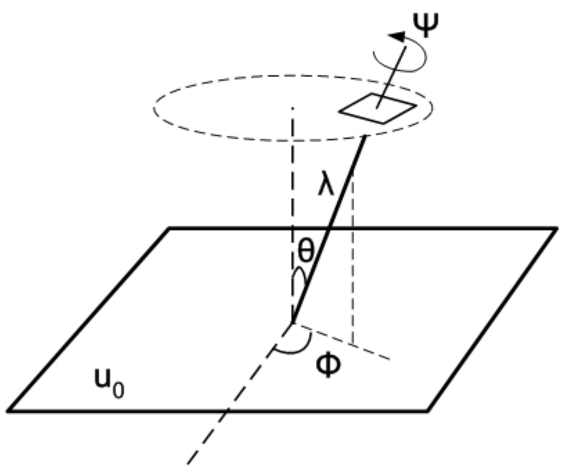


- ϕ: *longitude* angle between optical axis and a fixed vertical plane.
- $\theta = \arccos(1/t)$: *latitude* angle between optical axis and the normal to the image plane. **Tilt** $t > 1 \leftrightarrow \theta \in [0^\circ, 90^\circ]$.
- ψ: rotation angle of camera around optical axis.
- λ: *zoom* parameter.
- $\mathcal{T} = (e, f)^T$: translation, not presented here.

Geometric interpretation of the six affine parameters

$$\begin{pmatrix} x \\ y \end{pmatrix} \rightarrow \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$$

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \mathbf{H}_\lambda \mathbf{R}_1(\psi)\mathbf{T}_t\mathbf{R}_2(\phi) = \lambda \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$$

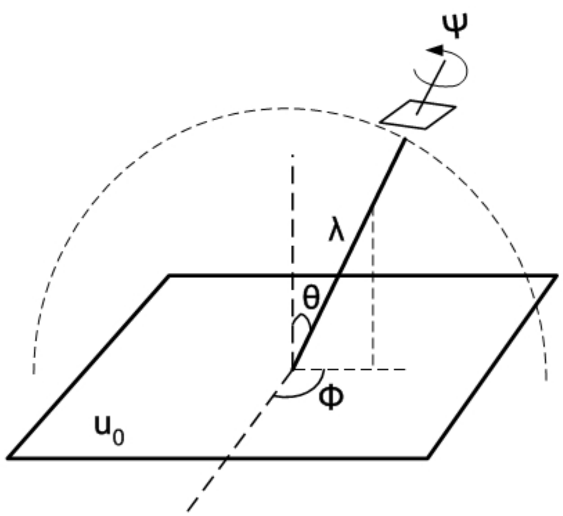


- **ϕ: *longitude* angle between optical axis and a fixed vertical plane.**
- $\theta = \arccos(1/t)$: *latitude* angle between optical axis and the normal to the image plane. **Tilt** $t > 1 \leftrightarrow \theta \in [0°, 90°]$.
- ψ: rotation angle of camera around optical axis.
- λ: *zoom* parameter.
- $\mathcal{T} = (e, f)^T$: translation, not presented here.

Geometric interpretation of the six affine parameters

$$\begin{pmatrix} x \\ y \end{pmatrix} \longrightarrow \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$$

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \mathbf{H}_\lambda \mathbf{R}_1(\psi) \mathbf{T}_t \mathbf{R}_2(\phi) = \lambda \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$$

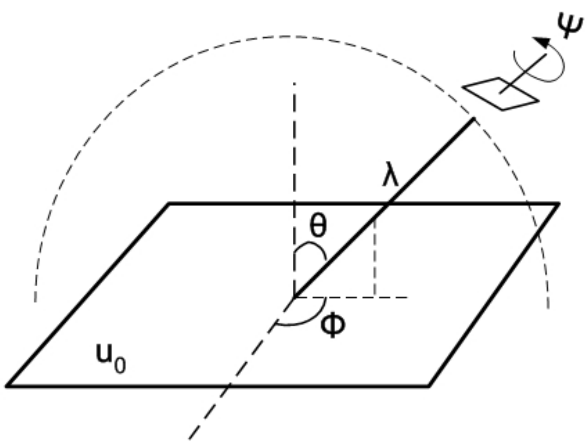


- ϕ: *longitude* angle between optical axis and a fixed vertical plane.
- $\theta = \arccos(1/t)$: ***latitude* angle between optical axis and the normal to the image plane.**
  Tilt $t > 1 \leftrightarrow \theta \in [0^\circ, 90^\circ]$.
- ψ: rotation angle of camera around optical axis.
- λ: *zoom* parameter.
- $\mathcal{T} = (e, f)^T$: translation, not presented here.

Geometric interpretation of the six affine parameters

$$\begin{pmatrix} x \\ y \end{pmatrix} \longrightarrow \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$$

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \mathbf{H}_\lambda \mathbf{R}_1(\psi) \mathbf{T}_t \mathbf{R}_2(\phi) = \lambda \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$$

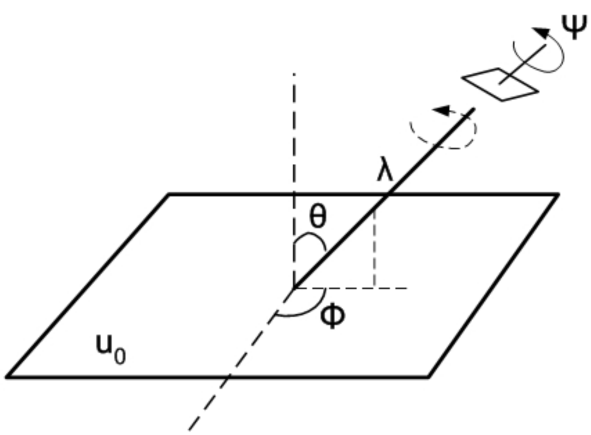


- ϕ: *longitude* angle between optical axis and a fixed vertical plane.
- $\theta = \arccos(1/t)$: ***latitude*** **angle between optical axis and the normal to the image plane.**
  Tilt $t > 1 \leftrightarrow \theta \in [0^\circ, 90^\circ]$.
- ψ: rotation angle of camera around optical axis.
- λ: *zoom* parameter.
- $\mathcal{T} = (e, f)^T$: translation, not presented here.

Geometric interpretation of the six affine parameters

$$\begin{pmatrix} x \\ y \end{pmatrix} \rightarrow \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$$

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \mathbf{H}_\lambda \mathbf{R}_1(\psi) \mathbf{T}_t \mathbf{R}_2(\phi) = \lambda \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$$

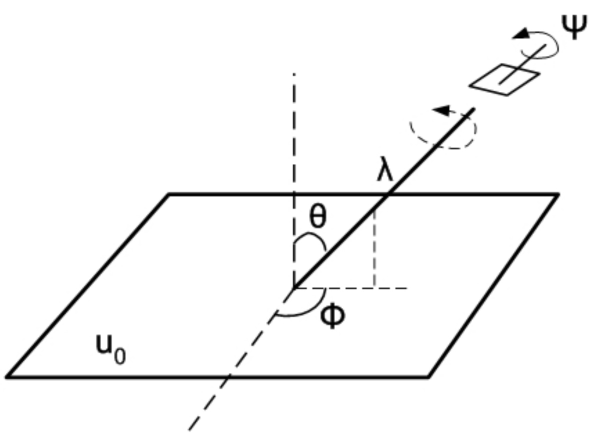


- ϕ: *longitude* angle between optical axis and a fixed vertical plane.
- $\theta = \arccos(1/t)$: *latitude* angle between optical axis and the normal to the image plane. **Tilt** $t > 1 \leftrightarrow \theta \in [0°, 90°]$.
- ψ: **rotation angle of camera around optical axis.**
- λ: *zoom* parameter.
- $\mathcal{T} = (e, f)^T$: translation, not presented here.

