Stable Affine Frames on Isophotes

Michal Perďoch Jiří Matas Štěpán Obdržálek
Center for Machine Perception, CTU in Prague, Czech Republic
{perdom1,matas,xobdrzal}@cmp.felk.cvut.cz

Abstract

We propose a new affine-covariant feature, the Stable Affine Frame (SAF). SAFs lie on the boundary of extremal regions, i.e. *on isophotes. But instead of requiring the whole isophote to be stable with respect to intensity perturbation as in maximally stable extremal regions (MSERs), stability is required only locally, for the primitives constituting the three-point frames. The primitives are extracted by an affine invariant process that exploits properties of bitangents and algebraic moments. Thus, instead of using closed stable isophotes,* i.e. *MSERs, and detecting affine frames on them, stable affine frames are attempted, on all, even unstable, extremal regions.*

We show experimentally on standard datasets that SAFs have repeatability comparable to the best affine covariant detectors tested in the state-of-the-art report [11] and consistently produce a significantly higher number of features per image. Moreover, the features cover images more evenly than MSERs, which facilitates robustness to occlusion. Without significant computational effort, it is possible to modify the detector to construct stable homography covariant frames.

1. Introduction

Affine-covariant region detectors have been used in many computer vision applications including wide baseline matching [16, 19, 10], object recognition [6, 9, 13, 17], categorization [4, 5, 15], and panorama building [1]. In a recent study of covariant detectors [11], the one of the best repeatabilities and region accuracies was achieved by the maximally stable extremal region (MSER) detector [10]. However, the evaluation also revealed its weaknesses: a comparatively small number of detected regions and a high sensitivity to blur.

Boundaries of MSERs are a subset of isophotes (iso-intensity curves, intensity contours), namely those isophotes whose shape is globally stable with respect to intensity perturbation. In this paper, we propose to drop the global stability requirement. Instead, primitives are detected in an affine-invariant way on any *stable part of an isophote*. From the primitives, affine-covariant frames (local coordinate systems) are formed. The frames, consisting of ordered triplets of points, are more useful than affine covariant regions as they directly facilitate affine invariant description of the image signal without any further processing such as detection of dominant gradient directions.

Experiments conducted on standard datasets confirm that the stable affine frames (SAFs) computed on isophotes have repeatability comparable to the best affine covariant detectors and produce a higher number of corresponding features than detector tested in [11]. SAFs also perform well on blurred images, thus overcoming both above-mentioned weaknesses of MSERs. The observation that local affine frames can be detected on the boundaries of extremal regions due to their data dependent shape has been made before and state of the art results on two object recognition problems have been reported [13, 14]. We show that detecting stable frames on all isophotes produces significantly higher quality of output, in terms of repeatability, number of features and coverage of the image, than detecting frames on globally stable isophotes, *i.e.* MSERs.

Isophotes are a complete representation of the image and any image can be fully reconstructed from a set of isophotes [3]. This work is an attempt towards recovering all affine-covariant structures contained in the nested sets of isophotes. The computational cost of the process is not unacceptable; isophotes can be enumerated in real-time on current CPUs using an efficient union-find algorithm. Analysis of the isophotes takes seconds in our implementation, which is not prohibitive in most applications.

The idea of finding covariant frames – structures covariant with affine or perspective transformation on digital curves – is not new. Lamdan *et al.* [8] pro-

Authors were supported by Grant Agency of the Czech Technical University in Prague under project CTU 0706313 and by Czech Science Foundation project 201/06/1821.

posed construction of affine covariant frame on a contour. A fully perspective canonical frame construction was proposed by Rothwell *et al*. [18]. However, all early approaches required a contour to be extracted a priori, *e.g*. by an edge detector or by thresholding, with all the associated problems like parameter setting and linking errors. In our approach, we check all isophotes as an integral part of the process and output any stable feature. The need for prior segmentation is obviated. The hard detection decision is based on geometric stability w.r.t. photometric changes, a quantity that has direct relevance in many matching applications.

The rest of the paper is structured as follows. Section 2 presents the process SAF detection. First, necessary definitions are introduced in 2.1. Two covariant frame constructions are described in Section 2.2. The procedure for finding stable frames on isophotes is explained in Section 2.3. In the experimental Section 3, the SAF detector is compared to the other state of the art detectors of affine covariant regions. The paper is concluded in Section 4.

2. Stable Affine Frames on Isophotes

2.1. Definitions

In the continuous domain, an isophote is defined as a curve of constant intensity. We will adapt this concept for a discrete image using the framework of extremal regions [10]. Let *image* I be a mapping $I : \mathcal{D} \subset \mathbb{Z}^2 \rightarrow \mathcal{S}$. Extremal regions are well defined on images if

1. Set $\mathcal{S}$ is totally ordered, i.e. reflexive, antisymmetric and transitive binary relation $\leq$ exists. In this paper only finite set of intensities $S = \{0, 1, \ldots, N\}$ is considered.
2. A binary spatial adjacency (neighbourhood) relation $\diamond \subset \mathcal{D} \times \mathcal{D}$ is defined. In this paper 4-neighbourhoods are used, i.e. $p, q \in \mathcal{D}$ are adjacent ($p \diamond q$) iff $\sum_{i=1}^{2} |p_i - q_i| = 1$.

