An Affine Invariant Salient Region Detector

Timor Kadir, Andrew Zisserman, and Michael Brady

Department of Engineering Science,
University of Oxford,
Oxford, UK.
{timork,az,jmb}@robots.ox.ac.uk

Abstract. In this paper we describe a novel technique for detecting salient regions in an image. The detector is a generalization to affine invariance of the method introduced by Kadir and Brady [10]. The detector deems a region salient if it exhibits unpredictability in both its attributes and its spatial scale.
The detector has significantly different properties to operators based on kernel convolution, and we examine three aspects of its behaviour: invariance to viewpoint change; insensitivity to image perturbations; and repeatability under intra-class variation. Previous work has, on the whole, concentrated on viewpoint invariance. A second contribution of this paper is to propose a performance test for evaluating the two other aspects. We compare the performance of the saliency detector to other standard detectors including an affine invariance interest point detector. It is demonstrated that the saliency detector has comparable viewpoint invariance performance, but superior insensitivity to perturbations and intra-class variation performance for images of certain object classes.

1 Introduction

The selection of a set of image regions forms the first step in many computer vision algorithms, for example for computing image correspondences [2,17,19,20, 22], or for learning object categories [1,3,4,23]. Two key issues face the algorithm designer: the subset of the image selected for subsequent analysis and the representation of the subset. In this paper we concentrate on the first of these issues. The optimal choice for region selection depends on the application. However, there are three broad classes of image change under which good performance may be required:
1. Global transformations. Features should be repeatable across the expected class of global image transformations. These include both geometric and photometric transformations that arise due to changes in the imaging conditions. For example, region detection should be covariant with viewpoint as illustrated in Figure 1. In short, we require the segmentation to commute with viewpoint change.
2. Local perturbations. Features should be insensitive to classes of semi-local image disturbances. For example, a feature responding to the eye of a human face should be unaffected by any motion of the mouth. A second class of disturbance is

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3021, pp. 228–241, 2004.

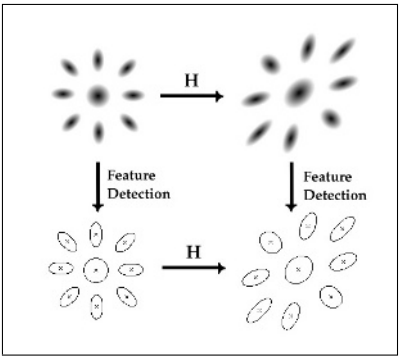


Fig. 1. Detected regions, illustrated by a centre point and boundary, should commute with viewpoint change – here represented by the transformation H.

where a region neighbours a foreground/background boundary. The detector can be required to detect the foreground region despite changes in the background.
3. Intra-class variations. Features should capture corresponding object parts under intra-class variations in objects. For example, the headlight of a car for different brands of car (imaged from the same viewpoint).

In this paper we make two contributions. First, in Section 2 we describe extensions to the region detector developed by Kadir and Brady [10]. The extensions include covariance to affine transformations (the first of the requirements above), and an improved implementation which takes account of anti-aliasing. The performance of the affine covariant region detector is assessed in Section 3 on standard test images, and compared to other state of the art detectors.

The second contribution is in specifying a performance measure for the two other requirements above, namely tolerance to local image perturbations and to intra-class variation. This measure is described in Section 4 and, again, performance is compared against other standard region operators.

Previous methods of region detection have largely concentrated on the first requirement. So-called corner features or interest points have had wide application for matching and recognition [7,21]. Recently, inspired by the pioneering work of Lindeberg [14], scale and affine adapted versions have been developed [2,18,19,20]. Such methods have proved to be robust to significant variations in viewpoint. However, they operate with relatively large support regions and are potentially susceptible to semi-local variations in the image; for example, movements of objects in a scene. They fail on criterion 2.

Moreover, such methods adopt a relatively narrow definition of saliency and scale; scale is usually defined with respect to a convolution kernel (typically a Gaussian) and saliency to an extremum in filter response. While it is certainly the case that there are many useful image features that can be defined in such a manner, efforts to generalise such methods to capture a broader range of salient image regions have had limited success.

