Report on Fully Affine Invariant Image Comparison

Jean-Michel Morel Guoshen Yu

May 22, 2008

Abstract

If a physical object has a smooth or piecewise smooth boundary, its images obtained by cameras in varying positions undergo smooth apparent deformations. These deformations are locally well approximated by affine transforms of the image plane.

In consequence the solid object recognition problem has often been led back to the computation of affine invariant image local features. Such invariant features could be obtained by normalization methods, but no fully affine normalization method exists for the time being. As a matter of fact, the scale invariance, which actually means invariance to blur, is only dealt with by methods inspired from the scale space theory, like the SIFT method. By simulating zooms out, this method normalizes the four translation, rotation and scale (blur) parameters, out of the six parameters of an affine transform. Affine normalization methods like MSER normalize with respect to all six parameters of the affine transform, but this normalization is imperfect, not dealing rigorously with blur.

The method proposed in this paper, affine SIFT (A-SIFT), simulates all image views obtainable by varying the two camera parameters left over by the SIFT method. Then it normalizes the other four parameters by simply using the SIFT method itself. The two additional parameters are the angles (a longitude and a latitude) defining the camera axis orientation. Mathematical arguments are developed to prove that the resulting method is fully affine invariant, up to an arbitrary precision.

Against any prognosis, simulating all views depending on the two camera orientation parameters is feasible with no dramatic computational load. The method permits to reliably identify features that have undergone tilts of large magnitude, up to 30 and more, while state-of-the-art methods do not exceed tilts of 2.5 (SIFT) or 4.5 (MSER). The report puts in evidence the role of high *transition tilts*: while a tilt from a frontal to an oblique view exceeding 6 is rare, higher transition tilts are common as soon as two oblique views of an object are compared. Thus, a fully affine invariance is required for 3D scene analysis. This fact is substantiated by many experiments.

Contents

1 Introduction

Image matching aims at establishing correspondences between similar objects that appear in different images. This is a fundamental step is many computer vision and image processing applications such as image recognition, 3D reconstruction, object tracking, robot localization and image registration (see, for example, [14]).

1.1 The shape recognition problem

The general shape recognition problem starts with several photographs of a physical object, possibly taken with different cameras and view points. These digital images are the *query* images. Given other digital images, the *search* images, the question is whether some of them contain, or not, a view of the object taken in the query image. This problem is by far more restrictive than the *categorization* problem, where the question is to recognize a class of objects, like chairs or cats, from some learning instances. This paper only deals with several instances of the very same object, or of copies of this object.

An object's view can deform from an image to another for two obvious reasons: First, because it underwent a physical deformation, and second, because the change of camera position induced an apparent deformation.

The most successful image matching algorithms usually first detect points of interest in the compared images, then select a region around each point of interest and finally associate an invariant descriptor or feature to each region. Correspondences may then be established by matching the descriptors. Detectors and descriptors should be as invariant as possible.

In recent years local image detectors have bloomed. All of them are translation invariant. The Harris point detector [20] is also rotation invariant. The Harris-Laplace and Hessian-Laplace region detectors [36, 38] are invariant to rotation and changes of scale. Some moment-based region detectors [29, 6] including the Harris-Affine and Hessian-Affine region detectors [37, 38], an edge-based region detector [65, 64], an intensity- based region detector [63, 64], an entropy-based region detector [23], and two independently developed level line-based region detectors MSER ("maximally stable extremal region") [33] and LLD ("level line descriptor") [46, 47, 48] are designed to be invariant to affine transformations. MSER, in particular, has been demonstrated to have often better performance than other affine invariant detectors [40]. (A short description of MSER is given in Section 8.1.) However, these detectors aren't fully affine invariant. As pointed out in [31], they start with initial feature scales and locations selected in a non-affine invariant manner. For instance, MSER and LLD are not fully scale invariant [48]. This is also the case for other image local descriptors, such as the distribution-based shape context [7], the geometric histogram [3] descriptors, the derivative-based complex filters [6, 57], and the moment invariants [67].

In his milestone paper [31], Lowe has proposed a scale-invariant feature transform (SIFT) descriptor that is invariant to image scaling and rotation and partially invariant to illumination and viewpoint changes. Although this detector is *a priori* less invariant to affine transforms than others, its performance turns out to be comparable. Furthermore, it really is scale invariant (see the math-

ematical analysis in [44]. Based on the scale-space theory [28], the SIFT procedure consists in normalizing local patches around robust scale covariant image key points. A number of SIFT variants and extensions, including PCA-SIFT [24] and gradient location-orientation histogram (GLOH) [39], that claim to have better robustness and distinctiveness with scaled-down complexity have been developed ever since [16, 27]. Demonstrated to be superior to other many descriptors [22, 39], SIFT and its variants have been popularly applied for scene recognition [13, 42, 56, 68, 18, 59, 72, 43] and detection [17, 49], robot localization [8, 60, 50, 21], image registration [71], image retrieval [19], motion tracking [66, 25], 3D modeling and reconstruction [54, 69], building panoramas [1, 9], photo management [70, 26, 61, 10], as well as symmetry detection [32].

In this impressive body of work, many methods achieve success in certain image matching applications. Nevertheless none of them is fully affine invariant. Being singled out for its sometimes superior performance [40, 39], MSER is only approximately special affine invariant, and misses the crucial zoom invariance. The mathematical analysis of the SIFT method proposed in [44] shows that with little approximation, the SIFT method is optimal in retrieving images up to a zoom, a rotation, and a translation. Thus, SIFT is fully similarity invariant, but it is not fully affine invariant.

The present paper proposes an image matching algorithm, Affine SIFT (A-SIFT) that is fully affine invariant (up to a prefixed precision). A fast multi-resolution version of this algorithm that has a reasonably small complexity of 1.5 to 5 times a single SIFT routine will be described. Experiments will show that A-SIFT achieves considerably better performance than SIFT and MSER under large viewpoint changes. To explain why and how, the notion of *transition tilt* from a view to another will be introduced (see Fig. 1). It will be proved that the transition tilts between compared images attainable with A-SIFT are ten times larger than the transition tilts attained by state-of-the art methods like SIFT or MSER.

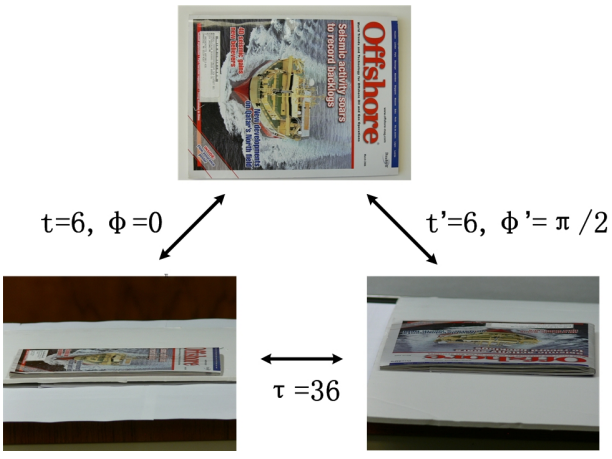


Figure 1: High transition tilts

The focus here is on affine invariance only. Other invariance requirements will not be discussed. For instance, photometric invariance of image matching methods is also required, because the lighting and observation conditions affect the color scales. In general, this photometric invariance

is correctly dealt with by state of the art methods like SIFT, MSER, or LLD, that retain only photometric invariants (image level lines, or gradient orientations).

Other issues are also at stake to achieve a final recognition decision: probably the most challenging is to define rejection and acceptation thresholds. This matter is also amply dealt with in recent literature. We refer to the empirical threshold of SIFT, that achieves acceptable results, and to the much accurate shape recognition thresholds used in the *a contrario* theory [53, 47, 12].

Section 2 describes the main ideas, the algorithm, and discusses precursors. Section 4 presents and discusses the crucial notion of *transition tilt*, that permits to evaluate the affine invariance of algorithms. It shows that very high tilts are likely, and can indeed be handled. Section 5 gives the mathematical formalism and a mathematical proof that A-SIFT is affine invariant. Section 6 addresses the critical sampling issues for the new-simulated parameters, and provides a complexity analysis and a fast version of the method. Section 7 is devoted to many comparative experiments.

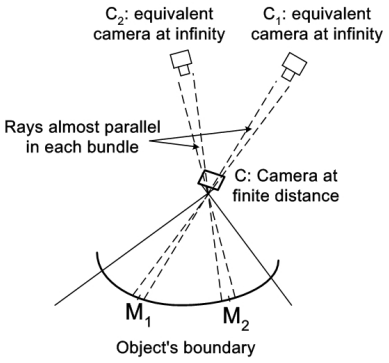


Figure 2: Local planar homographies are equivalent to multiple local cameras at infinity.

2 A fast presentation of the method

2.1 The affine simplification

Image distortions arising from viewpoint changes can be locally modeled by affine planar transforms, provided the object's boundaries are piecewise smooth [40]. In other terms, a perspective effect can be modeled by a combination of several different affine transforms in different image regions (see Fig. 2). The A-SIFT method described in this section simulates, up to some precision, all affine transforms of the image. It is in that sense invariant to all viewpoint changes. Indeed, by order 1 Taylor formula, any planar smooth deformation $(x, y) \to (X, Y) = (F_1(x, y), F_2(x, y))$ can be locally approximated around each point $(x_0, y_0) \to (X_0, Y_0)$ by the affine map

$$\begin{pmatrix} X - X_0 \\ Y - Y_0 \end{pmatrix} = \begin{bmatrix} \frac{\partial F_1}{\partial x}(x_0, y_0) & \frac{\partial F_1}{\partial y}(x_0, y_0) \\ \frac{\partial F_2}{\partial x}(x_0, y_0) & \frac{\partial F_2}{\partial y}(x_0, y_0) \end{bmatrix} \begin{pmatrix} x - x_0 \\ y - y_0 \end{pmatrix} + O \begin{pmatrix} (x - x_0)^2 + (y - y_0)^2 \\ (x - x_0)^2 + (y - y_0)^2) \end{pmatrix}.$$

Thus, all object deformations and all camera motions are locally approximated by affine transforms. For example, in the case of a flat object, the deformation induced by a camera motion is a planar homographic transform, which is smooth and therefore locally tangent to affine transforms. Conversely, *any affine transform with positive determinant can be interpreted as the apparent deformation induced on a planar object by a camera motion, the camera being assumed far away from the object.* Thus, under the local smoothness assumption of the object's boundary, the (local) deformation model of an image $u(x, y)$ under a deformation of the object or under a camera motion is

$$u(x, y) \to u(ax + by + e, cx + dy + f),$$

where the mapping

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

depicts any affine transform of the plane with positive determinant.

2.2 The affine challenge

How to recognize a portion of a planar image that has undergone an arbitrary affine transform? Since the affine transform depends upon six parameters, it is out of the question to just simulate all of them and compare the original image to all deformed images by all possible affine deformations. However, *simulation* can be a solution for a few parameters: the SIFT method actually simulates zooms out.

The other way that has been tried by many authors is *normalization*. Normalization is a magic method that, given a patch that has undergone an unknown affine transform, transforms the patch into a standardized one, where the effect of the affine transform has been eliminated (see Fig. 3). Normalization by translation is easily achieved: A patch around (x_0, y_0) is translated back to a patch around $(0, 0)$. A rotational normalization requires a circular patch. In this patch, a principal direction is found, and the patch is rotated so that this principal direction coincides with a fixed direction. Thus, of the six parameters in an affine transform, at least three are easily eliminated by normalization.

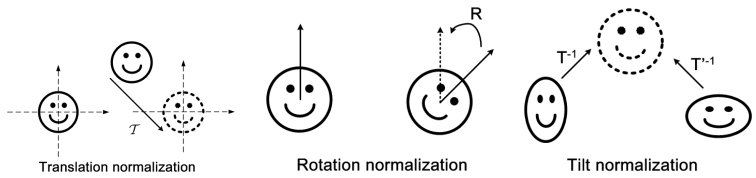


Figure 3: Normalization methods can eliminate the effect of a a class of affine transforms by associating the same standard patch to all transformed patches.

However, when it comes to the other three parameters, things get difficult and controversial. At least two methods have been recently proposed to perform a full affine normalization: MSER [33] and LLD [47]. Both of them apply to image level lines, or to image pieces of level lines, an

affine normalization in the spirit of the translation and rotation normalization explained above. Unfortunately, both of these methods miss a point, namely the fact that the three remaining parameters (zoom, camera axis longitude and latitude) cannot *stricto sensu* be normalized. Indeed, they depend on the irreversible image blur. The trouble comes from the fact that *affine transforms do not commute with the convolution by a radial blur kernel, with the only exception of rotations and translations.* Thus, the parameters in the affine transform do not play at all equivalent roles: some can be *normalized* and some cannot, and therefore must be *simulated*.

Thus, to make short a long story, this paper is dedicated to the proof that the three other parameters, which will be called *zoom, longitude and latitude (tilt)* parameters, can be simulated. The idea of combining simulation and normalization is not new. It is actually the main successful ingredient of the SIFT method. This method simulates all zooms out of the query and search images. Quoting D. Lowe [31]:

> Recently, there has been an impressive body of work on extending local features to be invariant to full affine transformations (Baumberg, 2000; Tuytelaars and Van Gool, 2000; Mikolajczyk and Schmid, 2002; Schaffalitzky and Zisserman, 2002; Brown and Lowe, 2002). This allows for invariant matching to features on a planar surface under changes in orthographic 3D projection, in most cases by resampling the image in a local affine frame. However, none of these approaches are yet fully affine invariant, as they start with initial feature scales and locations selected in a non-affine-invariant manner due to the prohibitive cost of exploring the full affine space. The affine frames are also more sensitive to noise than those of the scale-invariant features, so in practice the affine features have lower repeatability than the scale-invariant features unless the affine distortion is greater than about a 40 degree tilt of a planar surface (Mikolajczyk, 2002). Wider affine invariance may not be important for many applications, as training views are best taken at least every 30 degrees rotation in viewpoint (meaning that recognition is within 15 degrees of the closest training view) in order to capture non-planar changes and occlusion effects for 3D objects."

Lowe considers that *exploring the full affine space* would have a *prohibitive cost* –the aim here is to prove that it is not so. He soundly argues that the actual affine normalization methods are not really affine invariant, since they *start with initial feature scales and locations selected in a non-affine invariant manner*. Lowe also gives a cue on how to compensate for the lack of affine invariance of the SIFT method: by taking views of the object to be recognized every 30 degrees rotation in viewpoint. Notice that the talk is about real snapshots and not simulated views. David Pritchard's master thesis is actually a first step toward the method developed here. Quoting [51] in his 2003 master thesis on cloth parameters and motion capture:

> Cloth strongly resists stretching, but permits substantial bending; folds and wrinkles are a distinctive characteristic of cloth. This behaviour means that sections of the cloth are often seen at oblique angles, leading to large affine distortions of features in certain regions of the cloth. Unfortunately, SIFT features are not invariant to large affine distortions.(...) To compensate for this, we use an expanded set of reference features. We generate a new reference image by using a 2 x 2 transformation matrix T to scale the reference image by half horizontally. We repeat three more times, scaling vertically and along axes at 45 degrees, as shown in Figure 5.3. This simulates different oblique views

of the reference image. For each of these scaled oblique views, we collect a set of SIFT features. Finally, these new SIFT features are merged into the reference feature set. When performing this merge, we must adjust feature positions, scales and orientations by using T-1. This approach is compatible with the recommendations made by Lowe for correcting SIFT's sensitivity to affine change.

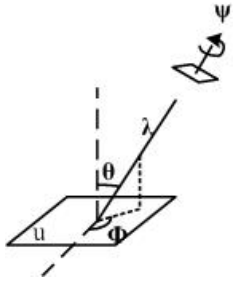


Figure 4: Geometric interpretation of the decomposition formula (1). This figure illustrates the four main parameters in the affine image deformation caused by a camera motion, starting from a frontal view u. The camera is assumed to stay far away from the image. The camera can first move parallel to the observed surface: this motion induces a translation $\mathcal{T}$ that is not represented here. Its optical axis can take a θ angle with respect to the normal to the image plane u. This parameter is called *latitude*. The plane containing the normal and the new position of the optical axis makes an angle ϕ with a fixed vertical plane. This angle is called *longitude*. The camera can also rotate around its optical axis (rotation parameter ψ). Last but not least, the camera can move forward or backward, or change focal length. This is the zoom parameter λ. If λ and μ are large with respect to the object's size, the image deformation of a frontal view λ, $t = 1$, $\phi = \psi = 0$ to a slanted view $\lambda', t', \phi', \psi'$ corresponds to an image deformation $\mathbf{u}(x, y) \rightarrow \mathbf{u}(A(x, y))$ as in Formula (1)

Given an affine map of the plane with positive determinant, consider its *unique decomposition*

$$A = H_\lambda R_1(\psi) T_t R_2(\phi) = \lambda \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \quad (1)$$

where $\lambda > 0$, λt is the determinant of A, R_i are rotations, $\phi \in [0, 180°[$, and T_t is a tilt, namely a diagonal matrix with a first eigenvalue equal to $t \geq 1$ and the second one equal to 1. Fig. 4 shows a camera motion interpretation of this affine decomposition: ϕ and $\theta = \arccos 1/t$ are the viewpoint angles and ψ parameterizes the camera spin. Thus, this figure illustrates the four main parameters in the affine image deformation caused by a camera motion, starting from a frontal view u. The camera is assumed to stay far away from the image. The camera can first move parallel to the object's plane: this motion induces a translation $\mathcal{T}$ that is not represented here. The camera can rotate around its optical axis (rotation parameter ψ). Its optical axis can take a θ angle with respect to the normal to the image plane u. This parameter is called *latitude*. The plane containing the normal and the new position of the optical axis makes an angle ϕ with a fixed vertical plane. This angle is called *longitude*. Last but not least, the camera can move forward or

backward. This is the zoom parameter λ. The motion of a frontal view $\lambda = 1$, $t = 1$, $\phi = \psi = 0$ to a slanted view corresponds to the image deformation $\mathbf{u}(x, y) \to \mathbf{u}(A(x, y))$ given by (1).

3 The algorithm

The proposed method simulates with enough accuracy *all* distortions caused by a variation of the direction of the optical axis of a camera (two parameters). Then it normalizes the other four by the SIFT method, or any other method that is rotation, translation, and scale invariant. More specifically, the method proceeds by the following steps. (See Fig. 5.)

1. Each image is transformed by simulating all possible affine distortions caused by the change of orientation of the camera axis of camera from a frontal position. These distortions depend upon two parameters: the longitude ϕ and the latitude θ. The images undergo ϕ-rotations followed by tilts with parameter $t = |\frac{1}{\cos\theta}|$ (a tilt by t in the direction of x is the operation $u(x, y) \to u(tx, y)$). For digital images, the tilt is performed as t-subsampling, and therefore requires the previous application of an antialiasing filter in the direction of x, namely the convolution by a gaussian with standard deviation $c\sqrt{t^2 - 1}$ (for good antialiasing, $c \simeq 0.6$, see [44].

2. These rotations and tilts are performed for a finite and small number of latitudes and longitudes, the sampling steps of these parameters ensuring that the simulated images keep close to any other possible view generated by other values of ϕ and θ.

3. All simulated images are compared to each other by a scale invariant, rotation invariant, and translation invariant algorithm (typically SIFT). SIFT normalizes the translation of the camera parallel to its focal plane, the rotation of the camera around its optical axis, and simulates the scale change, namely any camera motion that does not change the camera axis.

4. To be more specific, the latitudes θ are such that the associated tilts follow a geometric series $1, a, a^2, , \ldots, a^n$, with $a > 1$. Section 6.2 shows that $a = \sqrt{2}$ is a good compromise between accuracy and sparsity. The value n can go up to 6 or more, if the tilts are simulated on the query and the searched image, and up to 10 and more if the tilts are simulated on one image only. That way, transition tilts going up to 64 and more can be explored.

