A FULLY AFFINE INVARIANT IMAGE COMPARISON METHOD

Guoshen Yu

CMAP, Ecole Polytechnique,
91128 Palaiseau Cedex, France

*Jean-Michel Morel**

CMLA, ENS Cachan, 61 av. du
President Wilson, Cachan 94235, France

ABSTRACT

A fully affine invariant image comparison method, Affine-SIFT (ASIFT) is introduced. While SIFT is fully invariant with respect to only four parameters namely zoom, rotation and translation, the new method treats the two left over parameters : the angles defining the camera axis orientation. Against any prognosis, simulating all views depending on these two parameters is feasible. The method permits to reliably identify features that have undergone very large affine distortions measured by a new parameter, the *transition tilt*. State-of-the-art methods hardly exceed transition tilts of 2 (SIFT), 2.5 (Harris-Affine and Hessian-Affine) and 10 (MSER). ASIFT can handle transition tilts up 36 and higher (see Fig. 1).

Index Terms— image matching, affine invariance, scale invariance, affine normalization, SIFT.

1. INTRODUCTION

Local image detectors used for image comparison can be classified by their incremental invariance properties. All of them are translation invariant. The Harris point detector [3] is also rotation invariant. The Harris-Laplace, Hessian-Laplace and the DoG (Difference-of-Gaussian) region detectors [8, 10, 6, 2] are invariant to rotations and changes of scale. Some moment-based region detectors [5, 1] including the Harris-Affine and Hessian-Affine region detectors [9, 10], an edge-based region detector [17], an entropy-based region detector [4], and two level line-based region detectors MSER ("maximally stable extremal region") [7] and LLD ("level line descriptor") [15] are designed to be invariant to affine transformations. MSER, in particular, has been demonstrated to have often better performance than other affine invariant detectors, followed by Hessian-Affine and Harris-Affine [12, 8, 10]. These methods proceed by *normalizing* local patches, regions, or level lines that have undergone an unknown affine transform. Normalization transforms them into a standard object, where the effect of the affine transform has been eliminated. However, when a strong change of scale is present (in practice larger than 3), SIFT still beats all other methods [6]. Indeed, as proved mathematically [14], SIFT is fully scale invariant and, as pointed out in [6] none of the normalization methods is fully scale or affine invariant: *"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."*

*We thank ONR and CNES for their support.

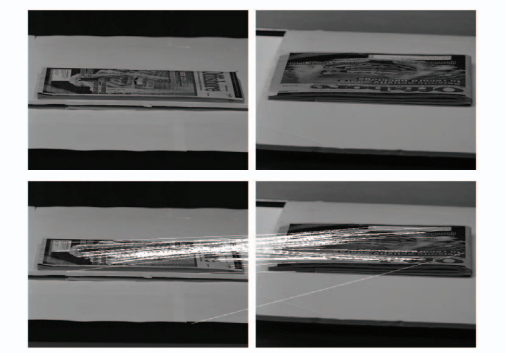

Fig. 1. Image pair with high transition tilt $t \approx 36$. Bottom: ASIFT finds 116 correct matches out of 120. SIFT, Harris-Affine, Hessian-Affine, and MSER fail completely.

2. THE AFFINE CAMERA MODEL

Image distortions arising from viewpoint changes can be locally modeled by affine planar transforms, provided the object's boundaries are piecewise smooth [12]. Thus, the (local) image deformation model under a camera motion is $u(x, y) \rightarrow u(ax + by + e, cx + dy + f)$, where $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is any linear planar map with positive determinant. Any such map has a

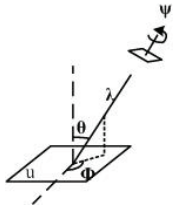

Fig. 2. Geometric interpretation of formula (1).

