SIAM J. IMAGING SCIENCES
Vol. 2, No. 2, pp. 438–469

ASIFT: A New Framework for Fully Affine Invariant Image Comparison*

Jean-Michel Morel† and Guoshen Yu‡

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. Even scale invariance is dealt with rigorously only by the scale-invariant feature transform (SIFT) method. By simulating zooms out and normalizing translation and rotation, SIFT is invariant to four out of the six parameters of an affine transform. The method proposed in this paper, affine-SIFT (ASIFT), simulates all image views obtainable by varying the two camera axis orientation parameters, namely, the latitude and the longitude angles, left over by the SIFT method. Then it covers the other four parameters by using the SIFT method itself. The resulting method will be mathematically proved to be fully affine invariant. Against any prognosis, simulating all views depending on the two camera orientation parameters is feasible with no dramatic computational load. A two-resolution scheme further reduces the ASIFT complexity to about twice that of SIFT. A new notion, the *transition tilt*, measuring the amount of distortion from one view to another, is introduced. While an *absolute tilt* from a frontal to a slanted view exceeding 6 is rare, much higher transition tilts are common when two slanted views of an object are compared (see Figure 1). The attainable transition tilt is measured for each affine image comparison method. The new method permits one to reliably identify features that have undergone transition tilts of large magnitude, up to 36 and higher. This fact is substantiated by many experiments which show that ASIFT significantly outperforms the state-of-the-art methods SIFT, maximally stable extremal region (MSER), Harris-affine, and Hessian-affine.

Key words. image matching, descriptors, affine invariance, scale invariance, affine normalization, scale-invariant feature transform (SIFT)

AMS subject classifications. 68T10, 68T40, 68T45, 93C85

DOI. 10.1137/080732730

1. Introduction.

Image matching aims at establishing correspondences between similar objects that appear in different images. This is a fundamental step in many computer vision and image processing applications such as image recognition, three-dimensional (3D) reconstruction, object tracking, robot localization, and image registration [11].

The general (solid) shape matching problem starts with several photographs of a physical object, possibly taken with different cameras and viewpoints. These digital images are the *query* images. Given other digital images, the *search* images, the question is whether or not some of them contain a view of the object taken in the query image. This problem is by far

*Received by the editors August 14, 2008; accepted for publication (in revised form) December 16, 2008; published electronically April 22, 2009.
http://www.siam.org/journals/siims/2-2/73273.html
†CMLA, ENS Cachan, 61 avenue du President Wilson, 94235 Cachan Cedex, France (Jean-Michel.Morel@cmla.ens-cachan.fr).
‡CMAP, Ecole Polytechnique, 91128 Palaiseau Cedex, France (yu@cmap.polytechnique.fr).

more restrictive than the *categorization* problem, where the question is to recognize a *class* of objects, like chairs or cats. In the shape matching framework several instances of the very *same* object, or of copies of this object, are to be recognized. The difficulty is that the change of camera position induces an apparent deformation of the object image. Thus, recognition must be invariant with respect to such deformations.

The state-of-the-art image matching algorithms usually consist of two parts: *detector* and *descriptor*. They first detect points of interest in the compared images and select a region around each point of interest, and then associate an invariant descriptor or feature to each region. Correspondences may thus be established by matching the descriptors. Detectors and descriptors should be as invariant as possible.

In recent years local image detectors have bloomed. They can be classified by their incremental invariance properties. All of them are translation invariant. The Harris point detector [17] is also rotation invariant. The Harris–Laplace, Hessian–Laplace, and difference-of-Gaussian (DoG) region detectors [34, 37, 29, 12] are invariant to rotations and changes of scale. Some moment-based region detectors [24, 3] including the Harris-affine and Hessian-affine region detectors [35, 37], an edge-based region detector [58, 57], an intensity-based region detector [56, 57], an entropy-based region detector [18], and two level line-based region detectors, MSER ("maximally stable extremal region") [32] and LLD ("level line descriptor") [44, 45, 8], are designed to be invariant to affine transforms. MSER, in particular, has been demonstrated to have often better performance than other affine invariant detectors, followed by Hessian-affine and Harris-affine [39].

In his milestone paper [29], Lowe has proposed a scale-invariant feature transform (SIFT) that is invariant to image scaling and rotation and partially invariant to illumination and viewpoint changes. The SIFT method combines the DoG region detector that is rotation, translation, and scale invariant (a mathematical proof of its scale invariance is given in [42]) with a descriptor based on the gradient orientation distribution in the region, which is partially illumination and viewpoint invariant [29]. These two stages of the SIFT method will be called, respectively, *SIFT detector* and *SIFT descriptor*. The SIFT detector is a priori less invariant to affine transforms than the Hessian-affine and the Harris-affine detectors [34, 37]. However, when combined with the SIFT descriptor [39], its overall affine invariance turns out to be comparable, as we shall see in many experiments.

The SIFT descriptor has been shown to be superior to many other descriptors [36, 38] such as the distribution-based shape context [5], the geometric histogram [2] descriptors, the derivative-based complex filters [3, 51], and the moment invariants [60]. A number of SIFT descriptor variants and extensions, including PCA (principal components analysis)-SIFT [19], GLOH (gradient location-orientation histogram) [38], and SURF (speeded-up robust features) [4], have been developed since [13, 22]. They claim more robustness and distinctiveness with scaled-down complexity. The SIFT method and its variants have been popularly applied for scene recognition [10, 40, 50, 61, 15, 52, 65, 41] and detection [14, 46], robot localization [6, 53, 47, 43], image registration [64], image retrieval [16], motion tracking [59, 20], 3D modeling and reconstruction [49, 62], building panoramas [1, 7], photo management [63, 21, 55, 9], and symmetry detection [30].

The mentioned state-of-the-art methods have achieved brilliant success. However, none of them is fully affine invariant. As pointed out in [29], Harris-affine and Hessian-affine

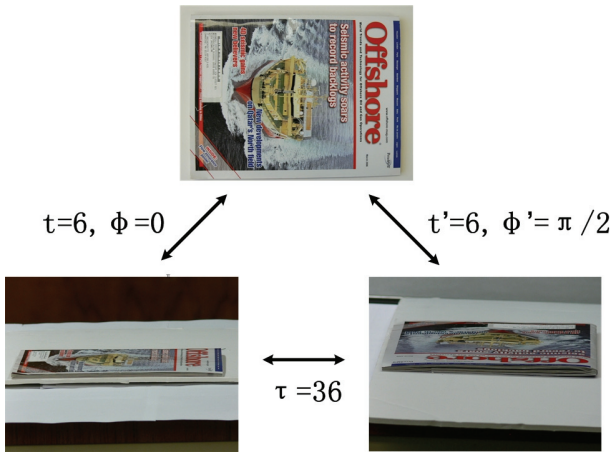

Figure 1. *The frontal image (above) is squeezed in one direction on the left image by a slanted view, and squeezed in an orthogonal direction by another slanted view. The compression factor or absolute tilt is about* 6 *in each view. The resulting compression factor, or transition tilt from left to right, is actually* 36. *See section* 2 *for the formal definition of these tilts. Transition tilts quantify the affine distortion. The aim is to detect image similarity under transition tilts as large as this one.*

start with initial feature scales and locations selected in a nonaffine invariant manner. The noncommutation between optical blur and affine transforms shown in section 3 also explains the limited affine invariance performance of the normalization methods MSER, LLD, Harris-affine, and Hessian-affine. As shown in [8], MSER and LLD are not even fully scale invariant: they do not cope with the drastic changes of the level line geometry due to blur. SIFT is actually the only method that is fully scale invariant. However, since it is not designed to cover the whole affine space, its performance drops quickly under substantial viewpoint changes.

The present paper proposes an affine invariant extension of SIFT (ASIFT) that is fully affine invariant. Unlike MSER, LLD, Harris-affine, and Hessian-affine which normalize all six affine parameters, ASIFT simulates three parameters and normalizes the rest. The scale and the changes of the camera axis orientation are the three simulated parameters. The other three, rotation and translation, are normalized. More specifically, ASIFT simulates the two camera axis parameters and then applies SIFT which simulates the scale and normalizes the rotation and the translation. A two-resolution implementation of ASIFT will be proposed that has about twice the complexity of a single SIFT routine. To the best of our knowledge the first work suggesting simulation of affine parameters appeared in [48], where the authors proposed simulating two tilt and two shear deformations in a cloth motion capture application.

The paper introduces a crucial parameter for evaluating the performance of affine recognition, the *transition tilt*. The transition tilt measures the degree of viewpoint change from one view to another. Figures 1 and 2 give a first intuitive approach to *absolute tilt* and *transition tilt*. They illustrate why simulating large tilts on both compared images proves necessary to obtain a fully affine invariant recognition. Indeed, transition tilts can be much larger than absolute tilts. In fact they can behave like the square of absolute tilts. The affine invariance performance of the state-of-the-art methods will be evaluated by their attainable transition tilts.

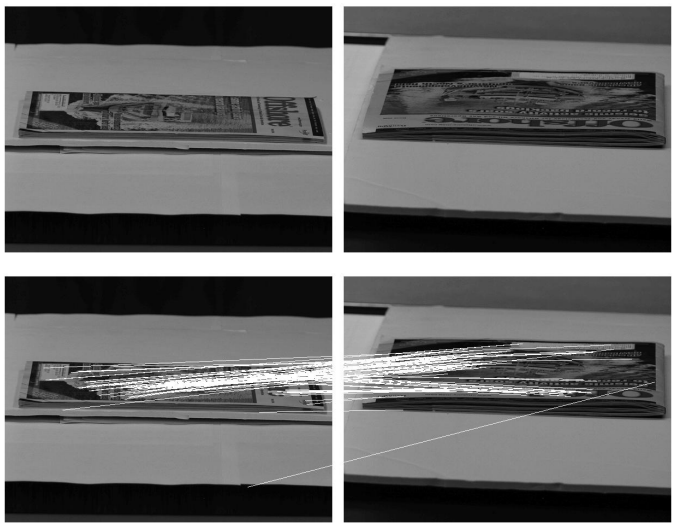

Figure 2. *Top: image pair with transition tilt* $t \approx 36$. *(SIFT, Harris-affine, Hessian-affine, and MSER fail completely.) Bottom: ASIFT finds* 120 *matches out of which* 4 *are false. See comments in text.*

The paper is organized as follows. Section 2 describes the affine camera model and introduces the transition tilt. Section 3 reviews the state-of-the-art image matching method SIFT, MSER, Harris-affine, and Hessian-affine and explains why they are not fully affine invariant. The ASIFT algorithm is described in section 4. Section 5 gives a mathematical proof that ASIFT is fully affine invariant, up to sampling approximations. Section 6 is devoted to extensive experiments where ASIFT is compared with the state-of-the art algorithms. Section 7 is the conclusion.

A website with an online demo is available at http://www.cmap.polytechnique.fr/~yu/research/ASIFT/demo.html. It allows the users to test ASIFT with their own images. It also contains an image dataset (for systematic evaluation of robustness to absolute and transition tilts) and more examples.

2. Affine camera model and tilts. As illustrated by the camera model in Figure 3, digital image acquisition of a flat object can be described as

$$\mathbf{u} = \mathbf{S}_1 \mathrm{G}_1 A \mathcal{T} u_0, \tag{2.1}$$

where $\mathbf{u}$ is a digital image and u_0 is an (ideal) infinite resolution frontal view of the flat object. $\mathcal{T}$ and A are, respectively, a plane translation and a planar projective map due to the camera motion. G_1 is a Gaussian convolution modeling the optical blur, and $\mathbf{S}_1$ is the standard sampling operator on a regular grid with mesh 1. The Gaussian kernel is assumed to be broad enough to ensure no aliasing by the 1-sampling, namely, $I\mathbf{S}_1\mathrm{G}_1 A\mathcal{T}u_0 = \mathrm{G}_1 A\mathcal{T}u_0$, where I denotes the Shannon–Whittaker interpolation operator. A major difficulty of the recognition problem is that the Gaussian convolution G_1, which becomes a broad convolution kernel when the image is zoomed out, does not commute with the planar projective map A.