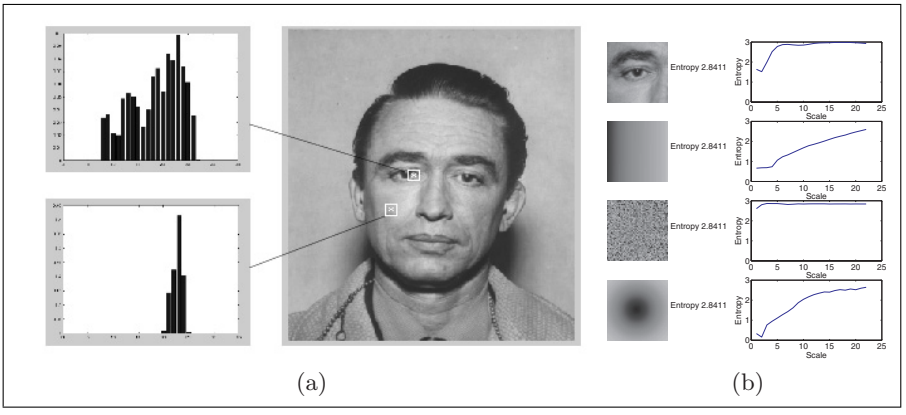


Fig. 2. (a) Complex regions, such as the eye, exhibit unpredictable local intensity hence high entropy. Image from NIST Special Database 18, Mugshot Identification Database. However, entropy is invariant to permutations of the local patch (b).

Other methods have extracted affine covariant regions by analysing the image isocontours directly [17,22] in a manner akin to watershed segmentation. Related methods have been used previously to extract features from mammograms [13]. Such methods have the advantage that they do not rely on excessive smoothing of the image and hence capture precise object boundaries. Scale here is defined in terms of the image isocontours rather than with respect to a convolution kernel or sampling window.

2 Information Theoretic Saliency

In this section we describe the saliency region detector. First, we review the approach of Kadir and Brady [10], then in Section 2.2 we extend the method to be affine invariant, and give implementation details in Sections 2.3 and 2.4.

2.1 Similarity Invariant Saliency

The key principle underlying the Kadir and Brady approach [10] is that salient image regions exhibit unpredictability, or 'surprise', in their local attributes *and* over spatial scale. The method consists of three steps: I. Calculation of Shannon entropy of local image attributes (e.g. intensity or colour) over a range of scales — $\mathcal{H}_D(s)$; II. Select scales at which the entropy over scale function exhibits a peak — $\mathbf{s_p}$; III. Calculate the magnitude change of the PDF as a function of scale at each peak — $\mathcal{W}_D(s)$. The final saliency is the product of $\mathcal{H}_D(s)$ and $\mathcal{W}_D(s)$ at each peak. The histogram of pixel values within a circular window of radius s, is used as an estimate of the local PDF. Steps I and III measure the feature-space and the inter-scale predictability respectively, while step II selects optimal scales. We discuss each of these steps next.

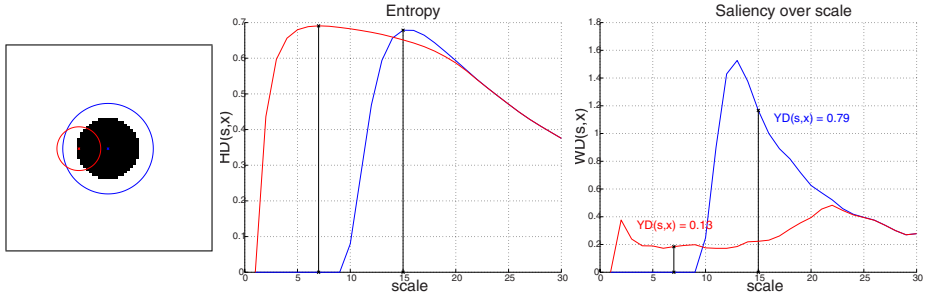


Fig. 3. The two entropy peaks shown in (a) correspond to the centre (in blue) and edge (in red) points in top image. Both peaks occur at similar magnitudes.

The entropy of local attributes measures the predictability of a region with respect to an assumed model of simplicity. In the case of entropy of pixel intensities, the model of simplicity corresponds to a piecewise constant region. For example, in Figure 2(a), at the particular scales shown, the PDF of intensities in the cheek region is peaked. This indicates that most of these pixels are highly predictable, hence entropy is low. However, the PDF in the eye region is flatter which indicates that here, pixel values are highly unpredictable and this corresponds to high entropy.