decomposition

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

which we note $A = \lambda R(\psi) T_t R(\phi)$, where $\lambda > 0$, λt is the determinant of A, $\phi \in [0, 180^\circ)$, $R(\psi)$ denotes the planar rotation with angle ψ, and T_t ($t \geq 1$) is called the *tilt*. Fig. 2 shows a camera motion interpretation of (1): ϕ and $\theta = \arccos 1/t$ are the camera viewpoint angles and ψ parameterizes the camera spin. In this affine model the camera stands far away from a planar object. Starting from a frontal position, a camera motion parallel to the object's plane induces an image translation. 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*. The tilt $t \geq 1$ is defined by $t \cos\theta = 1$. 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 short, (1) models the image deformation $\mathbf{u}(x, y) \rightarrow \mathbf{u}(A(x, y))$ induced by a camera motion from a frontal view $\lambda_0 = 1, t_0 = 1, \phi_0 = \psi_0 = 0$ to an oblique view characterized by λ, t, ϕ, and ψ.

3. HIGH TRANSITION TILTS

Equation (1) defines the *absolute* tilt, namely the image deformation ratio when the camera passes from a frontal view to an oblique view. But the compared images $\mathbf{u}_1(x, y) = \mathbf{u}(A(x, y))$ and $\mathbf{u}_2(x, y) = \mathbf{u}(B(x, y))$ are in general obtained from *two* oblique camera positions.

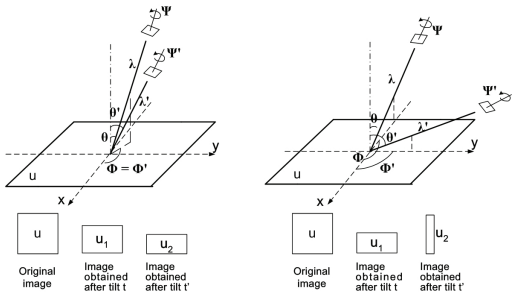

Fig. 3. Difference between *absolute* tilt and *transition* tilt. Left: longitudes $\phi = \phi'$, latitudes $\theta = 30°$, $\theta' = 60°$, absolute tilts $t = 1/\cos\theta = 2/\sqrt{3}$, $t' = 1/\cos\theta' = 2$, transition tilts $\tau(u_1, u_2) = t'/t = \sqrt{3}$. Right: $\phi = \phi' + 90°$, $\theta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

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* $BA^{-1} = H_\lambda R_1(\psi) T_\tau R_2(\phi)$*, with the notation of Formula* (1)*.*

Fig. 3 illustrates the affine transition between two images taken from different viewpoints, and in particular the difference between *absolute* tilt and *transition* tilt. With the two *absolute* tilts t and t' made in two orthogonal directions $\phi = \phi' + \pi/2$, one can verify that the *transition* tilt between $\mathbf{u}_1$ and $\mathbf{u}_2$ is the product $\tau = tt'$. Thus, *two moderate absolute tilts can lead to a large transition tilt!* Since in realistic cases the tilt can go up to 6 or even 8, it is easily understood that the *transition tilt* can go up to 36, 64, and more. Fig. 1 shows the ASIFT results for an image pair under orthogonal viewpoints (transition rotation $\phi = 90°$, absolute tilt $t \approx 6$) that leads to a transition tilt $\tau \approx 36$. This is not at all an exceptional situation. The relevance of the notion of transition tilt is corroborated by the fact that the highest transition tilt $\tau_{\max}$ permitting to match two images with absolute tilts t and t' is fairly independent from t and t'. It has been experimentally checked that SIFT works up to $\tau_{\max} \approx 2$. The attainable transition tilts for Harris-Affine and Hessian-Affine are close to 2.5. MSER is robust to transition tilts $\tau_{\max}$ between 5 and 10. But this performance is only verified when there is no substantial scale change between the images, and if the images contain highly contrasted objects. ASIFT attains regularly transition tilts larger than 36, and matches images beyond human performance (see Fig. 1).

4. THE ASIFT ALGORITHM

The idea of combining simulation and normalization is the main successful ingredient of the SIFT method. Indeed, scale changes amount to blur and cannot be normalized. Thus SIFT normalizes rotations and translations, but simulates all zooms out. David Pritchard's extension of SIFT [16] simulated four additional tilts. This is actually a first step toward the algorithm described below, which is also summarized in Fig. 4.