Geometric interpretation of the six affine parameters

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$$

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \mathbf{H}_\lambda \mathbf{R}_1(\psi)\mathbf{T}_t \mathbf{R}_2(\phi) = \lambda \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$$

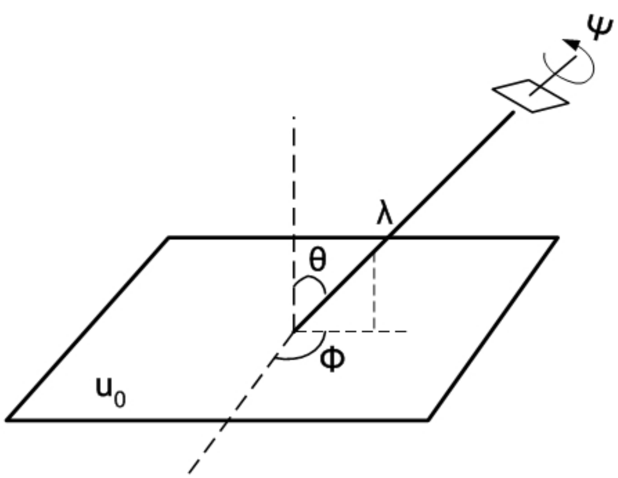


- ϕ: *longitude* angle between optical axis and a fixed vertical plane.
- $\theta = \arccos(1/t)$: *latitude* angle between optical axis and the normal to the image plane. **Tilt** $t > 1 \leftrightarrow \theta \in [0°, 90°]$.
- ψ: **rotation angle of camera around optical axis.**
- λ: *zoom* parameter.
- $\mathcal{T} = (e, f)^T$: translation, not presented here.

Geometric interpretation of the six affine parameters

$$\begin{pmatrix} x \\ y \end{pmatrix} \to \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$$

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \mathbf{H}_\lambda \mathbf{R}_1(\psi) \mathbf{T}_t \mathbf{R}_2(\phi) = \lambda \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$$

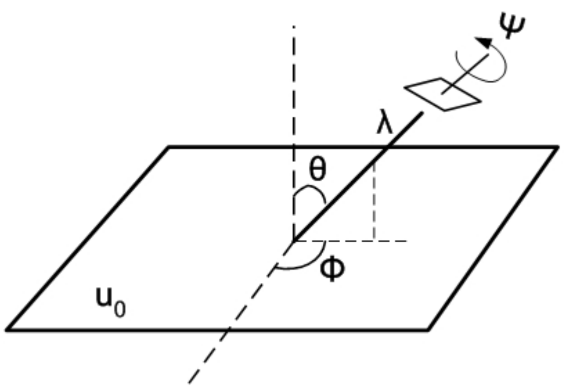


- ϕ: *longitude* angle between optical axis and a fixed vertical plane.
- $\theta = \arccos(1/t)$: *latitude* angle between optical axis and the normal to the image plane. **Tilt** $t > 1 \leftrightarrow \theta \in [0°, 90°]$.
- ψ: rotation angle of camera around optical axis.
- λ: ***zoom* parameter.**
- $\mathcal{T} = (e, f)^T$: translation, not presented here.

Geometric interpretation of the six affine parameters

$$\begin{pmatrix} x \\ y \end{pmatrix} \longrightarrow \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix}$$

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \mathbf{H}_\lambda \mathbf{R}_1(\psi) \mathbf{T}_t \mathbf{R}_2(\phi) = \lambda \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$$



- ϕ: *longitude* angle between optical axis and a fixed vertical plane.
- $\theta = \arccos(1/t)$: *latitude* angle between optical axis and the normal to the image plane. **Tilt** $t > 1 \leftrightarrow \theta \in [0°, 90°]$.
- ψ: rotation angle of camera around optical axis.
- λ: ***zoom* parameter.**
- $\mathcal{T} = (e, f)^T$: translation, not presented here.

For recognition, shapes must be described by local image features that are invariant to 8 parameters!

This leads to 4 Image comparison requirements and to

our PLAN

1. independence from sampling (interpolation required);
2. invariance from illuminance changes (at least 2 parameters);
3. independence from: scale (that means independence of blur!), rotation, translation: 4 parameters, SIFT method
4. independence from tilts: (slanted view angle): 2 parameters, A-SIFT method

If each one of the 8 parameters has 10 values (which is an underestimate) a single comparison of two images would require the simulation of 10^8 different possible views for a single image, followed by 10^8 comparisons.

Problem 1: Independence from sampling: From a digital image back to a continuous image by Shannon interpolation

- $\mathbf{S}_1$: the sampling operator at rate 1. The sampled digital image $u = \mathbf{S}_1\mathbf{u}$ is defined on $\mathbb{Z}^2$ by $u(n_1, n_2) = \mathbf{u}(n_1, n_2)$;
- If $u \in l^2 \cap l^1(\mathbb{Z}^2)$, the Shannon interpolate of u is the only $L^2(\mathbb{R}^2)$ function $\mathbf{u} = Iu$ having u as samples and with spectrum support contained in $(-\pi, \pi)^2$. Then $\mathbf{S}_1 Iu = u$.
- Conversely, if $\mathbf{u}$ is L^2 and band-limited in $(-\pi, \pi)^2$, then $I\mathbf{S}_1\mathbf{u} = \mathbf{u}$.

2 Solution of the second problem : Invariance to illumination conditions

- Contrast change: $g(s)$ increasing, smooth, $\mathbf{u} \rightarrow g(\mathbf{u})$
- Level lines of $\mathbf{u}$ and $g(\mathbf{u})$ are identical, (used in Mathematical Morphology, Matheron, Serra, and recently by the MSER and LLD methods shape recognition methods)
- The direction of the gradient $\frac{\nabla \mathbf{u}}{||\nabla u||}$ also invariant (SIFT method).

3 Solution of the third problem : independence from scale (BLUR !), rotation, and translation: The SIFT method

Figure 4: **Shapes change with distance: The level lines not stable by down-sampling**
This is the main problem with level lines methods (MSER)

Why the heat equation ?

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

$$\frac{\partial G_\sigma}{\partial \sigma} = \sigma \Delta G_\sigma, \quad G_\delta G_\beta = G_{\sqrt{\delta^2+\beta^2}}.$$

- Main assumption: the blur is gaussian

$$u = \mathbf{S}_1 G_s \mathbf{u}_0, \quad (\mathbf{G}_s * \mathbf{u}_0)(\mathbf{x}) =: \int_{\mathbb{R}^2} \mathbf{G}(\mathbf{y}) \mathbf{u}_0(\mathbf{x} - \mathbf{y}) d\mathbf{y}.$$

- If $s \geq 0.6$, Shannon's interpolation conditions are experimentally satisfied: $\mathbf{G}_1 \mathbf{u}_0 =: G_s * \mathbf{u}_0$ is "band-limited" ; $I \mathbf{S}_1 \mathbf{G}_1 \mathbf{u}_0 = \mathbf{G}_1 \mathbf{u}_0$.

- Key property : if $u_1 = \mathbf{S}_s \mathbf{G}_s \mathbf{u}_0$ and $u_2 = \mathbf{S}_t \mathbf{G}_t \mathbf{u}_0$, $t^2 = s^2 + \sigma^2$ then

$$\mathbf{u}_1 = \mathbf{G}_\sigma \mathbf{u}_2; \;\; u_1 = \mathbf{S}_s \mathbf{G}_\sigma I \mathbf{u}_2.$$

Why the heat equation and not other smart nonlinear PDE's ?