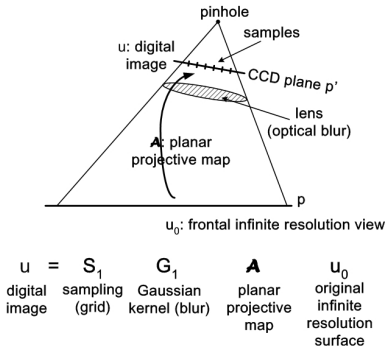

Figure 3. *The projective camera model* $\mathbf{u} = \mathbf{S}_1 \mathrm{G}_1 \mathcal{A} u_0$. $\mathcal{A}$ *is a planar projective transform (a homography).* G_1 *is an antialiasing Gaussian filtering.* $\mathbf{S}_1$ *is the CCD sampling.*

Figure 4. *The global deformation of the ground is strongly projective (a rectangle becomes a trapezoid), but the local deformation is affine: Each tile on the pavement is almost a parallelogram.*

2.1. The affine camera model. We shall proceed to a further simplification of the above model by reducing A to an affine map. Figure 4 shows one of the first perspectively correct Renaissance paintings by Paolo Uccello. The perspective on the ground is strongly projective: The rectangular pavement of the room becomes a trapezoid. However, each tile on the pavement is almost a parallelogram. This illustrates the local tangency of perspective deformations to affine maps. Indeed, by the first order Taylor formula, any planar smooth deformation can be approximated around each point by an affine map. The apparent deformation of a plane object induced by a camera motion is a planar homographic transform, which is smooth and therefore locally tangent to affine transforms. More generally, a solid object's apparent deformation arising from a change in the camera position can be locally modeled by affine planar transforms, provided that the object's facets are smooth. In short, all local perspective effects can be modeled by local affine transforms $u(x, y) \to u(ax + by + e, cx + dy + f)$ in each image region.

Figure 5 illustrates the same fact by interpreting the local behavior of a camera as equivalent to multiple cameras at infinity. These cameras at infinity generate affine deformations. In fact, a camera position change can generate any affine map with positive determinant. The next theorem formalizes this fact and gives a camera motion interpretation to affine deformations.

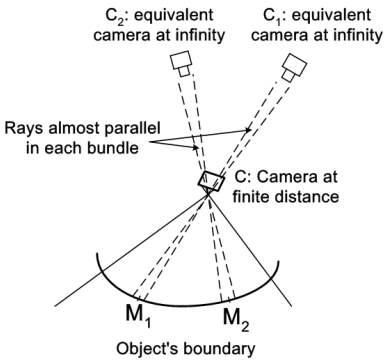

Figure 5. *A camera at finite distance looking at a smooth object is equivalent to multiple local cameras at infinity. These cameras at infinity generate affine deformations.*

Theorem 2.1. *Any affine map* $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ *with strictly positive determinant which is not a similarity has a unique decomposition*

$$A = H_\lambda R_1(\psi) T_t R_2(\phi) = \lambda \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix} \begin{bmatrix} t & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}, \tag{2.2}$$

where $\lambda > 0$, λt *is the determinant of* A, R_i *are rotations,* $\phi \in [0, \pi)$, *and* T_t *is a tilt, namely, a diagonal matrix with first eigenvalue* $t > 1$ *and second eigenvalue equal to* 1.

The theorem follows the singular value decomposition (SVD) principle. The proof is given in the appendix.

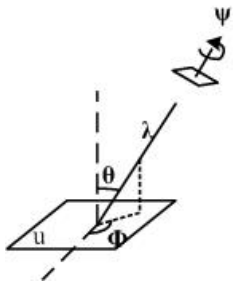

Figure 6. *Geometric interpretation of the decomposition* (2.2). *The image* u *is a flat physical object. The small parallelogram at the top right represents a camera looking at* u. *The angles* ϕ *and* θ *are, respectively, the camera optical axis longitude and latitude. A third angle* ψ *parameterizes the camera spin, and* λ *is a zoom parameter.*

Figure 6 shows a camera motion interpretation of the affine decomposition (2.2): ϕ and $\theta = \arccos 1/t$ are the viewpoint angles, ψ parameterizes the camera spin, and λ corresponds to the zoom. The camera is assumed to stay far away from the image and starts from a frontal view u, i.e., $\lambda = 1$, $t = 1$, $\phi = \psi = 0$. The camera can first move parallel to the object's plane: This motion induces a translation $\mathcal{T}$ that is eliminated by assuming (without loss of generality) that the camera axis meets the image plane at a fixed point. The plane containing the normal and the optical axis makes an angle ϕ with a fixed vertical plane. This angle is

called *longitude*. Its optical axis then makes a θ angle with the normal to the image plane u. This parameter is called *latitude*. Both parameters are classical coordinates on the *observation hemisphere*. The camera can rotate around its optical axis (rotation parameter ψ). Last but not least, the camera can move forward or backward, as measured by the zoom parameter λ.

In (2.2) the tilt parameter, which has a one-to-one relation to the latitude angle $t = 1/\cos\theta$, entails a strong image deformation. It causes a directional subsampling of the frontal image in the direction given by the longitude ϕ.

2.2. Transition tilts. The parameter t in (2.2) is called *absolute tilt*, since it measures the tilt between the *frontal* view and a *slanted* view. In real applications, both compared images are usually slanted views. The *transition tilt* is designed to quantify the amount of tilt between two such images.

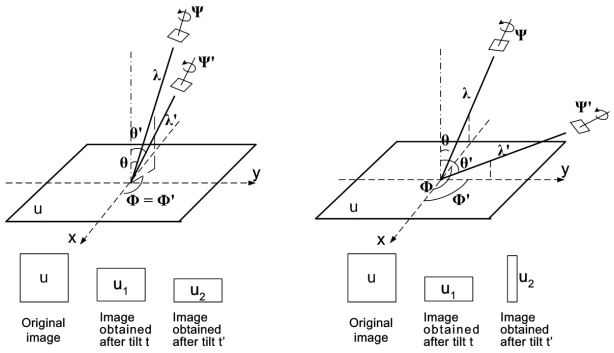

Figure 7. *Illustration of the difference between absolute tilt and transition tilt. Left: longitudes $\phi = \phi'$; latitudes $\theta = 30°$, $\theta' = 60°$; absolute tilts $t = 1/\cos\theta = 2/\sqrt{3}$, $t' = 1/\cos\theta' = 2$; transition tilts $\tau(u_1, u_2) = t'/t = \sqrt{3}$. Right: longitudes $\phi = \phi' + 90°$; latitudes $\theta = 60°$, $\theta' = 75.3°$; absolute tilts $t = 1/\cos\theta = 2$, $t' = 1/\cos\theta' = 4$; transition tilts $\tau(u_1, u_2) = t't = 8$.*

Definition 2.2. *Consider two views of a planar image, $u_1(x, y) = u(A(x, y))$ and $u_2(x, y) = u(B(x, y))$, where A and B are two affine maps such that BA^{-1} is not a similarity. With the notation of (2.2), we call transition tilt $\tau(u_1, u_2)$ and transition rotation $\phi(u_1, u_2)$ the unique parameters such that*

$$BA^{-1} = H_\lambda R_1(\psi) T_\tau R_2(\phi). \tag{2.3}$$

One can easily check the following structure properties for the transition tilt:

- The transition tilt is symmetric; i.e., $\tau(u_1, u_2) = \tau(u_2, u_1)$.
- The transition tilt depends only on the absolute tilts and on the longitude angle difference: $\tau(u_1, u_2) = \tau(t, t', \phi - \phi')$.
- One has $t'/t \le \tau \le t't$, assuming $t' = \max(t', t)$.
- The transition tilt is equal to the absolute tilt: $\tau = t'$ if the other image is in frontal view ($t = 1$).

Figure 7 illustrates the affine transition between two images taken from different viewpoints and in particular the difference between absolute tilt and transition tilt. On the left, the camera is first put in two positions corresponding to absolute tilts t and t' with the longitude

angles $\phi = \phi'$. The transition tilt between the resulting images u_1 and u_2 is $\tau(u_1, u_2) = t'/t$. On the right the tilts are made in two orthogonal directions: $\phi = \phi' + \pi/2$. A simple calculation shows that the transition tilt between u_1 and u_2 is the product $\tau(u_1, u_2) = tt'$. Thus, *two moderate absolute tilts can lead to a large transition tilt!* Since in realistic cases the absolute tilt can go up to 6, which corresponds to a latitude angle $\theta \approx 80.5°$, the *transition tilt* can easily go up to 36. The necessity of considering high transition tilts is illustrated in Figure 8.

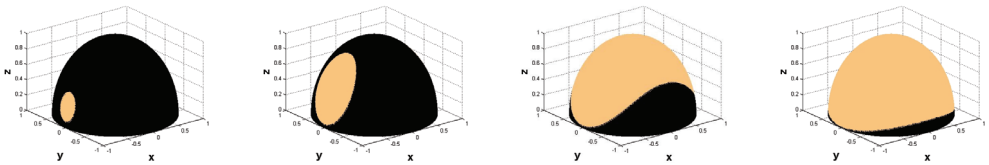

Figure 8. *This figure illustrates the necessity of considering high transition tilts to match to each other all possible views of a flat object. Two cameras take a flat object lying in the center of the hemisphere. Their optical axes point toward the center of the hemisphere. The first camera is positioned at the center of the bright region drawn on the first hemisphere. Its latitude is* $\theta = 80°$ *(absolute tilt* $t = 5.8$*). The black regions on the four hemispheres represent the positions of the second camera for which the transition tilt between the two cameras is, respectively, higher than* 2.5, 5, 10, *and* 40. *Only the fourth hemisphere is almost bright, but it needs a transition tilt as large as* 40 *to cover it well.*

3. State-of-the-art. Since an affine transform depends upon six parameters, it is prohibitive to simply *simulate* all of them and compare the simulated images. An alternative method that has been tried by many authors is *normalization*. As illustrated in Figure 9, normalization is a magic method that, given a patch that has undergone an unknown affine transform, transforms the patch into a standardized one that is independent of the affine transform.

Translation normalization can be 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, out of the six parameters in the affine transform, three are easily eliminated by normalization. Most state-of-the-art image matching algorithms adopt this normalization.

For the other three parameters, namely, the scale and the camera axis angles, things get more difficult. This section describes how the state-of-the-art image matching algorithms SIFT [29], MSER [32], LLD [44, 45, 8], and Harris-affine and Hessian-affine [35, 37] deal with these parameters.

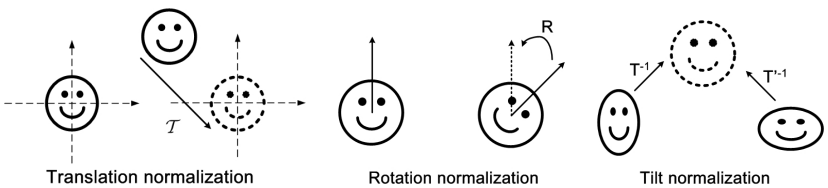

Figure 9. *Normalization methods seek to eliminate the effect of a class of affine transforms by associating the same standard patch to all transformed patches.*

3.1. Scale-invariant feature transform (SIFT). The initial goal of the SIFT method [29] is to compare two images (or two image parts) that can be deduced from each other (or from a common image) by a rotation, a translation, and a scale change. The method turns out to be robust also to rather large changes in viewpoint angle, which explains its success.

SIFT achieves the scale invariance by *simulating* the zoom in the scale-space. Following a classical paradigm, SIFT detects stable points of interest at extrema of the Laplacian of the image in the image scale-space representation. The scale-space representation introduces a smoothing parameter σ. Images u_0 are smoothed at several scales to obtain $w(\sigma, x, y) := (G_\sigma * u_0)(x, y)$, where

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

is the two-dimensional (2D) Gaussian function with integral 1 and standard deviation σ. The notation $*$ stands for the space 2D convolution.