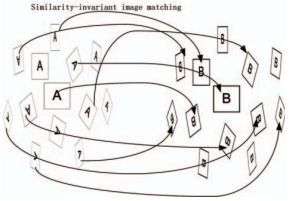

Fig. 4. Overview of ASIFT. Many pairs of rotated and tilted images obtained from images A and B are compared by SIFT.

1. Each image is transformed by simulating all possible linear distortions caused by the change of orientation of the camera axis. These distortions depend upon two parameters: the longitude ϕ and the latitude θ. The images undergo ϕ-rotations followed by tilts with parameter $t = |\frac{1}{\cos\theta}|$. For digital images, the tilt is performed as a 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 $c\sqrt{t^2 - 1}$, where $c = 0.8$ [14].
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 some scale invariant, rotation invariant, and translation invariant algorithm (typically SIFT). Since SIFT normalizes the translation of the camera parallel to its focal plane and the rotation of the camera around its optical axis, but simulates the scale change, all six camera parameters are either normalized or simulated by ASIFT.
4. The simulated latitudes θ correspond to tilts $t = 1$, a, a^2, ..., a^n, with $a > 1$. Taking $a = \sqrt{2}$ is a good compromise between accuracy and sparsity. The value n can go up to 5 or more. That way, all transition tilts from 1 to 32 and more are explored.
5. The longitudes ϕ follow for each t an arithmetic series 0, b/t, ..., kb/t where $b = 72°$ is a good compromise and k is the last integer such that $kb/t < 180°$.
6. Complexity: Each tilt is a t sub-sampling dividing the image area by t. The number of rotated images for each tilt is $(180/72)t = 2.5t$. Thus, *the method complexity is proportional to the number of tilts*. Controlling the total area of the simulated images is equivalent to controlling the algorithm complexity. Indeed, the SIFT search

time and memory size are proportional to the image area. This complexity can be further downgraded by a) subsampling the query and search images; b) identifying the successful pairs (t, ϕ); c) going back to the original resolution only for these pairs.

7. This description ends with a concrete example of how the multi-resolution search strategy can actually make the algorithm only twice slower than SIFT. Take $a = \sqrt{2}$, $n = 5$. The maximal absolute tilt for each image is 5.7 and the maximal transition tilt goes up to 32. The simulated image area is $5 \times 2.5 = 12.5$ times the original area. By a 3×3-subsampling of the original, this area is reduced to 1.4 times the one of the original image. If this reduction is applied to both the query and the search image, *the overall comparison complexity is equivalent to twice the SIFT complexity.* Fig. 5 shows the relatively sparse sampling of the longitude-latitude sphere needed to perform a fully affine recognition.

A mathematical proof that ASIFT is fully affine invariant (up to obvious precision issues) is given in [13].

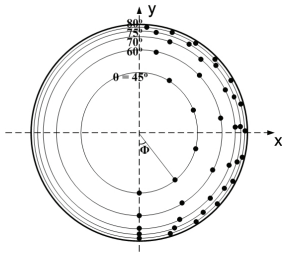

Fig. 5. Sampling (block dots) of the parameters $\theta = \arccos 1/t$ and ϕ in a zenith view of the observation half sphere.

5. EXPERIMENTS AND RESULTS

ASIFT is compared with the four state-of-the-art algorithms SIFT [6], Hessian-Affine, Harris-Affine [9, 10] and MSER [7] detectors, all coded with the SIFT descriptor [6]. The images used for the experiments are of size 600×450. [1]