- Scale space (Witkin, Koenderink): $\frac{\partial u}{\partial t} = \Delta u$
- Anisotropic diffusion (Perona-Malik): $\frac{\partial u}{\partial t} = div\left(\frac{\nabla u}{1+|\nabla u|^2}\right)$
- Mean curvature motion (Osher, Sethian), commutes with contrast changes! $\frac{\partial u}{\partial t} = |\nabla u| div\left(\frac{\nabla u}{|\nabla u|}\right)$
- Affine scale space (Sapiro, Tannenbaum, Alvarez, Guichard, Lions, M.): commutes with contrast changes and affine maps : $\frac{\partial u}{\partial t} = |\nabla u| div\left(\frac{\nabla u}{|\nabla u|}\right)^{\frac{1}{3}}$

But the reality is: images at different scales obey the heat equation.

SIFT scale invariant features transform

1. the initial digital image is $\mathbf{S}_1\mathbf{G}_1\mathbf{A}\mathbf{u}_0$, $\mathbf{A}$ is any SIMILARITY, $\mathbf{u}_0$ is the underlying infinite resolution planar image;

2. at all scales $\sigma > 0$, the SIFT method computes $\mathbf{u}(\sigma, \cdot) = \mathbf{G}_\sigma\mathbf{G}_1\mathbf{A}\mathbf{u}_0$ and 'key points" $(\sigma, \mathbf{x})$, namely scale and space extrema of $\Delta\mathbf{u}(\sigma, \cdot)$;

3. the blurred $\mathbf{u}(\sigma, \cdot)$ image is sampled around each key point at a pace proportional to $\sqrt{1+\sigma^2}$;

4. directions of the sampling axes are fixed by a dominant direction of $\nabla\mathbf{u}(\sigma, \cdot)$ in a σ-neighborhood of the key point;

5. this yields rotation, translation and scale invariant samples: the 4 parameters of $\mathbf{A}$ have been eliminated!;

6. the final SIFT descriptor keeps only orientations of the gradient to gain invariance w.r. light conditions.

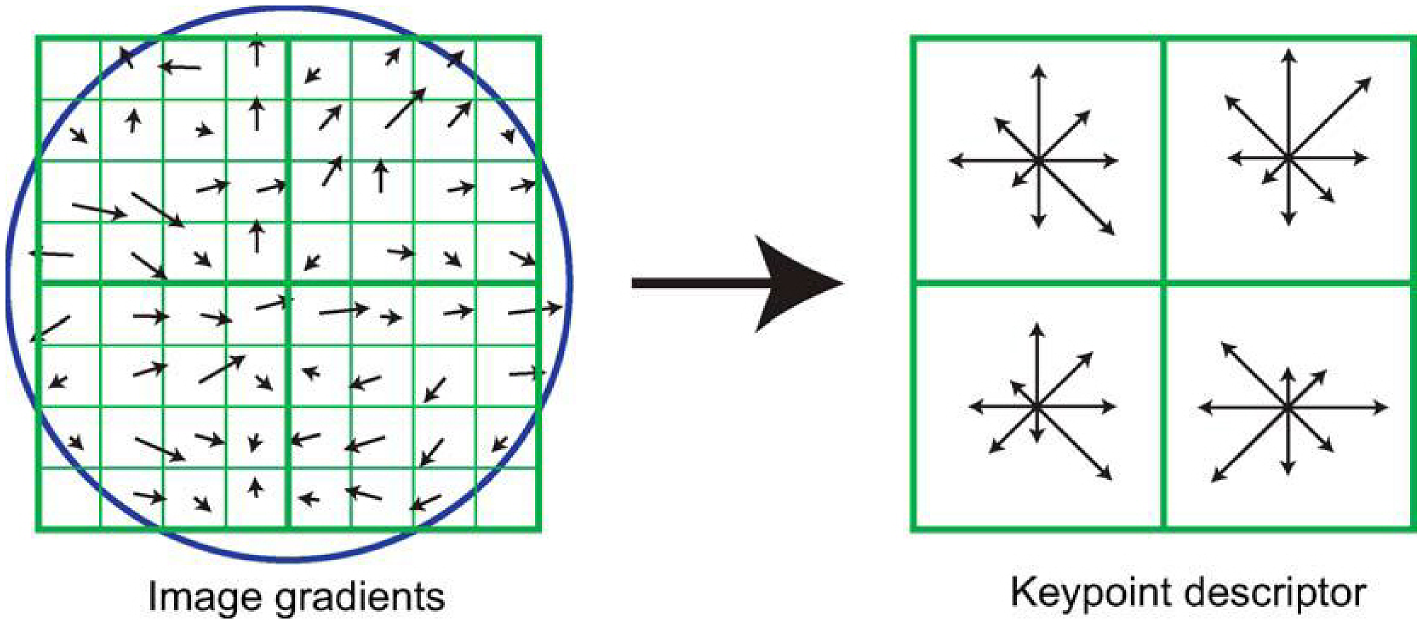


Figure 5: Each key-point is associated *a square image patch whose size is proportional to the scale and whose side direction is given by the assigned direction.* Example of a 2×2 descriptor array of orientation histograms (right) computed from an 8×8 set of samples (left). The orientation histograms are quantized into 8 directions and the length of each arrow corresponds to the magnitude of the histogram entry.

Figure 6: SIFT key points (scale and orientation). Each one is co-variant or invariant with respect to: translation, rotation, scale, and contrast changes (6 parameters out of 8)

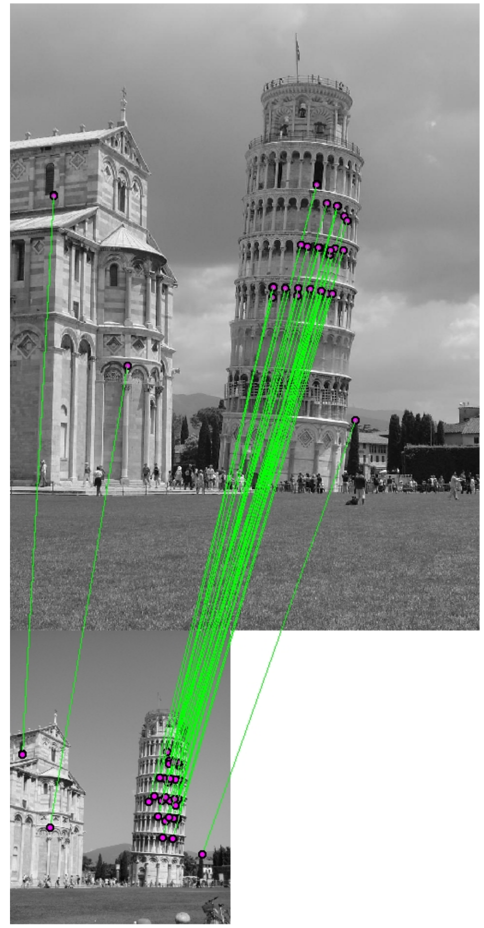
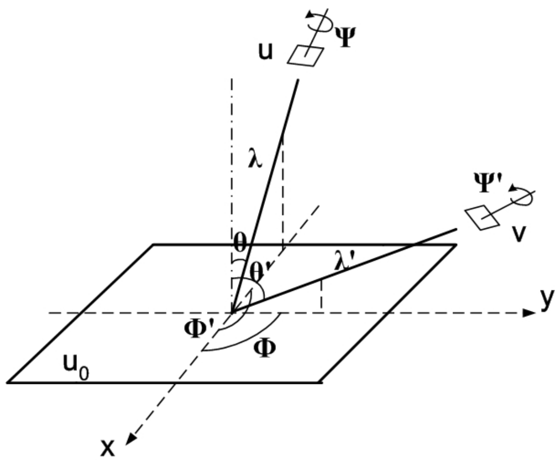
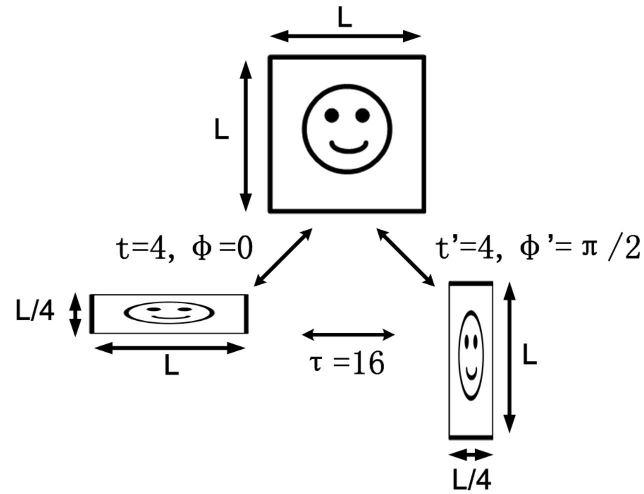