Let us identify an image I with a rectangular grid of pixels. Let $L = (V, E)$ be a planar graph where $V \subset \mathbb{Z}^2$ is a set of pixel corner coordinates in the grid and $E = \{\{a, b\} \colon a, b \in V, a \diamond b\}$ is a set of pixel edges. When refering to the pixel coordinates, we refer to its upper left corner. A *common edge* $e(p, q)$ of two neighbouring pixels $p, q \in \mathcal{D}$ with coordinates $p = (p_1, p_2), q = (q_1, q_2)$ is $e(p, q) = \{v_i, v_j\}$ such that

$$\begin{array}{llll} v_i = q, v_j = (q_1, q_2 + 1) & \text{if} & p_1 < q_1, p_2 = q_2, \\ v_i = p, v_j = (p_1, p_2 + 1) & \text{if} & p_1 > q_1, p_2 = q_2, \\ v_i = q, v_j = (q_1 + 1, q_2) & \text{if} & p_1 = q_1, p_2 < q_2, \\ v_i = p, v_j = (p_1 + 1, p_2) & \text{if} & p_1 = q_1, p_2 > q_2. \end{array} \quad (1)$$

Region $\mathcal{Q}$ is a subset of $\mathcal{D}$ such that for each $p, q \in \mathcal{Q}$ there is a sequence $p, a_1, a_2, \ldots, a_n, q$ and $p \diamond a_1, \ldots, a_i \diamond a_{i+1}, \ldots, a_n \diamond q$, i.e. region is a connected component in terms of adjacency relation $\diamond$.

Region Boundary. Let $\mathcal{Q}$ be a region and $L = (V, E)$ a rectangular grid of an image I. A simple closed path of vertices $\partial\mathcal{Q} = (v_1, \ldots, v_n)$, $e_i = \{v_i, v_{i+1}\} \in E$, is a *region boundary* iff for each $e_i = e(p, q) \in \partial\mathcal{Q}, p \in \mathcal{Q}, q \in \mathcal{D} \setminus \mathcal{Q}, q \diamond p$, i.e. region boundary consists of common edges between region pixels p and pixels q outside the region.

Extremal Region. Let $\mathcal{Q} \subset D$ be a region, and $\Omega_{\mathcal{Q}} = \{q : q \in \mathcal{D} \setminus \mathcal{Q}, \exists p \in \mathcal{Q}, p \diamond q\}$ set of pixels neighbouring with Q. We denote region $\mathcal{Q}$ an *extremal region* iff for all $p \in \mathcal{Q}, q \in \Omega_{\mathcal{Q}} : I(p) > I(q)$ (maximum intensity region) or $I(p) < I(q)$ (minimum intensity region). We denote a pair of extremal regions $\mathcal{Q}_1, \mathcal{Q}_2$ *nested extremal regions* iff $\mathcal{Q}_1 \subseteq \mathcal{Q}_2$.

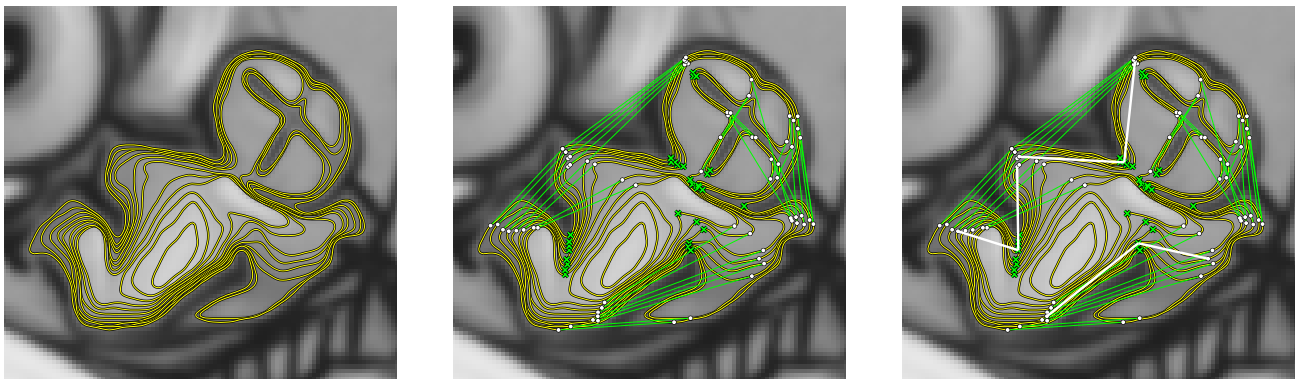
Outer Region Boundary. Let $\partial\mathcal{Q}$ be an extremal region boundary oriented (ordered) in a way that each pixel $p \in \mathcal{Q}$ is on the right-hand side of the path. $\partial\mathcal{Q}$ is an outer region boundary $B_{\mathcal{Q}}$ iff it has clockwise orientation in right-handed coordinate system. If $\partial\mathcal{Q}$ has counter-clockwise orientation we denote it a *region hole*.