In step II, scales are selected at which the entropy is peaked. Through searching for such extrema, the feature-space saliency is locally optimised. Moreover, since entropy is maximised when the PDF is flat, i.e. all present attribute values are in equal proportion, such peaks typically occur at scales where the statistics of two (or more) different pixel populations contribute equally to the PDF estimate. Figure 3(b) shows entropy as a function of scale for two points in Figure 3(a). The peaks in entropy occur at scales for which there are equal proportions of black and white pixels present. These significant, or salient scales, in the entropy function (analogous to the 'critical-points' in Gaussian scale-space [11,15]) serve as useful reference points since they are covariant with isotropic scaling, invariant to rotation and translation, and robust to small affine shears.

Note however, that the peaks for both points in Figure 3(b) attain an almost identical magnitude. This is to be expected since both patches contain almost identical proportions of black and white pixels. In fact, since histogramming destroys all local ordering information all permutations of the local patch do not affect its entropy. Figure 2(b) shows the entropy over scale function for an image patch taken from 2(a) and three permutations of its pixels: a linear ramp, a random reordering and a radial gradient. The entropy at the maximum scale (that of the whole patch) is the same for all permutations. However, the shape of the entropy function is quite different for each case.

The role of Step III, the inter-scale unpredictability measure $\mathcal{W}_D$, is to weight the entropy value such that some permutations are preferred over others. It is

defined as the magnitude change of the PDF as a function of scale, therefore those orderings that are statistically self-dissimilar over scale are ranked higher than those that exhibit stationarity.

Figure 3(c) shows $\mathcal{W}_D$ as a function of scale. It can be seen that the plot corresponding to the edge point has a much lower value than the one for the centre point at the selected scale value. In essence, it is a normalised measure of scale localisation. For example, in a noise image the pixel values are highly unpredictable at any one scale but over scale the statistics are stationary. However, a noise patch against a plain background would be salient due to the change in statistics.

In the continuous case, the saliency measure $\mathcal{Y}_D$, a function of scale s and position $\mathbf{x}$, is defined as:

$$\mathcal{Y}_D(\mathbf{s_p}, \mathbf{x}) \triangleq \mathcal{H}_D(\mathbf{s_p}, \mathbf{x})\, \mathcal{W}_D(\mathbf{s_p}, \mathbf{x}) \tag{1}$$

i.e. for each point $\mathbf{x}$ the set of scales $\mathbf{s_p}$, at which entropy peaks, is obtained, then the saliency is determined by weighting the entropy at these scales by $\mathcal{W}_D$. Entropy, $\mathcal{H}_D$, is given by:

$$\text{I} \qquad \mathcal{H}_D(s, \mathbf{x}) \triangleq -\int p(I, s, \mathbf{x}) \log_2 p(I, s, \mathbf{x})\, \mathrm{d}I \tag{2}$$

where $p(I, s, \mathbf{x})$ is the probability density of the intensity I as a function of scale s and position $\mathbf{x}$. The set of scales $\mathbf{s_p}$ is defined by:

$$\text{II} \qquad \mathbf{s_p} \triangleq \left\{ s : \frac{\partial \mathcal{H}_D(s, \mathbf{x})}{\partial s} = 0, \frac{\partial^2 \mathcal{H}_D(s, \mathbf{x})}{\partial s^2} < 0 \right\} \tag{3}$$

The inter-scale saliency measure, $\mathcal{W}_D(s, \mathbf{x})$, is defined by:

$$\text{III} \qquad \mathcal{W}_D(s, \mathbf{x}) \triangleq s \int \left| \frac{\partial}{\partial s} p(I, s, \mathbf{x}) \right| \mathrm{d}I \tag{4}$$

In this paper, entropy is measured for the grey level image intensity but other attributes, e.g. colour or orientation, may be used instead; see [8] for examples.

This approach has a number of attractive properties. It offers a more general model of feature saliency and scale compared to conventional feature detection techniques. Saliency is defined in terms of spatial unpredictability; scale by the sampling window and its parameterisation. For example, a blob detector implemented using a convolution of multiple scale Laplacian-of-Gaussian (LoG) functions [14], whilst responding to a number of different feature shapes, maximally responds only to LoG function itself (or its inverse); in other words, it acts as a matched filter[1]. Many convolution based approaches to feature detection exhibit the same bias, i.e. a preference towards certain features. This specificity has a detrimental effect on the quality of the features and scales selected. In

[1] This property is somewhat alleviated by the tendency of blurring to smooth image structures into LoG like functions.

contrast, the saliency approach responds equally to the LoG and all other permutations of its pixels provided that the constraint on $\mathcal{W}_D(s)$ is satisfied. This property enables the method to perform well over intra-class variations as is demonstrated in Section 4.