Figure 7: left : the Pisa tower, SIFT method
Outliers elimination method: Rabin, Gousseau, Delon

High Transition Tilts



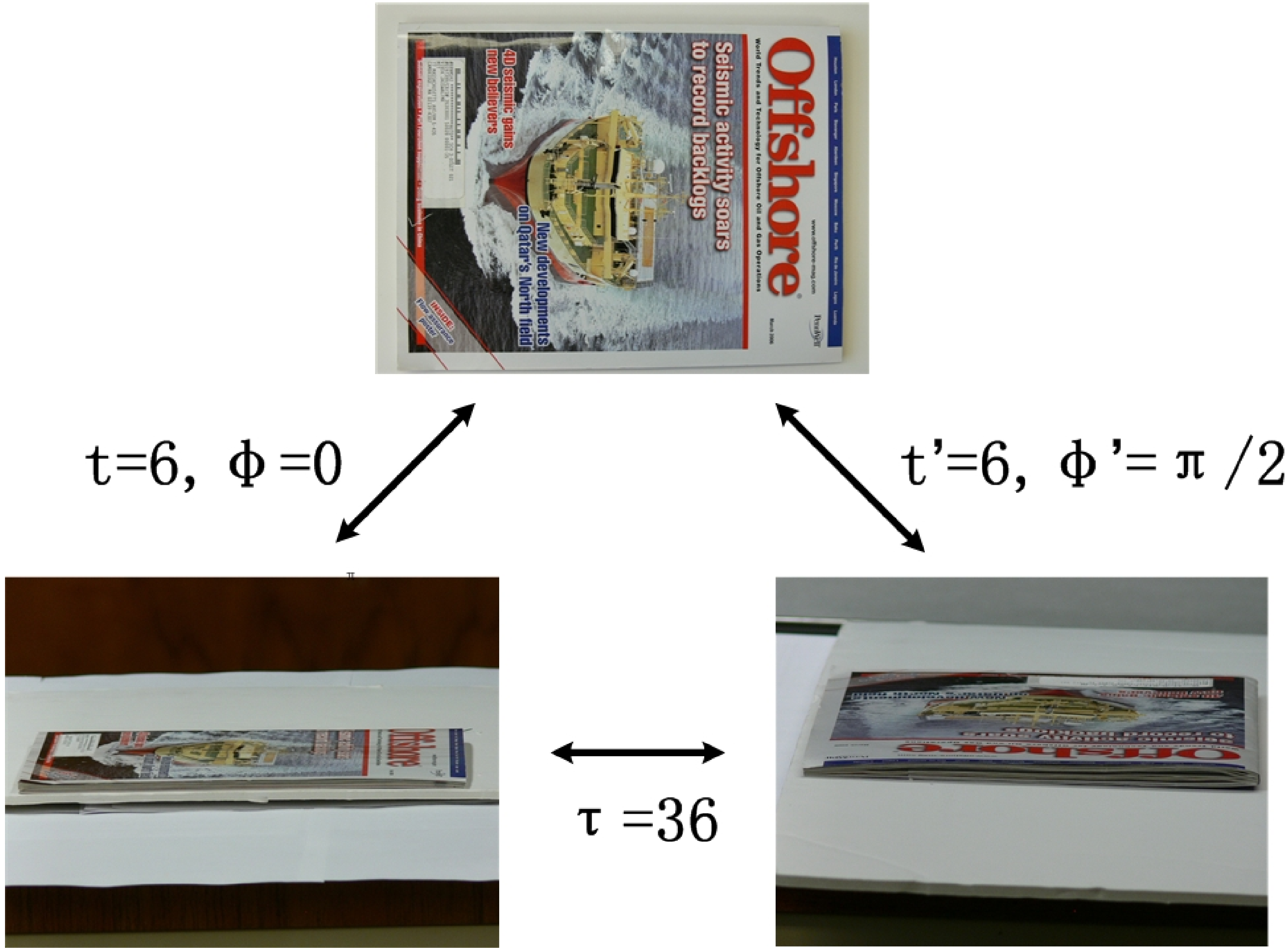


$$\mathbf{u} = \mathbf{G}_1\mathbf{H}_\lambda\mathbf{R}_1(\psi)\mathbf{T}_t\mathbf{R}_2(\phi)\mathbf{u}_0, \quad \mathbf{v} = \mathbf{G}_1\mathbf{H}_{\lambda'}\mathbf{R}_1(\psi')\mathbf{T}_{t'}\mathbf{R}_2(\phi')\mathbf{u}_0$$

$$\mathbf{v} = \mathbf{G}_1\mathbf{H}_\mu\mathbf{R}_1(\psi_1)\mathbf{T}_\tau\mathbf{R}_2(\phi_1)\mathbf{u}$$

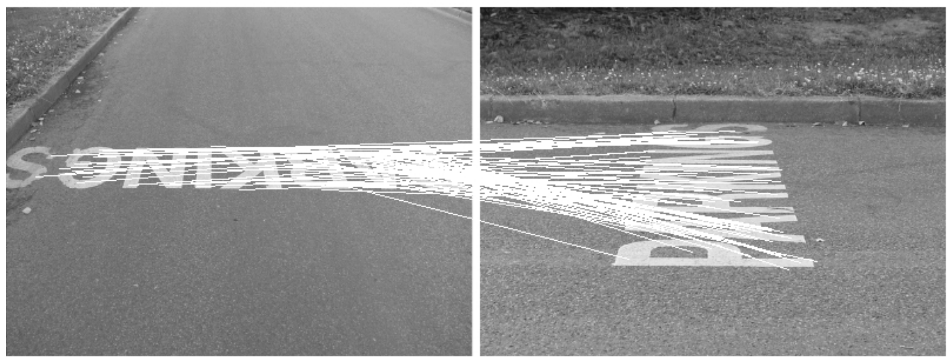
- Absolute tilts t: from $\mathbf{u}$ to $\mathbf{u}_0$.
- **Transition tilts** $\tau(t, t', \phi - \phi')$: the absolute tilt from $\mathbf{u}$ to $\mathbf{v}$ under the assumption that $\mathbf{u}$ is frontal.
- In contrast with absolute tilts, most transition tilts are **LARGE**.