5. The longitudes ϕ are for each tilt an arithmetic series $0, b/t, \ldots, kb/t$, where $b \simeq 72^\circ$ seems again a good compromise, and k is the last integer such that $kb/t < 180^\circ$.

6. Complexity: Each tilt is a t subsampling. Thus, the image area is divided by t. Counting all rotations associated with a tilt, the overall simulated image area for each tilt is $(180/72)t = 2.5t$. This implies that *the method complexity is proportional to the number of tilts and not to the number of generated views, that is much larger.* Controlling the overall simulated area means is equivalent to controlling the algorithm complexity. Indeed, the search time and the memory size of similarity invariant indicators are proportional to the image area. This complexity can be further downgraded by a) subsampling the query and search images; b) identifying the pairs (t, ϕ) that give positive results; c) going back to the original resolution only for these pairs.

7. This description ends with a concrete example of how the multiresolution search strategy can actually make the algorithm no slower than SIFT. Take $a = \sqrt{2}$, $n = 4$ (maximal tilt for each image is 4). Thus, the simulated image area is $4 \times 2.5 = 10$ times the original area. By making a 3-subsampling of the original, this area is reduced to $1,11$ times the one of the original image. If this trick is applied to both the query and the search image, *the overall complexity is equivalent to* 1.23 *the one of SIFT, but allows for a search with transition tilts going up to* $4 \times 4 = 16$. *This is to be compared to the 2.5 limit for the transition tilts attained in SIFT and the 4 limit for transition tilts in MSER.*

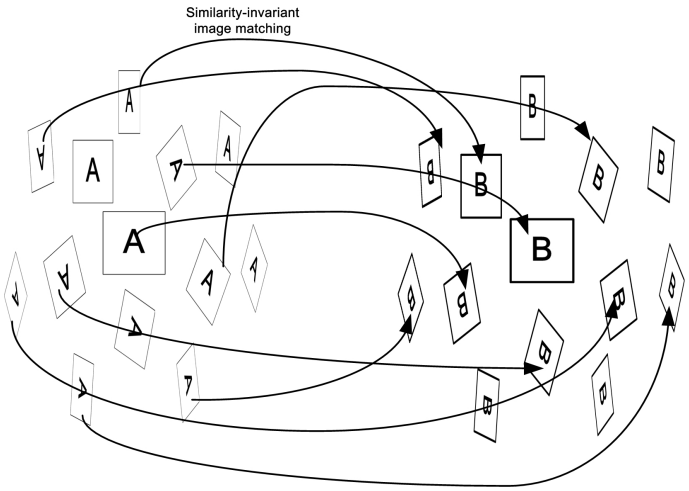


Figure 5: Overview of the A-SIFT algorithm. The square images A and B represent the compared images $\mathbf{u}$ and $\mathbf{v}$. A-SIFT simulates arbitrary camera changes of direction by applying rotations followed by a tilts to both images. The simulated images, represented by the parallelograms, are then compared with an image matching algorithm like SIFT, that is invariant to similarity transformations, i.e., invariant to scale change, rotation and translation.

4 High transition tilts

Equation (1) and its geometric interpretation in Fig. 4 are crucial to the scopes of this study. This last figure associates any linear map A with positive determinant with the planar deformation $\mathbf{u}(A(x, y))$ of a frontal view $\mathbf{u}(x, y)$, when the camera changes position. The parameter λ corresponds to a change of scale. The non critical translation parameter has been eliminated by assuming that the camera axis meets the image plane at a fixed point. Let us now consider the case where *two* camera positions, not necessarily frontal are at stake, corresponding to two different linear maps A and B. (Again, the translation parameter is left out of the discussion by fixing the intersection of the camera axis with image plane.) This physical situation is the generic one; when taking several snapshots of a scene, there is no particular reason why objects would be taken

frontally. The resulting images are $\mathbf{u}_1(x,y) = \mathbf{u}(A(x,y))$ and $\mathbf{u}_2(x,y) = \mathbf{u}(B(x,y))$. Let us now take one of these images as *reference image*, and the other one as search image.

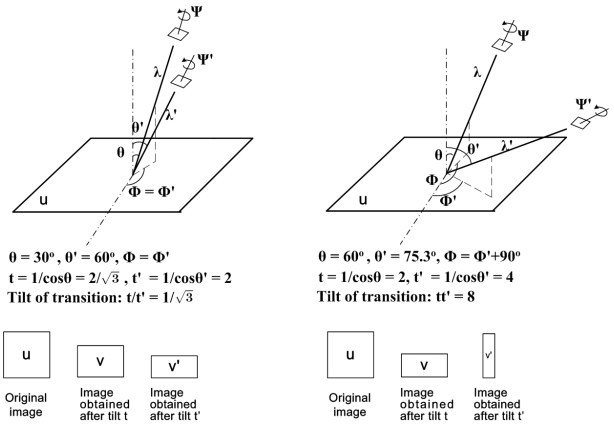


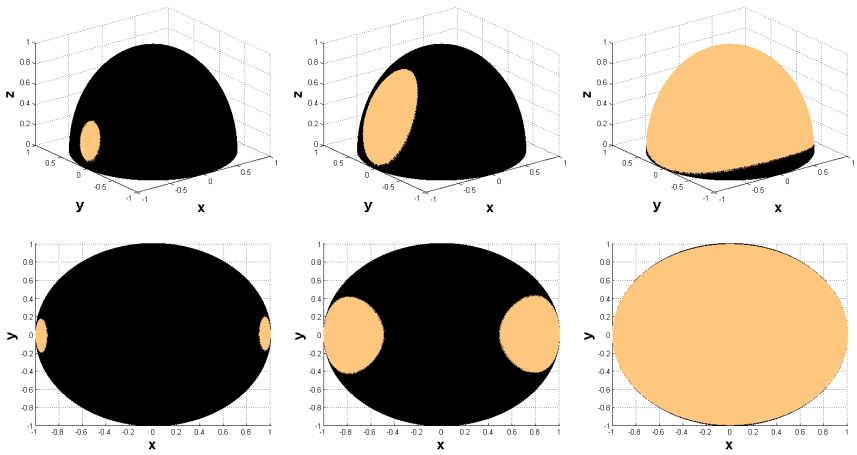
Figure 6: Illustration of the difference between absolute tilt and transition tilt. Left: The camera is put in two positions corresponding to tilts t and t', but with $\phi = \phi'$. The transition tilt between the resulting images $\mathbf{v}$ and $\mathbf{v}'$ is t'/t. Right: when $\phi = \phi' + \pi/2$, the transition tilt between $\mathbf{v}$ and $\mathbf{v}'$ is tt'. Thus, two moderate absolute tilts can lead to a huge transition tilt! In the left hand case, the transition tilt is $\sqrt{3}$ and therefore smaller than the absolute tilts. In the right hand case, the tilt is $tt' = 8$. The rectangles in the bottom show the image deformation after tilt. Compare $\mathbf{v}$ and $\mathbf{v}'$ in both cases.

Definition 1. *Given two views of a planar image,* $\mathbf{u}_1(x,y) = \mathbf{u}(A(x,y))$ *and* $\mathbf{u}_2(x,y) = \mathbf{u}(B(x,y))$*, we call transition tilt* $\tau(\mathbf{u}_1, \mathbf{u}_2)$ *and transition rotation* $\phi(\mathbf{u}_1, \mathbf{u}_2)$ *the unique parameters such that*

$$BA^{-1} = H_\lambda R_1(\psi) T_\tau R_2(\phi), \tag{2}$$

with the notation of Formula (1)*.*

It is an easy check that the transition tilt is symmetric, namely $\tau(\mathbf{u}_1, \mathbf{u}_2) = \tau(\mathbf{u}_2, \mathbf{u}_1)$. Fig. 6 illustrates the affine transition between two images taken from different viewpoints, and in particular the difference between absolute tilt and transition tilt. The camera is first put in two positions corresponding to absolute tilts t and t', but with $\phi = \phi'$. The transition tilt between the resulting images $\mathbf{v}$ and $\mathbf{v}'$ is $\tau = t'/t$, assuming $t' = \max(t', t)$. On the second illustration of Fig. 6, the tilts are made in two orthogonal directions: $\phi = \phi' + \pi/2$. Then an easy calculation shows that the transition tilt between $\mathbf{v}$ and $\mathbf{v}'$ is the product $\tau(\mathbf{v}, \mathbf{v}') = tt'$. Thus, *two moderate absolute tilts can lead to a large transition tilt!* In the first case considered in the figure, the transition tilt is $\sqrt{3}$ and therefore smaller than the absolute tilts. In the second case, the tilt is $tt' = 8$. Since in realistic cases the tilt can go up to 6 or even 8, is easily understood that the *transition tilt* can go up to 36, 84, and more. Fig. 7 shows the regions of the observation half sphere that can be attained with a given transition tilt from a fixed viewpoint with latitude 80° (absolute tilt $t = 5.85$). From this strong latitude, it needs a $\tau = 40$ transition tilt to attain most other viewpoints. SIFT and MSER only attain small regions.



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

$\theta = 80°$

Figure 7: High transition tilts. The bright regions are the parts of the observation half sphere that can be attained with a given transition tilt from a fixed viewpoint with latitude 80° (absolute tilt $t = 5.85$). From this strong latitude, it needs a $\tau = 40$ transition tilt to attain most other viewpoints. SIFT and MSER only attain small regions.

Fig. 8 shows the A-SIFT results for a pair of images under orthogonal viewpoints (transition rotation $\phi = 90°$) that leads to an extreme transition tilt $t \approx 37$. This is not at all an exceptional situation. It just so happens that the object's planar surface is observed at the same latitude by both views with a tilt $t \simeq t' \simeq 6$. This figure shows two snapshots of a magazine lying on a table, not even really flat, and with a non lambertian surface plagued with reflections. The difference of longitudes being about 90 degrees, the transition tilt between both images is surprisingly high: $\tau = tt' \simeq 37$. Thus, it is many times larger than the transition tilt attainable with SIFT or MSER. A-SIFT finds 120 matches out of which only 4 are wrong.

The relevance of the notion of transition tilt is corroborated by the fact that the highest transition tilt τ_{max} permitting to match two images with absolute tilts t and t' is fairly independent from t and t'. It has been experimentally checked that for SIFT $\tau_{max} \simeq 2.5$ and for MSER $\tau_{max} \simeq 4$.

To demonstrate this for SIFT, the transition tilts attainable by SIFT have been explored by systematic tilt simulations and tests. The experiments have been performed in the most favorable conditions for SIFT. The seed image $\mathbf{u}_0$ is a high quality frontal view of the Graffiti series. Tilted views from this frontal view were simulated by subsampling the image in one direction by a factor $\sqrt{t}$, and oversampling the image in the orthogonal direction by the same factor. That way, the absolute tilt is t, but the image area is not decreased. A set of tilted-rotated images $\mathbf{u}_1 = \mathbf{u}_0(t_1, 0)$ and $\mathbf{u}_2 = \mathbf{u}(t_2, \phi)$ was generated by this method from $\mathbf{u}$ with absolute tilts $t_1 = (\sqrt{2})^k$, $k = 1$, $2, \ldots, 5$, $t_2 = (2^{\frac{1}{4}})^l$, $l = 1, 2, \ldots, 14$, and ϕ_2 in a dense subset of $[0,\ 90°]$. The table shows for each pair t_1, t_2 the maximal longitude ϕ_{max} ensuring that $\mathbf{u}_1(t_1, 0°)$ and $\mathbf{u}_2(t_2, \phi_{max})$ match. On the right of ϕ_{max}, the table displays in each box the corresponding transition tilt $\tau(t_1, 0, t_2, \varphi_{max})$.

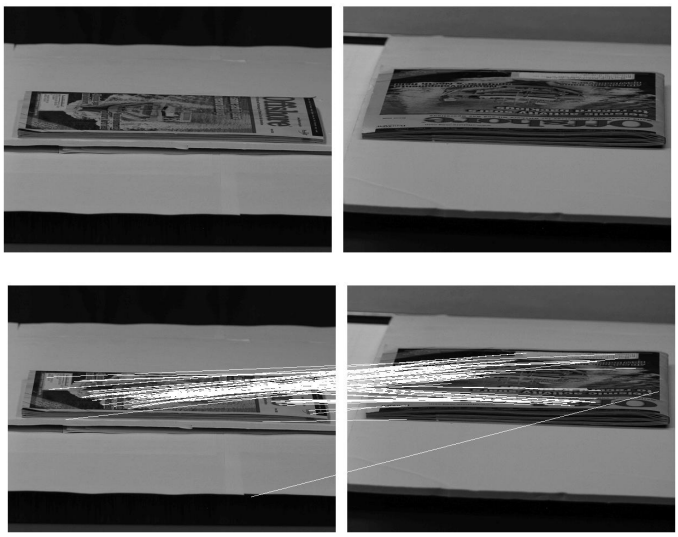


Figure 8: Top: Image pair with transition tilt $t \approx 37$. (SIFT and MSER fail completely.) Bottom: A-SIFT finds 120 matches out which 4 are false. See comments in text.

Conspicuously enough, τ_{max} is most of the time close to 2.5. This experiment, and other similar ones, substantiate the empirical law that *SIFT works for comparing images with transition tilts smaller than 2.5.* In all of these tests, success with SIFT means that at least 20 correct SIFT descriptors, or SIFs, have been found. For a short description of SIFs, see Section 8.2.

SIFT is not affine invariant, however, it is robust to moderate affine transform as SIFT descriptors encode local patches which may be covariant to small affine transform. The examples illustrated in Fig. 10 show that SIFT continues to work well when there is a tilt of $t = 2$ between the two images but the performance drops dramatically when $t > 2$.

5 The mathematical justification

5.1 Image operators formalizing A-SIFT

All continuous image operators applied to a continuous image $\mathbf{u}$, including the sampling operator that gives back a digital image, will be written in bold capital letters $\mathbf{A}$, $\mathbf{B}$. For a sake of simplicity, the operator composition $\mathbf{A} \circ \mathbf{B}$ will be written as a mere juxtaposition $\mathbf{AB}$. Given a transformation of the image plane A, define the associated transform of $\mathbf{u}$ by $\mathbf{Au}(\mathbf{x}) =: \mathbf{u}(A\mathbf{x})$. For instance, if $H_\lambda(x, y) = (\lambda x, \lambda y)$ is a homothety $H_\lambda u(\mathbf{x}) = \mathbf{u}(\lambda \mathbf{x})$, $\mathbf{H}_\lambda \mathbf{u}(\mathbf{x}) = \mathbf{u}(\lambda \mathbf{x})$ is the corresponding expansion of $\mathbf{u}$ by a λ^{-1} factor. In the same way if R is a rotation, $\mathbf{Ru} = \mathbf{u} \circ R$ is the image rotation by R^{-1}, and so on.

The next list gives the main notations for images and image operators.

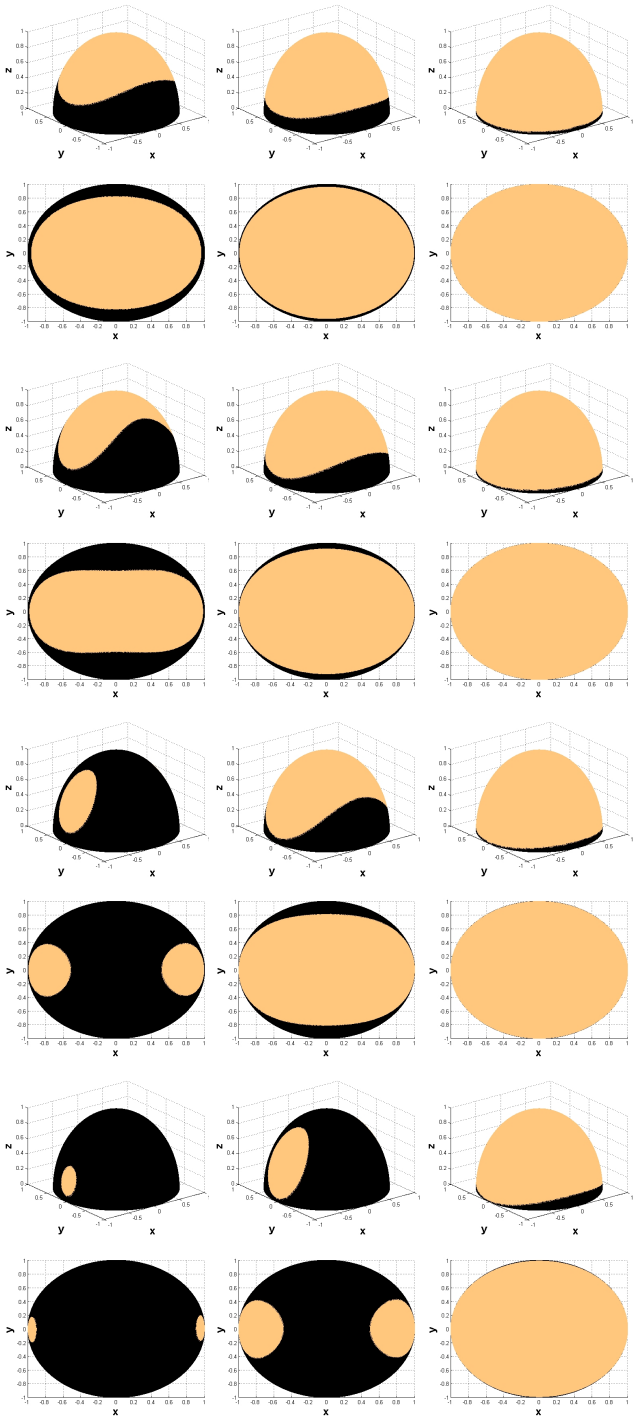


Figure 9: Transition tilt values, both perspective view (odd rows) and view from zenith (even rows) are shown. From top to bottom: latitude angle of the first image $\theta = 45,\ 60,\ 70,\ 80°$ that correspond to respectively absolute tilts $t = \sqrt{2},\ 2,\ 2.9,\ 5.8$. From left to right: transition tilt $< 2.5,\ 5,\ 40$ (gray part).

Figure 10: Top and bottom: SIFT detects respectively 234 and 28 matches between a frontal image and two images with tilts $t \approx 2$ and $t \approx 2.3$. This latter value is close the limiting tilt for SIFT to work.

	$t_1 = \sqrt{2}$	$t_1 = 2$	$t_1 = 2\sqrt{2}$	$t_1 = 4$	$t_1 = 4\sqrt{2}$
$t_2 = 2^{1/4}$	90°/1.7	60°/2.2	0°/2.4		
$t_2 = 2^{1/2}$	90°/2.0	56°/2.4	11°/2.1		
$t_2 = 2^{3/4}$	90°/2.4	50°/2.6	20°/2.1	0°/2.4	
$t_2 = 2$	63°/2.6	36°/2.4	20°/2.1	9°/2.2	
$t_2 = 2 \times 2^{1/4}$	37°/2.4	30°/2.3	23°/2.3	9°/1.9	
$t_2 = 2 \times 2^{1/2}$	18°/2.6	22°/2.2	24°/2.6	12°/2.0	0°/1.4
$t_2 = 2 \times 2^{3/4}$	6°/2.4	16°/2.2	21°2.6	16°/2.5	5°/1.4
$t_2 = 4$	0°/2.8	9°/2.2	18°/2.6	14°/2.4	9°/1.8
$t_2 = 4 \times 2^{1/4}$		4°/2.4	11°/2.2	12°/2.3	8°/2.0
$t_2 = 4 \times 2^{1/2}$			6°/2.2	7°/1.9	8°/2.3
$t_2 = 4 \times 2^{3/4}$			0°/2.4	5°/2.0	8°/2.5
$t_2 = 8$				0°/2.0	7°/2.5
$t_2 = 8 \times 2^{1/4}$					4°/2.6
$t_2 = 8 \times 2^{1/2}$					3°/2.9

Table 1: m/n in each entry means the longitude angle ϕ between the two viewpoints / the transition tilt $\tau(t_1, t_2, \phi)$. This Table shows that SIFT covers a transition tilt $\tau \approx 2.5$.