2.2 Affine Invariant Saliency

In the original formulation of [10], the method was invariant to the similarity group of geometric transformations and to photometric shifts. In this section, we develop the method to be fully affine invariant to geometric transformations. In principle, the modification is quite straightforward and may be achieved by replacing the circular sampling window by an ellipse: under an affine transformation, circles map onto ellipses. The scale parameter s is replaced by a vector $\mathbf{s} = (s, \rho, \theta)$, where ρ is the axis ratio and θ the orientation of the ellipse. Under such a scheme, the major and minor axes of the ellipse are given by $s/\sqrt{\rho}$ and $s\sqrt{\rho}$ respectively.

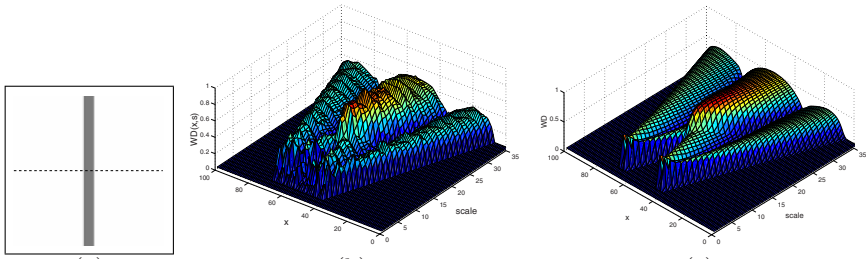
Increasing the dimensionality of the sampling window creates the possibility of degenerate cases. For example, in the case of a dark circle against a white background (see Figure 3(a)) any elliptical sampling window that contains an equal number of black and white pixels ($\mathcal{H}_D$ constraint) but does not exclude any black pixels at the previous scale ($\mathcal{W}_D$ constraint) will be considered equally salient. Such cases are avoided by requiring that the inter-scale saliency, $\mathcal{W}_D$, is smooth across a number of scales. A simple way to achieve this is to apply a 3-tap averaging filter to $\mathcal{W}_D$ over scale.

2.3 Local Search

The complexity of a full search can be significantly reduced by adopting a local strategy in the spirit of [2,19,20]. Our approach is to start the search only at seeds points (positions and scales) found by applying the original similarity invariant search. Each seed circle is then locally adapted in order to maximise two criteria, $\mathcal{H}_D$ (entropy) and $\mathcal{W}_D$ (inter-scale saliency). $\mathcal{W}_D$ is maximised when the ratio and orientation match that of the local image patch [9] at the correct scale, defined by a peak in $\mathcal{H}_D$. Therefore, we adopt an iterative refinement approach. The ratio and orientation are adjusted in order to maximise $\mathcal{W}_D$, then the scale is adjusted such that $\mathcal{H}_D$ is peaked. The search is stopped when neither the scale nor shape change (or a maximum iteration count is exceeded).

The final set of regions are chosen using a greedy clustering algorithm which operates from the most salient feature down (highest value of $\mathcal{Y}_D$) and clusters together all features within the support region of the current feature. A global threshold on value or number is used.

The performance of this local method is compared to exhaustive search in Section 3.



(a) (b) (c)

Fig. 4. $\mathcal{W}_D$ as a function of x and scale for image shown in (a) at the y-position indicated by the dashed line using standard sampling (b), and anti-aliased sampling (c).

2.4 Anti-aliased Sampling

The simplest method for estimating local PDFs from images is to use histogramming over a local neighbourhood, for example a circular region; pixels inside the region are counted whilst those outside are not. However, this binary approach gives rise to step changes in the histogram as the scale is increased. $\mathcal{W}_D$ is especially sensitive to this since it measures the difference between two concentric sampling windows. For example, Figure 4(b) shows the variation of $\mathcal{W}_D$ as a function of x and scale for the image shown in 4(a). The surface is taken at a point indicated by the dashed line. Somewhat surprisingly the surface is highly irregular and noisy even for this ideal noise-free image, consequently, so is the saliency space. Intuitively, the solution to this problem lies with a smoother transition between the pixels that are included in the histogram and the ones that are not.