High Transition Tilts



$$\tau = 36 \Rightarrow \theta = \mathbf{88.41}^{\circ}$$

High Transition Tilts: 79 matches (A-SIFT)

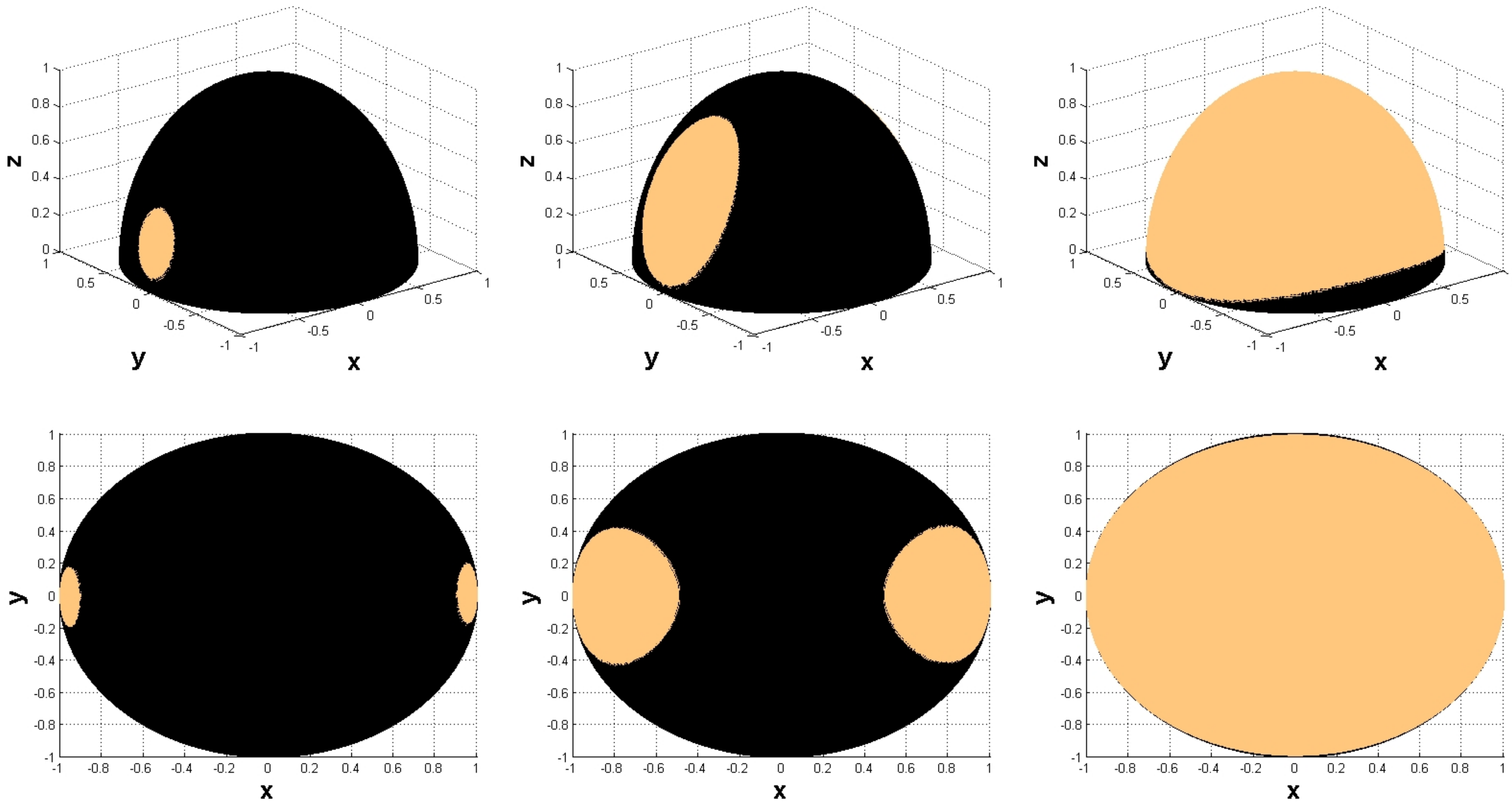


High Transition Tilts: 50 matches (A-SIFT)

State-of-the-art

- SIFT (Scale-Invariant Feature Transform) [Lowe 99, 04]:
  - Rotation and translation are *normalized*.
  - Zoom is *simulated* in the scale space.
  - Modest robustness to tilt: $\tau_{\max} < 2.5$.
- MSER (Maximally Stable Extremal Region) [Matas et al. 02]
  - Zoom and tilt are inverted by *normalization* (but only an approximation: normalization does not commute with blur).
  - Rotation invariance is used (rotation is normalizable).
  - Weakness: few features, limited affine invariance $\tau_{\max} < 5$.

Transition tilts attainable with each method

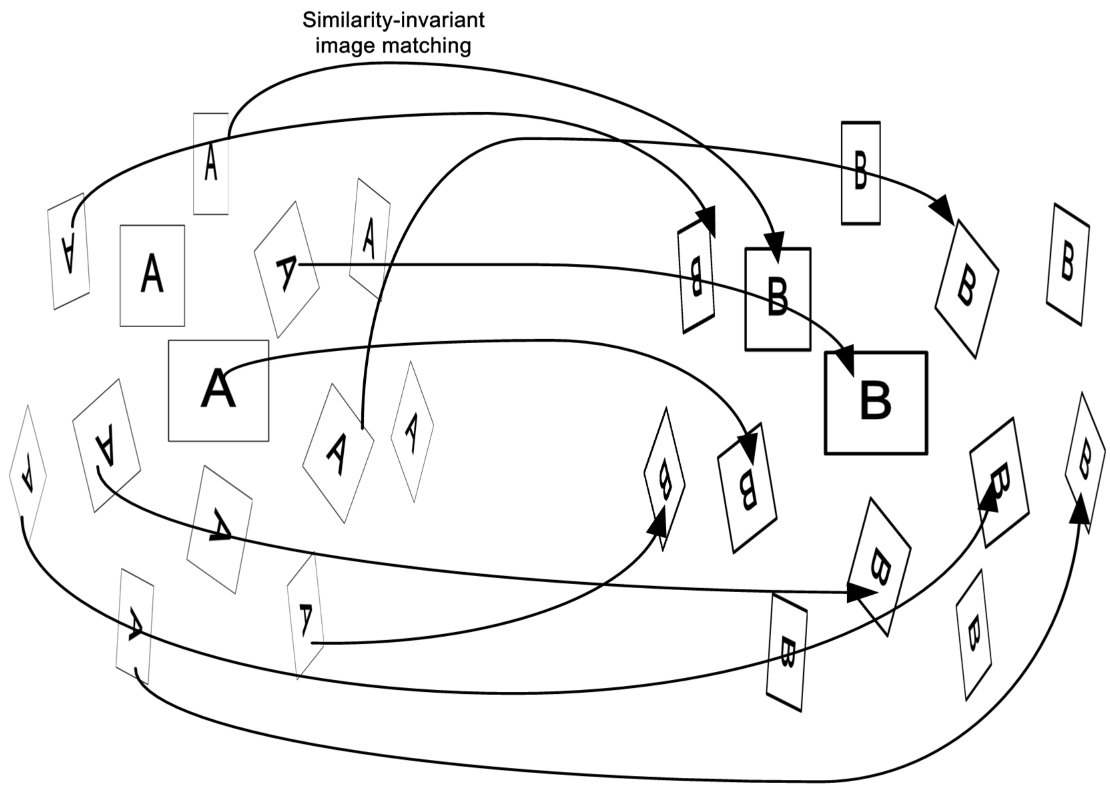


$\tau < 2.5$ (SIFT) $\qquad$ $\tau < 5$ (MSER) $\qquad$ $\tau < 40$ (A-SIFT)