Intensity Adjacency Relation. Let $\mathcal{Q}_1 \subset \mathcal{Q}_2$ be two nested extremal regions. $\mathcal{Q}_1, \mathcal{Q}_2$ are *intensity adjacent*, denoted $\mathcal{Q}_1 \lhd \mathcal{Q}_2$ iff $\nexists \mathcal{Q}, \mathcal{Q}_1 \subseteq \mathcal{Q} \subseteq \mathcal{Q}_2, \mathcal{Q} \neq \mathcal{Q}_1, \mathcal{Q} \neq \mathcal{Q}_2$. Thus there exists exactly one intensity value $s \in I(\mathcal{Q}_2), s \notin I(\mathcal{Q}_1)$, where $I(\mathcal{X}) \subset \mathcal{S}$ is set of intensities in region $\mathcal{X}$.

Discrete Isophote. Let $\mathcal{Q}$ be an extremal region. We denote its *outer region boundary* $B_{\mathcal{Q}}$ a discrete isophote. Thus, discrete isophotes are enumerated using an effective algorithm for enumeration of extremal regions introduced in [10].

2.2. Construction of Affine Frames

The geometric transformation between two corresponding planar patches acquired by a perspective camera can be locally approximated by an affine transformation [7]. A two-dimensional affine transformation possesses six degrees of freedom, six independent constraints are required to determine it. Numerous affine covariant frame constructions have been proposed in the literature [8, 18, 13]. We use only two constructions, one based on properties of bitangents, the second exploiting covariant properties of first and second algebraic moments. We chose the two constructions since (i) they typically cover different parts of a contour and are thus not redundant, (ii) together they provide more features than any method tested in [11] and (iii) one depends only on local properties of the



(a) (b) (c)

Figure 1. Example of stable affine frame construction: (a) each 10^{th} isophote on a part of an image, (b) entry and exit points (white) and the farthest point (green cross) from a bitangent (green hair lines) constructed on isophotes, (c) SAFs; white lines connecting points $(1,0)^T$, $(0,0)^T$ and $(0,1)^T$ in the frame coordinate system.

isophote whereas the other one on quantities derived from the whole isophote.

Detection of bitangent lines on concavities (see Fig. 1(b)), exploits the fact that affine transformation preserves tangency. First, curvature sign is estimated and inflection points located (the curvature sign of the curvature is preserved by affine transformations with positive determinant). A local descent to the nearest bitangent, as in [2], is employed to find all bitangents on the concavity. Each local concavity is endowed with entry, exit and farthest point from the bitangent.

Affine covariant frames are represented as a matrix of the affine transformation which maps points $(1,0)^\top, (0,0)^\top, (0,1)^\top$ from a normalized coordinate system into image coordinates. The construction produces three points p, q, r – entry, exit points and farthest point on a concavity. The affine transformation A representing this covariant frame is computed as

$$\mathsf{A} = \begin{pmatrix} \mathsf{M}_{21} & \mathsf{M}_{22} & \mathsf{M}_{23} \\ \mathsf{M}_{31} & \mathsf{M}_{32} & \mathsf{M}_{33} \\ 0 & 0 & 1 \end{pmatrix}^{-1}, \mathsf{M} = \begin{pmatrix} p_1 & q_1 & r_1 \\ p_2 & q_2 & r_2 \\ 1 & 1 & 1 \end{pmatrix}^{-1} \tag{2}$$

Note that the construction can be easily extended to efficiently compute a perspective covariant canonical frame of Rothwell *et al.* [18] where all steps, including selection of primitives, are performed invariantly to a homography.

Our second construction combines covariance matrix $\Sigma(\mathcal{Q})$ with two points – the centre of gravity $\mu(\mathcal{Q})$ of region $\mathcal{Q}$ and a point of extremal curvature q_κ

$$\mu(\mathcal{Q}) = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} q, \quad \Sigma(\mathcal{Q}) = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} (q-\mu)(q-\mu)^\top. \tag{3}$$

The center of gravity $\mu(\mathcal{Q})$ provides two constraints, *i.e.* resolving translation. The symmetric 2×2 matrix $\Sigma(\mathcal{Q})$ of second central algebraic moments gives three constraints. Together, the centre of gravity and the covariance matrix fix the affine transformation up to an unknown rotation. Normalization by the covariance matrix therefore allows affine-invariant measurement of distances, angles and curvatures.

To find local extrema of curvature q_κ we need to estimate curvature at each point of the discrete curve – isophote. Curvature estimation via curvature scale-space [12] is time consuming. Rather an approximation of the continuous curve is computed by smoothing the discrete curve with a Gaussian kernel, as in [20]

$$C_{\mathcal{Q}}(t) = B_{\mathcal{Q}}(t) * G(t,\sigma), \tag{4}$$

where $B_{\mathcal{Q}}$ is a discrete curve, t a curve parameter; σ is set to a small value to filter quantisation effects. The approximated continuous isophote $C_{\mathcal{Q}}$, is normalized using $\mu(\mathcal{Q})$ and $\Sigma(\mathcal{Q})$ to allow affine-invariant measurement of angles

$$N_{\mathcal{Q}}(t) = (C_{\mathcal{Q}}(t) - \mu(\mathcal{Q}))\Sigma^{-1/2}(\mathcal{Q}). \tag{5}$$

The local curvature $\kappa(t)$ is then computed from the angle between vectors $\mathbf{l}$, $\mathbf{r}$ of a fixed length cast in opposite directions from point $N_{\mathcal{Q}}(t)$ along the curve $N_{\mathcal{Q}}$. The curvature κ is estimated from the angle $\cos\alpha(t) = (l.r)/(|\,l||\,r|)$, as follows

$$\begin{aligned} \kappa(t) &= \delta(t)\frac{1+\cos\alpha(t)}{2}, \text{ where} \\ \delta(t) &= \begin{cases} 1 & \text{if } l_1 r_2 - l_2 r_1 > 0 \\ -1 & \text{otherwise} \end{cases} \end{aligned} \tag{6}$$

Each of the local maxima q_κ of $\kappa(t)$ provides the remaining constraint for the affine transformation and thus fixes one affine covariant frame.