- $\mathbf{u}(\mathbf{x})$: 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

$$\mathbf{S}_\delta \mathbf{u}(n_1, n_2) = \mathbf{u}(n_1\delta, n_2\delta); \tag{3}$$

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

$$Iu(x, y) =: \sum_{(n_1,n_2)\in\mathbb{Z}^2} u(n_1, n_2)\text{sinc}(x - n_1)\text{sinc}(y - n_2),$$

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

$$I\mathbf{S}_1\mathbf{u} = \mathbf{u}. \tag{4}$$

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

- $\mathbf{G}_\delta(x_1, x_2) = \frac{1}{2\pi(\mathbf{c}\delta)^2} e^{-\frac{x_1^2+x_2^2}{2(\mathbf{c}\delta)^2}}$, : a gaussian kernel scaled by δ, that is, a function on $\mathbb{R}^2$ with integral 1 and standard deviation proportional to δ. Thus, $\int_{\mathbb{R}^2} \mathbf{G}_\delta(\mathbf{x})d\mathbf{x} = 1$ and $\mathbf{G}_\delta(\mathbf{x}) = \frac{1}{\delta^2} G_1(\frac{\mathbf{x}}{\delta})$. $\mathbf{G}_\delta$ also denotes the associated convolution operator $\mathbf{G}_\delta\mathbf{u}(\mathbf{x}) =: (\mathbf{G} * \mathbf{u})(\mathbf{x}) = \int_{\mathbb{R}^2} \mathbf{G}(\mathbf{y})\mathbf{u}(\mathbf{x}-\mathbf{y})d\mathbf{y}$. By the classical semigroup property:

$$\mathbf{G}_\delta \mathbf{G}_\beta = \mathbf{G}_{\sqrt{\delta^2+\beta^2}} \tag{5}$$

5.2 The affine camera model

The whole image comparison process, based on local features, can proceed as though images where (locally) obtained by using digital cameras far away (at infinity). The geometric deformations induced by the motion of such cameras are affine maps. A model is also needed for the two main camera parameters not deducible from its position, namely sampling and blur. The digital image is defined on the camera CCD plane. The pixel width can be taken as length unit, and the origin and axes chosen so that the camera pixels are indexed by $\mathbb{Z}^2$. The associated image sampling operator will be denoted by $\mathbf{S}_1$. The digital initial image is always assumed well-sampled and obtained by a gaussian blur with standard deviation 0.6. (See [44] for a detailed analysis of why this model is sufficient and coherent for most digital images, and compatible with the SIFT method.) In all that follows, $\mathbf{u}_0$ denotes the (theoretical) infinite resolution image that would be obtained by a frontal snapshot of a plane object with infinitely many pixels. The digital image obtained by any camera at infinity is $u = \mathbf{S}_1\mathbf{G}_1\mathbf{A}\mathcal{T}\mathbf{u}_0$, where $\mathbf{A}$ is *any* linear map with positive determinant and $\mathcal{T}$ any plane translation. Thus we can summarize the general image formation model with cameras at infinity as follows.

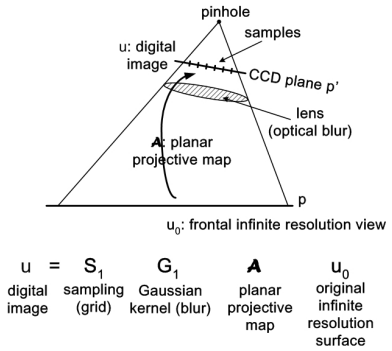


Figure 11: The projective camera model $u = \mathbf{S}_1\mathbf{G}_1\mathcal{A}\mathbf{u}_0$. $\mathcal{A}$ is a planar projective transform (a homography). $\mathbf{G}_1$ is an anti-aliasing gaussian filtering. $\mathbf{S}_1$ is the CCD sampling.

Definition 2. Image formation model. *Digital images of a planar object whose frontal infinite resolution image is* $\mathbf{u}_0$*, obtained by a digital camera far away from the object, satisfy*

$$u =: \mathbf{S}_1\mathbf{G}_1\mathbf{A}\mathcal{T}\mathbf{u}_0 \tag{6}$$

where $\mathbf{A}$ *is any linear map and* $\mathcal{T}$ *any plane translation.* $\mathbf{G}_1$ *denotes a gaussian kernel broad enough to ensure no aliasing by 1-sampling, namely* $I\mathbf{S}_1\mathbf{G}_1\mathbf{A}\mathcal{T}\mathbf{u}_0 = \mathbf{G}_1\mathbf{A}\mathcal{T}\mathbf{u}_0$.

The formal description of A-SIFT will be by far simpler if sampling issues do not interfere. All operations and all reasoning will be made with continuous well sampled images. It is easy to deduce afterwards the discrete operators acting on samples. $\mathcal{T}$ denotes an arbitrary translation, $\mathbf{R}$ an arbitrary rotation, $\mathbf{H}_\lambda$ an arbitrary homothety, and $\mathbf{G}$ an arbitrary gaussian convolution, all applied to continuous images. In the particular case in the digital image formation model (6) where $\mathbf{A}$ is a frontal view of $\mathbf{u}_0$, $\mathbf{A} = \mathbf{HR}\mathcal{T}$ is the composition of a translation $\mathcal{T}$, a homothety $\mathbf{H}$, and a rotation $\mathbf{R}$. Thus the digital image is $u = \mathbf{S}_1\mathbf{G}_1\mathbf{H}\mathcal{T}\mathbf{R}\mathbf{u}_0$. The following lemma is easily proven for the SIFT method (see [44] or the appendix, section 8.4).

Lemma 1. *For any rotation* $\mathbf{R}$ *and any translation* $\mathcal{T}$*, the SIFT descriptors of* $\mathbf{S}_1\mathbf{G}_1\mathbf{R}\mathcal{T}\mathbf{u}_0$ *are identical to those of* $\mathbf{S}_1\mathbf{G}_1\mathbf{u}_0$*. For any rotation* $\mathbf{R}$*, translation* $\mathcal{T}$ *and homothety* $\mathbf{H}$*, the SIFT descriptors of* $\mathbf{u} = \mathbf{S}_1\mathbf{G}_1\mathbf{u}_0$ *with scales larger than* $\sqrt{\lambda^2 - 1}$ *are identical to those of* $\mathbf{v} = \mathbf{S}_1\mathbf{G}_1\mathbf{H}_\lambda\mathbf{R}\mathcal{T}\mathbf{u}_0$.

5.3 Inverting tilts

We shall denote by $*_y$ the 1-D convolution convolution operator in the y-direction. When we write $\mathbf{G}*_y$, we mean that $\mathbf{G}$ is a one-dimensional gaussian, depending on y, and the 1-D convolution means

$$\mathbf{G} *_y \mathbf{u}(x, y) =: \int \mathbf{G}(z)\mathbf{u}(x, y - z)dz.$$

There are *three different notions of tilt*, that we must carefully distinguish.

Definition 3. *Given* $t > 1$*, the tilt factor, define*

- *the absolute tilt :* $\mathbf{T}_t^x\mathbf{u}_0(x, y) =: \mathbf{u}_0(tx, y)$*. In case this tilt is made in the* y *direction. It will be denoted by* $\mathbf{T}_t^y\mathbf{u}_0(x, y) =: \mathbf{u}_0(x, ty)$;
- *the simulated tilt (taking into account camera blur):* $\mathbb{T}_t^x\mathbf{v} =: \mathbf{T}_t^x\mathbf{G}_{\sqrt{t^2-1}} *_x \mathbf{v}$*. In case the simulated tilt is done in the* x *direction, it is denoted* $\mathbb{T}_t^y\mathbf{v} =: \mathbf{T}_t^y\mathbf{G}_{\sqrt{t^2-1}} *_y \mathbf{v}$.
- *the digital tilt (transforming a digital image* u *into a digital image) :* $u \to \mathbf{S}_1\mathbb{T}_t^x Iu$*. This is the one that is used in the algorithm. It is correct because, as we shall see, the simulated tilt yields a blur permitting* $\mathbf{S}_1$*-sampling.*

If $\mathbf{u}_0$ is an infinite resolution image observed with a t camera tilt in the x direction, the observed image is $\mathbf{G}_1\mathbf{T}_t^x\mathbf{u}_0$. Our main problem is to reverse such tilts. This operation is in principle impossible, because absolute tilts do not commute with blur. However, the next lemma shows that $\mathbb{T}_t^y$ is actually a pseudo inverse to $\mathbf{T}_t^x$.

Lemma 2. *One has*

$$\mathbb{T}_t^y = \mathbf{H}_t\mathbf{G}_{\sqrt{t^2-1}} *_y (\mathbf{T}_t^x)^{-1}.$$

Proof Since $(\mathbf{T}_t^x)^{-1}\mathbf{u}(x,y) = \mathbf{u}(\frac{x}{t}, y)$,

$$\left(\mathbf{G}^y_{\sqrt{t^2-1}} *_y (\mathbf{T}_t^x)^{-1}\mathbf{u}\right)(x,y) = \int \mathbf{G}_{\sqrt{t^2-1}}(z)\mathbf{u}(\frac{x}{t}, y-z)dz.$$

Thus

$$\mathbf{H}_t\left(\mathbf{G}^y_{\sqrt{t^2-1}} *_y (\mathbf{T}_t^x)^{-1}\mathbf{u}\right)(x,y) = \int \mathbf{G}_{\sqrt{t^2-1}}(z)\mathbf{u}(x, ty-z)dz =$$

$$\left(\mathbf{G}_{\sqrt{t^2-1}} *_y \mathbf{u}\right)(x,ty) = \left(\mathbf{T}_t^y\mathbf{G}_{\sqrt{t^2-1}} *_y \mathbf{u}\right)(x,y).$$

□

The meaning of the next result is that a tilted digital image $\mathbf{G}_1\mathbf{T}_t^x\mathbf{u}$ can be tilted back by tilting in the orthogonal direction. The price to pay is a t zoom out. The second relation in the theorem means that the application of the simulated tilt to an image that can be well sampled by $\mathbf{S}_1$ yields an image that keeps that well sampling property.

Theorem 1. *Let $t \geq 1$. Then*

$$\mathbb{T}_t^y(\mathbf{G}_1\mathbf{T}_t^x) = \mathbf{G}_1\mathbf{H}_t; \tag{7}$$

$$\mathbb{T}_t^y\mathbf{G}_1 = \mathbf{G}_1\mathbf{T}_t^y. \tag{8}$$

Proof By Lemma 2,

$$\mathbb{T}_t^y(\mathbf{G}_1\mathbf{T}_t^x) = \mathbf{H}_t\mathbf{G}_{\sqrt{t^2-1}} *_y ((\mathbf{T}_t^x)^{-1}\mathbf{G}_1\mathbf{T}_t^x). \tag{9}$$

By a variable change in the integral defining the convolution, it is an easy check that

$$(\mathbf{T}_t^x)^{-1}\mathbf{G}_1\mathbf{T}_t^x\mathbf{u} = \left(\frac{1}{t}\mathbf{G}_1(\frac{x}{t}, y)\right) * \mathbf{u}, \tag{10}$$

and by the separability of the 2D gaussian in two 1D gaussians,

$$\frac{1}{t}\mathbf{G}_1(\frac{x}{t}, y) = \mathbf{G}_1(x)\mathbf{G}_1(y). \tag{11}$$

From (10) and (11) one obtains

$$(\mathbf{T}^x)^{-1}\mathbf{G}_1\mathbf{T}_t^x\mathbf{u} = ((\mathbf{G}_t(x)\mathbf{G}_1(y)) * \mathbf{u} = \mathbf{G}_t(x) *_x \mathbf{G}_1(y) *_y \mathbf{u},$$

which implies

$$\mathbf{G}_{\sqrt{t^2-1}} *_y (\mathbf{T}^x)^{-1}\mathbf{G}_1\mathbf{T}_t^x\mathbf{u} = \mathbf{G}_{\sqrt{t^2-1}} *_y (\mathbf{G}_t(x) *_x \mathbf{G}_1(y) *_y \mathbf{u}) = \mathbf{G}_t\mathbf{u}.$$

Indeed, the 1D convolutions in x and y commute and $\mathbf{G}_t * \mathbf{G}_{\sqrt{t^2-1}} = \mathbf{G}_t$ by the Gaussian semigroup property (5). Substituting the last proven relation in (9) yields

$$\mathbb{T}_t^y\mathbf{G}_1\mathbf{T}_t^x\mathbf{u} = \mathbf{H}_t\mathbf{G}_t\mathbf{u} = \mathbf{G}_1\mathbf{H}_t\mathbf{u}.$$

The second relation (8) follows immediately by noting that $\mathbf{H}_t = \mathbf{T}_t^y\mathbf{T}_t^x$. □

5.4 Proof that A-SIFT works

The meaning of Theorem 1 is that we can design an exact algorithm that simulates all inverse tilts for comparing two images. After interpolation, A-SIFT handles two images $\mathbf{u} = \mathbf{G}_1\mathbf{A}\mathcal{T}_1\mathbf{u}_0$ and $\mathbf{v} = \mathbf{G}_1\mathbf{B}\mathcal{T}_2\mathbf{u}_0$ that are two snapshots from different view points of a flat object whose front infinite resolution image is denoted by $\mathbf{u}_0$. For a sake of simplicity, we break the symmetry, and set $\tilde{\mathbf{u}}_0 =: \mathbf{A}\mathcal{T}_1\mathbf{u}_0$, so that $\mathbf{u} = \mathbf{G}_1\tilde{\mathbf{u}}_0$ and $\mathbf{v} = \mathbf{G}_1\mathbf{B}\mathcal{T}_2\mathcal{T}_1^{-1}\mathbf{A}^{-1}\tilde{\mathbf{u}}_0 = \mathbf{G}_1\mathbf{B}\mathbf{A}^{-1}\mathcal{T}\tilde{\mathbf{u}}_0$ for a translation $\mathcal{T}$ that depends on $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathbf{A}$. Let us use the decomposition given by (1),

$$\mathbf{B}\mathbf{A}^{-1} = \mathbf{R}_1\mathbf{T}_t^x\mathbf{H}_\lambda\mathbf{R}_2,$$

where $\mathbf{R}_1$, $\mathbf{R}_2$ are rotations, $\mathbf{H}_\lambda$ a zoom, and $\mathbf{T}_t^x(x, y) = (tx, y)$ is the transition tilt from $\mathbf{u}$ to $\mathbf{v}$. In summary A-SIFT has to compare the interpolated images

$$\mathbf{v} = \mathbf{G}_1\mathbf{R}_1\mathbf{T}_t^x\mathbf{H}_\lambda\mathbf{R}_2\mathcal{T}\tilde{\mathbf{u}}_0 \text{ and } \mathbf{u} = \mathbf{G}_1\tilde{\mathbf{u}}_0.$$

5.4.1 The A-SIFT formal algorithm

The following algorithm, where image sampling issues are eliminated by interpolation, is actually a proof that A-SIFT manages to compare $\mathbf{u}$ and $\mathbf{v}$ obtained from $\mathbf{u}_0$ by arbitrary camera positions at infinity. In this ideal algorithm, a "'dense enough" set of rotations and tilts is applied to $\mathbf{v}$, so that each one of the simulated rotation-tilts is "close enough" to any other rotation-tilt. In the mathematical setting, this approximation must be infinitesimal. In the practical empirical setting, we'll have to explore how dense the sets of rotations and tilts must be (see Section 6).

A-SIFT Algorithm (formal)

1. Apply a dense set of all possible rotations (and therefore also a rotation close to $\mathbf{R}_1^{-1}$) to $\mathbf{v}$. Thus, some of the simulated images will be arbitrary close to $\mathbf{v} \to \mathbf{R}_1^{-1}\mathbf{G}_1\mathbf{R}_1\mathbf{H}_\lambda\mathbf{T}_t^x\mathbf{R}_2\mathcal{T}\tilde{\mathbf{u}}_0 = \mathbf{G}_1\mathbf{T}_t^x\mathbf{H}_\lambda\mathbf{R}_2\mathcal{T}\tilde{\mathbf{u}}_0$;

2. apply in continuation a dense set of simulated tilts $\mathbb{T}_t^y$, and therefore also one arbitrary close to the right one $\mathbb{T}_t^y =$, to $\mathbf{R}_1^{-1}\mathbf{v} = \mathbf{G}_1\mathbf{T}_t^x\mathbf{H}_\lambda\mathbf{R}_2\mathcal{T}\tilde{\mathbf{u}}_0$. By Theorem 1, this yields

$$\mathbb{T}_t^y\mathbf{R}_1^{-1}\mathbf{v} = \mathbf{G}_1\mathbf{H}_t\mathbf{H}_\lambda\mathbf{R}_2\mathcal{T}\tilde{\mathbf{u}}_0 = \mathbf{G}_1\mathbf{H}_{t\lambda}\mathbf{R}_2\mathcal{T}\tilde{\mathbf{u}}_0;$$

3. perform a SIFT comparison of $\mathbf{G}_1\mathbf{H}_{t\lambda}\mathbf{R}_2\mathcal{T}\tilde{\mathbf{u}}_0$, which is a frontal view of $\tilde{\mathbf{u}}_0$, with $\mathbf{u} = \mathbf{G}_1\tilde{\mathbf{u}}_0$ which also is a frontal view of $\tilde{\mathbf{u}}_0$. By Lemma 1, the application of SIFT to both images detects their common SIFs for scales larger than $\sqrt{(t\lambda)^2 - 1}$.

The above algorithm description is also a proof of the following consistency theorem.

Theorem 2. *Let* $\mathbf{u} = \mathbf{G}_1\mathbf{A}\mathcal{T}_1\mathbf{u}_0$ *and* $\mathbf{v} = \mathbf{B}\mathcal{T}_2\mathbf{u}_0$ *be two images obtained from an infinite resolution image* $\mathbf{u}_0$ *by cameras at infinity with arbitrary position and focal lengths. Then A-SIFT, applied with a dense set of tilts and longitudes, simulates two views of* $\mathbf{u}$ *and* $\mathbf{v}$ *that are obtained from each other by a translation, a rotation, and a camera zoom. As a consequence, these images match by the SIFT algorithm.*

Remark 1. *Even if the above proof, and the statement of Lemma 1, deal with asymptotic statements when the sampling steps tend to infinity or when the SIFT scales tend to infinity, the rate approximation is very quick, a fact that can only be checked experimentally. This fact is actually extensively verified by the huge amount of experimental evidence on SIFT, that shows first that the recognition of scale invariant features (SIFs) is robust to a substantial variation of latitude and longitude, and second that the scale invariance is quite robust to moderate errors on scale. The next section evaluates the adequate sampling rates and ranges for tilts and longitudes.*

5.4.2 Simulating midway tilts

The algorithm of Section 5.4.1 can be implemented in several ways. In the above description, the transition tilt $\mathbf{T}_t^x$ is directly inverted on one of the images. This strategy is consistent, but not optimal. As we have seen, the transition tilt can be very large. It is preferable to simulate moderate tilts on two images that large tilts on one of them. To this aim a *midway image* can be reached from both images by applying a $\sqrt{t}$ tilt to one of them and a $\sqrt{t}$ tilt to the other one. The only change to the formal algorithm will be that rotations and tilts are applied to both images, not just to one of them.

Midway A-SIFT (formal)

1. Apply a dense set of all possible rotations to both images, and therefore $\mathbf{R}_2$ to $\mathbf{u}$ and $\mathbf{R}_1^{-1}$ to $\mathbf{v}$;
2. apply in continuation a dense set of simulated tilts $\mathbb{T}_t^x$ in a fixed $[0, t_{max}]$ range;
3. perform a SIFT comparison of all pairs of resulting images.