The underlying problem is, in fact, an instance of aliasing. Restated from a sampling perspective, the binary representation of the window is sampling without pre-filtering. Evidently, this results in severe aliasing. This problem has long been recognised in the Computer Graphics community and numerous methods have been devised to better represent primitives on a discrete display [5].

To overcome this problem we use a smooth sampling window (i.e. a filtered version of the ideal sampling window). However, in contrast to the CG application, here, the window weights the contributions of the pixels to the histogram not the pixel values themselves; pixels near the edge contribute less to the count than ones near the centre. It does not blur the image.

Griffin [6] and Kœnderink and van Doorn [12] have suggested weighting histogram counts using a Gaussian window, but not in relation to anti-aliasing. However, for our purposes, the Gaussian poorly represents the statistics of the underlying pixels towards the edges due to the slow drop-off. Its long tails cause a slow computation since more pixels will have to be considered and also results in poor localisation. The traditional 'pro-Gaussian' arguments do not seem appropriate here.

Analytic solutions for the optimal sampling window are, in theory at least, possible to obtain. However, empirically we have found the following function works well:

$$SW(z) = \frac{1}{1 + \left(\frac{z}{s}\right)^n} \qquad z = \sqrt{\left(\frac{x'}{\sqrt{\rho}}\right)^2 + (y'\sqrt{\rho})^2}. \tag{5}$$

with $n = 42$ where $x' = x\cos\theta + y\sin\theta$ and $y' = y\cos\theta - x\sin\theta$ achieves the desired rotation. We truncate for small values of $SW(z)$. This sampling window gives scalar values as a function of distance, z, from the window centre, which are used to build the histogram. Figure 4(c) shows the same slice through $\mathcal{W}_D$ space but generated using Equation 5 for the sampling weights. Further implementation details and analysis may be found in [9,10].

3 Performance under Viewpoint Variations

The objective here is to determine the extent to which detected regions commute with viewpoint. This is an example of the global transformation requirement discussed in the introduction.

For these experiments, we follow the testing methodology proposed in [18, 19]. The method is applied to an image set[2] comprising different viewpoints of the same (largely planar) scene for which the inter-image homography is known. Repeatability is determined by measuring the area of overlap of corresponding features. Two features are deemed to correspond if their projected positions differ by less than 1.5 pixels. Results are presented in terms of error in overlapping area between two ellipses μ_a, μ_b:

$$\epsilon_S = 1 - \frac{\mu_a \cap \left(A^T \mu_b A\right)}{\mu_a \cup \left(A^T \mu_b A\right)} \tag{6}$$

where A defines a locally linearized affine transformation of the homography between the two images and $\mu_a \cap (A^T \mu_b A)$ and $\mu_a \cup (A^T \mu_b A)$ represent the area of intersection and union of the ellipses respectively.

Figure 5(a) shows the repeatability performance as a function of viewpoint of three variants of the affine invariant salient region detector: exhaustive search without anti-aliasing (FS Affine ScaleSal), exhaustive search with anti-aliasing (AA FS Affine ScaleSal), and local search with anti-aliasing (AA LS Affine ScaleSal). The performance is compared to the detector of Mikolajczyk and Schmid [19], denoted Affine MSHar. Results are shown for $\epsilon_S < 0.4$.

It can be seen that the full search Affine Saliency and Affine MSHar features have a similar performance over the range of viewpoints. However, from 40° the anti-aliased sampling provides some gains, though curiously diminishes performance at 20°. The local search anti-aliased Affine Saliency performs reasonably well compared to the full search methods but of course takes a fraction of the time to compute.