In this construction the shape-normalizing transformation M does not scale the regions (we are ignoring one additional constraint – the scale of the covariance matrix)

$$\mathsf{M} = \frac{\Sigma^{1/2}(\mathcal{Q})}{\sqrt{|\Sigma(\mathcal{Q})|}}, \quad \text{and} \quad u = \mathsf{M}^{-1}(q_\kappa - \mu(\mathcal{Q})) \tag{7}$$

where $|\Sigma(\mathcal{Q})|$ is the determinant of $\Sigma(\mathcal{Q})$. The rotation angle $\phi = \mathrm{tg}^{-1}(u_2/u_1)$ is computed and the shape normalization matrix M is combined with the rotation and scaled by $|u|$:

$$\mathsf{N} = |u|\,\mathsf{M} \begin{pmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{pmatrix}. \tag{8}$$

Finally, the transformation A is computed

$$\mathsf{A} = \begin{pmatrix} \mathsf{N}_{11} & \mathsf{N}_{12} & p_1 \\ \mathsf{N}_{21} & \mathsf{N}_{22} & p_2 \\ 0 & 0 & 1 \end{pmatrix}. \tag{9}$$

2.3. Stability of the Affine Frames

To identify stable affine frames, we need to form sequences of related frames on intensity adjacent isophotes and to define similarity of affine covariant frames. From a feature matching perspective, the natural choice for a similarity measure would be a function of descriptors computed in the normalized frame coordinates such as correlation of normalized patches or distance of SIFT descriptors. However, evaluation of this function would be time consuming. Hence, we adopt the following "geometric" similarity.

Let $\mathsf{A}_1, \mathsf{A}_2$ be transformation matrices of two frames and $\mathcal{P} = \{(1,0)^\top, (0,0)^\top, (0,1)^\top\}$ be points in the normalized coordinate system. Similarity of the frames $d(\mathsf{A}_1, \mathsf{A}_2)$ is then

$$d(\mathsf{A}_1, \mathsf{A}_2) = \max_{p \in \mathcal{P}} \|p - \mathsf{A}_1^{-1}\mathsf{A}_2 p\|. \tag{10}$$

In other words, each point p is expressed in the other frame coordinates and the maximum distance over all points is the geometric similarity of the frames[1].

The *intensity adjacency relation*, the partial ordering introduced in section 2.1, allows sorting of isophotes in terms of spatial and intensity adjacency. Two extremal regions $\mathcal{Q}_1 \lhd \mathcal{Q}_2$ determining two isophotes $B_{\mathcal{Q}_1}, B_{\mathcal{Q}_2}$ are contained one in the other $\mathcal{Q}_1 \subset \mathcal{Q}_2$ and they differ exactly in pixels of one intensity. Let us now define a *sequence of corresponding affine frames*. Let extremal regions $\mathcal{Q}_1, \ldots, \mathcal{Q}_n$ be intensity adjacent $\mathcal{Q}_1 \lhd \mathcal{Q}_2, \mathcal{Q}_2 \lhd \mathcal{Q}_3, \ldots, \mathcal{Q}_{n-1} \lhd \mathcal{Q}_n$. Let $A_1, \ldots, A_n$ be local affine frames of one type of construction, constructed on isophotes $B_{\mathcal{Q}_1}, \ldots, B_{\mathcal{Q}_n}$ respectively. A sequence of affine frames $(\mathsf{A}_1, \ldots, \mathsf{A}_n)$ is a *sequence of corresponding affine frames* iff

$$\forall i, j, \mathcal{Q}_i \lhd \mathcal{Q}_j : d(\mathsf{A}_i, \mathsf{A}_j) < \theta_L, \tag{11}$$

where θ_L is the parameter of the method determining the maximum allowed displacement of pair of frames between two intensity adjacent isophotes.

The stability of an affine frame A_s is established as follows. The longest subsequence $(\mathsf{A}_k, \ldots, \mathsf{A}_x, \ldots, \mathsf{A}_l), 1 \le k < l \le n$ of corresponding frames $(\mathsf{A}_1, \ldots, \mathsf{A}_n)$ on intensity adjacent isophotes satisfying

$$\max_{i=k,\ldots,l} d(\mathsf{A}_i, \mathsf{A}_x) < \theta_S \tag{12}$$

is found. θ_S is the parameter of the method – spatial displacement allowed within subsequence $(\mathsf{A}_k, \ldots, \mathsf{A}_l)$. *Stability* of the frame A_x is defined as the length of the stable subsequence $S_x = l - k$. A frame is stable if $S_x > \Delta$, where Δ is the *stability threshold* and S_x is locally maximal in the sequence of corresponding frames. The sensitivity of the SAF detector is thus controlled by a single parameter Δ. For intensity perturbations of size Δ, positional variation of a stable frame is smaller than θ_S.