Taking apart all sampling issues and several thresholds eliminating unreliable features, the SIFT detector can be summarized in one single sentence: *The SIFT method computes scale-space extrema (σ_i, x_i, y_i) of the spatial Laplacian of $w(\sigma, x, y)$ and then samples for each of these extrema a square image patch whose origin is (x_i, y_i), whose x-direction is one of the dominant gradients around (x_i, y_i), and whose sampling rate is $\sqrt{\sigma_i^2 + \mathbf{c}^2}$, where the constant $\mathbf{c} = 0.8$ is the tentative standard deviation of the initial image blur.*

The resulting samples of the digital patch at scale σ_i are encoded by the SIFT descriptor based on the gradient direction, which is invariant to nondecreasing contrast changes. This accounts for the robustness of the method to illumination changes. The fact that only local histograms of the direction of the gradient are kept explains the robustness of the descriptor to moderate tilts. The following theorem proved in [42] confirms the experimental evidence that SIFT is almost perfectly similarity invariant.

Theorem 3.1. *Let u and v be two images that are arbitrary frontal snapshots of the same continuous flat image u_0, $u = \mathbf{G}_\beta H_\lambda \mathcal{T} R u_0$ and $v = \mathbf{G}_\delta H_\mu u_0$, taken at different distances, with different Gaussian blurs and different zooms, and up to a camera translation and rotation around its optical axis. Without loss of generality, assume $\lambda \leq \mu$. Then if the blurs are identical ($\beta = \delta = \mathbf{c}$), all SIFT descriptors of u are identical to SIFT descriptors of v. If $\beta \neq \delta$ (or $\beta = \delta \neq \mathbf{c}$), the SIFT descriptors of u and v become (quickly) similar when their scales grow, namely, as soon as $\frac{\sigma_1}{\max(\mathbf{c},\beta)} \gg 1$ and $\frac{\sigma_2}{\max(\mathbf{c},\delta)} \gg 1$, where σ_1 and σ_2 are the scales associated with the two descriptors.*

The extensive experiments in section 6 will show that SIFT is robust to transition tilts smaller than $\tau_{\max} \approx 2$, but fails completely for larger tilts.

3.2. Maximally stable extremal regions (MSER). MSER [32] and LLD [44, 45, 8] try to be affine invariant by an affine normalization of the most robust image level sets and level lines. Both methods *normalize* all of the six parameters in the affine transform. We shall focus on MSER, but the discussion applies to LLD as well.

Extremal regions is the name given by the authors to the connected components of upper or lower level sets. MSERs are defined as maximally contrasted regions in the following way. Let $Q_1, \ldots, Q_{i-1}, Q_i, \ldots$ be a sequence of nested extremal regions $Q_i \subset Q_{i+1}$, where Q_i is

Figure 10. *Top: the same shape at different scales. Bottom: their level lines (shown at the same size). The level line shape changes with scale (in other terms, it changes with the camera distance to the object).*

defined by a threshold at level i. In other terms, Q_i is a connected component of 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 $|Q|$. Once MSERs are computed, an affine normalization is performed on the MSERs before they can be compared. Affine normalization up to a rotation is achieved by diagonalizing each MSER's second order moment matrix and by applying the linear transform that performs this diagonalization to the MSER. Rotational invariants are then computed over the normalized region.

As pointed out in [8], MSER is not fully scale invariant. This fact is illustrated in Figure 10. In MSER the scale normalization is based on the size (area) of the detected extremal regions. However, scale change is not just a homothety: It involves a blur followed by subsampling. The blur merges the regions and changes their shape and size. In other terms, the limitation of the method is the noncommutation between the optical blur and the affine transform. As shown in the image formation model (2.1), the image is blurred *after* the affine transform A. The normalization procedure does not exactly eliminate the affine deformation, because $A^{-1}\mathbf{G}_1 A u_0 \neq \mathbf{G}_1 u_0$. Their difference can be considerable when the blur kernel is broad, i.e., when the image is taken with a big zoom-out or with a large tilt. This noncommutation issue is actually a limitation of all the normalization methods.

The feature sparsity is another weakness of MSER. MSER uses only highly contrasted level sets. Many natural images contain few such features. However, the experiments in section 6 show that MSER is robust to transition tilts $\tau_{\max}$ between 5 and 10, a performance much higher than SIFT. But this performance is verified only when there is no substantial scale change between the images and if the images contain highly contrasted objects.

3.3. Harris-affine and Hessian-affine. Like MSER, Harris-affine and Hessian-affine *normalize* all six parameters in the affine transform. Harris-affine [35, 37] first detects Harris key points in the scale-space using the approach proposed by Lindeberg [23]. Then affine normalization is realized by an iterative procedure that estimates the parameters of elliptical regions and normalizes them to circular ones: At each iteration the parameters of the elliptical regions are estimated by minimizing the difference between the eigenvalues of the second order moment matrix of the selected region; the elliptical region is normalized to a circular region; and the position of the key point and its scale in scale-space are estimated. This iterative procedure due to [25, 3] finds an isotropic region, which is covariant under affine transforms.

The eigenvalues of the second moment matrix are used to measure the affine shape of the point neighborhood. The affine deformation is determined up to a rotation factor. This factor can be recovered by other methods, for example, by a normalization based on the dominant gradient orientation as in the SIFT method.

The Hessian-affine is similar to the Harris-affine, but the detected regions are blobs instead of corners. Local maximums of the determinant of the Hessian matrix are used as base points, and the remainder of the procedure is the same as for Harris-affine.

As pointed out in [29], in both methods the first step, namely, the multiscale Harris or Hessian detector, is clearly not affine covariant. The features resulting from the iterative procedure should instead be fully affine invariant. The experiments in section 6 show that the Harris-affine and Hessian-affine are robust to transition tilts of maximal value $\tau_{\max} \approx 2.5$. This disappointing result may be explained by the failure of the iterative procedure to capture large transition tilts.

4. Affine-SIFT (ASIFT). The idea of combining simulation and normalization is the main ingredient of the SIFT method. The SIFT detector normalizes rotations and translations and simulates all zooms out of the query and of the search images. Because of this feature, it is the only fully scale-invariant method.

As described in Figure 11, ASIFT simulates with enough accuracy all distortions caused by a variation of the camera optical axis direction. Then it applies the SIFT method. In other words, ASIFT simulates three parameters: the scale, the camera longitude angle, and the latitude angle (which is equivalent to the tilt) and normalizes the other three (translation and rotation). The mathematical proof that ASIFT is *fully* affine invariant will be given in section 5. The key observation is that, although a tilt distortion is irreversible due to its noncommutation with the blur, it can be compensated up to a scale change by digitally simulating a tilt of same amount in the orthogonal direction. As opposed to the *normalization* methods that suffer from this noncommutation, ASIFT *simulates* and thus achieves the full affine invariance.

Contrary to what has been formerly claimed, simulating the whole affine space is not prohibitive at all with the proposed affine space sampling. A two-resolution scheme will further reduce the ASIFT complexity to about twice that of SIFT.

4.1. ASIFT algorithm. ASIFT proceeds by the following steps.

1. Each image is transformed by simulating all possible affine distortions caused by the change of camera optical axis orientation 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 by a directional t-subsampling. It requires the previous application of an antialiasing filter in the direction of x, namely, the convolution by a Gaussian with standard deviation $\mathbf{c}\sqrt{t^2 - 1}$. The value $\mathbf{c} = 0.8$ is the value chosen by Lowe for the SIFT method [29]. As shown in [42], it ensures a very small aliasing error.
2. These rotations and tilts are performed for a finite and small number of latitude and longitude angles, 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 by a similarity invariant matching algorithm (SIFT).

The sampling of the latitude and longitude angles is specified below and will be explained in detail in section 4.2.

- The latitudes θ are sampled so that the associated tilts follow a geometric series $1, a, a^2, \ldots, a^n$, with $a > 1$. The choice $a = \sqrt{2}$ is a good compromise between accuracy and sparsity. The value n can go up to 5 or more. In consequence transition tilts going up to 32 and more can be explored.
- The longitudes ϕ are for each tilt an arithmetic series $0, b/t, \ldots, kb/t$, where $b \simeq 72°$ seems again a good compromise, and k is the last integer such that $kb/t < 180°$.

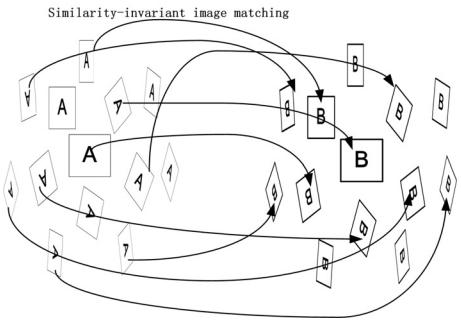

Figure 11. *Overview of the ASIFT algorithm. The square images A and B represent the compared images* $\mathbf{u}$ *and* $\mathbf{v}$*. ASIFT simulates all distortions caused by a variation of the camera optical axis direction. The simulated images, represented by the parallelograms, are then compared by SIFT, which is invariant to scale change, rotation, and translation.*

4.2. Latitude and longitude sampling. The ASIFT latitude and the longitude sampling will be determined experimentally.

Sampling ranges. The camera motion illustrated in Figure 6 shows ϕ varying from 0 to 2π. But, by Theorem 2.1, simulating $\phi \in [0, \pi)$ is enough to cover all possible affine transforms.

The sampling range of the tilt parameter t is more critical. Object recognition under any slanted view is possible only if the object is perfectly planar and Lambertian. Since this is never the case, a practical physical upper bound $t_{\max}$ must be experimentally obtained by using image pairs taken from indoor and outdoor scenes, each image pair being composed of a frontal view and a slanted view. Two case studies were performed. The first was a magazine placed on a table with the artificial illumination coming from the ceiling as shown in Figure 12. The outdoor scene was a building façade with some graffiti as illustrated in Figure 13. The images have 600×450 resolution. For each image pair, the true tilt parameter t was obtained by on-site measurements. ASIFT was applied with very large parameter sampling ranges and small sampling steps, thus ensuring that the actual affine distortion was accurately approximated. The ASIFT matching results of Figures 12 and 13 show that the physical limit is $t_{\max} \approx 4\sqrt{2}$ corresponding to a view angle $\theta_{\max} = \arccos 1/t_{\max} \approx 80°$. The sampling range $t_{\max} = 4\sqrt{2}$ allows ASIFT to be invariant to transition tilt as large as $(4\sqrt{2})^2 = 32$. (With higher resolution images, larger transition tilts would definitely be attainable.)

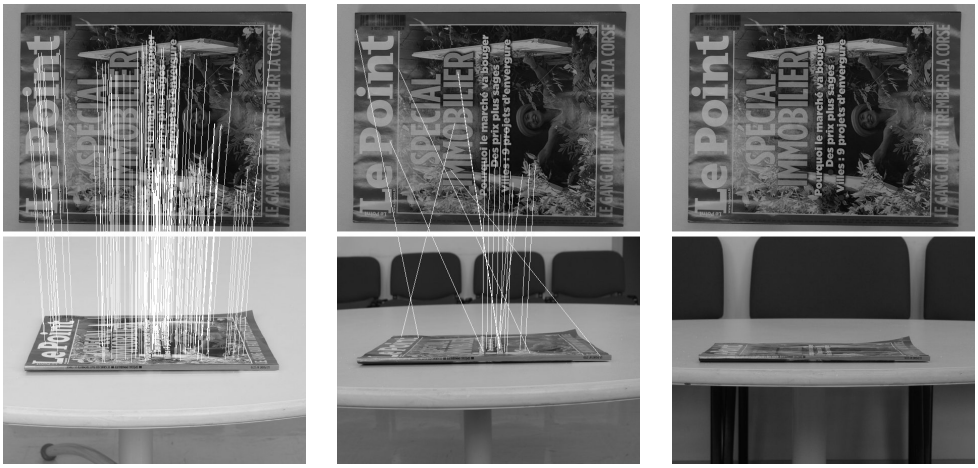

Figure 12. *Finding the maximal attainable absolute tilt. From left to right, the tilt* t *between the two images is, respectively,* $t \approx 3, 5.2, 8.5$. *The number of correct ASIFT matches is, respectively,* 151, 12, *and* 0.