[2] Graffiti6 from http://www.inrialpes.fr/lear/people/Mikolajczyk/

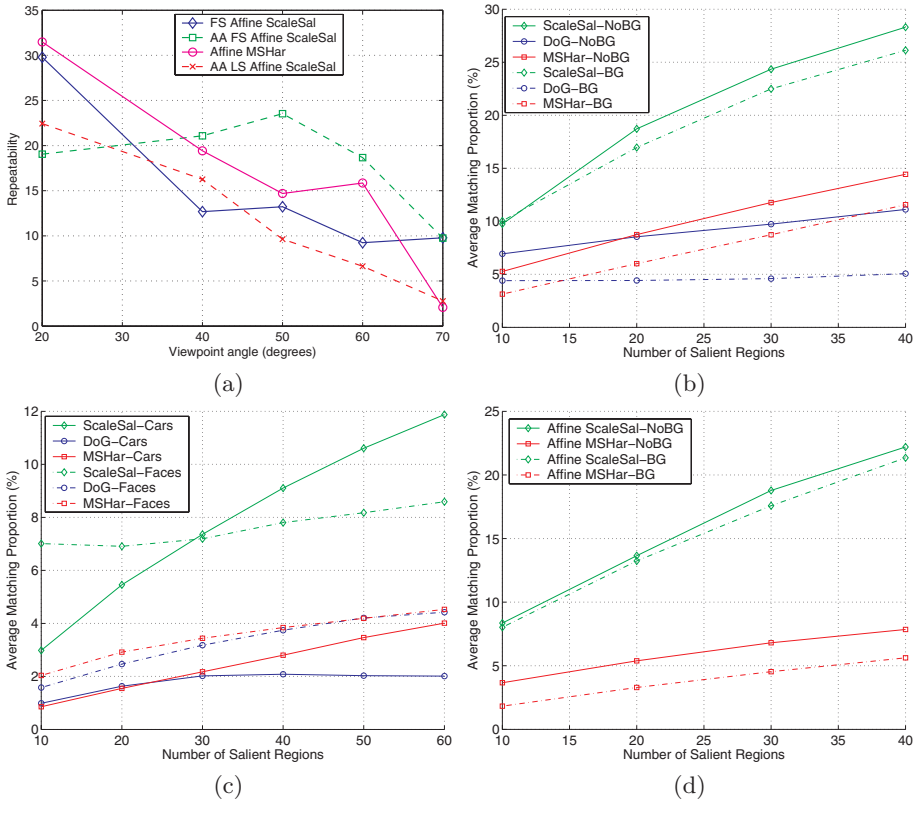


Fig. 5. Repeatability results under (a) viewpoint changes, (b,d) background perturbations and intra-class variations for Bike images, (c) intra-class variations for car and face images. Plots (b,c) are for similarity invariant and (d) for affine invariant detectors.

4 Performance under Intra-class Variation and Image Perturbations

The aim here is to measure the performance of a region detector under intra-class variations and image perturbations – the other two requirements specified in the introduction. In the following subsections we develop this measure and then compare performance of the salient region detector to other region operators. In these experiments we used similarity invariant versions of three detectors: similarity Saliency (ScaleSal), Difference-Of-Gaussian (DoG) blob detector [16] and the multi-scale Harris (MSHar) with Laplacian scale selection — this is Affine MSHar without the affine adaptation [19]. We also used affine invariant detectors Affine ScaleSal and Affine MSHar. An affine invariant version of the DoG detector was not available.

4.1 The Performance Measure

We will discuss first measuring repeatability over intra-class variation. Suppose we have a set of images of the same object class, e.g. motorbikes. A region detection operator which is unaffected by intra-class variation will reliably select regions on *corresponding parts* of all the objects, say the wheels, engine or seat for motorbikes. Thus, we assess performance by measuring the (average) number of correct correspondences over the set of images.

The question is: what constitutes a correct corresponding region? To determine this, we use a proxy to the true intra-class transformation by assuming that an affinity approximately maps one imaged object instance to another. The affinities are estimated here by manually clicking on corresponding points in each image, e.g. for motorbikes the wheels and seat/petrol tank join. We consider a region to match if it fulfils three requirements: its position matches within 10 pixels; its scale is within 20% and normalised mutual information[3] between the appearances is > 0.2. For the affine invariant detectors, the scale test is replaced with the overlap error, $\epsilon_s < 0.4$ (Eq. 6), and the mutual information is applied to elliptical patches transformed to circles. These are quite generous thresholds since the objects *are* different and the geometric mapping approximate.

In detail we measure the average correspondence score S as follows. N regions are detected on each image of the M images in the dataset. Then for a particular reference image i the correspondence score S_i is given by the proportion of corresponding to detected regions for all the other images in the dataset, i.e.:

$$S_i = \frac{\text{Total number of matches}}{\text{Total number of detected regions}} = \frac{N_M^i}{N(M-1)} \tag{7}$$

The score S_i is computed for $M/2$ different selections of the reference image, and averaged to give S. The score is evaluated as a function of the number of detected regions N. For the DoG and MSHar detectors the features are ordered on Laplacian (or DoG) magnitude strength, and the top N regions selected.