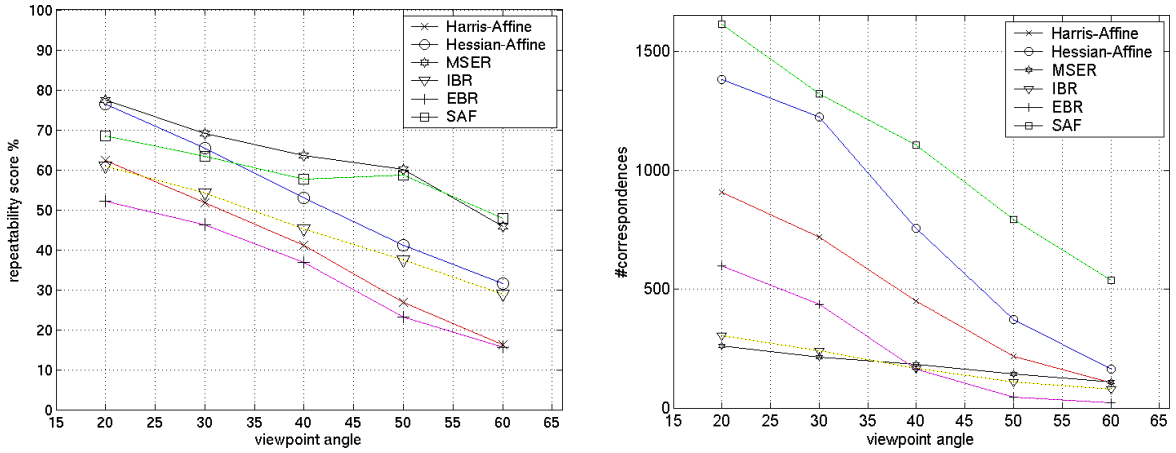
3. Experiments

The repeatability of the proposed SAF detector is evaluated in Section (3.1). The evaluation follows a standard protocol introduced in [11]. In Section 3.2, SAFs are compared with the MSER+LAF method [13] that computes local affine frames on MSERs. The test highlights the difference between detection of affine frames on stable isophotes and detection of stable frames on (possibly unstable) isophotes. We show that the "commutation" of the stability operation leads to higher repeatability for the same number of detected structures. Finally in Section (3.3), we show that SAFs are detected in areas where the MSER detector has no responses. The experiment demonstrates that areas with unstable isophotes (thus without MSERs), but with stable *parts* of isophotes are common – *e.g.* in natural scenes, blured images or images with smooth gradients. All experiments are conducted on standard, publicly available image sets[2] used in [11].

Parameter setting. In all experiments, parameters of the SAF detector were set as follows: the frame geometric localisation stability threshold $\theta_S = 0.25$ (Eq. 12) and inter-intensity frame similarity threshold $\theta_L = 1.2\,\theta_S$ (Eq. 11).

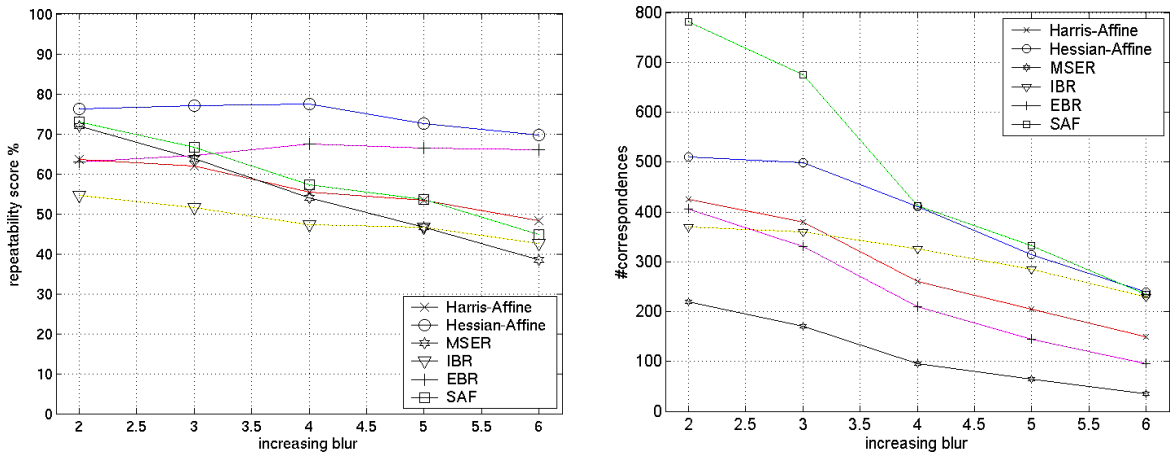
[1] We do not use the term "distance' as $d(\mathsf{A}_1, \mathsf{A}_2)$ is not symmetric.

[2] `http://www.robots.ox.ac.uk/~vgg/research/affine/`



(a) (b)

Figure 2. The GRAFFITI scene: (a) repeatability score ρ, (b) number of correspondences.



(a) (b)

Figure 3. The BIKES scene: (a) repeatability score ρ, (b) number of correspondences.