Let us now prove that this algorithm works, namely that two of the simulated images are deduced from each other by a similarity. The query and target images are $\mathbf{u} = \mathbf{G}_1 A \mathcal{T}_1 \mathbf{u}_0$ and $\mathbf{v} = \mathbf{G}_1 B \mathcal{T}_2 \mathbf{u}_0$. By the usual decomposition of a linear map (1),

$$BA^{-1} = R_1 T_t^x R_2 H_\lambda = (R_1 T_{\sqrt{t}^x})(T_{\sqrt{t}}^x R_2 H_\lambda).$$

Notice that by the relation

$$\mathbb{T}_t^x \mathbf{R}(-\frac{\pi}{2}) = \mathbf{R}(\frac{\pi}{2}) \mathbb{T}_t^y, \tag{12}$$

the algorithm also simulates tilts in the y direction, up to $\mathbf{R}(\frac{\pi}{2})$ rotation. In particular, the above algorithm applies:

1. $\mathbb{T}_{\sqrt{t}}^x \mathbf{R}_2$ to $\mathbf{G}_1 \mathbf{A} \mathcal{T}_1 \mathbf{u}_0$, which by (8) yields $\tilde{\mathbf{u}} = \mathbf{G}_1 \mathbf{T}_{\sqrt{t}}^x \mathbf{R}_2 \mathbf{A} \mathcal{T}_1 \mathbf{u}_0 =: \mathbf{G}_1 \tilde{\mathbf{A}} \mathcal{T}_1 \mathbf{u}_0$;
2. $\mathbf{R}(\frac{\pi}{2}) \mathbb{T}_{\sqrt{t}}^y \mathbf{R}_1^{-1}$ to $\mathbf{G}_1 \mathbf{B} \mathcal{T}_2 \mathbf{u}_0$, which by (8) yields $\mathbf{G}_1 \mathbf{R}(\frac{\pi}{2}) \mathbf{T}_{\sqrt{t}}^y \mathbf{R}_1^{-1} \mathbf{B} \mathcal{T}_2 \mathbf{u}_0 =: \mathbf{G}_1 \tilde{\mathbf{B}} \mathcal{T}_2 \mathbf{u}_0$.

Let us show that $\tilde{A}$ and $\tilde{B}$ only differ by a similarity. Indeed,

$$\tilde{B}^{-1} R(\frac{\pi}{2}) H_{\sqrt{t}} \tilde{A} = B^{-1} R_1 T_{\sqrt{t}^{-1}}^y T_{\sqrt{t}}^x H_{\sqrt{t}} R_2 A =$$

$$B^{-1}R_1T^y_{\sqrt{t}^{-1}}T^x_{\sqrt{t}}H_{\sqrt{t}}R_2A = B^{-1}R_1T^x_tR(\frac{\pi}{2})R_2A = B^{-1}(BA^{-1})A = I,$$

where I is the identity. It follows that $\tilde{B} = R(\frac{\pi}{2})H_{\sqrt{t}}\tilde{A}$. Thus,

$$\tilde{\mathbf{u}} = \mathbf{G}_1\tilde{\mathbf{A}}\mathcal{T}_1\mathbf{u}_0 \text{ and } \tilde{\mathbf{v}} = \mathbf{G}_1\mathbf{R}(\frac{\pi}{2})\mathbf{H}_{\sqrt{t}}\tilde{\mathbf{A}}\mathcal{T}_2\mathbf{u}_0,$$

that are two of the simulated images, are deduced from each other by a rotation and a $\sqrt{t}$ zoom. It follows that their SIFs are identical as soon as the scale of the SIF exceeds $\sqrt{t}$.

5.5 Conclusion on the algorithms

The above descriptions have neglected the sampling issues, but care was taken that input images and output images be always written in the $\mathbf{G}_1\mathbf{u}$ form. For the digital input images, that always have the form $u = \mathbf{S}_1\mathbf{G}_1\mathbf{u}_0$, the Shannon interpolation algorithm is I is first applied, to give back $I\mathbf{S}_1\mathbf{G}_1\mathbf{u}_0 = \mathbf{G}_1\mathbf{u}_0$. For the output images, that always have the form $\mathbf{G}_1\mathbf{v}$, the sampling $\mathbf{S}_1$ gives back a digital image.

Thus, the descriptions of the formal algorithm A-SIFT and of its "midway" version are changed into a digital algorithm by:

- replacing everywhere the inputs $\mathbf{G}_1\mathbf{u}$ by their digital version $\mathbf{S}_1\mathbf{G}_1\mathbf{u}$;
- by applying digital rotations to digital images : $u \to \mathcal{R}u =: \mathbf{S}_1\mathbf{R}Iu$;
- by applying digital tilts as defined in Def. 3, namely $u \to \mathbf{S}_1\mathcal{T}^x_tIu$.

That way, the formal algorithms are transformed into digital algorithms. The proofs need not be repeated, since by Shannon interpolation and sampling, it is equivalent to talk about $\mathbf{S}_1\mathbf{G}_1\mathbf{u}_0$ or about $\mathbf{G}_1\mathbf{u}_0$.

Clearly the midway algorithm is better, because it only needs simulating tilts that are square roots of the real transition tilts. Thus, all of the forthcoming discussion will focus on the midway version, that we'll simply call A-SIFT.

6 Parameter sampling and complexity

6.1 Sampling ranges

The camera motion depicted in Fig. 4 shows that ϕ should naturally cover all the directions from 0 to 2π. Under the affine camera model, the images taken at ϕ and $\phi + \pi$ are identical up to a rotation of π, i.e., $R_1(\psi)T_tR_2(\phi + \pi) = R_1(\psi + \pi)T_tR_2(\phi)$. Therefore it is enough to simulate ϕ from 0 to π as the rotation-invariant SIFT is invariant to $R_1(\psi)$ and $R_1(\psi + \pi)$.

The sampling range of the tilt parameter t determines the degree of the tilt invariance the algorithm can achieve. Image recognition under a remarkable viewpoint change in practice requires that the scene is planar and Lambertian and its structures are not squashed when observed from an oblique viewpoint. Due to these physical limitations, affine image recognition is impractical

under too big a tilt t. The physical upper bound $t_{\max}$ can be obtained experimentally using some images taken from indoor and outdoor scenes, each image pair being composed of a frontal view and an oblique view.

The images used in the experiments satisfy as much as possible the physical conditions mentioned above. The indoor scene is a magazine placed on a table with the artificial illumination coming from the ceiling as shown in Fig. 12. The outdoor scene is a building façade with some graffiti as illustrated in Fig. 13. For each pair of images, the true tilt parameter t between them is obtained by manual measurement. A-SIFT is applied with very large parameter sampling ranges and small sampling steps, so that the simulated views cover accurately the true affine distortion. The A-SIFT matching results depicted in Figs. 12 and 13 show that the limit is $t_{\max} \approx 5.6$ that corresponds to a view angle $\theta_{\max} = \arccos 1/t_{\max} \approx 80°$. A-SIFT finds a large number of matches when the tilt between the frontal image and the oblique image is smaller than about 5.6. Therefore we set the tilt simulation range $t_{\max} = 4\sqrt{2}$.

Let us emphasize that when the two images under comparison are taken from orthogonal longitude angles (see Fig. 8 as an example), i.e., $\phi = \phi' + \pi/2$, the maximum tilt invariance A-SIFT with $t_{\max} = 4\sqrt{2}$ can achieve in theory is about $t^2_{\max} = 32$.

However, these experiments only fix reasonable bounds for all purpose algorithms. For high resolution images, for very flat lambertian surfaces, larger tilts might be recognizable.

6.2 Sampling steps

In order to have A-SIFT invariant to any affine transform, one needs to sample the tilt t and angle ϕ with a high enough precision. The sampling steps $\triangle t$ and $\triangle\phi$ will be fixed experimentally by testing several natural images.

The camera motion model illustrated in Fig. 4 indicates that the sampling precision of the latitude angle $\theta = \arccos 1/t$ should increase with θ. A geometric sampling for t satisfies this requirement. Naturally, the sampling ratio $\triangle t = t_{k+1}/t_k$ should be independent of the angle ϕ. Fig. 14-a illustrates the number $\tilde{N}_S(t_1, t_2)$ of SIFT matches (dark pixels represent large values) between $\mathbf{v}_1 = \mathbf{T}_{t_1}\mathbf{u}$ and $\mathbf{v}_2 = \mathbf{T}_{t_2}\mathbf{u}$, with $t_1, t_2 \in [1, 8]$, where $\mathbf{u}$ is a natural image. When $|t_1 - t_2|$ increases, $\tilde{N}_S(t_1, t_2)$ decreases. This decay is faster when t_1 and t_2 are closer to 1. Fig. 14-b shows a 1/4 threshold of $\tilde{N}_S(t_1, t_2)/N_S(t_1)$, where $N_S(t_1) = \tilde{N}_S(t_1, t_1)$ is the number of SIFT matches between $\mathbf{v}_1 = \mathbf{T}_{t_1}\mathbf{u}$ and itself. (The values of $N_S(t_1)$ appear on the diagonal of the $\tilde{N}_S$ image shown in Fig. 14-a.) Thus, in Fig. 14-b, the white pixels correspond to $\tilde{N}_S/N_S > 1/4$. The dashed lines in Fig. 14-b mark a geometrically sampled t grid, with sampling step $\triangle t = \sqrt{2}$. Following the arrows in Fig. 14-b, one can verify that this sampling step ensures that $\tilde{N}_S/N_S > 1/4$, which means that with $\triangle t = \sqrt{2}$, the number of A-SIFT matches exceeds 1/4 time the maximum number of matches. This number is large enough to have A-SIFT algorithm robust to any tilt change. Thus, in the sequel, the tilt sampling step will be $\triangle t = \sqrt{2}$.

As can be observed from the camera motion model in Fig. 4, one needs a finer ϕ sampling when $\theta = \arccos 1/t$ increases: the image distortion caused by a fixed longitude angle displacement $\triangle\phi$, is much more drastic when the latitude angle θ increases. This fact is confirmed by the experiments shown in Fig. 15. The curves plot for different tilt parameters t, different test images and different tilt implementations the values $\tilde{N}_S(\phi, t)/N_S(t)$ versus ϕ, where $\tilde{N}_S(\phi, t)$ is the number of SIFT matches between $\mathbf{v}_1 = \mathbf{R}_1(\phi)\mathbf{T}_t\mathbf{u}$ and $\mathbf{v}_2 = \mathbf{T}_t\mathbf{u}$ and $N_S(t) = \tilde{N}_S(0, t)$ is the number of

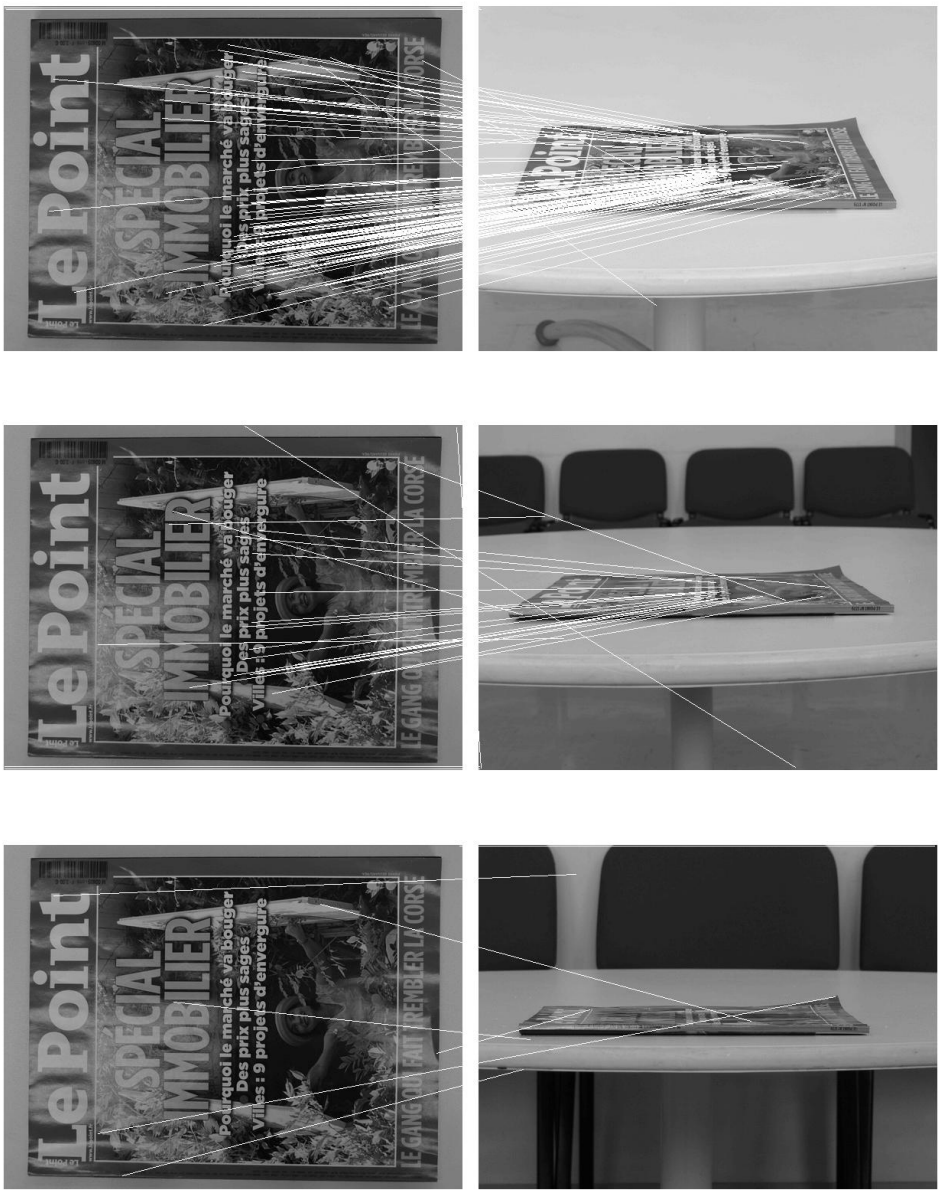


Figure 12: A-SIFT on an indoor scene. From top to bottom: tilt distortion t between the two images are respectively $t \approx 3, 5.2, 8.5$; the number of matches are respectively 107 (3 false), 25 (7 false), 7 (all false).

Figure 13: A-SIFT on an outdoor scene. From top to bottom: tilt distortion t between the two images are respectively $t \approx 3.8, 5.6, 8$; the number of matches are respectively 71 (4 false), 33 (4 false), 6 (all false).

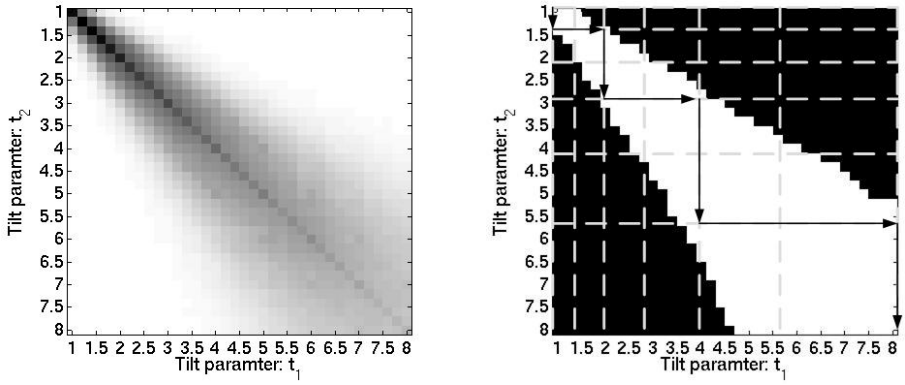


Figure 14: Left: the number $\tilde{N}_S(t_1, t_2)$ of SIFT matches between $\mathbf{v}_1 = T_{t_1}\mathbf{u}$ and $\mathbf{v}_2 = T_{t_2}\mathbf{u}$ (black pixels represent big values). Right: a thresholding on $\tilde{N}_S(t_1, t_2)/N_S$, where N_S is the number of SIFT matches between $\mathbf{v}_1 = T_{t_1}\mathbf{u}$ and itself that corresponds to a perfect latitude normalization, with white pixels corresponding to $\tilde{N}_S/N_S > 1/4$.

SIFT matches between $\mathbf{v}_1 = \mathbf{T}_t\mathbf{u}$ and itself. One observes that the larger t, the faster the decay of $N_S(\phi, t)$ versus ϕ. Thresholding at $\tilde{N}_S/N_S > 1/5$, one observes from Table 2 that for all the images and tilt implementations under test, ϕ in first approximation order proportional to $1/t$. More precisely, with $\phi \approx 18^\circ \times \frac{2}{t}$ one obtains $\tilde{N}_S/N_S > 1/5$, which means that longitude angle sampling step $\triangle\phi = \frac{36^\circ}{t}$ ensures that the number of A-SIFT matches exceeds $1/5$ time the maximum number of matches. This number is large enough to have A-SIFT algorithm robust to any longitude angle change. Since the quantification projects a value to its nearest sampled grid, the longitude sampling step in the sequel will be $\triangle\phi = 2 \times \frac{36^\circ}{t} = \frac{72^\circ}{t}$.

Fig. 16 illustrates the sampling of the parameters $\theta = \arccos 1/t$ and ϕ. At bigger θ the sampling of θ as well as the sampling of ϕ are denser.

6.3 Acceleration with multi-resolution

The multi-resolution procedure accelerates A-SIFT by selecting the transforms that yield SIFT matches on low-resolution (LR) versions of the compared images. In case of success only, the procedure simulates the identified affine transforms on the query, and applies SIFT to compare them to the targets.

The multi-resolutions A-SIFT is summarized as follows.

1. Down-sample all compared images $\mathbf{u}$ and $\mathbf{v}$ by a $K \times K$ factor: $\mathbf{u}' = \mathbf{S}_K\mathbf{G}_K\mathbf{u}$ and $\mathbf{v}' = \mathbf{S}_K\mathbf{G}_K\mathbf{v}$, where $\mathbf{G}_K$ is an anti-aliasing gaussian filtering.
2. Low-resolution (LR) A-SIFT: perform A-SIFT between $\mathbf{u}'$ and $\mathbf{v}'$.
3. Identify the M affine transforms yielding the biggest numbers of matches between $\mathbf{u}'$ and

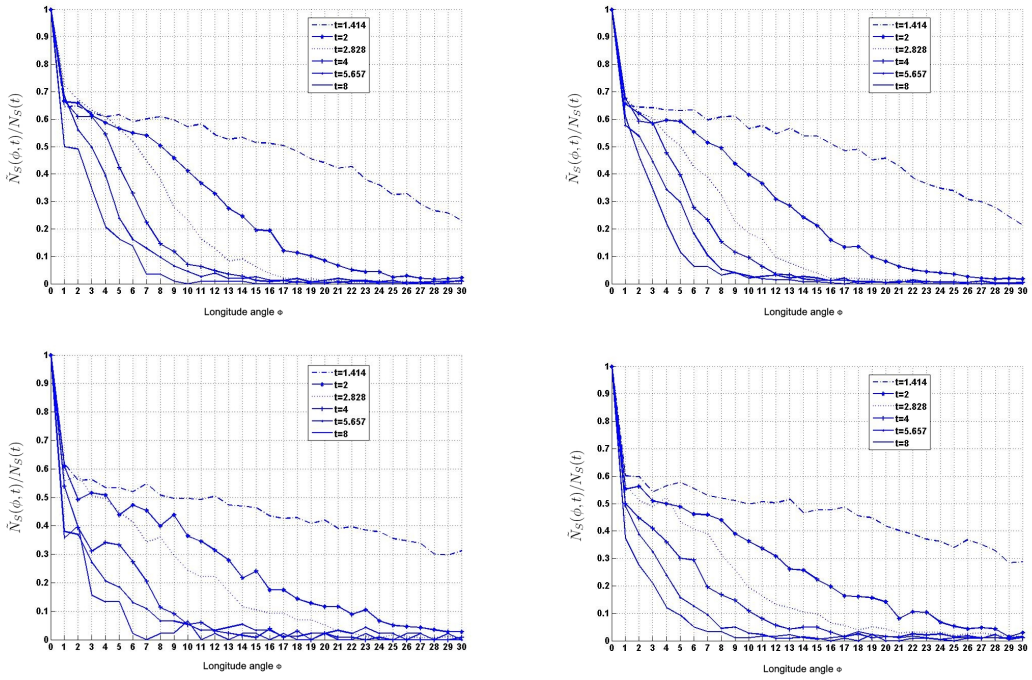


Figure 15: The curves plot for different tilt parameters t, different test images and different tilt implementations the values $\tilde{N}_S(\phi,t)/N_S(t)$ versus ϕ, where $\tilde{N}_S(\phi,t)$ is the number of SIFT matches between $\mathbf{v}_1 = \mathbf{R}_1(\phi)\mathbf{T}_t\mathbf{u}$ and $\mathbf{v}_2 = \mathbf{T}_t\mathbf{u}$ and $N_S(t) = \tilde{N}_S(0,t)$ is the number of SIFT matches between $\mathbf{v}_1 = \mathbf{T}_t\mathbf{u}$ and itself, which corresponds to a perfect longitude normalization. Top and bottom rows: the images under test are respectively "Graffiti" and "Leuven". Left and right columns: a tilt of t is implemented respectively by t times subsampling in x direction and by $\sqrt{t}$ times subsampling in x direction together with $\sqrt{t}$ times oversampling in y direction.