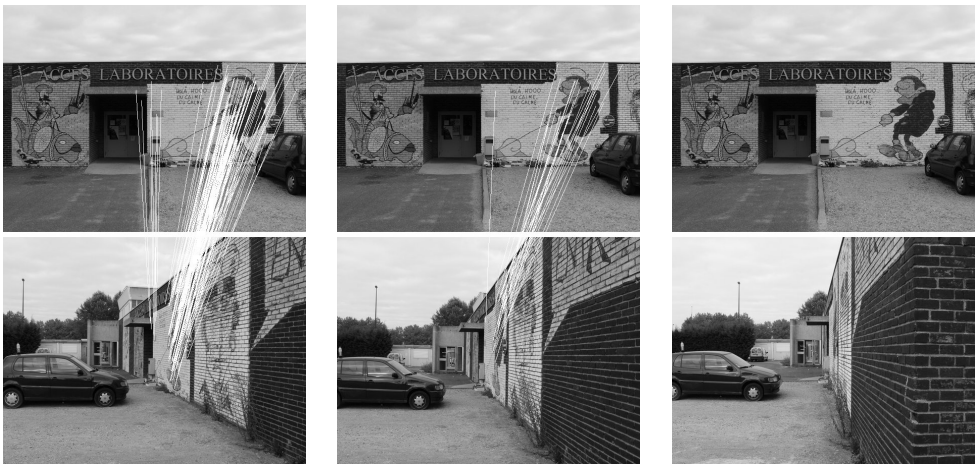

Figure 13. *Finding the maximal attainable absolute tilt. From left to right, the absolute tilt* t *between the two images is, respectively,* $t \approx 3.8, 5.6, 8$. *The number of correct ASIFT matches is, respectively,* 116, 26, *and* 0.

Sampling steps. In order to have ASIFT 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 $\triangle\phi$ must be fixed experimentally by testing several natural images.

The camera motion model illustrated in Figure 6 indicates that the sampling precision of the latitude angle $\theta = \arccos 1/t$ should increase with θ: The image distortion caused by a fixed latitude angle displacement $\triangle\theta$ is more drastic at larger 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. In what follows, the tilt sampling step is experimentally fixed to $\triangle t = \sqrt{2}$.

Similarly to the latitude sampling, one needs a finer longitude ϕ sampling when $\theta = \arccos 1/t$ increases: The image distortion caused by a fixed longitude angle displacement $\triangle\phi$

is more drastic at larger latitude angle θ. The longitude sampling step in what follows will be $\triangle\phi = \frac{72^\circ}{t}$.

The sampling steps $\triangle t = \sqrt{2}$ and $\triangle\phi = \frac{72^\circ}{t}$ were validated by successfully applying SIFT between images with simulated tilt and longitude variations equal to the sampling step values. The extensive experiments in section 6 justify the choice as well. Figure 14 illustrates the resulting irregular sampling of the parameters $\theta = \arccos 1/t$ and ϕ on the observation hemisphere: The samples accumulate near the equator.

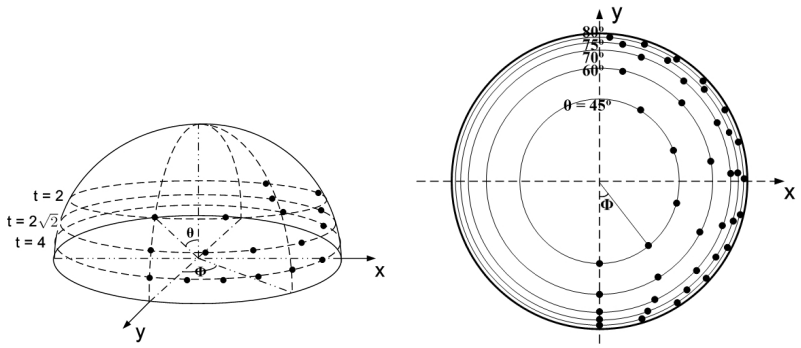

Figure 14. *Sampling of the parameters $\theta = \arccos 1/t$ and ϕ. The samples are the black dots. Left: perspective illustration of the observation hemisphere (only $t = 2, 2\sqrt{2}, 4$ are shown). Right: zenith view of the observation hemisphere. The values of θ are indicated on the figure.*

4.3. Acceleration with two resolutions. The two-resolution procedure accelerates ASIFT by applying the ASIFT method described in section 4.1 on a low-resolution version of the query and the search images. In case of success, the procedure selects the affine transforms that yielded matches in the low-resolution process, then simulates the selected affine transforms on the original query and search images, and finally compares the images simulated by SIFT. The two-resolution method is summarized as follows.

1. Subsample the query and the search 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 antialiasing Gaussian discrete filter and $\mathbf{S}_K$ is the $K \times K$ subsampling operator.
2. Low-resolution ASIFT: Apply ASIFT as described in section 4.1 to $\mathbf{u}'$ and $\mathbf{v}'$.
3. Identify the M affine transforms yielding the largest numbers of matches between $\mathbf{u}'$ and $\mathbf{v}'$.
4. High-resolution ASIFT: Apply ASIFT to $\mathbf{u}$ and $\mathbf{v}$, but simulate only the M affine transforms.

Figure 15 shows an example. The low-resolution ASIFT that is applied on the $K \times K = 3 \times 3$ subsampled images finds 19 correspondences and identifies the $M = 5$ best affine transforms. The high-resolution ASIFT finds 51 correct matches.

4.4. ASIFT complexity. The complexity of the ASIFT method will be estimated under the recommended configuration: The tilt and angle ranges are $[t_{\min}, t_{\max}] = [1, 4\sqrt{2}]$ and $[\phi_{\min}, \phi_{\max}] = [0^\circ, 180^\circ]$, and the sampling steps are $\triangle t = \sqrt{2}$, $\triangle\phi = \frac{72^\circ}{t}$. A t tilt is simulated by t times subsampling in one direction. The query and the search images are subsampled

Figure 15. *Two-resolution ASIFT. Left: Low-resolution ASIFT applied on the* 3×3 *subsampled images finds* 19 *correct matches. Right: High-resolution ASIFT finds* 51 *matches.*

by a $K \times K = 3 \times 3$ factor for the low-resolution ASIFT. Finally, the high-resolution ASIFT simulates the M best affine transforms that are identified, but only in case they lead to enough matches. In real applications where a query image is compared with a large database, the likely result for the low-resolution step is failure. The final high-resolution step counts only when the images match at low resolution.

Estimating the ASIFT complexity boils down to calculating the image area simulated by the low-resolution ASIFT. Indeed the complexity of the image matching *feature computation* is proportional to the input image area. One can verify that the total image area simulated by ASIFT is proportional to the number of simulated tilts t: The number of ϕ simulations is proportional to t for each t, but the t subsampling for each tilt simulation divides the area by t. More precisely, the image area input to low-resolution ASIFT is

$$\frac{1 + (|\Gamma_t| - 1)\frac{180^\circ}{72^\circ}}{K \times K} = \frac{1 + 5 \times 2.5}{3 \times 3} = 1.5$$

times as large as that of the original images, where $|\Gamma_t| = |\{1, \sqrt{2}, 2, 2\sqrt{2}, 4, 4\sqrt{2}\}| = 6$ is the number of simulated tilts and $K \times K = 3 \times 3$ is the subsampling factor. Thus the complexity of the low-resolution ASIFT feature calculation is 1.5 times as much as that of a single SIFT routine. The ASIFT algorithm in this configuration is invariant to transition tilts up to 32. Higher transition tilt invariance is attainable with larger $t_{\max}$. The complexity growth is *linear* and thus marginal with respect to the *exponential* growth of transition tilt invariance.

Low-resolution ASIFT simulates 1.5 times the area of the original images and generates in consequence about 1.5 times more features on both the query and the search images. The complexity of low-resolution ASIFT *feature comparison* is therefore $1.5^2 = 2.25$ times as much as that of SIFT.

If the image comparisons involve a large database where most comparisons will be failures, ASIFT stops essentially at the end of the low-resolution procedure, and the overall complexity is about twice the SIFT complexity, as argued above.

If the comparisons involve a set of images with high matching likelihood, then the high-resolution step is no longer negligible. The overall complexity of ASIFT depends on the number M of the identified good affine transforms simulated in the high-resolution procedure as well as on the simulated tilt values t. However, in case the images match, ASIFT ensures many more detections than SIFT, because it explores many more viewpoint angles. In that case the *complexity rate per match detection* is in practice equal to or smaller than the per match detection complexity of a SIFT routine.

The SIFT subroutines can be implemented in parallel in ASIFT (for both the low-resolution and the high-resolution ASIFT). Recently many authors have investigated SIFT accelerations [19, 13, 22]. A realtime SIFT implementation has been proposed in [54]. Obviously all the SIFT acceleration techniques directly apply to ASIFT.

5. The mathematical justification. This section proves mathematically that ASIFT is fully affine invariant, up to sampling errors. The key observation is that a tilt can be compensated up to a scale change by another tilt of the same amount in the orthogonal direction.

The proof is given in a continuous setting which is by far simpler because the image sampling does not interfere. Since the digital images are assumed to be well sampled, the Shannon interpolation (obtained by zero-padding) paves the way from discrete to continuous.

To lighten the notation, G_σ will also denote the convolution operator on $\mathbb{R}^2$ with the Gauss kernel $\mathrm{G}_\sigma(x,y) = \frac{1}{2\pi(\mathbf{c}\sigma)^2} e^{-\frac{x^2+y^2}{2(\mathbf{c}\sigma)^2}}$, namely, $\mathrm{G}u(x,y) := (\mathrm{G} * u)(x,y)$, where the constant $\mathbf{c} = 0.8$ is chosen for good antialiasing [29, 42]. The one-dimensional (1D) Gaussians will be denoted by $\mathrm{G}^x_\sigma(x,y) = \frac{1}{\sqrt{2\pi}\mathbf{c}\sigma} e^{-\frac{x^2}{2(\mathbf{c}\sigma)^2}}$ and $\mathrm{G}^y_\sigma(x,y) = \frac{1}{\sqrt{2\pi}\mathbf{c}\sigma} e^{-\frac{y^2}{2(\mathbf{c}\sigma)^2}}$. G_σ satisfies the semigroup property

$$\mathrm{G}_\sigma \mathrm{G}_\beta = \mathrm{G}_{\sqrt{\sigma^2+\beta^2}} \tag{5.1}$$

and commutes with rotations:

$$\mathrm{G}_\sigma R = R \mathrm{G}_\sigma. \tag{5.2}$$

We shall denote by $*_y$ the 1D convolution operator in the y-direction. In the notation $\mathrm{G}*_y$, G is a 1D Gaussian depending on y and

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

5.1. Inverting tilts. Let us distinguish the following two tilting procedures.

Definition 5.1. *Given $t > 1$, the tilt factor, define the following:*

- *The geometric tilt: $T^x_t u_0(x,y) := u_0(tx, y)$. In case this tilt is made in the y-direction, it will be denoted by $T^y_t u_0(x,y) := u_0(x, ty)$.*

- *The simulated tilt (taking into account camera blur):* $\mathbb{T}_t^x v := T_t^x \mathrm{G}^x_{\sqrt{t^2-1}} *_x v$. *In case the simulated tilt is done in the y-direction, it is denoted* $\mathbb{T}_t^y v := T_t^y \mathrm{G}^y_{\sqrt{t^2-1}} *_y v$.

As described by the image formation model (2.1), an infinite resolution scene u_0 observed from a slanted view in the x-direction is distorted by a *geometric* tilt before it is blurred by the optical lens; i.e., $u = \mathrm{G}_1 T_t^x u_0$. Reversing this operation is in principle impossible, because of the tilt and blur noncommutation. However, the next lemma shows that a *simulated* tilt $\mathbb{T}_t^y$ in the orthogonal direction actually provides a pseudoinverse to the *geometric* tilt T_t^x.

Lemma 5.2. $\mathbb{T}_t^y = H_t \mathrm{G}^y_{\sqrt{t^2-1}} *_y (T_t^x)^{-1}$.

Proof. Since $(T_t^x)^{-1} u(x,y) = u(\frac{x}{t}, y)$,

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

Thus

$$\begin{aligned} H_t \left(\mathrm{G}_{\sqrt{t^2-1}} *_y (T_t^x)^{-1} u\right)(x,y) &= \int \mathrm{G}_{\sqrt{t^2-1}}(z) u(x, ty - z) dz \\ &= \left(\mathrm{G}^y_{\sqrt{t^2-1}} *_y u\right)(x, ty) = \left(T_t^y \mathrm{G}^y_{\sqrt{t^2-1}} *_y u\right)(x,y). \end{aligned}$$ ∎

By the next lemma, a tilted image $\mathrm{G}_1 T_t^x u$ can be tilted back by tilting in the orthogonal direction. The price to pay is a t zoom-out. The second relation in the lemma means that the application of the *simulated* tilt to a well-sampled image yields an image that keeps the image well sampled. This fact is crucial in simulating tilts on digital images.