Testing absolute tilts

Fig. 6 shows the setting adopted for evaluating the maximum absolute tilt and transition tilt attained by each algorithm. A magazine and a poster were photographed for the experiments. Unlike SIFT and ASIFT, the Hessian-Affine, Harris-Affine and MSER detectors are not robust to scale changes. Thus, to focus on tilts, the pairs of images under comparison were chosen free of scale changes. The poster shown in Fig. 7 was photographed with a reflex camera with viewpoint angles between the camera axis and the normal to the poster varying from $\theta = 0^\circ$ (frontal view) to $\theta = 80^\circ$. It seems physically unrealistic to insist on larger latitudes. Table 1 compares ASIFT with the performance of the other algorithms in terms of number of correct matches. One of these matching results is illustrated in Fig. 7. For these images SIFT works with angles smaller than 45°. The performance of Harris-Affine and Hessian-Affine plummets when the angle goes from 45 to 65°. Beyond this value, they fail completely. MSER struggles at the angle of 45° and fails at 65° degrees. ASIFT functions until 80°.

[1] A website with an online demo is available. 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 and more examples.

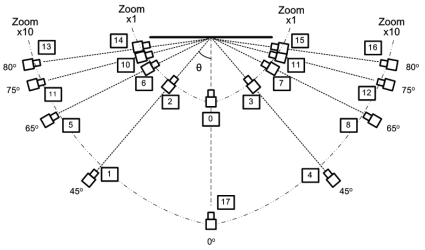

Fig. 6. Camera positions for systematic comparison.

θ/t	SIFT	HarAff	HesAff	MSER	ASIFT
80°/5.8	3	0	0	2	110
75°/3.9	2	1	0	4	152
65°/2.4	5	12	5	6	468
45°/1.4	171	54	26	15	707

Table 1. Absolute tilt invariance comparison for viewpoint angles between 45 and 80°. The latitude angles and the absolute tilts are listed in the left column.

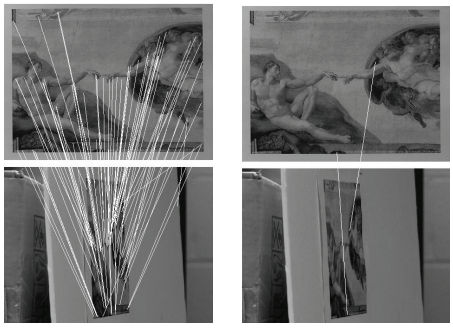

Fig. 7. Correspondences between the poster at frontal view and at 80° angle, absolute tilt $t = 5.8$. ASIFT (shown), SIFT, Harris-Affine, Hessian-Affine and MSER (shown) find respectively 110, 3, 0, 0 and 2 correct matches.

The above experiments and other many lead to the following conclusion for maximal absolute tilts. SIFT hardly exceeds a $t_{\max} = 2$ absolute tilt. The limit is $t_{\max} \approx 2.5$ for Harris-Affine and Hessian-Affine. The performance of MSER depends heavily on the type of image. For images with highly contrasted regions, MSER reaches an absolute tilt $t \approx 4$. However, if the images do not contain highly contrasted regions or if the scale change is larger than 3, the performance of MSER decays strongly, even under small tilts. For ASIFT, an absolute tilt of $t_{\max} \approx 5.8$ corresponding to the extreme viewpoint angle of 80° is always attained.

Transition Tilt Tests

Fig. 8 shows SIFT, Harris-Affine and Harris-Affine failing on a seemingly easy example. Indeed, the small absolute tilts $t_1 = t_2 = 2$ combined with the longitude angles $\phi_1 = 0^\circ$ and $\phi_2 = 50^\circ$ yield a moderate transition tilt $\tau \approx 3$ that is out of reach for these methods. ASIFT works perfectly. MSER works well under these optimal conditions: highly contrasted images and no scale change.

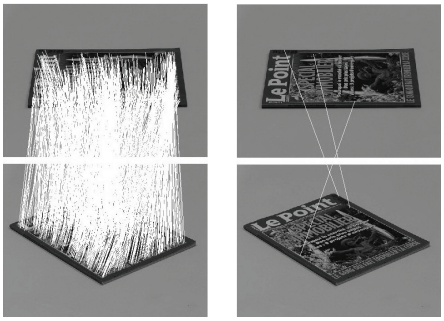

Fig. 8. Correspondences between the magazine images taken with absolute tilts $t_1 = t_2 = 2$ with longitude angles $\phi_1 = 0°$ and $\phi_2 = 50°$, transition tilt $\tau = 3$. ASIFT (shown), SIFT (shown), Harris-Affine, Hessian-Affine and MSER find respectively 745, 3, 1, 3, 87 correct matches.