	$t=2$	$t=2\sqrt{2}$	$t=4$	$t=4\sqrt{2}$	$t=8$
Graf, $\mathbf{T}^c$	32	30	30	31	32
Graf, $\mathbf{T}^b$	30	28	30	33	32
Leuven, $\mathbf{T}^c$	31	33	29	27	24
Leuven, $\mathbf{T}^b$	32	28	28	26	24

Table 2: The table shows for several values of the tilt parameter t, for several different images, and for several tilt implementations the value of $\phi \times t$ obtained when ϕ is the maximal value satisfying $\tilde{N}_S(\phi,t)/N_S(t) > 1/5$. Here, $\tilde{N}_S(\phi,t)$ is the number of SIFT matches between $\mathbf{v}_1 = \mathbf{R}_1(\phi)\mathbf{T}_t\mathbf{u}$ and $\mathbf{v}_2 = \mathbf{T}_t\mathbf{u}$ and $N_S(t) = \tilde{N}_S(0,t)$ is the number of SIFT matches between $\mathbf{v}_1 = \mathbf{T}_t\mathbf{u}$ and itself. "Graf" and "Leuven" stand for the tested images "Graffiti" and "Leuven". $\mathbf{T}^c$ and $\mathbf{T}^b$ symbolize two different tilt implementations. $\mathbf{T}^c$ implements the tilt by a t-sub-sampling in the x direction. $\mathbf{T}^b$ implements the tilt by a $\sqrt{t}$-sub-sampling in the x direction together with a $\sqrt{t}$-over-sampling in the y direction. The fact that the $t \times \phi$ entries are roughly constant yields the simple empirical law $\phi = C/t$ for the longitude step, where C is a constant.

$\mathbf{v}'$. They are retained only if the matches are meaningful. In practice, it is enough to put a threshold on the number k of matches, and $k = 15$ seems to be a good choice.

4. High-resolution (HR) A-SIFT: apply A-SIFT between $\mathbf{u}$ and $\mathbf{v}$ by simulating only the affine transforms previously identified.

Fig. 17 shows an example. The low-resolution A-SIFT that is applied on the 3×3 sub-sampled images finds 26 correspondences and identifies the 5 best affine transforms. The high-resolution A-SIFT finds 245 matches.

6.4 A-SIFT Complexity

The complexity of the A-SIFT algorithm will be estimated under the recommended baseline configuration: The tilt and angle ranges are $[t_{\min}, t_{\max}] = [1, 4\sqrt{2}]$ and $[\phi_{\min}, \phi_{\max}] = [0°, 180°]$, and the sampling steps are $\triangle t = \sqrt{2}$, $\triangle\phi = 36° \times \frac{t}{2}$. Each t tilt is simulated by image sub-sampling in one direction by a t factor. All images are sub-sampled by a $K \times K = 3 \times 3$ factor for the low-resolution A-SIFT. Finally, the high-resolution A-SIFT simulates the $M = 5$ best affine transformations that are identified, but only in case they contain enough matches. When matching an image to a large database, the most common event is failure. Thus, the final high-resolution step is only to be taken into account when comparing images of the same scene.

The complexity of the descriptor computation is proportional to the input image area. This area is proportional to the number of simulated tilts t. Indeed, the number of ϕ simulations is proportional to t for each t, but the t sub-sampling for each tilt simulation divides the area by t. More precisely, the image area input to low-resolution A-SIFT is

$$\frac{1 + (|\Gamma_t| - 1)\frac{180°}{2\times 36°}}{K^2} = \frac{1 + 5 \times 2.5}{9} = 1.5$$

times as large as that of the original images, where $|\Gamma_t|$ is the number of tilt simulations. Thus,

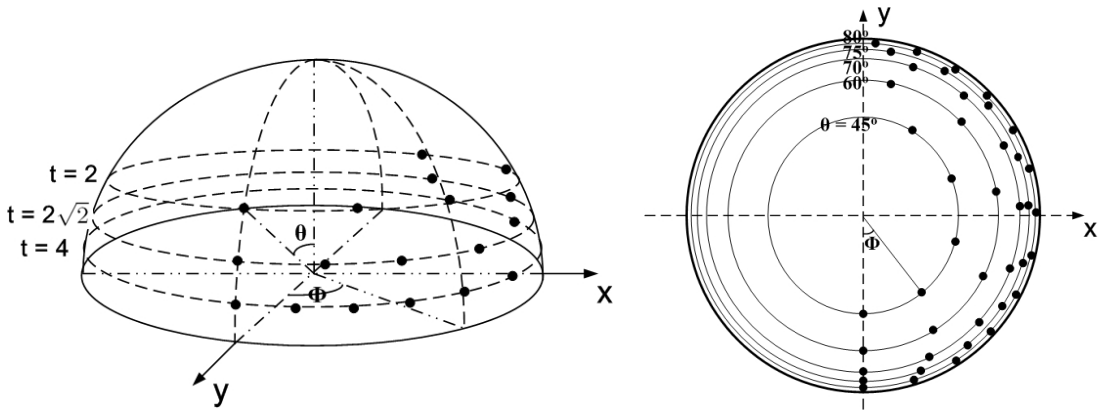


Figure 16: Sampling of the parameters $\theta = \arccos 1/t$ and ϕ. Black dots represent the sampling. Left: perspective illustration (only $t = 2, 2\sqrt{2}, 4$ are shown). Right: zenith view of the observation half sphere. The values of θ are indicated on the figure.

the complexity of the low-resolution A-SIFT is 1.5 times as much as that of a single SIFT routine, and generates 1.5 as many SIFs. Here we must distinguish two cases:

1. If the comparisons involve a large database (where most comparisons will be failures), the complexity is propositional to the number of SIFs in the queries multiplied by the number of SIFs in the targets. Since A-SIFT introduces a a 1.5 area factor, the final complexity is simply $1.5^2 = 2.25$ times the SIFT complexity.

2. If the comparisons involve a set of images with high match likeliness, then the high resolution step is no more negligible. Then, it can only be asserted that the complexity will be less than $6.5 + 2.5 = 9$ times a SIFT routine on the same images. However, in that case, A-SIFT ensures many more detections than SIFT, because it explores many more viewpoint angles. Thus, the *complexity rate per detected SIF* might be much closer to, or even smaller than the per detection complexity in a SIFT routine.

For the high-resolution A-SIFT, this factor is $M = 5$. Therefore the total complexity of the A-SIFT is 6.5 times a SIFT routine.

The SIFT subroutines can be implemented in parallel in A-SIFT (for both the low-resolution and the high-resolution A-SIFT). Recently many authors have investigated SIFT accelerations [24, 16, 27]. A realtime SIFT implementation has been proposed in [62]. Obviously, all of these accelerations directly apply to A-SIFT.

7 Experiments

A-SIFT image matching performance will be compared with the state-of-the-arts approaches SIFT [31] and MSER [33]. SIFT is probably the most popular image matching algorithm for

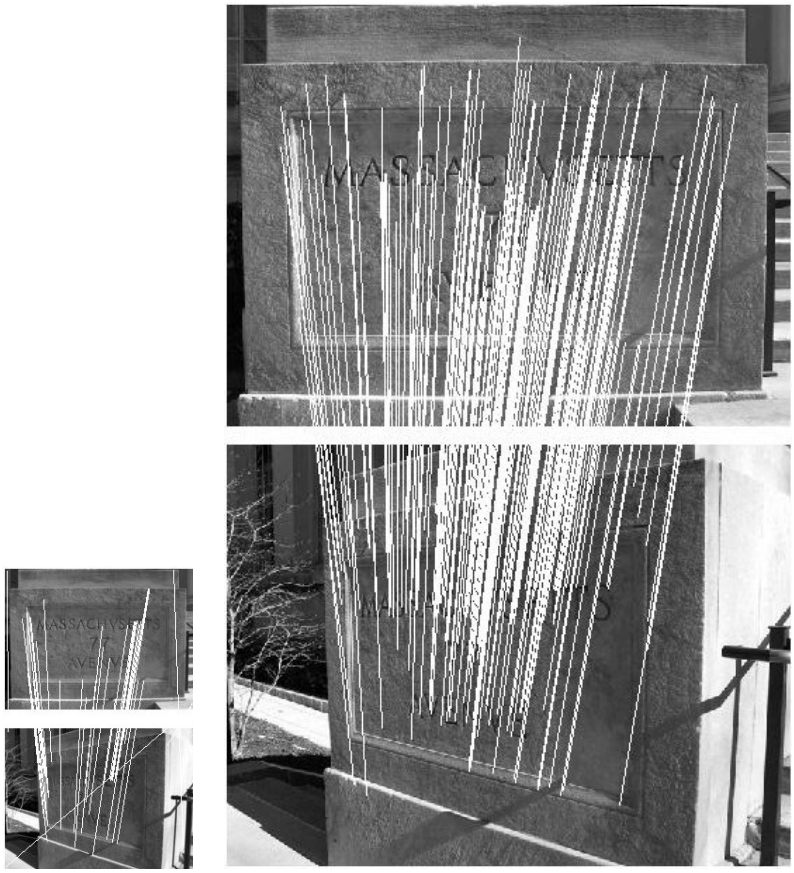


Figure 17: "77 Mass Ave". Left: low-resolution A-SIFT. Right: high-resolution A-SIFT.

its robust performance under image scale changes, rotation, translation, moderate illumination changes and viewpoint changes [39]. MSER is the most robust to large viewpoint changes [40]. The Lowe [30] SIFT reference software and the Matas et al. [34] MSER online demo were used. Applications of A-SIFT and comparisons with SIFT will also be performed for video matching, object tracing and symmetry detection.

7.1 Image matching

The experiments will show images taken from different viewpoints with varying tilts, zooms, and transition tilts. In the image pairs compared by A-SIFT or SIFT, correspondences will be connected by white segments. Note that the parallelism or coherent directions of the connecting lines usually indicates that most correspondences are correct. On the MSER results, the correspondences are numerated. The blue lines represent the epipolar geometry and are not correspondences.

7.1.1 A systematic comparison with SIFT and MSER

	$t_1 = 2$	$t_1 = 2\sqrt{2}$	$t_1 = 4\sqrt{2}$
$t_2 = 2$	$45°/2.8$	$45°/3.7$	
$t_2 = 2\sqrt{2}$	$45°/3.7$	$45°/4.8$	
$t_2 = 4\sqrt{2}$			

Table 3: m/n in each entry with: m = longitude angle ϕ between the two viewpoints /n= transition tilt $\tau(t_1, t_2, \phi)$. This table illustrates the fact that MSER covers transition tilts up to $\tau \approx 4.5$. The experiments were performed with the real images Magazine with $t = 2,\ 2.8,\ 5.6$, and $\phi = 0°,\ 45°,\ 90°$. In each entry the largest τ ensuring matching success is given. The blank entries mean that MSER fails.

Fig. 18 shows the setting that we have adopted to make a systematic comparison between A-SIFT, SIFT and MSER. A poster illustrated in Fig. 19 is photographed with a reflex camera, with distances varying between $\times 1$ and $\times 10$, which is the maximum focal distance change, and with viewpoint angles between the camera axis and the normal to the poster that varies from 0 degree (frontal view) to 80 degrees. It is clear that beyond 80 degrees, to establish a correspondence between the frontal image and the extreme viewpoint becomes absolutely haphazard. Even when the photo acquisition conditions and the image resolution are excellent, with such a big view angle change the observed surface becomes in general reflective, and the image in the resulting photo is totally different from the frontal view. Nevertheless, A-SIFT works until 80 degrees, and it would be unrealistic to insist on bigger angles.

Fig. 20 shows the thumbnails of the images taken from different viewpoints with a setting as shown in Fig. 18. The triple of the numbers c/s/m placed above or below each image gives the number of correct correspondences obtained by A-SIFT (c) compared with those obtained by SIFT (s) and MSER (m). Table 4 summarizes in more detail the total number of correspondences and the number of correct ones achieved by each approach. Some matching results are illustrated in Figs. 22 to 25.

One remarks first that MSER, which uses maximal stay level lines as features, obtains systematically much less correspondences than SIFT and A-SIFT whose features are based on local maxima in the scale-space. This has been confirmed by LLD, a novel image matching approach independently developed at ENS Cachan that applies also level lines as features [48, 47]. Let us recall that robust image matching requires a sufficiently big number of correspondences.

For images taken at short distance, the number of SIFT correspondences drops dramatically when the angle is bigger than 65 degrees (that corresponds to a tilt $t \approx 2.3$) and it fails completely when the angle exceeds 75 degrees (tilt $t \approx 3.8$); the MSER correspondences remain rather stable at a small number until 75 degrees and tends to fail completely under bigger angles; A-SIFT works perfectly until 80 degrees (tilt $t \approx 5.8$). Images taken at a camera-object distance multiplied by 10 exhibits less perspective effects but contains less meaningful pixels at big angles. For these images the SIFT performance drops considerably: recognition is possible only with angles smaller than 45 degrees; MSER struggles at the angle of 45 degrees and fails at 65 degrees; A-SIFT again functions perfectly until 80 degrees.

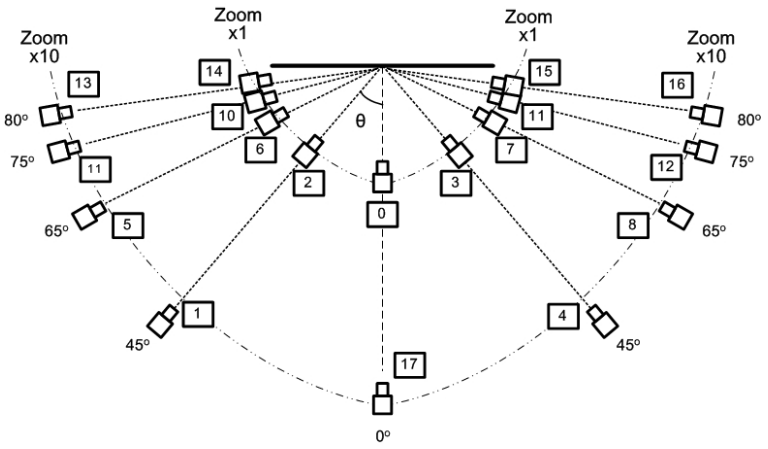


Figure 18: This figure and the following one show the setting adopted to compare SIFT, the most popular method, with A-SIFT. A poster is photographed with a reflex camera from distances varying between 1 and 10, which is the maximum focal distance change, and with a viewpoint angle between the camera axis and the normal to the poster that varies from 0 degree (frontal view) to 80 degrees.

7.1.2 Comparison with SIFT and MSER

Figs. 26-31 compare the A-SIFT image matching results with SIFT and MSER on various types of images. On non flat or non coplanar objects, the absolute or transition tilts in the same scene may considerably change. Thus it is important to allow for large transition tilts. Actually the absolute and transition tilts can vary on the very same flat object, as illustrated in Fig. 21.

Fig. 26 illustrates the performance on an object that is almost flat but taken with a very oblique view. A-SIFT works perfectly, SIFT fails completely, and MSER, the affine-invariant

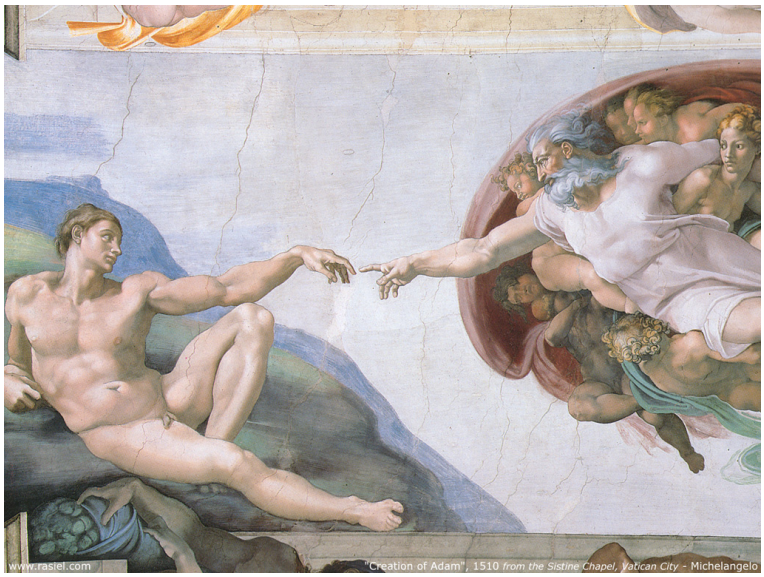


Figure 19: The poster that is photographed in the experiments.

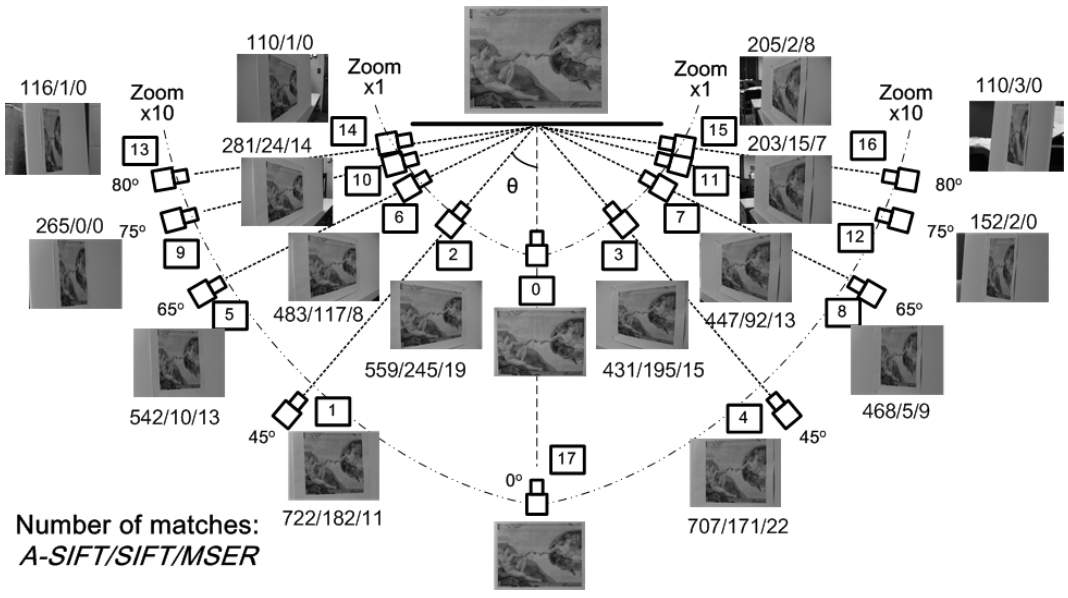


Figure 20: This figure shows several images of the same flat object taken from different viewpoints with the setting shown in Fig. 18. The triple of the numbers A/S/M placed below each image gives the number of correct correspondences obtained by A-SIFT (A) compared with those obtained by SIFT (S) and MSER (M).