Lemma 5.3. *Let* $t \geq 1$. *Then*

$$\mathbb{T}_t^y (\mathrm{G}_1 T_t^x) = \mathrm{G}_1 H_t, \tag{5.3}$$

$$\mathbb{T}_t^y \mathrm{G}_1 = \mathrm{G}_1 T_t^y. \tag{5.4}$$

Proof. By Lemma 5.2, $\mathbb{T}_t^y = H_t \mathrm{G}^y_{\sqrt{t^2-1}} *_y (T_t^x)^{-1}$. Thus,

$$\mathbb{T}_t^y (\mathrm{G}_1 T_t^x) = H_t \mathrm{G}^y_{\sqrt{t^2-1}} *_y ((T_t^x)^{-1} \mathrm{G}_1 T_t^x). \tag{5.5}$$

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

$$(T_t^x)^{-1} \mathrm{G}_1 T_t^x u = \left(\frac{1}{t} \mathrm{G}_1\left(\frac{x}{t}, y\right)\right) * u, \tag{5.6}$$

and by the separability of the 2D Gaussian in two 1D Gaussians,

$$\frac{1}{t} \mathrm{G}_1\left(\frac{x}{t}, y\right) = \mathrm{G}_t(x) \mathrm{G}_1(y). \tag{5.7}$$

From (5.6) and (5.7) one obtains

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

which implies

$$\mathrm{G}^y_{\sqrt{t^2-1}} *_y (T^x)^{-1} \mathrm{G}_1 T^x_t u = \mathrm{G}^y_{\sqrt{t^2-1}} *_y (\mathrm{G}^x_t(x) *_x \mathrm{G}^y_1(y) *_y u) = \mathrm{G}_t u.$$

Indeed, the 1D convolutions in x and y commute and $\mathrm{G}^y_{\sqrt{t^2-1}} * \mathrm{G}^y_1 = \mathrm{G}^y_t$ by the Gaussian semigroup property (5.1). Substituting the last proven relation in (5.5) yields

$$\mathbb{T}^y_t \mathrm{G}_1 T^x_t u = H_t \mathrm{G}_t u = \mathrm{G}_1 H_t u.$$

The second relation (5.4) follows immediately by noting that $H_t = T^y_t T^x_t$. ∎

5.2. Proof that ASIFT works. The meaning of Lemma 5.3 is that we can design an exact algorithm that simulates all inverse tilts, up to scale changes.

Theorem 5.4. *Let $u = \mathrm{G}_1 A \mathcal{T}_1 u_0$ and $v = \mathrm{G}_1 B \mathcal{T}_2 u_0$ be two images obtained from an infinite resolution image u_0 by cameras at infinity with arbitrary position and focal lengths. (A and B are arbitrary affine maps with positive determinants and $\mathcal{T}_1$ and $\mathcal{T}_2$ arbitrary planar translations.) Then ASIFT, applied with a dense set of tilts and longitudes, simulates two views of u and 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.*

Proof. We start by giving a formalized version of ASIFT using the above notation.

(Dense) ASIFT

1. Apply a dense set of rotations to both images u and v.
2. Apply in continuation a dense set of *simulated* tilts $\mathbb{T}^x_t$ to all rotated images.
3. Perform a SIFT comparison of all pairs of resulting images.

Notice that by the relation

$$\mathbb{T}^x_t R\left(\frac{\pi}{2}\right) = R\left(\frac{\pi}{2}\right) \mathbb{T}^y_t, \tag{5.8}$$

the algorithm also simulates tilts in the y-direction, up to an $R(\frac{\pi}{2})$ rotation.

By the affine decomposition (2.2),

$$BA^{-1} = H_\lambda R_1 T^x_t R_2. \tag{5.9}$$

The *dense* ASIFT applies in particular

1. $\mathbb{T}^x_{\sqrt{t}} R_2$ to $\mathrm{G}_1 A \mathcal{T}_1 u_0$, which by (5.2) and (5.4) yields $\tilde{u} = \mathrm{G}_1 T^x_{\sqrt{t}} R_2 A \mathcal{T}_1 u_0 := \mathrm{G}_1 \tilde{A} \mathcal{T}_1 u_0$;
2. $R(\frac{\pi}{2}) \mathbb{T}^y_{\sqrt{t}} R_1^{-1}$ to $\mathrm{G}_1 B \mathcal{T}_2 u_0$, which by (5.2) and (5.4) yields $\mathrm{G}_1 R(\frac{\pi}{2}) T^y_{\sqrt{t}} R_1^{-1} B \mathcal{T}_2 u_0 := \mathrm{G}_1 \tilde{B} \mathcal{T}_2 u_0$.

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

$$\tilde{B}^{-1} R\left(\frac{\pi}{2}\right) H_{\sqrt{t}} \tilde{A} = B^{-1} R_1 T^y_{\sqrt{t}^{-1}} T^x_{\sqrt{t}} H_{\sqrt{t}} R_2 A = B^{-1} R_1 T^x_t R_2 A = B^{-1} (H_{\frac{1}{\lambda}} B A^{-1}) A = H_{\frac{1}{\lambda}}.$$

It follows that $\tilde{B} = R(\frac{\pi}{2}) H_{\lambda\sqrt{t}} \tilde{A}$. Thus,

$$\tilde{u} = \mathrm{G}_1 \tilde{A} \mathcal{T}_1 u_0 \quad \text{and} \quad \tilde{v} = \mathrm{G}_1 R\left(\frac{\pi}{2}\right) H_{\lambda\sqrt{t}} \tilde{A} \mathcal{T}_2 u_0$$

are two of the images simulated by ASIFT, and are deduced from each other by a rotation and a $\lambda\sqrt{t}$ zoom. It follows from Theorem 3.1 that their descriptors are identical as soon as the scale of the descriptors exceeds $\lambda\sqrt{t}$. ∎

Remark 1. The above proof gives the value of the simulated tilts achieving success: If the transition tilt between u and v is t, then it is enough to simulate a $\sqrt{t}$ tilt on both images.

5.3. Algorithmic sampling issues. Although the above proof deals with asymptotic statements when the sampling steps tend to zero or when the SIFT scales tend to infinity, the approximation rate is quick, a fact that can only be checked experimentally. This fact is actually extensively verified by the huge amount of experimental evidence on SIFT, which shows first that the recognition of scale-invariant features is robust to a rather large latitude and longitude variation, and second that the scale invariance is quite robust to moderate errors on scale. Section 4.2 has evaluated the adequate sampling rates and ranges for tilts and longitudes.

The above algorithmic description has neglected the image sampling issues, but care was taken that input images and output images were always written in the $\mathrm{G}_1 u$ form. For the digital input images, which always have the form $\mathbf{u} = \mathbf{S}_1 \mathrm{G}_1 u_0$, the Shannon interpolation algorithm I is first applied, to give back $I\mathbf{S}_1 \mathrm{G}_1 u_0 = \mathrm{G}_1 u_0$. For the output images, which always have the form $\mathrm{G}_1 u$, the sampling $\mathbf{S}_1$ gives back a digital image.

6. Experiments. ASIFT image matching performance will be compared with the state-of-the-art approaches using the detectors SIFT [29], MSER [32], Harris-affine, and Hessian-affine [34, 37], all combined with the most popular SIFT descriptor [29]. The MSER detector combined with the correlation descriptor as proposed in the original work [32] was initially included in the comparison, but its performance was found to be slightly inferior to that of the MSER detector combined with the SIFT descriptor, as indicated in [36]. Thus only the latter will be shown. In the following, the methods will be named after their detectors, namely, ASIFT, SIFT, MSER, Harris-affine, and Hessian-affine.

The experiments include extensive tests with the standard Mikolajczyk database [33], a systematic evaluation of methods' invariance to absolute and transition tilts and other images of various types (resolution 600×450).

In the experiments the Lowe [28] reference software was used for SIFT. For all the other methods we used the binaries of the MSER, the Harris-affine, and the Hessian-affine detectors and the SIFT descriptor provided by the authors, all downloadable from [33].

The low-resolution ASIFT applied a 3×3 image subsampling. ASIFT may detect repeated matches from the image pairs simulated with different affine transforms. All the redundant matches have been removed. (A match between two points p_1 and p_2 was considered redundant with a match between p_3 and p_4 if $\mathbf{d}^2(p_1, p_3) < 3$ and $\mathbf{d}^2(p_2, p_4) < 3$, where $\mathbf{d}(p_i, p_j)$ denotes the Euclidean distance between p_i and p_j.)

6.1. Standard test database. The standard Mikolajczyk database [33] was used to evaluate the methods' robustness to four types of distortions, namely blur, similarity, viewpoint change, and jpeg compression. Five image pairs (image 1 vs. images 2 to 6) with increasing amount of distortion were used for each test. Figure 16 illustrates the number of correct matches achieved by each method. For each method, the number of image pairs m on which more than 20 correct matches are detected and the average number of matches n over these m pairs are shown for each test. Among the methods under comparison, ASIFT is the only one that works well for the entire database. It also systematically finds more correct matches. More precisely, we have the following:

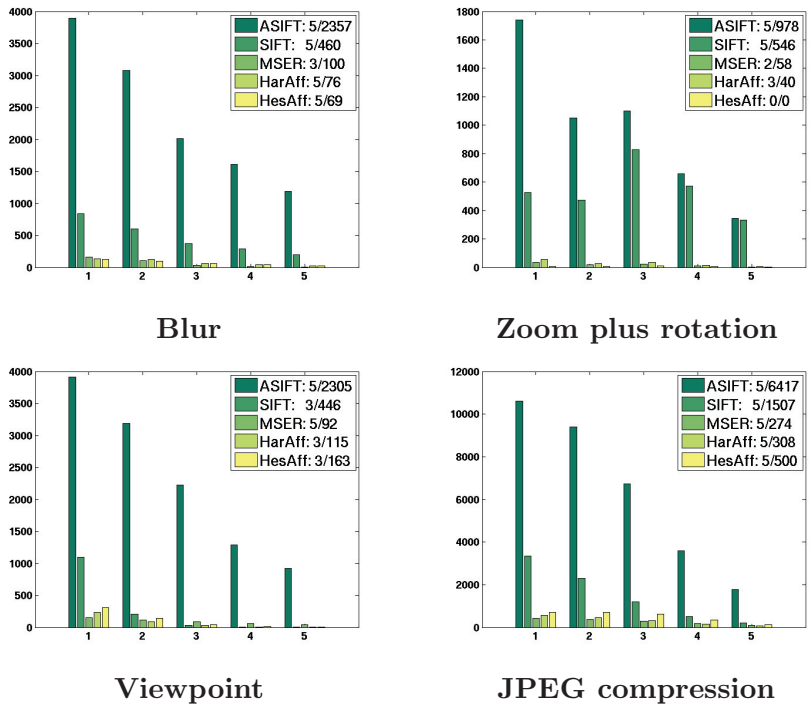

Figure 16. *Number of correct matches achieved by ASIFT, SIFT, MSER, Harris-affine, and Hessian-affine under four types of distortions, namely, blur, zoom plus rotation, viewpoint change, and jpeg compression, in the standard Mikolajczyk database. In the top right corner of each graph m/n gives for each method the number of image pairs m on which more than 20 correct matches were detected, and the average number of matches n over these m pairs.*

- *Blur.* ASIFT and SIFT are very robust to blur, followed by Harris-affine and Hessian-affine. MSER is not robust to blur.
- *Zoom plus rotation.* ASIFT and SIFT are very robust to zoom plus rotation, while MSER, Harris-affine, and Hessian-affine have limited robustness, as explained in section 3.
- *Viewpoint change.* ASIFT is very robust to viewpoint change, followed by MSER. On average ASIFT finds 20 times more matches than MSER. SIFT, Harris-affine, and Hessian-affine have comparable performance: They fail when the viewpoint change is substantial.