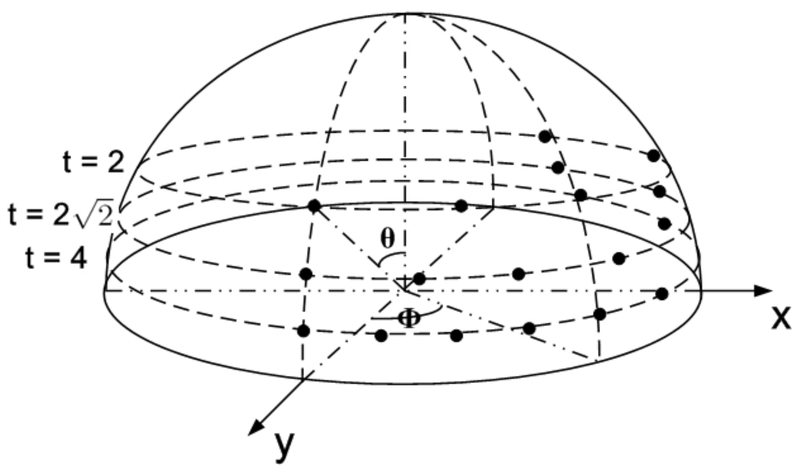
$$\theta = 80^\circ$$

Affine-SIFT (A-SIFT) Overview

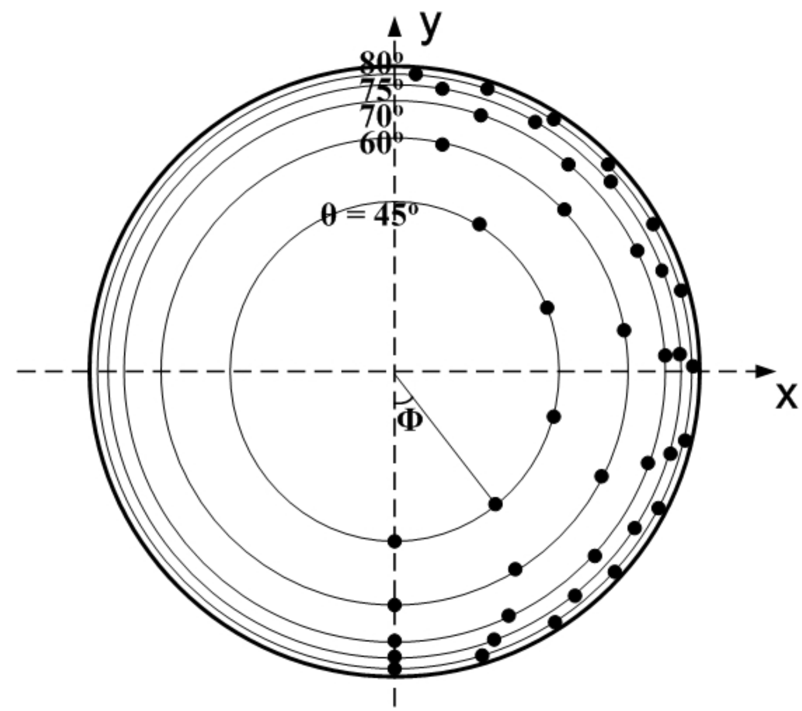
- Simulate the tilts (two parameters).
- Simulated images are compared by a rotation-, translation- and zoom-invariant algorithm, e.g., SIFT. (SIFT normalizes translation and rotation and simulates zoom.)



Sampling the observation sphere



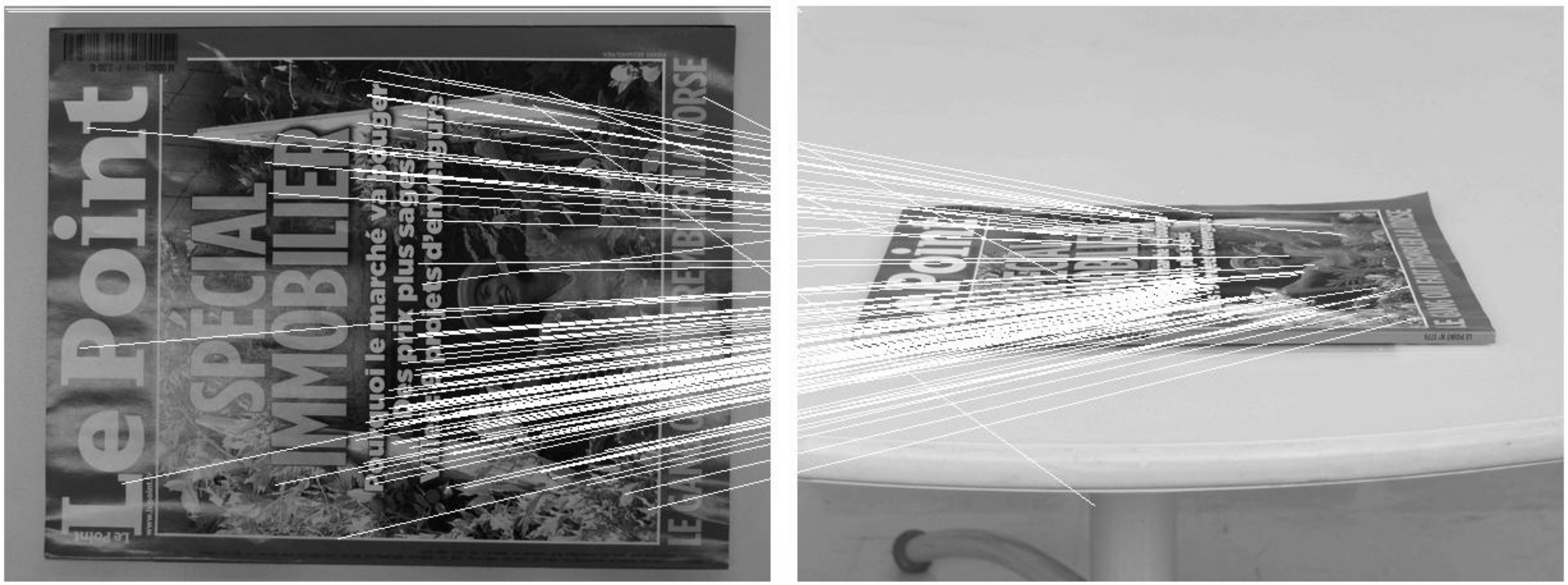
Perspective view



View from the zenith

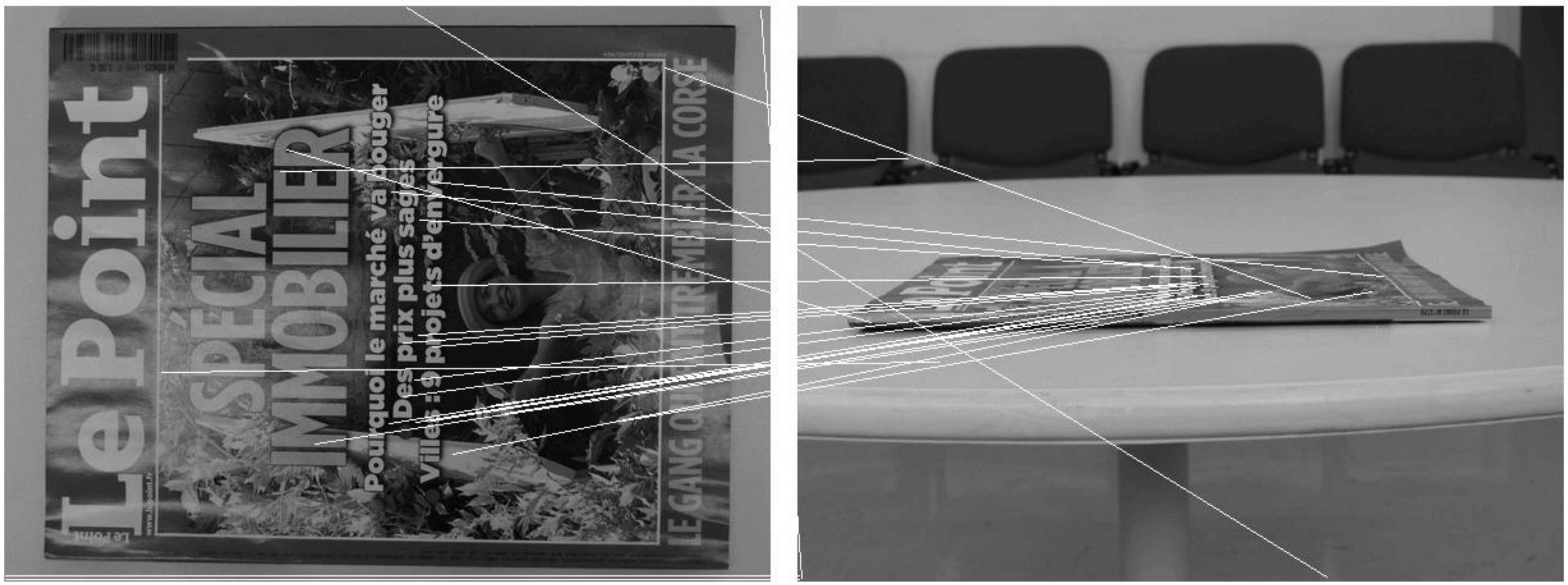
Figure 8: It is enough to simulate five tilts $\sqrt{2}, 2, 2\sqrt{2}, 4, 4\sqrt{2}$ and a growing but moderate number of longitudes per tilt. The overall simulated image area is five times the original.