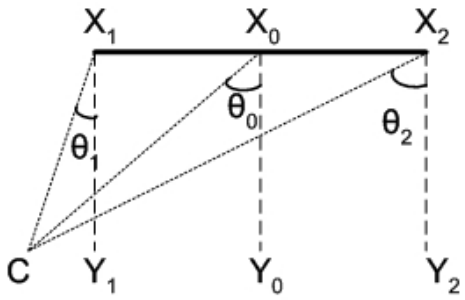


Figure 21: When the camera view angle is large, the absolute tilt of a plane object can vary considerably in the same image. The performance of the recognition should be maintained anyway.

Angle:	−80°	−75°	−65°	−45°	45°	65°	75°	80°
SIFT Zx1	2/1	24/24	117/117	246/245	197/195	92/92	15/15	3/2
MSER Zx1	0/0	17/14	15/8	23/19	22/15	15/13	10/7	11/8
A-SIFT Zx1	111/110	281/281	484/483	562/559	431/428	447/444	203/203	205/204
SIFT Zx10	2/1	2/0	10/10	183/182	171/171	5/5	2/2	5/3
MSER Zx10	0/0	7/0	16/13	28/11	23/22	9/9	0/0	0/0
A-SIFT Zx10	118/116	268/265	542/542	723/722	708/707	469/468	154/152	110/110

Table 4: Summary of the results of the experiments that compare A-SIFT with SIFT and MSER for viewpoint angles between 45 and 80 degrees. m/n: number of matches/number of correct matches.

method, functions reasonably well for two reasons: first the existence of the highly contrasted shapes meets the working condition of this method, and second the camera-object distances of the two images are close. With a bigger angle on the same object (tilt about 2.5), as shown in Fig. 27, SIFT fails logically, and MSER fails as well, since the shapes presented under such a big angle tend to be mixed and therefore no longer provide "MSERs" (*Maximal Stable Extremal Regions*). Fig. 28 presents an image pair with a considerable viewpoint change, on a desk supporting many non-coplanar objects. A-SIFT finds 62 correspondences out of which 58 are correct. SIFT fails completely. MSER finds 13 correspondences out of which only 2 are correct. Fig. 29 shows images of a building façade taken from very different viewpoints. The transformation of the rectangle façade on the left to a trapezia on the right indicates that the transformation is not affine, but strongly perspective. Nevertheless, since a projective transformation can be locally modeled by affine transforms, a large number of correspondences is established by A-SIFT. Fig. 30 shows the results of the standard test pair Graffiti 1 and Graffiti 6 proposed by Mikolajczyk [35]. A-SIFT finds 724 correspondences, out of which 3 are false. SIFT finds 6 correspondences: the $\tau = 3.2$ transition tilt is just a bit too large. MSER finds 127 correspondences out of which 50 are correct. Proposed by Matas et al. in their online demo [34] as a standard image to test MSER [33], the

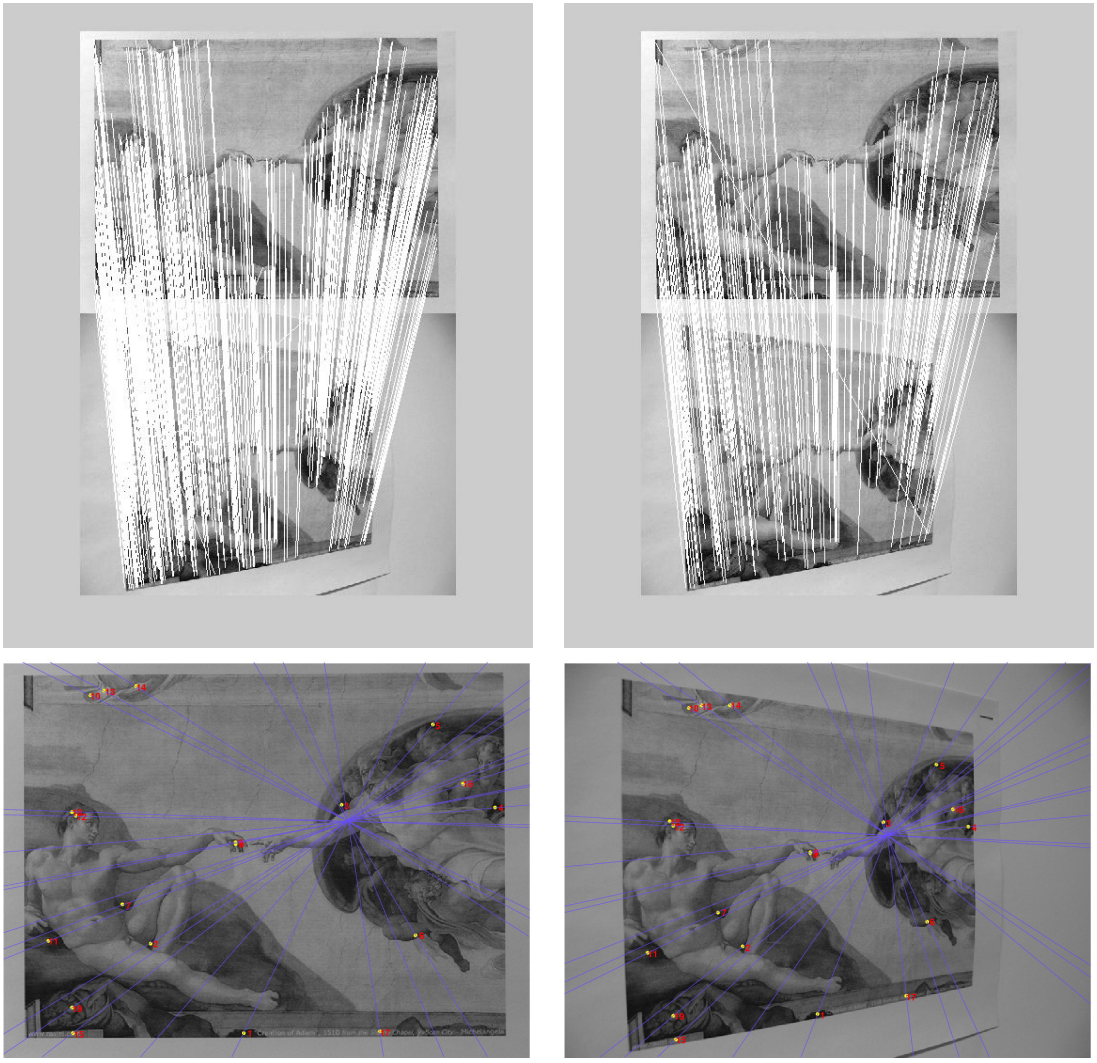


Figure 22: Correspondences between the poster images taken from short distance (zoom $\times 1$) at frontal view and at -45° angle. The absolute tilt varies: $t = 2$ (middle), $t \prec 2$ (left part), $t \succ 2$ (right part). Top left: A-SIFT finds 562 correspondences out which 559 are correct. Top right: SIFT finds 246 correspondences out which 245 are correct. Bottom: MSER finds 29 correspondences, out of which 13 are correct.

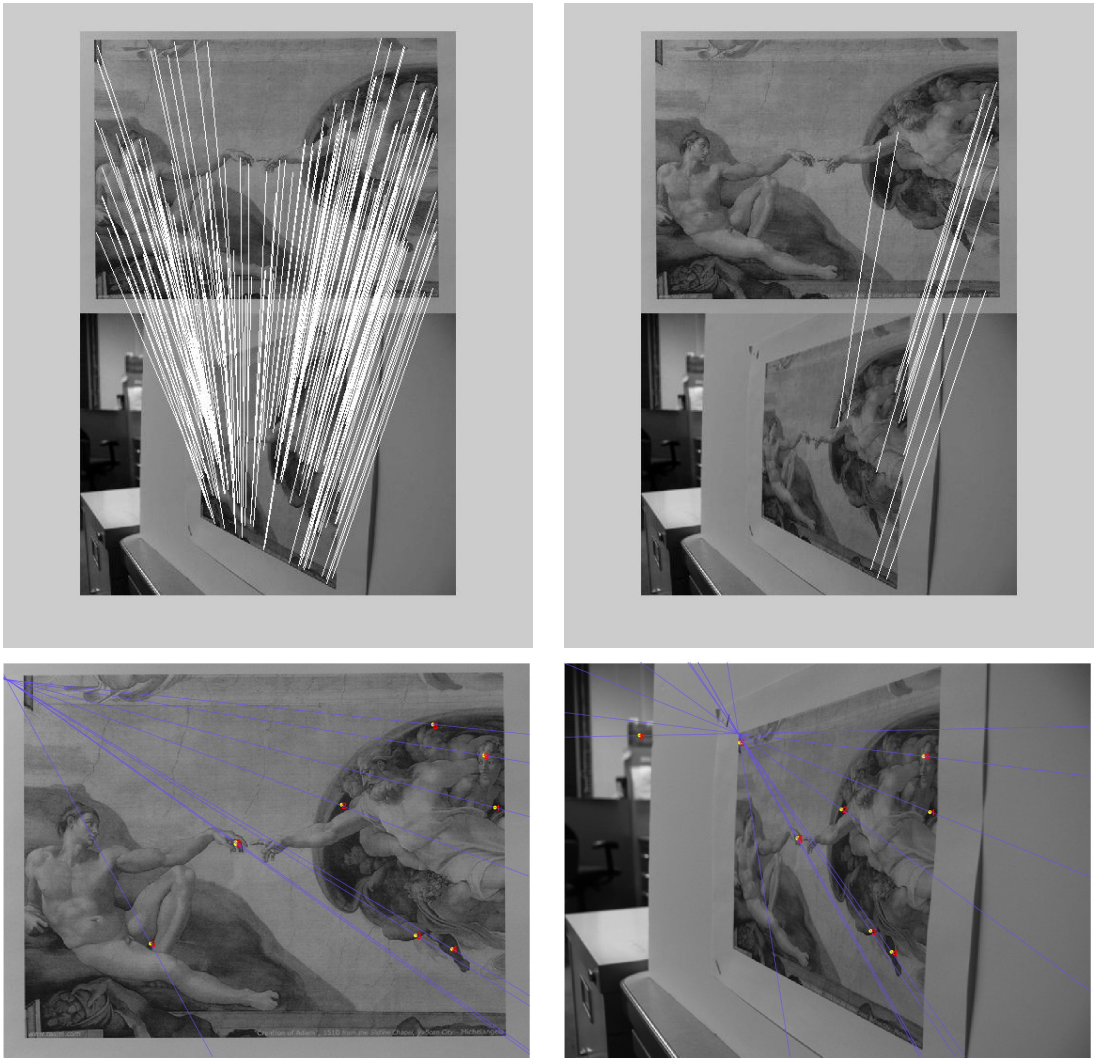


Figure 23: Correspondences between the poster images taken from short distance (zoom $\times 1$) at frontal view and at 75° angle. The local absolute tilt varies: $t = 4$ (middle), $t \prec 4$ (right part), $t \succ 4$ (left part). Top left: A-SIFT finds 203 correspondences, all correct. Top right: SIFT finds 15 correspondences, all correct and all on the right part, where the tilt is lower. Bottom: MSER finds 10 correspondences, out of which 7 are correct.

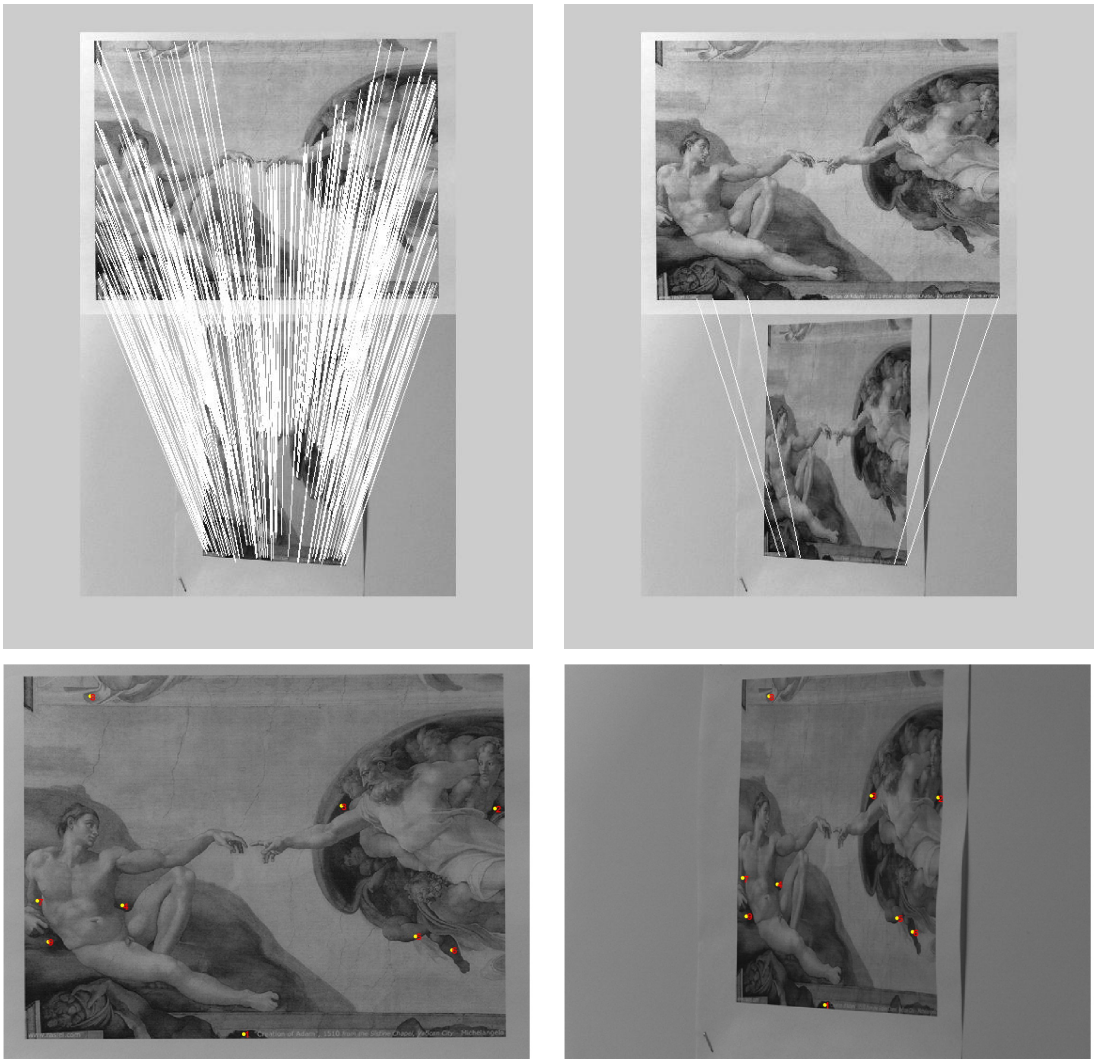


Figure 24: Correspondences between the poster images taken from long distance (zoom $\times 10$) at frontal view and at 65° angle, absolute tilt $t = 2.4$. Top left: A-SIFT finds 469 correspondences, out of which 468 are correct. Top right: SIFT finds 5 correspondences, all correct. Bottom: MSER finds 9 correspondences, all correct.

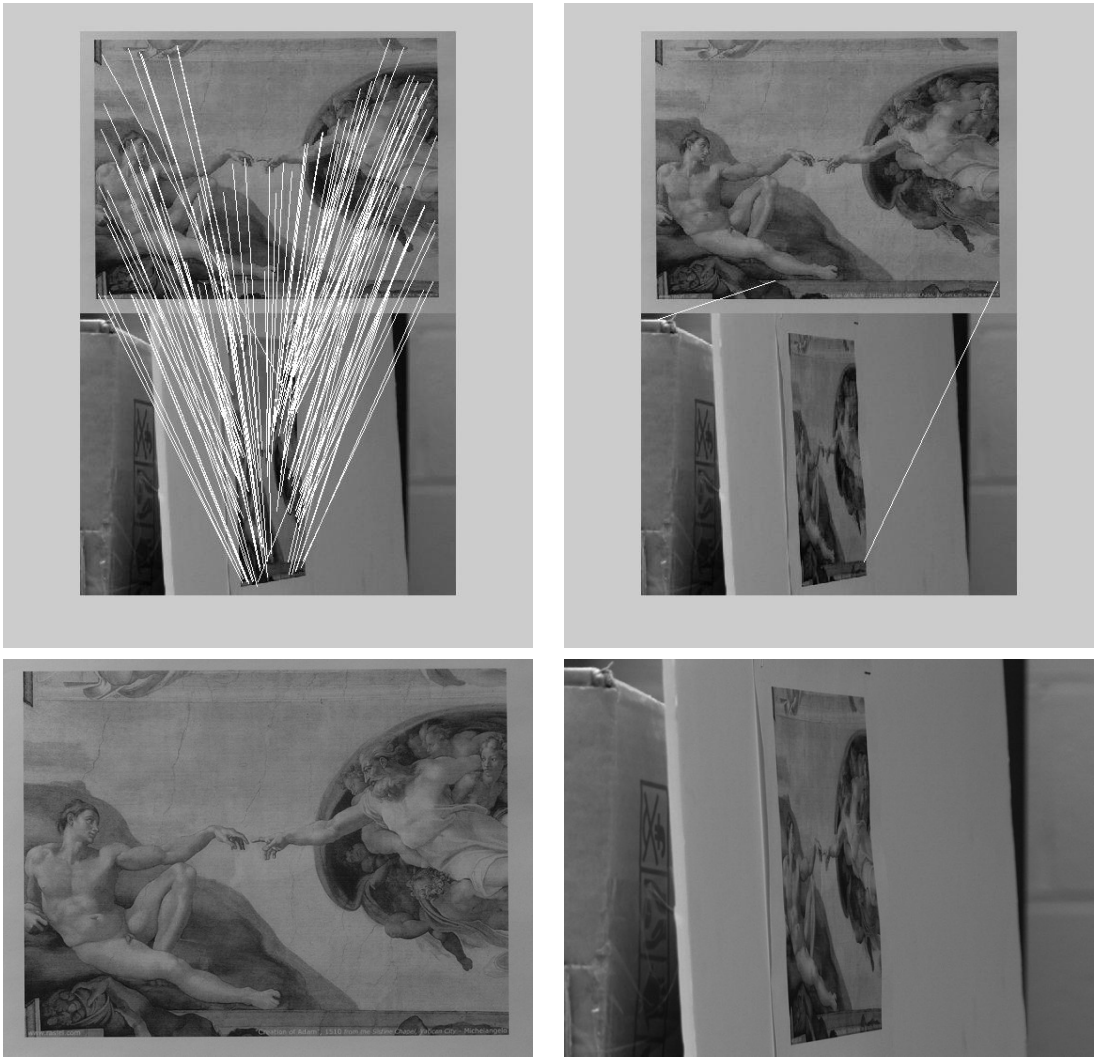


Figure 25: Correspondences between the poster images taken from long distance (zoom $\times 10$) at frontal view and at $-80°$ angle, absolute tilt $t = 5.6$. Top left: A-SIFT finds 118 correspondences, out of which 116 are correct. Top right: SIFT finds 5 correspondences, all correct. Bottom: MSER finds 0 correspondence.

images in Fig. 31 show a number of containers placed on a desktop [1]. A-SIFT finds 196 matches out of which 2 are false. MSER find 70 matches out of which 50 are inliers. Let us note that images in Figs. 30 and 31 provide optimal conditions for MSER: the camera-object distances are similar and well contrasted shapes are present. But let us recall that MSER fails under large scale changes or when well contrasted shapes are not present.