  The test images (see Figure 17) provided optimal conditions for MSER: The camera-object distances are similar, and well-contrasted shapes are always present.
- *Compression.* All considered methods are very robust to jpeg compression.

Figure 17 shows the classic image pair Graffiti 1 and 6. ASIFT finds 925 correct matches. SIFT, Harris-affine, and Hessian-affine find, respectively, 0, 3, and 1 correct matches: The $\tau \approx 3.2$ transition tilt is just a bit too large for these methods. MSER finds 42 correct correspondences.

The next sections describe more systematic evaluations of the robustness to absolute and transition tilts of the compared methods. The normalization methods MSER, Harris-affine,

Figure 17. *Two Graffiti images with transition tilt $\tau \approx 3.2$. ASIFT (shown), SIFT (shown), Harris-affine, Hessian-affine, and MSER (shown) find* 925, 2, 3, 1, *and* 42 *correct matches.*

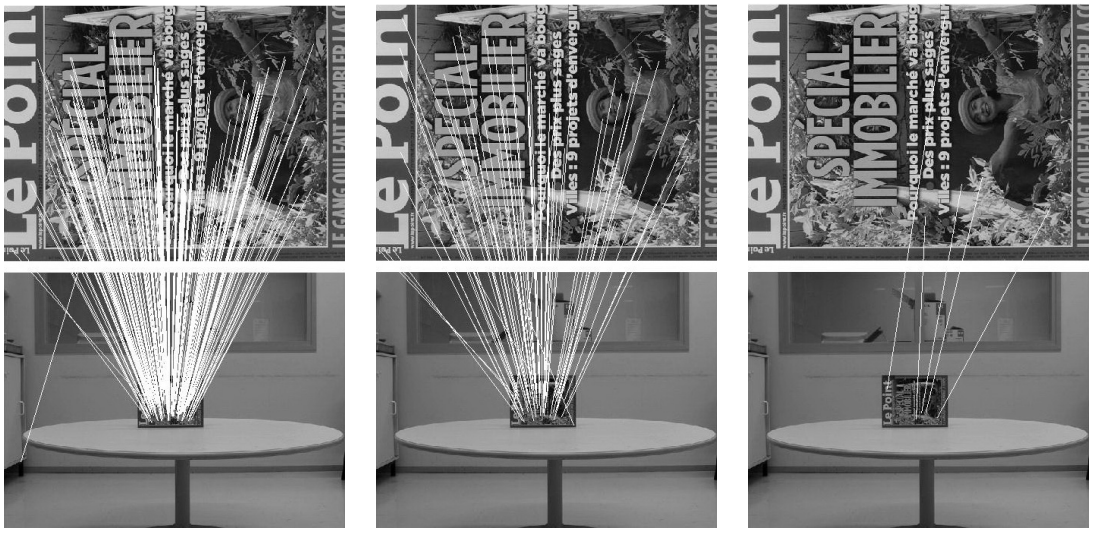

Figure 18. *Robustness to scale change. ASIFT (shown), SIFT (shown), Harris-affine (shown), Hessian-affine, and MSER find, respectively,* 221, 86, 4, 3, *and* 4 *correct matches. Harris-affine, Hessian-affine, and MSER are not robust to scale change.*

and Hessian-affine have been shown to fail under large-scale changes (see another example in Figure 18). To focus on tilt invariance, the experiments will therefore take image pairs with similar scales.

6.2. Absolute tilt tests. Figure 19(a) illustrates the experimental setting. The painting illustrated in Figure 20 was photographed with an optical zoom varying between $\times 1$ and $\times 10$

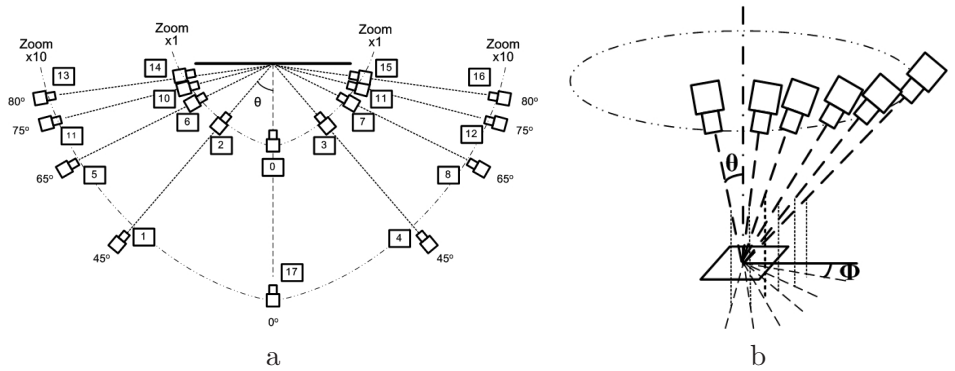

Figure 19. *The settings adopted for systematic comparison. Left: absolute tilt test. An object is photographed with a latitude angle varying from* 0° *(frontal view) to* 80°*, from distances varying between* 1 *and* 10*, which is the maximum focus distance change. Right: transition tilt test. An object is photographed with a longitude angle* ϕ *that varies from* 0° *to* 90°*, from a fixed distance.*

Figure 20. *The painting (left) and the magazine cover (right) that were photographed in the absolute and transition tilt tests.*

and with viewpoint angles between the camera axis and the normal to the painting varying from 0° (frontal view) to 80°. It is clear that, beyond 80°, establishing a correspondence between the frontal image and the extreme viewpoint becomes haphazard. With such a big change of view angle on a reflective surface, the image in the slanted view can be totally different from that in the frontal view.

Table 1 summarizes the performance of each algorithm in terms of number of correct matches. Some matching results are illustrated in Figures 22–23. MSER, which uses maximally stable level sets as features, obtains most of the time many fewer correspondences than the methods whose features are based on local maxima in the scale-space. As depicted in Figure 21, for images taken at a short distance (zoom ×1) the tilt varies on the same flat object because of the perspective effect, an example being illustrated in Figure 22. The number of SIFT correspondences drops dramatically when the angle is larger than 65° (tilt $t \approx 2.3$), and it fails completely when the angle exceeds 75° (tilt $t \approx 3.8$). At 75°, as shown in Figure 22, most SIFT matches are located on the side closer to the camera where the actual tilt is actually smaller. The performance of Harris-affine and Hessian-affine decays considerably when the angle goes over 75° (tilt $t \approx 3.8$). The MSER correspondences are always fewer and show a noticeable decline over 65° (tilt $t \approx 2.4$). ASIFT works until 80° (tilt $t \approx 5.8$).

Table 1

Absolute tilt invariance comparison with photographs of the painting in Figure 20. Number of correct matches of ASIFT, SIFT, Harris-affine (HarAff), Hessian-affine (HesAff), and MSER for viewpoint angles between 45° and 80°. Top: images taken with zoom ×1. Bottom: images taken with zoom ×10. The latitude angles and the absolute tilts are listed in the far left column. For the ×1 zoom, strong perspective effect is present, and the tilts shown are average values.

Z × 1					
θ/t	SIFT	HarAff	HesAff	MSER	ASIFT
−80°/5.8	1	16	1	4	110
−75°/3.9	24	36	7	3	281
−65°/2.3	117	43	36	5	483
−45°/1.4	245	83	51	13	559
45°/1.4	195	86	26	12	428
65°/2.4	92	58	32	11	444
75°/3.9	15	3	1	5	202
80°/5.8	2	6	6	5	204
Z × 10					
θ/t	SIFT	HarAff	HesAff	MSER	ASIFT
−80°/5.8	1	1	0	2	116
−75°/3.9	0	3	0	6	265
−65°/2.3	10	22	16	10	542
−45°/1.4	182	68	45	19	722
45°/1.4	171	54	26	15	707
65°/2.4	5	12	5	6	468
75°/3.9	2	1	0	4	152
80°/5.8	3	0	0	2	110

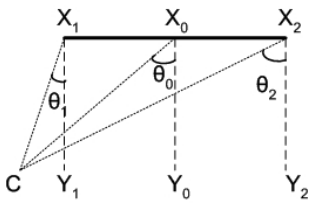

Figure 21. *When the camera focus distance is small, the absolute tilt of a plane object can vary considerably in the same image due to the strong perspective effect.*

Consider now images taken at a camera-object distance multiplied by 10, as shown in Figure 23. For these images the SIFT performance drops considerably: Recognition is possible only with angles smaller than 45°. The performance of Harris-affine and Hessian-affine declines steeply when the angle goes from 45° to 65°. Beyond 65° they fail completely. MSER struggles at the angle of 45° and fails at 65°. ASIFT functions perfectly until 80°.

Rich in highly contrasted regions, the magazine shown in Figure 20 is more favorable to MSER. Table 2 shows the result of a similar experiment performed with the magazine, with the latitude angles from 50° to 80° on one side and with the camera focus distance ×4. Figure 24 shows the result with 80° angle. The performance of SIFT, Harris-affine, and Hessian-affine drops steeply with the angle going from 50° to 60° (tilt t from 1.6 to 2). Beyond 60° (tilt $t = 2$) they fail completely. MSER finds many correspondences until 70° (tilt $t \approx 2.9$). The

Figure 22. *Correspondences between the painting images taken from short distance (zoom* $\times 1$*) at frontal view and at* 75° *angle. The local absolute tilt varies:* $t \approx 4$ *(middle),* $t < 4$ *(right),* $t > 4$ *(left). ASIFT (shown), SIFT (shown), Harris-affine, Hessian-affine, and MSER (shown) find, respectively,* 202, 15, 3, 1, *and* 5 *correct matches.*

Figure 23. *Correspondences between long distance views (zoom* $\times 10$*), frontal view and* 80° *angle, absolute tilt* $t \approx 5.8$*. ASIFT (shown), SIFT, Harris-affine (shown), Hessian-affine, and MSER (shown) find, respectively,* 116, 1, 1, 0, *and* 2 *correct matches.*

number of correspondences drops when the angle exceeds 70° and becomes too small at 80° (tilt $t \approx 5.8$) for robust recognition. ASIFT works until 80°.

The above experiments suggest an estimate of the maximal absolute tilts for the method under comparison. For SIFT, this limit is hardly above 2. The limit is about 2.5 for Harris-affine and Hessian-affine. The performance of MSER depends on the type of image. For images

Table 2

Absolute tilt invariance comparison with photographs of the magazine cover (Figure 20). Number of correct matches of ASIFT, SIFT, Harris-affine (HarAff), Hessian-affine (HesAff), and MSER for viewpoint angles between 50° and 80°. The latitude angles and the absolute tilts are listed in the far left column.

θ/t	SIFT	HarAff	HesAff	MSER	ASIFT
50°/1.6	267	131	144	150	1692
60°/2.0	20	29	39	117	1012
70°/2.9	1	2	2	69	754
80°/5.8	0	0	0	17	349

Figure 24. *Correspondences between magazine images taken with zoom ×4, frontal view and 80° angle, absolute tilt $t \approx 5.8$. ASIFT (shown), SIFT (shown), Harris-affine, Hessian-affine, and MSER (shown) find, respectively,* 349, 0, 0, 0, *and* 17 *correct matches.*

with highly contrasted regions, MSER reaches a 5 absolute tilt. However, if the images do not contain highly contrasted regions, the performance of MSER can drop under small tilts. For ASIFT, a 5.8 absolute tilt that corresponds to an extreme viewpoint angle of 80° is easily attainable.

6.3. Transition tilt tests. The magazine shown in Figure 20 was placed face-up and photographed to obtain two sets of images. As illustrated in Figure 19(b), for each image set the camera with a fixed latitude angle θ corresponding to $t = 2$ and 4 circled around, the longitude angle ϕ growing from 0° to 90°. The camera focus distance and the optimal zoom were ×4. In each set the resulting images have the same absolute tilt $t = 2$ or 4, while the transition tilt τ (with respect to the image taken at $\phi = 0°$) goes from 1 to $t^2 = 4$ or 16 when ϕ goes from 0° to 90°. To evaluate the maximum invariance to transition tilt, the images taken at $\phi \neq 0$ were matched against the one taken at $\phi = 0$.