3.1. Repeatability Evaluation

To allow comparison of performance of the detector with the state of the art methods, we used the test protocol proposed by Mikolajczyk *et al.* [11]. The test sets contain sequences of six images with increasing effects of given transformation (viewpoint change GRAFFITI, scale change BOAT, blur BIKES, intensity change CARS and compression UBC). A groundtruth homography transformations are provided between first image of the sequence – *reference image* and other five *test images*. Mikolajczyk *et al.* defined a repeatability score based on *overlap error* ϵ

$$\epsilon(R_{\mathsf{E}_1}, R_{\mathsf{E}_2}) = 1 - \frac{R_{\mathsf{E}_1} \cap R_{H_{12}^{-\top}\mathsf{E}_2 H_{12}^{-1}}}{R_{\mathsf{E}_1} \cup R_{H_{12}^{-\top}\mathsf{E}_2 H_{12}^{-1}}}, \qquad (13)$$

where R_{E} represents the elliptic region defined by $x^\top R_{\mathsf{E}} x = 1$ and H_{12} is a homography between reference and test image. To compensate for different sizes of regions from different detectors, a scale factor is computed, that transforms region R_{E_1} it into normalized size (equivalent to radius 30 pixels). Region R_{E_2} is rescaled using the same factor before evaluating the overlap error. One-to-one correspondences are established by finding minimum overlap errors between reference and test image regions and pairs with $\epsilon < 0.6$ are kept. The *repeatability score* ρ is then the ratio between number of corresponding pairs and the number of all regions detected in common part of the scene *i.e.* the part visible in both images. Two factors are evaluated: (i) the repeatability score, which estimates the probability that a region will be detected in both images and (ii) the *number of correspondences* which is related to region density. Both factors are important and closely related, setting sharper thresholds for many detectors results in smaller number of features and higher repeatability and vice versa lower thresh-

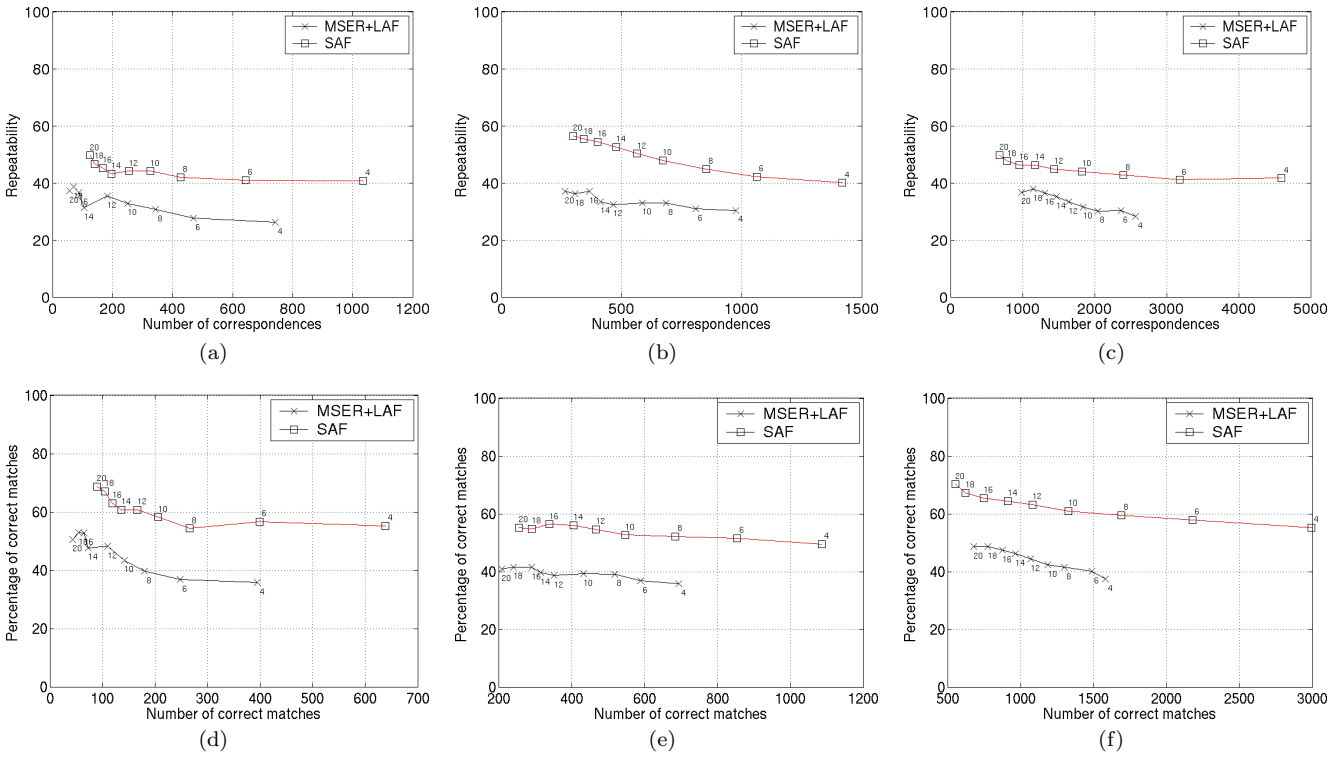


Figure 4. Repeatability vs. number of correspondences for different stability thresholds: (a) BIKES, (b) GRAFITTI, (c) BOAT. Percentage of correct matches in all tentative matches vs. number of correct matches: (d) BIKES, (e) GRAFITTI, (f) BOAT.