What is the maximal absolute tilt



t = 3 ($\theta = 70.5°$), 107 A-SIFT matches (3 false).

What is the maximal absolute tilt

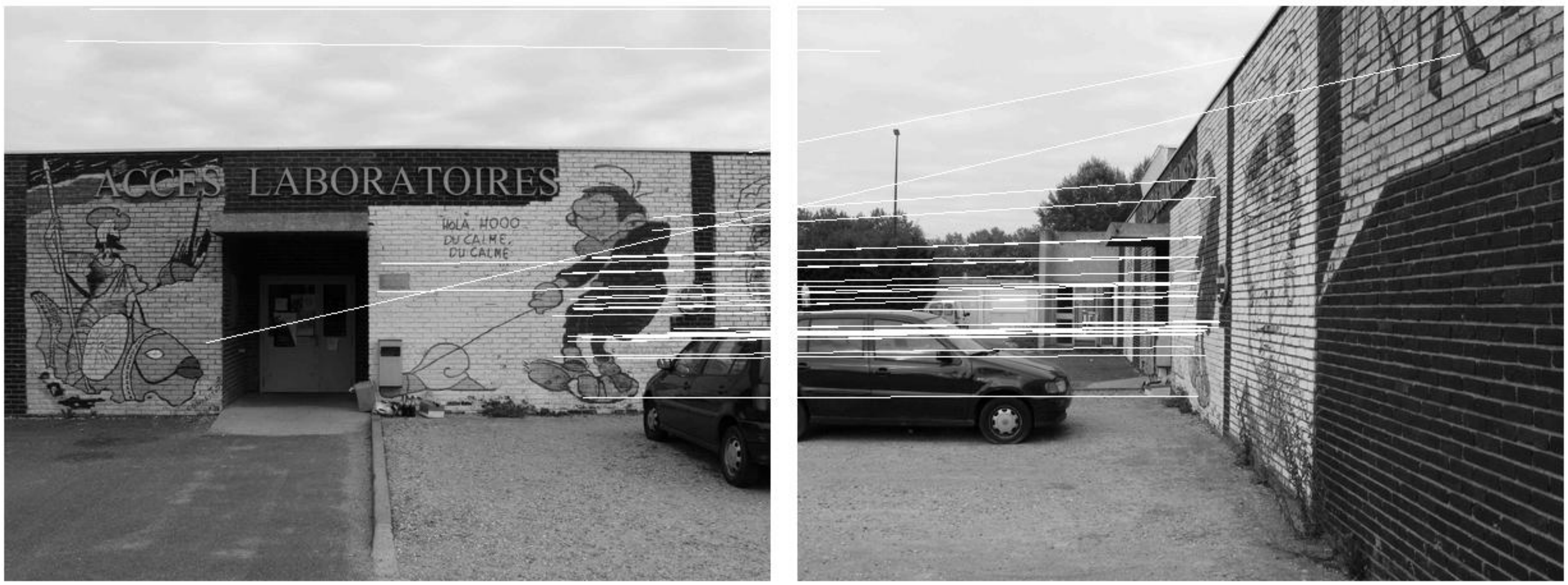


t = 5.2 ($\theta = 78.9°$), 25 A-SIFT matches (7 false).

What is the maximal absolute tilt

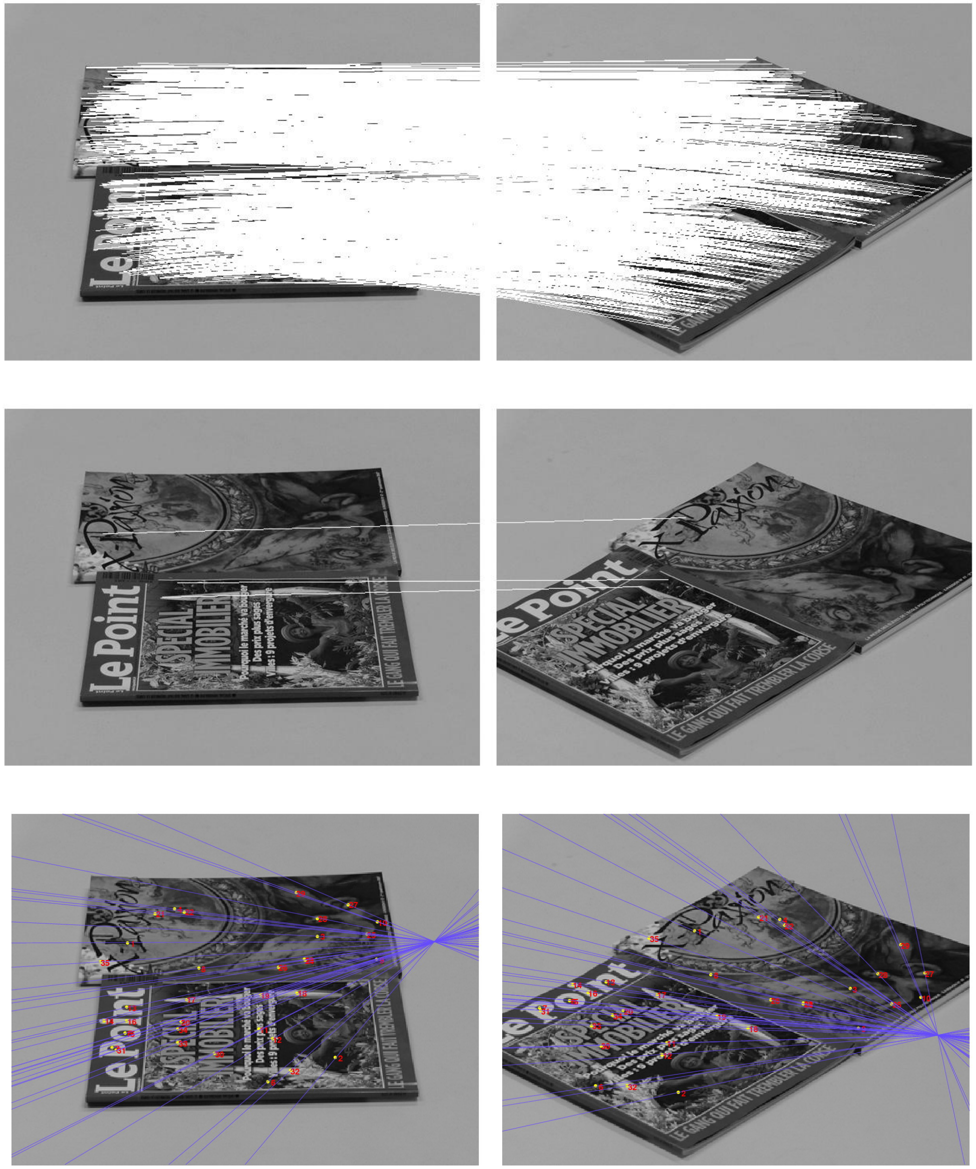
t = 3.8 ($\theta = 74.7°$), 71 A-SIFT matches (4 false).