Table 2 compares the performance of the algorithms for a set of magazine images that all have a $t = 4$ absolute tilt . The maximal transition tilt is therefore 16. For these images, SIFT, Harris-Affine and Hessian-Affine struggle with a 1.9 transition tilt. They fail completely over this value. MSER works stably up to a $\tau \approx 7.7$ transition tilt. Over this value, the number of correspondences is too small for reliable recognition. ASIFT works perfectly up to $\tau = 16$. As shown in Fig. 1, ASIFT actually attains transition tilts as large as 36.

Fig. 9 illustrates a round building. After a viewpoint change, the left and right sides sustain big transition tilts. ASIFT finds 123 correspondences covering the graffiti on all the left, central and right parts of the building. The other methods either fail or find a small number of matches in the central part.

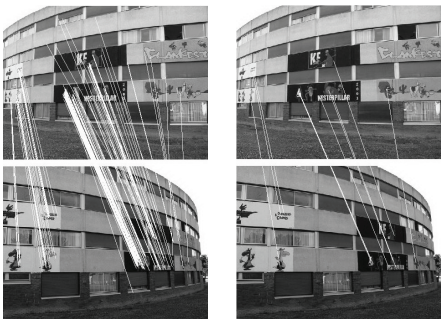

Fig. 9. Round building, transition tilt $\tau \in [1.8, \infty)$. ASIFT (shown), SIFT, Harris-Affine, Hessian-Affine and MSER (shown) find 123, 19, 5, 7 and 13 correct matches.

6. CONCLUSION

Fig. 10 shows a last image pair with moderate transition tilts where all methods fail except ASIFT. This is because *normalization* methods, ideal in principle, do not deal in practice correctly with small shapes, large absolute tilts, and low contrast. *Simulation* methods are by far more extensive. At first sight prohibitive, they turn out to be feasible, thanks to the very sparse sampling of the observation sphere shown in Fig. 5. The robustness of the SIFT method to moderate transition tilts is key to this sparse sampling.

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

Table 2. Transition tilt performance. For the first image $\phi_1 = 0°$ and for both images the absolute tilt is $t_1 = t_2 = 4$. The longitude of the second image ϕ_2 and the resulting transition tilt τ are given in the first column.

Fig. 10. Image matching: road signs. Transition tilt $\tau \approx 2.6$. ASIFT (shown), SIFT, Harris-Affine, Hessian-Affine and MSER find respectively 50, 0, 0, 0 and 1 correct matches.

7. REFERENCES

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

[2] L. Fevrier. A wide-baseline matching library for Zeno. *Technical report*, 2007.

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

[4] T. Kadir, A. Zisserman, and M. Brady. An Affine Invariant Salient Region Detector. *ECCV*, 228–241, 2004.

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

[6] D.G Lowe. Distinctive image features from scale-invariant key points. *IJCV*, 60(2):91–110, 2004.

[7] 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.

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

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

[10] K. Mikolajczyk and C. Schmid. Scale and Affine Invariant Interest Point Detectors. *IJCV*, 60(1):63–86, 2004.

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

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

[13] J.M. Morel and G. Yu. ASIFT: A New Framework for Fully Affine Invariant Image Comparison. to appear in SIAM Journal on Imaging Sciences, 2009.

[14] J.M. Morel and G. Yu. On the consistency of the SIFT method. to appear in Inverse Problems and Imaging (IPI), 2008.

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

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

[17] T. Tuytelaars and L. Van Gool. Matching Widely Separated Views Based on Affine Invariant Regions. *IJCV*, 59(1):61–85, 2004.