Table 3 compares the performance of the algorithms. When the absolute tilt is $t = 2$, the SIFT performance drops dramatically when the transition tilt goes from 1.3 to 1.7. With a transition tilt over 2.1, SIFT fails completely. Similarly, a considerable performance decline

Table 3

Transition tilt invariance comparison (object photographed: the magazine cover shown in Figure 20). Number of correct matches of ASIFT, SIFT, Harris-affine (HarAff), Hessian-affine (HesAff), and MSER for viewpoint angles between 50° and 80°. The affine parameters of the two images are $\phi_1 = 0°$, $t_1 = t_2 = 2$ (above), $t_1 = t_2 = 4$ (below). ϕ_2 and the transition tilts τ are in the far left column.

$t_1 = t_2 = 2$					
ϕ_2/τ	SIFT	HarAff	HesAff	MSER	ASIFT
10°/1.3	408	233	176	124	1213
20°/1.7	49	75	84	122	1173
30°/2.1	5	24	32	103	1048
40°/2.5	3	13	29	88	809
50°/3.0	3	1	3	87	745
60°/3.4	2	0	1	62	744
70°/3.7	0	0	0	51	557
80°/3.9	0	0	0	51	589
90°/4.0	0	0	1	56	615
$t_1 = t_2 = 4$					
ϕ_2/τ	SIFT	HarAff	HesAff	MSER	ASIFT
10°/1.9	22	32	14	49	1054
20°/3.3	4	5	1	39	842
30°/5.3	3	2	1	32	564
40°/7.7	0	0	0	28	351
50°/10.2	0	0	0	19	293
60°/12.4	1	0	0	17	145
70°/14.3	0	0	0	13	90
80°/15.6	0	0	0	12	106
90°/16.0	0	0	0	9	88

is observed for Harris-affine and Hessian-affine when the transition tilt goes from 1.3 to 2.1. Hessian-affine slightly outperforms Harris-affine, but both methods fail completely when the transition tilt goes above 3. Figure 25 shows an example that SIFT, Harris-affine, and Hessian-affine fail completely under a moderate transition tilt $\tau \approx 3$. MSER and ASIFT work stably up to a 4 transition tilt. ASIFT finds 10 times as many correspondences as MSER covering a much larger area.

Under an absolute tilt $t = 4$, SIFT, Harris-affine, and Hessian-affine struggle at a 1.9 transition tilt. They fail completely when the transition tilt gets bigger. MSER works stably until a 7.7 transition tilt. Over this value, the number of correspondences is too small for reliable recognition. ASIFT works perfectly up to the 16 transition tilt. The above experiments show that the maximum transition tilt, about 2 for SIFT and 2.5 for Harris-affine and Hessian-affine, is by far insufficient. This experiment and others confirm that MSER ensures a reliable recognition until a transition tilt of about 10, but this is only true when the images under comparison are free of scale change and contain highly contrasted regions. The experimental limit transition tilt of ASIFT goes easily up to 36 (see Figure 2).

6.4. Other test images. ASIFT, SIFT, MSER, Harris-affine, and Hessian-affine will now be tried with various classic test images and some new ones. Used by Matas et al. in their online demo [31] as a standard image to test MSER [32], the images in Figure 26 [57] show a number

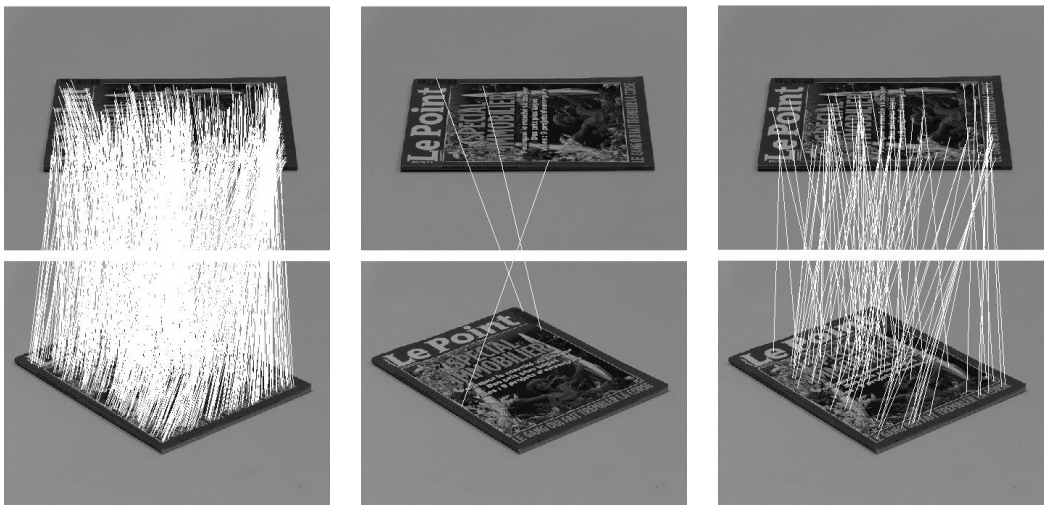

Figure 25. *Correspondences between the magazine images taken with absolute tilts* $t_1 = t_2 = 2$ *with longitude angles* $\phi_1 = 0^\circ$ *and* $\phi_2 = 50^\circ$*, transition tilt* $\tau \approx 3$*. ASIFT (shown), SIFT (shown), Harris-affine, Hessian-affine, and MSER (shown) find, respectively,* 745, 3, 1, 3, 87 *correct matches.*

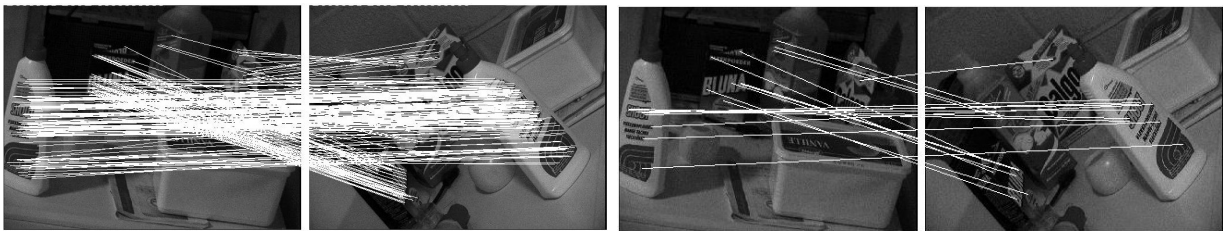

Figure 26. *Image matching (images used by Matas et al.* [31]*). Transition tilt:* $\tau \in [1.6, 3.0]$*. ASIFT (shown), SIFT, Harris-affine, Hessian-affine, and MSER (shown) find, respectively,* 254, 10, 23, 11, *and* 22 *correct matches. Background images reprinted from* [*T. Tuytelaars and L. Van Gool,* Matching widely separated views based on affine invariant regions*, Int. J. Comput. Vis.,* 59 (2004)*, pp.* 61–85]*. Copyright* © 2004 *Springer Science+Business Media. Reprinted with permission of the publisher and authors. All rights reserved.*

of containers placed on a desktop.[1] ASIFT, SIFT, Harris-affine, Hessian-affine, and MSER find, respectively, 254, 10, 23, 11, and 22 correct correspondences. Figure 27 contains two orthogonal road signs taken under a view change that makes a transition tilt $\tau \approx 2.6$. ASIFT successfully matches the two signs, finding 50 correspondences while all the other methods totally fail. The pair of aerial images of the Pentagon shown in Figure 28 show a moderate transition tilt $\tau \approx 2.5$. ASIFT works perfectly by finding 378 correct matches, followed by MSER that finds 17. Harris-affine, Hessian-affine, and SIFT fail by finding, respectively, 6, 2, and 8 matches. The Statue of Liberty shown in Figure 29 presents a strong relief effect. ASIFT finds 22 good matches. The other methods fail completely. Figure 30 shows some deformed cloth (images from [26, 27]). ASIFT significantly outperforms the other methods by finding, respectively, 141 and 370 correct matches, followed by SIFT that finds 31 and 75 matches. Harris-affine, Hessian-affine, and MSER do not get a significant number of matches.

[1]We thank Tinne Tuytelaars for kindly providing us with the images.

Figure 27. *Image matching: road signs. Transition tilt* $\tau \approx 2.6$. *ASIFT (shown), SIFT, Harris-affine, Hessian-affine, and MSER (shown) find, respectively,* 50, 0, 0, 0, *and* 1 *correct matches.*

Figure 28. *Pentagon, with transition tilt* $\tau \approx 2.5$. *ASIFT (shown), SIFT (shown), Harris-affine, Hessian-affine, and MSER (shown) find, respectively,* 378, 6, 2, 8, *and* 17 *correct matches.*

Figure 29. *Statue of Liberty, with transition tilt* $\tau \in [1.3, \infty)$. *ASIFT (shown), SIFT (shown), Harris-affine, Hessian-affine, and MSER find, respectively,* 22, 1, 0, 0, *and* 0 *correct matches.*

Figure 30. *Image matching with object deformation. Left: flag. ASIFT (shown), SIFT, Harris-affine, Hessian-affine, and MSER find, respectively,* 141, 31, 15, 10, *and* 2 *correct matches. Right: SpongeBob. ASIFT (shown), SIFT, Harris-affine, Hessian-affine, and MSER find, respectively,* 370, 75, 8, 6, *and* 4 *correct matches.*

7. Conclusion. This paper has attempted to prove, by mathematical arguments, by a new algorithm, and by careful comparisons with state-of-the art algorithms, that a fully affine invariant image matching is possible. The proposed ASIFT image matching algorithm extends the SIFT method to a fully affine invariant device. It simulates the scale and the camera optical direction, and normalizes the rotation and the translation. The search for a full invariance was motivated by the existence of large transition tilts between two images taken from different viewpoints. As the tables of results showed, the notion of transition tilt has proved efficient to quantify the distortion between two images due to the viewpoint change, and also to give a fair and new evaluation criterion of the affine invariance of classic algorithms. In particular, SIFT and Hessian-affine are characterized by transition tilts of 2 and 2.5, respectively. In the case of MSER, however, the transition tilt varies strongly between 2 and 10, depending on image contrast and scale. ASIFT was shown to cope with transition tilts up to 36. Future research will focus on remaining challenges, such as the recognition under drastic illumination changes.

Appendix. Proof of Theorem 2.1.

Proof. Consider the real symmetric positive semidefinite matrix A^tA, where A^t denotes the transposed matrix of A. By classic spectral theory there is an orthogonal transform O such that $A^tA = ODO^t$, where D is a diagonal matrix with ordered eigenvalues $\lambda_1 \geq \lambda_2$. Set $O_1 = AOD^{-\frac{1}{2}}$. Then $O_1O_1^t = AOD^{-\frac{1}{2}}D^{-\frac{1}{2}}O^tA^t = AOD^{-1}O^tA^t = A(A^tA)^{-1}A^t = I$. Thus, there are orthogonal matrices O_1 and O such that

$$A = O_1 D^{\frac{1}{2}} O^t. \tag{A.1}$$

Since the determinant of A is positive, the product of the determinants of O and O_1 is positive. If both determinants are positive, then O and O_1 are rotations and we can write $A = R(\psi)DR(\phi)$. If ϕ is not in $[0, \pi)$, changing ϕ into $\phi - \pi$ and ψ into $\psi + \pi$ ensures

that $\phi \in [0, \pi)$. If the determinants of O and O_1 are both negative, replacing O and O_1, respectively, by $\left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right) O$ and $\left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right) O_1$ makes them into rotations without altering (A.1), and we can as above ensure $\phi \in [0, \pi)$ by adapting ϕ and ψ. The final decomposition is obtained by taking for λ the smaller eigenvalue of $D^{\frac{1}{2}}$. ■

REFERENCES

[1] A. Agarwala, M. Agrawala, M. Cohen, D. Salesin, and R. Szeliski, *Photographing long scenes with multi-viewpoint panoramas*, ACM Trans. Graph., 25 (2006), pp. 853–861.

[2] A. Ashbrook, N. Thacker, P. Rockett, and C. Brown, *Robust recognition of scaled shapes using pairwise geometric histograms*, in BMVC '95: Proceedings of the 6th British Conference on Machine Vision (Vol. 2), BMVA Press, Surrey, UK, 1995, pp. 503–512.