Scene	SAF	MSER	HesAff	HarAff	IBR	EBR
Grafitti	2	1	3	5	4	6
Bikes	3	5	1	4	2	6
Boat	3	2	1	4	6	5
Cars	2	1	3	4	6	5

Table 1. Ranking of affine detectors[11] on standard scenes[2] based on the repeatability score.

Scene	SAF	MSER	HesAff	HarAff	IBR	EBR
Grafitti	1	4	2	3	6	5
Bikes	1	6	2	4	3	5
Boat	1	5	2	3	6	4
Cars	1	5	2	3	6	4

Table 2. Ranking of affine detectors[11] on standard scenes[2] based on number of correspondences.

olds provide more features but the repeatability often drops as more spurious detections are obtained [11].

The stability threshold Δ of the proposed SAF detector was set to $\Delta = 10$. With this setting, the number of features and the repeatability score is comparable to other feature detectors. On the GRAFFITI scene Fig. 2, the SAF detector outperforms most of the state of the art methods both in terms of repeatability and number of corresponding features. On the blur scene - BIKES, the transformation, the MSER method is sensitive to, SAF detector performs reasonably well in terms of repeatability score and more important provide significantly higher number of corresponding features.

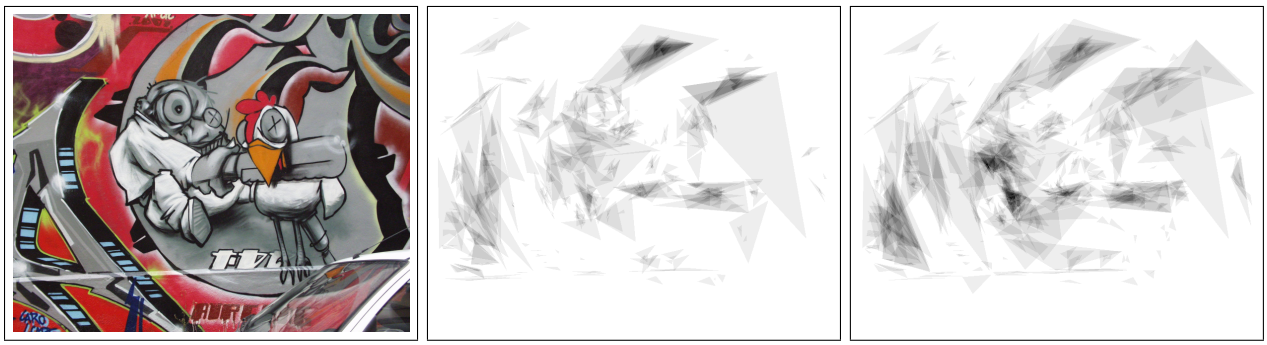
Results on some other scenes from [11] are compared in Tables 1 and 2. The rank is the average rank of a detector on all image pairs of given sequence. Table 1 shows that proposed detector delivers repeatability score comparable to the state of the art methods. Moreover, Table 2 shows that the SAF detector provides the highest number of correspondences on all scenes, while maintaining comparable repeatability score.

3.2. Comparison with the MSER+LAF method[13]

Although the method for evaluating repeatability used in previous experiment is well established for comparing affine regions, fully affine structures do not fit well into the framework. Mikolajczyk *et al.* [11] characterize affine covariant regions by the center of gravity and an ellipse, which fix only five degrees of freedom of the affine transformation. Both SAFs and MSER+LAF detectors provide full affine covariant frames, *i.e.* they resolve also the "orientation" of the ellipse. The computation of repeatability score was thus modified as follows. Let $\mathsf{A}_1, \mathsf{A}_2$ be transformation matrices of two frames and H_{12} the groundtruth homography matrix. Let $\mathcal{P} = \{(1,0)^\top, (0,0)^\top, (0,1)^\top\}$ be points in the normalized coordinate system. "Overlap

(a) (b) (c)

Figure 5. Image coverage, BIKES scene images 1 and 4. White areas are not covered, darker areas are covered with higher number of frames. (a) reference image 1, (b) MSER+LAF method: 251 repeated (32.85% of detected) frames, (c) SAF method: 319 repeated (44.74% of detected) frames.



(a) (b) (c)

Figure 6. Image coverage, GRAFFITI scene images 1 and 5. White areas are not covered, darker areas are covered with higher number of frames. (a) reference image 1, (b) MSER+LAF method: 586 repeated (33.01% of detected) frames, (c) SAF method: 665 repeated (47.94% of detected) frames.

error" of the frames $\delta(\mathsf{A}_1, \mathsf{A}_2)$ is then

$$\delta(\mathsf{A}_1, \mathsf{A}_2) = \max_{p \in \mathcal{P}} \|p - \mathsf{A}_1^{-1}\mathsf{H}_{12}^{-1}\mathsf{A}_2 p\|. \qquad (14)$$

Two frames are considered corresponding if their "overlap error" $\delta(\mathsf{A}_1, \mathsf{A}_2) < 0.3$. Both compared detectors produce structures of similar size and thus there is no reason for the rescaling to a common region size. Frames computed using different constructions do not interfere. One-to-one frame correspondences are therefore computed separately for each type of frame construction.