7.2 Video matching and object tracking

Object tracking in video is a challenging problem. Difficulties in tracking objects can arise due to abrupt object motion, object pose change, object-to-object and object-to-scene occlusions, and camera motion. Applications of object tracking include motion-based recognition, automated surveillance, video indexing, human-computer interaction, vehicle and robot navigation, 3D reconstruction, image registration, etc. (see [73] for a survey). Due to its invariance to similarity transformations and to the large number of feature points it usually detects, SIFT has been recently applied in object tracking and video matching [25, 5, 62, 52, 55]. However, as SIFT is robust only to tilts $t < 2.5$, its performance drops significantly if the object changes pose moderately or deforms. In this subsection, A-SIFT is compared with SIFT in video matching and object tracking.

A video sequence at VGA resolution (640×480) and frame rate 30 fps was used in the experiments. Some frames are shown in Figs. 32 and 33. The camera position was fixed while the object kept changing its pose and slightly deforming, thus inducing rotation, translation, change of scale and oblique views of itself in the video.

For video matching, a query image was to be matched with the video frames. Fig. 32 shows some result samples obtained with A-SIFT compared to SIFT. While SIFT fails completely because of the object's pose change and deformation, A-SIFT matches successfully the query images in all video frames.

Instead of finding correspondences between a query image and the video frames, object tracking in video is realized by finding correspondences between video frames. Fig. 33 illustrates some samples of object tracking results obtained with A-SIFT and SIFT. Again A-SIFT succeeds and SIFT fails, due to the pose changes and deformations of the object.

7.3 Symmetry detection in perspective

Symmetry detection has drawn considerable attention in computer vision and has been used for numerous applications such as image indexing, completion of occluded shapes, object detection, facial image analysis and visual attention (see, for example, [11] for a survey). The image projection is usually approximated by plane affine transforms for symmetry detection in perspective [45]. Some recent works apply SIFT, MSER and other affine-invariant detectors and descriptors to detect bilateral symmetry [32, 11]. Conversely, symmetry has been used to extract affine-invariant image features [2].

A-SIFT can be used to detect bilateral symmetry in an image $\mathbf{u}$, by simply looking for correspondences between $\mathbf{u}(x, y)$ and its flipped version $\mathbf{u}(-x, y)$. After being flipped, symmetric structures become either identical if taken in frontal view, or identical up to an oblique view otherwise. A correspondence between $\mathbf{u}(x, y)$ and $\mathbf{u}(-x, y)$ therefore connects a pair of bilateral symmetrical

[1] We thank Michal Perdoch for having kindly provided us with the images.

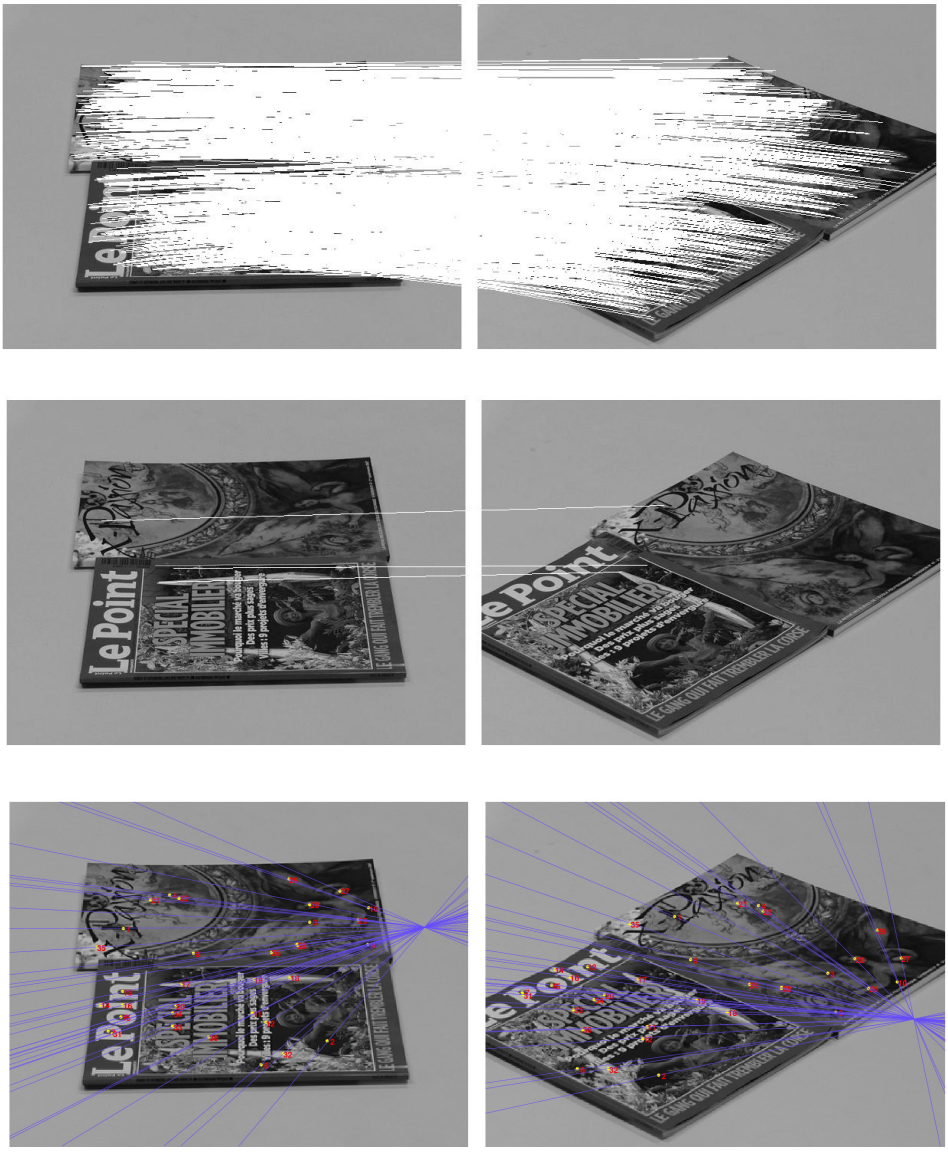


Figure 26: Correspondences between the images Magazine 1 and Magazine 2. The absolute tilts for these objects are $t = t' \simeq 2.1$. The transition tilt is larger, $\tau = 3.0$. Top: A-SIFT finds 1667 correspondences, all correct. Middle: SIFT finds 3 correspondences (SIFT usually fails for transition tilts larger than 2.5). Bottom: MSER finds 46 correspondences, out of which 35 are correct.

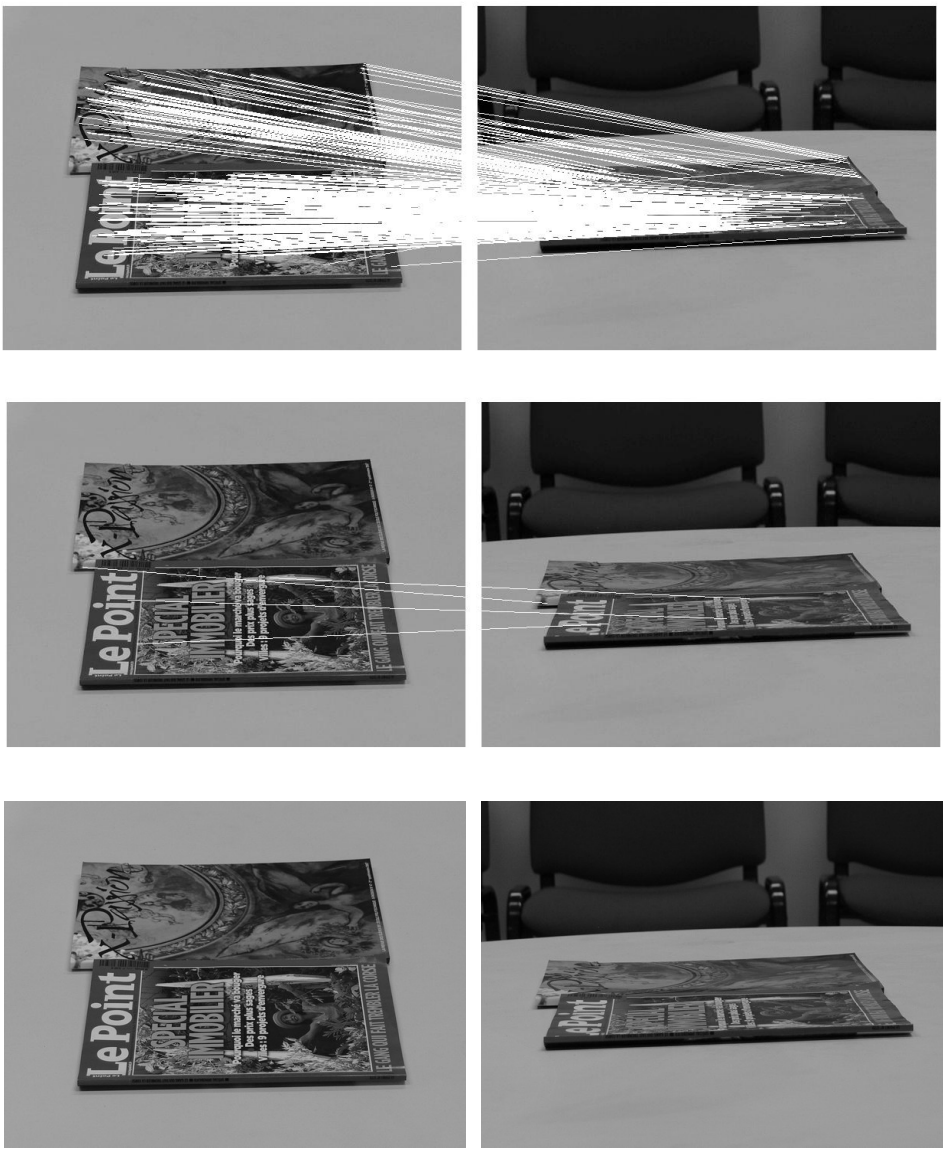


Figure 27: Correspondences between the images Magazine 1 and Magazine 3. Absolute tilts $t = 2.1$ (left), $t' = 6.0$ (right). transition tilt: $\tau = 2.9$. Top: A-SIFT finds 338 correspondences, out of which 2 are false. Middle: SIFT finds 5 correspondences. Bottom: MSER finds 3 false correspondences in total that have been rejected. MSER fails because the shapes, submitted to high absolute tilts, mix together. The "Maximally Stable Extremal Regions" are no more stable under such tilts.

Figure 28: Correspondences between the images Bureau 1 and Bureau 8. Transition tilt: $\tau \approx 3$. Top: A-SIFT finds 62 correspondences, out of which 4 are false. Middle: SIFT finds 0 correspondences. Bottom: MSER finds 13 correspondences, out of which 2 are correct.

Figure 29: Correspondences between the images Facade 1 and Facade 8. Absolute (and transition) tilt $t = 3.8$ ($\theta = 74.7°$) Top: A-SIFT finds 71 correspondences, out of which 4 are false. Middle: SIFT finds 1 correspondence. Bottom: MSER finds 34 correspondences, out of which only 2 are correct.

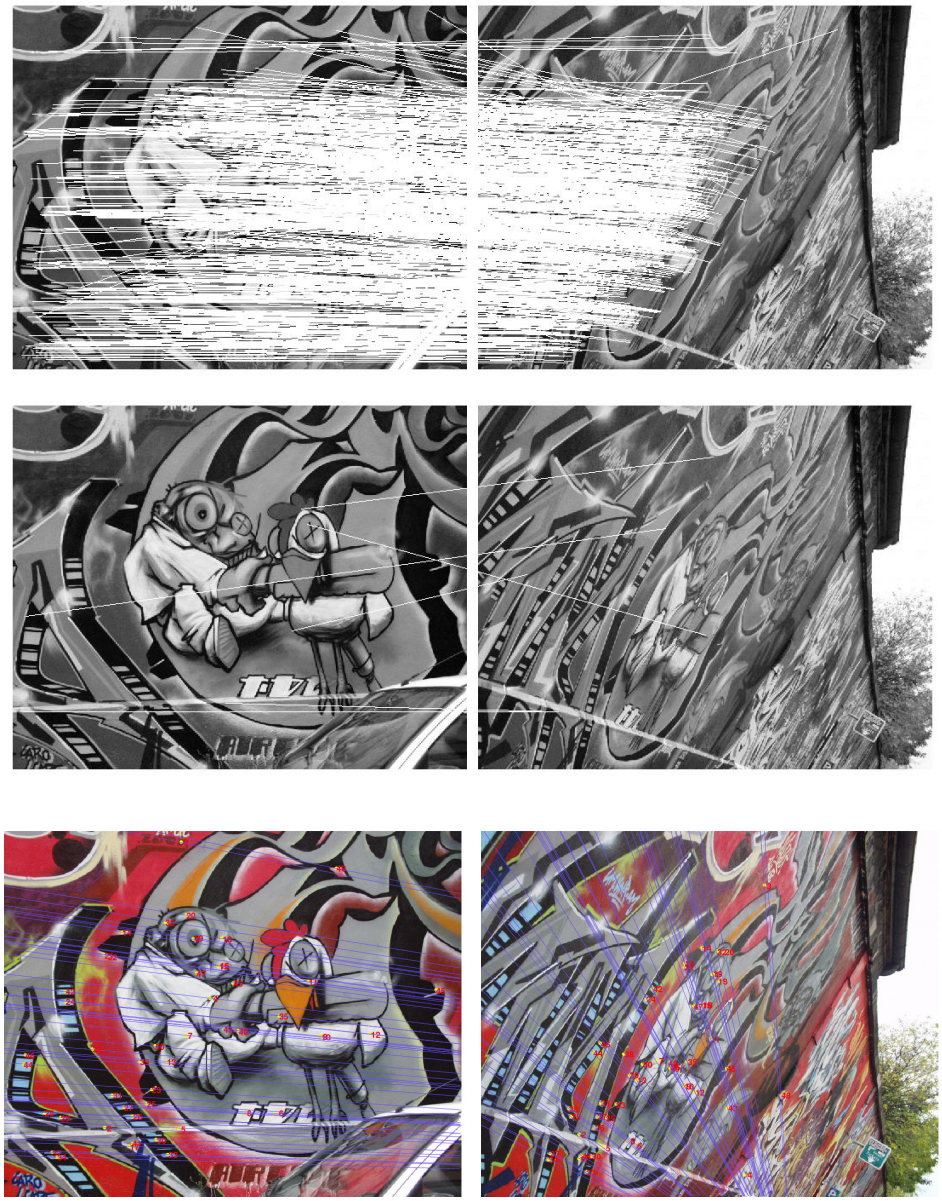


Figure 30: Correspondences between the images Graffiti 1 and Graffiti 6. Transition tilt: $\tau \approx 3.2$. Top: A-SIFT finds 724 correspondences, out of which 3 are false. Middle: SIFT finds 6 correspondence. Bottom: MSER finds 127 correspondences, out of which 50 are correct. In this example, MSER is successful. But MSER fails to work under big image zoom or if the shapes in the images are not well contrasted.

Figure 31: Image matching (images proposed by Matas et al [34]). 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.

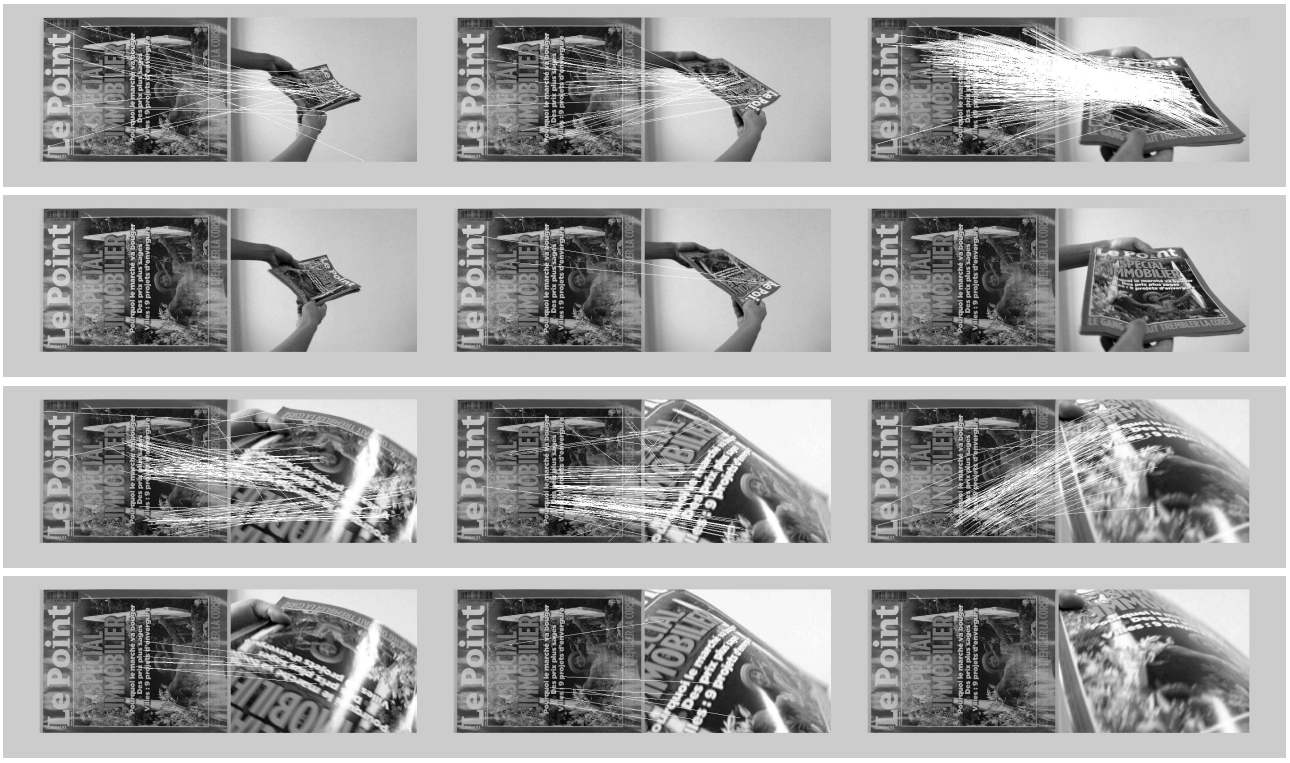


Figure 32: Video matching: one looks for correspondences between a query image and video frames. First and third row: A-SIFT succeeds. (23, 34, 412, 94, 107, 89). Second and fourth row: SIFT fails.