[3] A. Baumberg, *Reliable feature matching across widely separated views*, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, 2000, pp. 774–781.

[4] H. Bay, T. Tuytelaars, and L. Van Gool, *Surf: Speeded up robust features*, in Computer Vision—ECCV 2006, Springer-Verlag, Berlin, Heidelberg, 2006, pp. 404–417.

[5] S. Belongie, J. Malik, and J. Puzicha, *Shape matching and object recognition using shape contexts*, IEEE Trans. Pattern Anal. Mach. Intell., 24 (2002), pp. 509–522.

[6] M. Bennewitz, C. Stachniss, W. Burgard, and S. Behnke, *Metric localization with scale-invariant visual features using a single perspective camera*, in Proceedings of the European Robotics Symposium, Springer Tracts in Advanced Robotics (STAR) 22, Springer-Verlag, New York, 2006, pp. 143–157.

[7] M. Brown and D. Lowe, *Recognising panoramas*, in Proceedings of the 9th IEEE International Conference on Computer Vision, 2003, pp. 1218–1225.

[8] F. Cao, J.-L. Lisani, J.-M. Morel, P. Musé, and F. Sur, *A Theory of Shape Identification*, Springer-Verlag, New York, 2008.

[9] E. Chang, *EXTENT: Fusing context, content, and semantic ontology for photo annotation*, in Proceedings of the 2nd International Workshop on Computer Vision Meets Databases, ACM, New York, 2005, pp. 5–11.

[10] Q. Fan, K. Barnard, A. Amir, A. Efrat, and M. Lin, *Matching slides to presentation videos using SIFT and scene background matching*, in Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, 2006, pp. 239–248.

[11] O. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*, MIT Press, Cambridge, MA, 1993.

[12] L. Février, *A Wide-baseline Matching Library for Zeno*, Internship report, http://www.di.ens.fr/~fevrier/papers/2007-InternsipReportILM.pdf (2007).

[13] J. Foo and R. Sinha, *Pruning SIFT for scalable near-duplicate image matching*, in Proceedings of the 18th Conference on Australasian Database, Vol. 63, Australian Computer Society, Darlinghurst, Australia, 2007, pp. 63–71.

[14] G. Fritz, C. Seifert, M. Kumar, and L. Paletta, *Building detection from mobile imagery using informative SIFT descriptors*, in Proceedings of the 14th Scandinavian Conference on Image Analysis, Lecture Notes in Comput. Sci. 3540, Springer-Verlag, Berlin, pp. 629–638.

[15] I. Gordon and D. Lowe, *What and where: 3D object recognition with accurate pose*, in Toward Category-Level Object Recognition, Lecture Notes in Comput. Sci. 4170, Springer-Verlag, Berlin, 2006, pp. 67–82.

[16] J. Hare and P. Lewis, *Salient regions for query by image content*, in Image and Video Retrieval, Lecture Notes in Comput. Sci. 3115, Springer-Verlag, Berlin, 2004, pp. 317–325.

[17] C. Harris and M. Stephens, *A combined corner and edge detector*, in Proceedings of the Fourth Alvey Vision Conference, Vol. 15, 1988, pp. 147–151.

[18] T. Kadir, A. Zisserman, and M. Brady, *An affine invariant salient region detector*, in Proceedings of the 8th European Conference on Computer Vision, 2004, pp. 228–241.

[19] Y. Ke and R. Sukthankar, *PCA-SIFT: A more distinctive representation for local image descriptors*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, 2004, pp. 506–513.

[20] J. Kim, S. Seitz, and M. Agrawala, *Video-based document tracking: Unifying your physical and electronic desktops*, in Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology, 2004, pp. 99–107.
[21] B. Lee, W. Chen, and E. Chang, *Fotofiti: Web service for photo management*, in Proceedings of the 14th Annual ACM International Conference on Multimedia, 2006, pp. 485–486.
[22] H. Lejsek, F. Ásmundsson, B. Jónsson, and L. Amsaleg, *Scalability of local image descriptors: A comparative study*, in Proceedings of the 14th Annual ACM International Conference on Multimedia, 2006, pp. 589–598.
[23] T. Lindeberg, *Scale-space theory: A basic tool for analyzing structures at different scales*, J. Appl. Stat., 21 (1994), pp. 225–270.
[24] T. Lindeberg and J. Gårding, *Shape-adapted smoothing in estimation of 3-D depth cues from affine distortions of local 2-D brightness structure*, in Proceedings of the Third European Conference on Computer Vision, Springer-Verlag, New York, 1994, pp. 389–400.
[25] T. Lindeberg and J. Gårding, *Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure*, Image Vision Comput., 15 (1997), pp. 415–434.
[26] H. Ling and D. Jacobs, *Deformation invariant image matching*, in Proceedings of the Tenth IEEE International Conference on Computer Vision, 2005, pp. 1466–1473.
[27] H. Ling and K. Okada, *Diffusion distance for histogram comparison*, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006, pp. 246–253.
[28] D. Lowe, *Demo Software: SIFT Keypoint Detector*, http://www.cs.ubc.ca/~lowe/keypoints/ (July 2005).
[29] D. Lowe, *Distinctive image features from scale-invariant key points*, Int. J. Comput. Vis., 60 (2004), pp. 91–110.
[30] G. Loy and J. Eklundh, *Detecting symmetry and symmetric constellations of features*, in Proceedings of the 9th European Conference on Computer Vision (ECCV), Lecture Notes in Comput. Sci. 3952, Springer-Verlag, Berlin, 2006, pp. 508–521.
[31] J. Matas, O. Chum, M. Urban, and T. Pajdla, *WBS Image Matcher*, Online demo, http://cmp.felk.cvut.cz/~wbsdemo/demo/.
[32] J. Matas, O. Chum, M. Urban, and T. Pajdla, *Robust wide-baseline stereo from maximally stable extremal regions*, Image Vision Comput., 22 (2004), pp. 761–767.
[33] K. Mikolajczyk, *Affine Covariant Features*, http://www.robots.ox.ac.uk/~vgg/research/affine/ (2007).
[34] K. Mikolajczyk and C. Schmid, *Indexing based on scale invariant interest points*, in Proceedings of the 8th International Conference on Computer Vision (ICCV), 2001, pp. 525–531.
[35] K. Mikolajczyk and C. Schmid, *An affine invariant interest point detector*, in Proceedings of the Seventh European Conference on Computer Vision (ECCV), Springer-Verlag, London, 2002, pp. 128–142.
[36] K. Mikolajczyk and C. Schmid, *A performance evaluation of local descriptors*, in Proceedings of the International Conference on Computer Vision and Pattern Recognition, Vol. 2, 2003, pp. 257–263.
[37] K. Mikolajczyk and C. Schmid, *Scale and affine invariant interest point detectors*, Int. J. Comput. Vis., 60 (2004), pp. 63–86.
[38] K. Mikolajczyk and C. Schmid, *A performance evaluation of local descriptors*, IEEE Trans. Pattern Anal. Mach. Intell., (2005), pp. 1615–1630.
[39] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Gool, *A comparison of affine region detectors*, Int. J. Comput. Vis., 65 (2005), pp. 43–72.
[40] P. Moreels and P. Perona, *Common-frame model for object recognition*, in Proceedings of Neural Information Processing Systems, 2004.
[41] P. Moreels and P. Perona, *Evaluation of features detectors and descriptors based on 3D objects*, Int. J. Comput. Vis., 73 (2007), pp. 263–284.
[42] J. Morel and G. Yu, *On the Consistency of the SIFT Method*, Technical report, CMLA, ENS Cachan, Cachan, France, 2008.
[43] A. Murarka, J. Modayil, and B. Kuipers, *Building local safety maps for a wheelchair robot using vision and lasers*, in Proceedings of the 3rd Canadian Conference on Computer and Robot Vision, IEEE Computer Society, Washington, DC, 2006.
[44] P. Musé, F. Sur, F. Cao, and Y. Gousseau, *Unsupervised thresholds for shape matching*, in Proceedings of the International Conference on Image Processing, 2003, pp. 647–650.

[45] P. Musé, F. Sur, F. Cao, Y. Gousseau, and J. Morel, *An a contrario decision method for shape element recognition*, Int. J. Comput. Vis., 69 (2006), pp. 295–315.
[46] A. Negre, H. Tran, N. Gourier, D. Hall, A. Lux, and J. Crowley, *Comparative study of people detection in surveillance scenes*, in Structural, Syntactic, and Statistical Pattern Recognition, Lecture Notes in Comput. Sci. 4109, Springer-Verlag, Berlin, 2006, pp. 100–108.
[47] D. Nister and H. Stewenius, *Scalable recognition with a vocabulary tree*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 2161–2168.
[48] D. Pritchard and W. Heidrich, *Cloth motion capture*, Computer Graphics Forum, 22 (2003), pp. 263–271.
[49] F. Riggi, M. Toews, and T. Arbel, *Fundamental matrix estimation via TIP-transfer of invariant parameters*, in Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), IEEE Computer Society, Washington, DC, 2006, pp. 21–24.
[50] 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*, in Advances in Image and Video Technology, Lecture Notes in Comput. Sci. 4872, Springer-Verlag, Berlin, 2007, pp. 586–599.
[51] F. Schaffalitzky and A. Zisserman, *Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?,"* in Proceedings of the Seventh European Conference on Computer Vision (ECCV), Springer-Verlag, London, 2002, pp. 414–431.
[52] P. Scovanner, S. Ali, and M. Shah, *A 3-dimensional SIFT descriptor and its application to action recognition*, in Proceedings of the 15th International Conference on Multimedia, ACM, New York, 2007, pp. 357–360.
[53] S. Se, D. Lowe, and J. Little, *Vision-based mobile robot localization and mapping using scale-invariant features*, in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2001, pp. 2051–2058.
[54] S. Sinha, J. Frahm, M. Pollefeys, and Y. Genc, *GPU-based video feature tracking and matching*, in Proceedings of EDGE 2006, Workshop on Edge Computing Using New Commodity Architectures, 2006.
[55] N. Snavely, S. Seitz, and R. Szeliski, *Photo tourism: Exploring photo collections in 3D*, ACM Trans. Graphics, 25 (2006), pp. 835–846.
[56] T. Tuytelaars and L. Van Gool, *Wide baseline stereo matching based on local, affinely invariant regions*, in Proceedings of the British Machine Vision Conference, 2000, pp. 412–425.
[57] T. Tuytelaars and L. Van Gool, *Matching widely separated views based on affine invariant regions*, Int. J. Comput. Vis., 59 (2004), pp. 61–85.
[58] T. Tuytelaars and L. Van Gool, *Content-based image retrieval based on local affinely invariant regions*, in Proceedings of the Third International Conference on Visual Information and Information Systems, 1999, pp. 493–500.
[59] L. Vacchetti, V. Lepetit, and P. Fua, *Stable real-time 3D tracking using online and offline information*, IEEE Trans. Pattern Anal. Mach. Intell., 26 (2004), pp. 1385–1391.
[60] L. Van Gool, T. Moons, and D. Ungureanu, *Affine/photometric invariants for planar intensity patterns*, in Proceedings of the 4th European Conference on Computer Vision-Volume I, Springer-Verlag, London, 1996, pp. 642–651.
[61] M. Veloso, F. von Hundelshausen, and P. Rybski, *Learning visual object definitions by observing human activities*, in Proceedings of the IEEE-RAS International Conference on Humanoid Robots, 2005, pp. 148–153.
[62] M. Vergauwen and L. Van Gool, *Web-based 3D reconstruction service*, Mach. Vision Appl., 17 (2006), pp. 411–426.
[63] K. Yanai, *Image collector* III: *A web image-gathering system with bag-of-keypoints*, in Proceedings of the 16th International Conference on World Wide Web, ACM, New York, 2007, pp. 1295–1296.
[64] G. Yang, C. Stewart, M. Sofka, and C. Tsai, *Alignment of challenging image pairs: Refinement and region growing starting from a single keypoint correspondence*, IEEE Trans. Pattern Anal. Mach. Intell., 23 (2007), pp. 1973–1989.
[65] J. Yao and W. Cham, *Robust multi-view feature matching from multiple unordered views*, Pattern Recognition, 40 (2007), pp. 3081–3099.