The detectors are compared on three pairs of images selected from the defocused blur sequence BIKES (image 1 vs. 4), the viewpoint change sequence GRAFFITI (1 vs. 5), and the scale change sequence BOAT (1 vs. 3). In Fig. 4, the trade-off between the number of correspondences and repeatability for different stability thresholds $\Delta \in \{4, 6, 8, \ldots, 20\}$ is compared. Results clearly show the substantial improvement of repeatability of SAFs w.r.t. MSER+LAF.

To verify the performance in a matching experiment, we adopted the descriptor used in [14]. The descriptors were computed on square spanning from $(-1, -1)^\top$ to $(2, 2)^\top$ in the frame coordinate system. Tentative matches were established as mutually nearest neighbours and verified by the groundtruth homography, *i.e.* a pair of frames $\mathsf{A}_1, \mathsf{A}_2$ is a correct match if $\delta(\mathsf{A}_1, \mathsf{A}_2) < 0.3$. Hence for each threshold, we obtained the number of correct matches and the ratio of the number of correct to the number of all tentative matches. Fig. 4(d-f) shows that descriptors computed on SAFs are more discriminative and more reliable than descriptors computed on MSER+LAF frames.

3.3. Image Coverage

Previous experiments focus on evaluation of repeatability and the number of responses of affine covariant detectors. However, there are other important properties of a detector, such as speed and the property we call *coverage* – the spatial distribution of detector responses. Computational complexity of the SAF algorithm is higher than of MSER+LAF method, e.g. on the images of the BIKES scene (800x600) SAF detector runs about five seconds vs. 600ms for MSER+LAF

method. On the otherside, features of a fast detector are fruitless if all responses are concentrated in a small area. This is a well known problem encountered in tracking and narrow-baseline stereo applications. There are many scenes where almost all features are located on a prominent textured object, such as a tree or a bush. Parts of the image without any detected features are "invisible" to the higher level algorithms. We are not aware of a qualitative test for the coverage – our evaluation is therefore rather informal.

The test is carried out on two image pairs (from experiment 3.2): BIKES and GRAFFITI. Affine frames are detected using MSER+LAF and the SAF methods. Frames detected on the reference image that have a correspondence in the test image were identified using known groundtruth homography as in 3.2. Each frame that is detected in both images is visualised in the reference image *i.e.* it increases all pixel values inside the triangle it covers by 1. The final "coverage" is shown in Figures 5 and 6 for MSER+LAF (b) and SAFs (c). We see clearly that uncovered area is much smaller for the SAF method than for the MSER+LAF method.

4. Conclusions

A new affine-covariant detector of the Stable Affine Frame (SAF) was proposed. We showed how to construct SAFs on isophotes by maximisation of their geometric and photometric stability. Instead of using only stable isophotes (MSERs), stable affine frames are sought on all, even unstable, isophotes.

We showed experimentally on standard data that SAFs have a repeatability comparable to the best affine covariant detectors [11] and consistently produce a significantly higher number of features per image. Compared with MSERs, SAF have the following two advantages: (i) they cover images more evenly, which might positively affect robustness to occlusion and precision of multi-view geometry estimation, and (ii) perform well on images that are blurred. Overall, SAFs provide a strong alternative to MSERs (combined with local affine frame constructions) in applications where the longer running time is not an issue.

References

[1] M. Brown and D. Lowe. Recognising panoramas. In *ICCV*. IEEE, 2003.

[2] D. Buesching. Efficiently finding bitangents. In *ICPR*. IEEE, 1996.

[3] V. Caselles, B. Coll, and J.-M. Morel. Topographic maps and local contrast changes in natural images. *IJCV*, 33(1):5–27, 1999.

[4] G. Dorko and C. Schmid. Selection of scale-invariant parts for object class recognition. In *ICCV03*, pages 634–640, 2003.

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, pages 264–271, 2003.

[6] V. Ferrari, T. Tuytelaars, and L. V. Gool. Simultaneous object recognition and segmentation by image exploration. In *ECCV*, pages 40–54, 2004.

[7] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge, 2003.

[8] Y. Lamdan, J. Schwartz, and H. Wolfson. Object recognition by affine invariant matching. In *CVPR*, pages 335–344, 1988.

[9] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[10] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, pages 384–393, 2002.

[11] 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(1-2):43–72, 2005.

[12] F. Mokhtarian and A. K. Mackworth. A theory of multiscale, curvature-based shape representation for planar curves. *PAMI*, 14(8):789–805, 1992.

[13] S. Obdrzalek and J. Matas. Object recognition using local affine frames on distinguished regions. In *BMVC*, pages 113–122, 2002.

[14] S. Obdrzalek and J. Matas. Sub-linear indexing for large scale object recognition. In *BMVC*, 2005.

[15] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *ECCV*, 2004.

[16] P. Pritchett and A. Zisserman. Wide baseline stereo matching. In *ICCV*, pages 754–760, 1998.

[17] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In *CVPR*, 2003.

[18] C. Rothwell, A. Zisserman, D. Forsyth, and J. Mundy. Canonical frames for planar object recognition. In *ECCV*, LNCS 588. Springer, 1992.

[19] T. Tuytelaars and L. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *BMVC*, 2000.

[20] D. Wuescher and K. Boyer. Robust contour decomposition using a constant curvature criterion. *PAMI*, 13(1):41–51, 1991.