What is the maximal absolute tilt



t = 5.6 ($\theta = 79.7°$), 33 A-SIFT matches (4 false).

Experiments: Image Matching



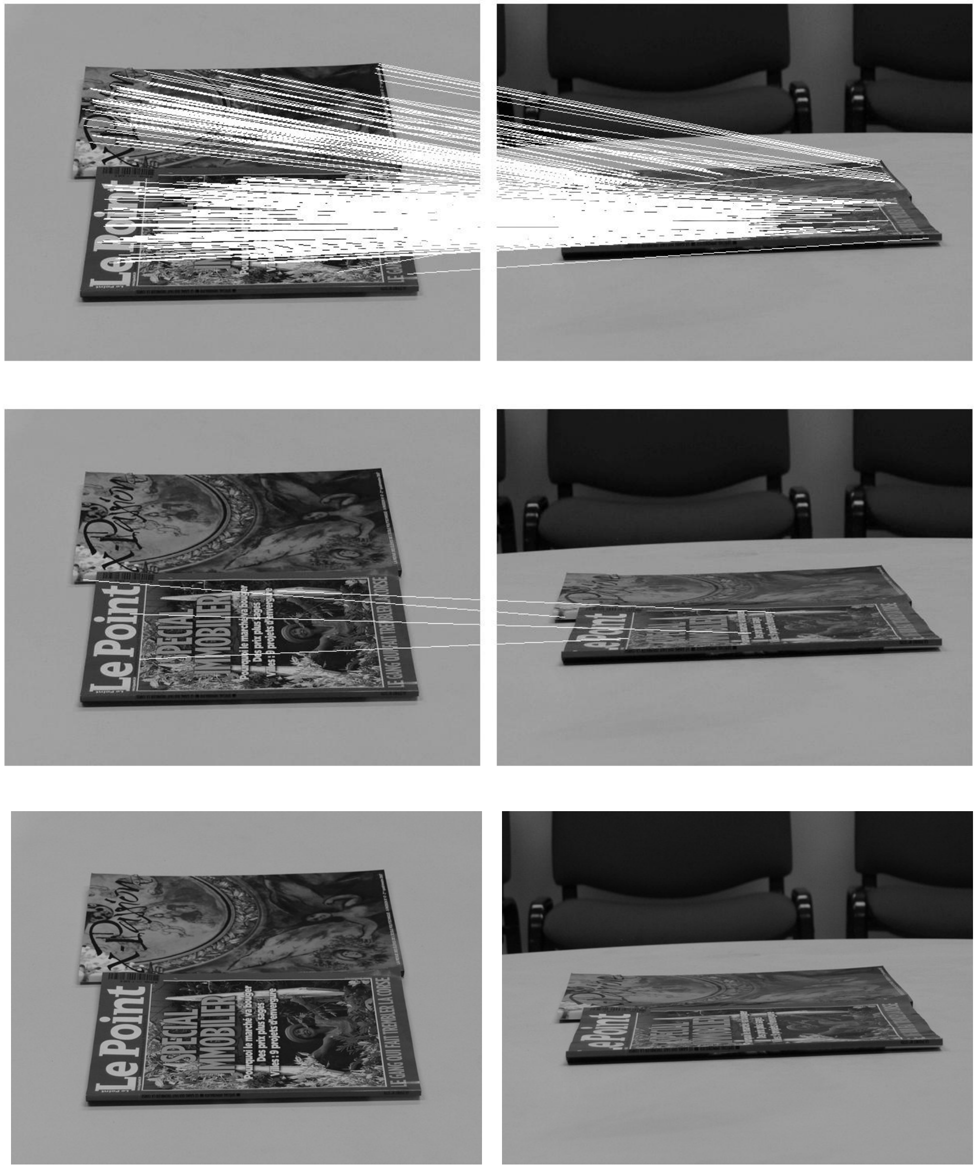
Absolute tilts $t = t' = 2.1$. transition tilt: $\tau = 3.0$.

Top: A-SIFT finds 1667 correspondences, all correct.

Middle: SIFT finds 3 correspondences.

Bottom: MSER finds 46 correspondences, out of which 35 are correct.

Experiments: Image Matching



Absolute tilts $t = 2.1$ (left), $t' = 6.0$ (right). transition tilt: $\tau = 2.9$.

Top: A-SIFT finds 338 correspondences, out of which 2 are false.

Middle: SIFT finds 5 correspondences.

Bottom: MSER finds 3 false correspondences in total that have been rejected.

Experiments: Image Matching

Transition tilt: $\tau \approx 3.2$.

Top: A-SIFT finds 724 correspondences, out of which 3 are false.

Middle: SIFT finds 6 correspondences.

Bottom: MSER finds 127 correspondences, out of which 50 are correct.

Experiments: Image Matching

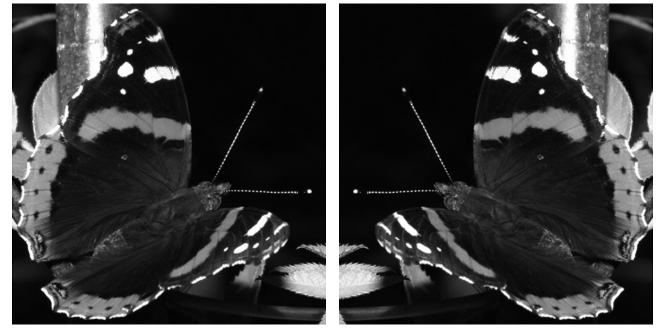
Top: A-SIFT finds 255 matches out of which 1 is false.

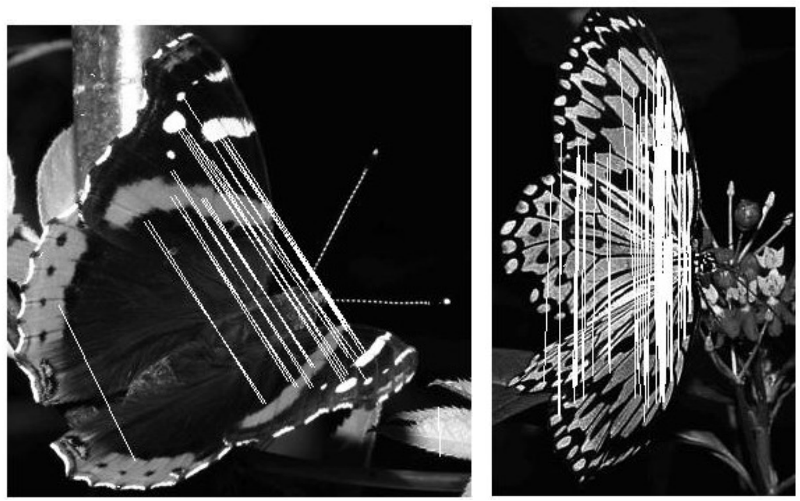
Middle: SIFT finds 16 matches out of which 6 are false.

Bottom: MSER finds 70 tentative correspondences out of which there are 51 inliers.

Symmetry Detection in Perspective

- Symmetry detection = image comparison with the flipped version.
- Symmetric object not in frontal view $\Rightarrow$ a viewpoint change between the flipped images.





— SIFT fails: big viewpoint change.

— MSER fails: two wings confused.

- D. G. Lowe, Object recognition from local scale-invariant features IJCV, 2, 1150–1157, 1999
- D. G. Lowe, Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision, 60(2):91-110, 2004
- D. Pritchard and W. Heidrich. Cloth Motion Capture. Computer Graphics Forum, 22(3):263 271, 2003.
- 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. M.: Report on Fully Affine Invariant Image Comparison, CMLA website