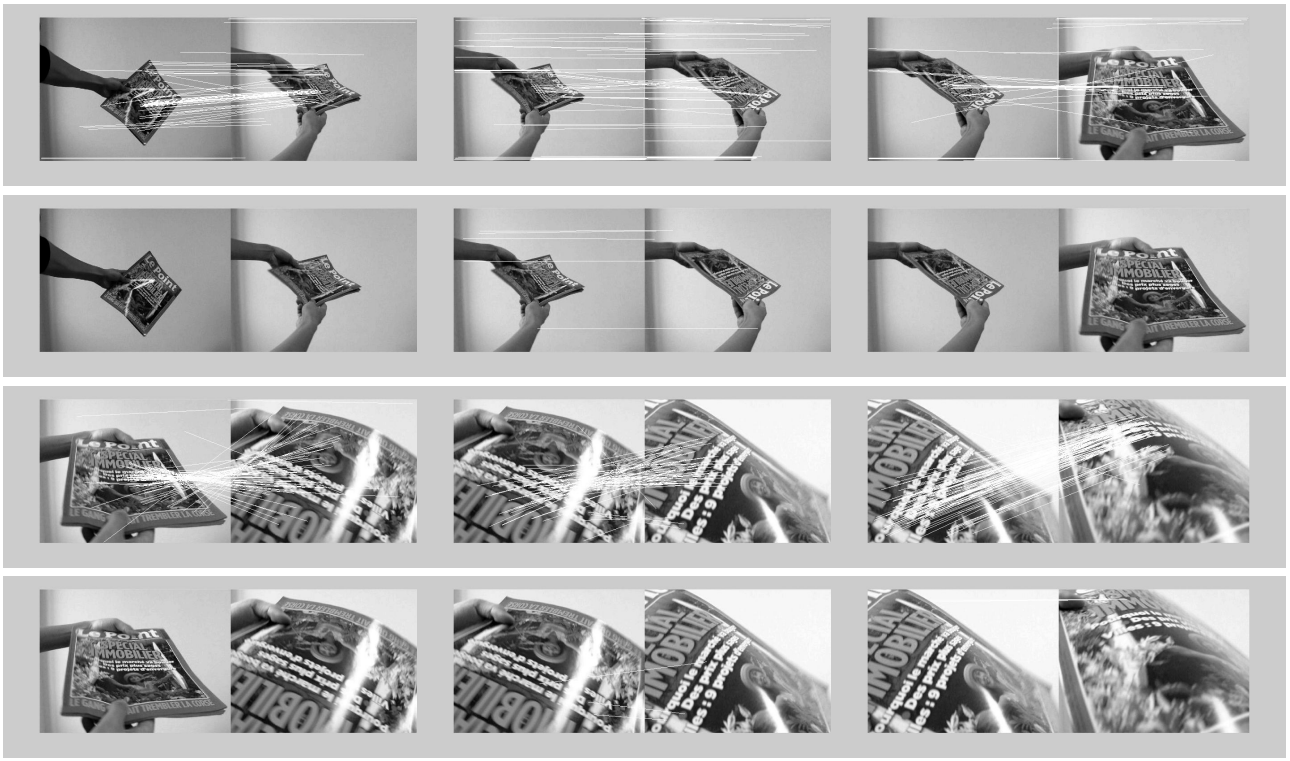


Figure 33: Object tracking: one looks for correspondences between video frames. First and third row: A-SIFT succeeds. Second and fourth row: SIFT fails.

points in $\mathbf{u}(x, y)$. Fig. 34 shows some examples of bilateral symmetry detection. A-SIFT, invariant to view point change, results in significantly better symmetry detection in perspective with respect to SIFT.

Figure 34: Symmetry detection with A-SIFT by finding reflective correspondences. SIFT fails to detect symmetry on these images where objects are not in frontal position.

8 Key notes

8.1 Maximally Stable Extremal Regions (MSER)

The MSER method introduced by Matas et al. [33] achieves the affine invariance by selecting the most robust connected components of upper and lower level sets as image features.

Extremal regions is the name given by the authors to the connected components of upper or lower level sets. Maximally stable extremal regions, or MSERs, are defined as maximally contrasted regions in the following way. let $Q_1, ..., Q_{i-1}, Q_i, ...$ be a sequence of nested extremal regions, i.e. $Q_i \subset Q_{i+1}$ where Q_i is defined by a threshold at level i or, in other terms, Q_i is an upper (resp. lower) level set at level i. An extremal region in the list Q_{i_0} is said to be maximally stable if the area variation $q(i) =: |Q_{i+1} \setminus Q_{i-1}|/|Q_i|$ has a local minimum at i_0, where $|Q|$ denotes the area of a region $|Q|$. Clearly the above measure is a measure of contrast along the boundary ∂Q_i of Q_i. Indeed, assuming that u is C^1 and that the grey level increment between i and $i+1$ is infinitesimal, the area $|Q_{i+1} \setminus Q_{i-1}|$ varies least when $\int_{\partial Q_i} |\nabla u|$ is maximal. It is a straightforward consequence of their definition that the MSERs possess the following robustness and invariance properties:

- invariance to every affine transformation of image intensities;
- covariance to all special affine transforms of the image domain if their "tilt" is not too large (otherwise, the contrast on the regions boundary is affected, or the region's boundaries mix);
- stability, since only extremal regions whose support is virtually unchanged over a range of thresholds are selected;

The MSER extraction is a first step of image matching. Once MSERs are computed, an affine normalization is performed on the MSERs before they can be compared. The affine normalization proposed in [33] is based on moment methods. Affine invariance up to a rotation is achieved by diagonalizing the covariance matrix and then applying the linear transform that performs its diagonalization to each region. Rotational invariants are then applied over the normalized region. This procedure is affine invariant and yields potential candidates to a match. [33] use invariant descriptions only as a preliminary test. The final check in the original method is made by using correlation. The normalized circular regions are correlated (for all relative rotations).

8.2 Scale Invariant Features (SIFs) and other descriptors

SIFT interest points (or SIFs) are obtained as the maxima of the Laplacian of the image (approximated by a difference of Gaussians) through a Gaussian pyramid. Many variations exist on the computation of interest points, following the pioneering work of Harris and Stephens [20]. In particular, recent methods are affine invariant. In [40], an overview and a comparison between the main affine invariant region detectors is presented. One of the conclusions is that no method dramatically outperforms all the other ones, although the highest score is obtained by the MSER detector [33].

SIFT descriptors, or SIFs, are basically local histograms of the gradient direction, weighted by the gradient norm, in the vicinity of the key point. These histograms are invariant to rotations of the image domain and thresholding and normalization of image gradients is used in order to achieve some invariance to illumination changes. In the recent years, several other local descriptors have been proposed, incorporating further invariance to changes in viewing conditions. In particular, MSER [33] uses moment invariants to describe the vicinity of the interest points. This approach was also used by Monasse in [41]. A recent paper [39] aims at comparing the different descriptors. Performance is evaluated by examining the so-called ROC curves plotting the number of false positive detections as a function of false negative detections. While on one of the methods, the gradient location and orientation histograms (GLOH [39]) seems slightly better than the other ones, the difference (in particular with SIFT) is not that large. Let us remark that there are two ways to achieve geometrical invariance: either descriptors are computed in invariant regions, or they have a group invariance by themselves. For instance, in [6], skew and stretch are corrected in the neighborhood computation. An affine contrast change is first applied. Then, descriptors are rotation invariant gray level moments.

8.3 Matching and Grouping

Simple procedures, such as the thresholding of ratios between the best and second best matches in SIFT, have been used in SIFT and A-SIFT. MSER [33] uses a voting procedure over the nearest measurements comparing a set of invariants that form the descriptors. In [12, 53], some improvements on the SIFT descriptors definition and on the matching step have been proposed, based on *a contrario* techniques. The use of a grouping step improves the matching results. In SIFT, a Hough transform procedure [4] is proposed, but other methods [58] use greedy procedures based on RANSAC [15].

8.4 Appendix: Scale and SIFT: consistency of the method

We denote by $\mathcal{T}$ an arbitrary translation, by $\mathbf{R}$ an arbitrary rotation, by $\mathbf{H}$ an arbitrary homothety, and by $\mathbf{G}$ an arbitrary gaussian convolution, all applied to continuous images.

In the particular case of the digital image formation model (6) where $\mathbf{A}$ is a frontal view of $\mathbf{u}_0$, $\mathbf{A} = \mathbf{HR}\mathcal{T}$ is the composition of a translation $\mathcal{T}$, a homothety $\mathbf{H}$, and a rotation $\mathbf{R}$. Thus the digital image is $u = \mathbf{S}_1\mathbf{G}_1\mathbf{H}\mathcal{T}\mathbf{R}\mathbf{u}_0$, for some $\mathbf{H}$, $\mathcal{T}$, $\mathbf{R}$ as above. The following lemma is easily proven for the SIFT method (see [44]).

Lemma 3. *For any rotation $\mathbf{R}$ and any translation $\mathcal{T}$, the SIFT descriptors of $\mathbf{S}_1\mathbf{G}_\delta\mathbf{HR}\mathcal{T}\mathbf{u}_0$ are identical to those of $\mathbf{S}_1\mathbf{G}_\delta\mathbf{H}\mathbf{u}_0$.*

By Lemma 3 the only involved invariance claimed by the SIFT method, is the scale invariance. The SIFT method deals with the space extrema of the Laplacian of the scale space of $\mathbf{u}_0$, $\mathbf{u}(\sigma, \mathbf{x}) := (\mathbf{G}_\sigma\mathbf{u}_0)(\mathbf{x})$.

Proposition 1. *Let u and v be two digital images that are frontal snapshots of the same continuous flat image $\mathbf{u}_0$, $u = \mathbf{S}_1\mathbf{G}_\beta\mathbf{H}_\lambda\mathbf{u}_0$ and $v =: \mathbf{S}_1\mathbf{G}_\delta\mathbf{H}_\mu\mathbf{u}_0$, taken at different distances, with different gaussian blurs and possibly different sampling rates. Let $\mathbf{w}(\sigma, \mathbf{x}) = (\mathbf{G}_\sigma\mathbf{u}_0)(\mathbf{x})$ denote the scale space of $\mathbf{u}_0$. Then the scale spaces of $\mathbf{u}$ and $\mathbf{v}$ are*

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

If $(s_0, \mathbf{x}_0)$ is a key point of $\mathbf{w}$ satisfying $s_0 \geq \max(\lambda\beta, \mu\delta)$, then it corresponds to a key point of $\mathbf{u}$ at the scale σ_1 such that $\lambda\sqrt{\sigma_1^2 + \beta^2} = s_0$, whose SIFT descriptor is sampled with mesh σ_1. 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 σ_2.

Proposition 2. *Let u and v be two digital images obtained by frontal snapshots of a planar surface. Then u and v have the same SIFT descriptors if and only if one of them is obtained by over-sampling the other one, which also means that their blurs are identical.*

Theorem 3. *Let u and v be two digital images that are frontal snapshots of the same continuous flat image $\mathbf{u}_0$, $u = \mathbf{S}_1\mathbf{G}_\beta\mathbf{H}_\lambda\mathbf{u}_0$ and $v =: \mathbf{S}_1\mathbf{G}_\delta\mathbf{H}_\mu\mathbf{u}_0$, taken at different distances, with different gaussian blurs and possibly different sampling rates. If $\lambda\beta \neq \mu\delta$, the SIFT descriptors of u and v are actually different. Indeed, the key points are in good correspondence, but their associated scales should satisfy $\lambda\sigma_1 = \mu\sigma_2$ and satisfy instead the scale approximate relation*

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

Thus, if $\sigma_1 \gg \beta$ and $\sigma_2 \gg \delta$, then the SIFT descriptors of u and v are similar if not identical.

References

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

[2] A. Anjulan and N. Canagarajah. Affine Invariant Feature Extraction Using Symmetry. *Advanced Concepts for Intelligent Vision Systems: 7th International Conference, ACIVS 2005, Antwerp, Belgium, September 20-23, 2005: Proceedings*, 2005.

[3] AP Ashbrook, NA Thacker, PI Rockett, and CI Brown. Robust recognition of scaled shapes using pairwise geometric histograms. *Proc. BMVC*, pages 503–512, 1995.

[4] DH Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.

[5] A. Banno, K. Hasegawa, and K. Ikeuchi. Motion estimation of a moving range sensor by image sequences and distorted range data. *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 1618–1623, 2005.

[6] A. Baumberg. Reliable feature matching across widely separated views. *Proc. IEEE CVPR*, 1:774–781, 2000.

[7] S. Belongie, J. Malik, and J. Puzicha. Shape Matching and Object Recognition Using Shape Contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002.

[8] M. Bennewitz, C. Stachniss, W. Burgard, and S. Behnke. Metric Localization with Scale-Invariant Visual Features Using a Single Perspective Camera. *European Robotics Symposium 2006*, 2006.

[9] M. Brown and D.G. Lowe. Recognising panoramas. *Proc. ICCV*, 1(2):3, 2003.

[10] E.Y. Chang. EXTENT: fusing context, content, and semantic ontology for photo annotation. *Proceedings of the 2nd international workshop on Computer vision meets databases*, pages 5–11, 2005.

[11] H. Cornelius and G. Loy. Detecting Bilateral Symmetry in Perspective. *5th Workshop on Perceptual Organization in Computer Vision*, pages 191–198, 2006.

[12] A. Desolneux, L. Moisan, and J.-M. Morel. *Gestalt Theory and Image Analysis, a Probabilistic Approach*. Interdisciplinary Applied Mathematics series, Springer Verlag, 2007. Preprint available at http://www.cmla.ens-cachan.fr/Utilisateurs/morel/lecturenote.pdf.

[13] Q. Fan, K. Barnard, A. Amir, A. Efrat, and M. Lin. Matching slides to presentation videos using SIFT and scene background matching. *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 239–248, 2006.

[14] O. Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. Mit Press, 1993.

[15] M.A. Fischler and R.C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[16] J.J. Foo and R. Sinha. Pruning SIFT for scalable near-duplicate image matching. *Proceedings of the eighteenth conference on Australasian database-Volume 63*, pages 63–71, 2007.

[17] G. Fritz, C. Seifert, M. Kumar, and L. Paletta. Building detection from mobile imagery using informative SIFT descriptors. *Lecture notes in computer science*, pages 629–638.

[18] I. Gordon and D.G. Lowe. What and Where: 3D Object Recognition with Accurate Pose. *Lecture Notes in Computer Science*, 4170:67, 2006.

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

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

[21] Aniket Murarka Joseph. Building local safety maps for a wheelchair robot using vision and lasers.

[22] Mikolajczyk K and Schmid C. A Performance Evaluation of Local Descriptors. In *International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 257–263, June 2003.

[23] T. Kadir, A. Zisserman, and M. Brady. An Affine Invariant Salient Region Detector. In *European Conference on Computer Vision*, pages 228–241, 2004.

[24] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. *Proc. CVPR*, 2:506–513, 2004.

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

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

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

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

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

[30] D.G. Lowe. SIFT Keypoint Detector: online demo http://www.cs.ubc.ca/∼lowe/keypoints/.

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

[32] G. Loy and J.O. Eklundh. Detecting symmetry and symmetric constellations of features. *Proceedings of ECCV*, 2:508–521, 2006.

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

[34] J. Matas, O. Chum, M. Urban, and T.g Pajdla. Wbs image matcher: online demo http://cmp.felk.cvut.cz/∼wbsdemo/demo/.

[35] K Mikolajczyk. http://www.robots.ox.ac.uk/∼vgg/research/affine/.

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

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

[38] K. Mikolajczyk and C. Schmid. Scale and Affine Invariant Interest Point Detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[39] K. Mikolajczyk and C. Schmid. A Performance Evaluation of Local Descriptors. *IEEE Trans. PAMI*, pages 1615–1630, 2005.

[40] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L.V. Gool. A Comparison of Affine Region Detectors. *International Journal of Computer Vision*, 65(1):43–72, 2005.

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

[42] P. Moreels and P. Perona. Common-frame model for object recognition. *Proc. NIPS*, 2004.

[43] P. Moreels and P. Perona. Evaluation of Features Detectors and Descriptors based on 3D Objects. *International Journal of Computer Vision*, 73(3):263–284, 2007.

[44] J.M. Morel and G. Yu. On the consistency of the SIFT method. Technical Report Prepublication, CMLA, ENS Cachan, 2008.

[45] D. Mukherjee, A. Zisserman, and J. Brady. Shape from symmetry—detecting and exploiting symmetry in affine images. *Philosophical Transactions: Physical Sciences and Engineering*, 351(1695):77–106, 1995.

[46] P. Musé, F. Sur, F. Cao, and Y. Gousseau. Unsupervised thresholds for shape matching. *Image Processing, 2003. Proceedings. 2003 International Conference on*, 2, 2003.

[47] P. Musé, F. Sur, F. Cao, Y. Gousseau, and J.M. Morel. An A Contrario Decision Method for Shape Element Recognition. *International Journal of Computer Vision*, 69(3):295–315, 2006.

[48] P. Musé, F. Sur, F. Cao, J.L. Lisani, and J.M. Morel. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. Mit Press, 2007.

[49] A. Negre, H. Tran, N. Gourier, D. Hall, A. Lux, and JL Crowley. Comparative study of People Detection in Surveillance Scenes. *Structural, Syntactic and Statistical Pattern Recognition, Proceedings Lecture Notes in Computer Science*, 4109:100–108, 2006.

[50] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. *Proc. CVPR*, pages 2161–2168, 2006.

[51] D. Pritchard and W. Heidrich. Cloth Motion Capture. *Computer Graphics Forum*, 22(3):263–271, 2003.

[52] A. Amir A. Efrat Q. Fan, K. Barnard and M. Lin. Matching slides to presentation videos using SIFT and scene background matching. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 239–248, New York, NY, USA, 2006. ACM.

[53] J. Rabin, Y. Gousseau, and J. Delon. A contrario matching of local descriptors. Technical Report hal-00168285, Ecole Nationale Supérieure des Télécommunications, Paris, France, 2007.

[54] F. Riggi, M. Toews, and T. Arbel. Fundamental Matrix Estimation via TIP-Transfer of Invariant Parameters. *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)-Volume 02*, pages 21–24, 2006.

[55] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. Segmenting, Modeling, and Matching Video Clips Containing Multiple Moving Objects. *IEEE Trans. PAMI*, pages 477–491, 2007.

[56] J. Ruiz-del Solar, P. Loncomilla, and C. Devia. A New Approach for Fingerprint Verification Based on Wide Baseline Matching Using Local Interest Points and Descriptors. *Lecture Notes in Computer Science*, 4872:586, 2007.

[57] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or How do I organize my holiday snaps?. *Proc. ECCV*, 1:414–431, 2002.

[58] C. Schmid, G. Dorko, S. Lazebnik, K. Mikolajczyk, and J. Ponce. Pattern recognition with local invariant features. *Handbook of Pattern Recognition and Computer Vision, World Scientific*, 3.

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

[60] S. Se, D. Lowe, and J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, 2, 2001.

[61] N. Snavely, S.M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Transactions on Graphics (TOG)*, 25(3):835–846, 2006.

[62] Marc Pollefeys Sudipta N Sinha, Jan-Michael Frahm and Yakup Genc. Gpu-based video feature tracking and matching. *EDGE 2006, workshop on Edge Computing Using New Commodity Architectures, Chapel Hill*, 2006.

[63] T. Tuytelaars and L. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. *British Machine Vision Conference*, pages 412–425, 2000.

[64] T. Tuytelaars and L. Van Gool. Matching Widely Separated Views Based on Affine Invariant Regions. *International Journal of Computer Vision*, 59(1):61–85, 2004.

[65] T. Tuytelaars, L. Van Gool, et al. Content-based image retrieval based on local affinely invariant regions. *Int. Conf. on Visual Information Systems*, pages 493–500, 1999.

[66] L. Vacchetti, V. Lepetit, and P. Fua. Stable Real-Time 3D Tracking Using Online and Offline Information. *IEEE Trans PAMI*, pages 1385–1391, 2004.

[67] L.J. Van Gool, T. Moons, and D. Ungureanu. Affine/Photometric Invariants for Planar Intensity Patterns. *Proceedings of the 4th European Conference on Computer Vision-Volume I-Volume I*, pages 642–651, 1996.

[68] M. Veloso, F. von Hundelshausen, and PE Rybski. Learning visual object definitions by observing human activities. *Humanoid Robots, 2005 5th IEEE-RAS International Conference on*, pages 148–153, 2005.

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

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

[71] G. Yang, CV Stewart, M. Sofka, and CL Tsai. Alignment of challenging image pairs: Refinement and region growing starting from a single keypoint correspondence. *IEEE Trans. Pattern Anal. Machine Intell*, 2007.

[72] J. Yao and W.K. Cham. Robust multi-view feature matching from multiple unordered views. *Pattern Recognition*, 40(11):3081–3099, 2007.

[73] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys (CSUR)*, 38(4), 2006.