In order to test insensitivity to image perturbation the data set is split into two parts: the first contains images with a uniform background and the second, images with varying degrees of background clutter. If the detector is robust to background clutter then the average correspondence score S should be similar for both subsets of images.

4.2 Intra-class Variation Results

The experiments are performed on three separate data-sets, each containing different instances from an object class: 200 images from Caltech Motorbikes (Side), 200 images from Caltech Human face (Front), and all 126 Caltech Cars (Rear) images. Figure 6 shows examples from each data set[4].

[3] $MI(A,B) = 2(H(A) + H(B) - H(A,B))/(H(A) + H(B))$

[4] Available from http://www.robots.ox.ac.uk/~vgg/data/.

(a) (b)

Fig. 6. Example images from (a) the two parts of the Caltech motorbike data set without background clutter (top) and with background clutter (bottom), and (b) Caltech cars (top) and Caltech faces (bottom).

The average correspondence score S results for the similarity invariant detectors are shown in Figure 5(b) and (c). Figure 5(d) shows the results for the affine detectors on the motorbikes. For all three data sets and at all thresholds the best results are consistently obtained using the saliency detector. However, the repeatability for all the detectors is lower for the face and cars compared to the motorbike case. This could be due to the appearances of the different object classes; motorbikes tend to appear more complex than cars and faces.

Figure 7 shows smoothed maps of the locations at which features were *detected* in all 200 images in the motorbike image set. All locations have been back projected onto a reference image. Bright regions are those at which detections are more frequent. The map for the saliency detector indicates that most detections are near the object with a few high detection points near the engine, seats wheel centres, headlamp. In contrast, the DoG and MSHar maps show a much more diffuse pattern over the entire area caused by poor localisation and false responses to background clutter.

4.3 Image Perturbation Results

The motorbike data set is used to assess insensitivity to background clutter. There are 84 images with a uniform background, and 116 images with varying degrees of background clutter; see Figure 6(a).

Figure 5(b) shows separate plots for motorbike images with and without background clutter at N=10 to 40. The saliency detector finds, on average, approximately 25% of 30 features within the matching constraints; this corresponds to about 7 features per image on average. In contrast, the MSHar and DoG detectors select 2-3 object features per image at this threshold. Typical examples of the matched regions selected by the saliency detector on this data set are shown in Figure 4.3.

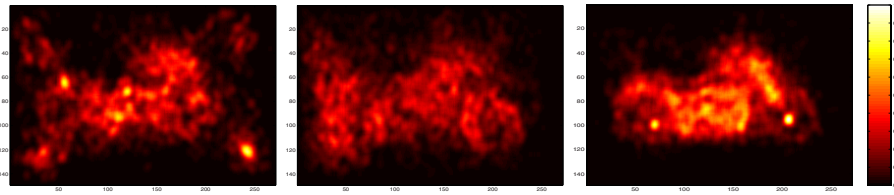


Fig. 7. Smoothed map of the detected features over all 200 images in the motorbike set back projected onto one image. The colour indicates the normalised number of detections in a given area (white is highest). Note the relative 'tightness' of the bright areas of the saliency detector compared to the DoG and MSHarr.

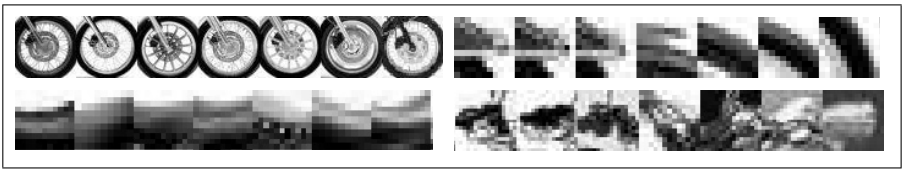


Fig. 8. Examples of the matched regions selected by the similarity saliency detector from the motorbike images: whole front wheels; front mud-guard/wheel corner; seat; headlamp.

There is also a marked difference in the way the various detectors are affected by clutter. It has little effect on the ScaleSal detector whereas it significantly reduces the DoG performance and similarly that of MSHar. Similar trends are obtained for the affine invariant detectors applied to the motorbikes images, shown in Figure 5(d).

Local perturbations due to changes in the scene configuration, background clutter or changes within in the object itself can be mitigated by ensuring compact support of any probing elements. Both the DoG and MSHar methods rely on relatively large support windows which cause them to be affected by non-local changes in the object and background; compare the two cluttered and uncluttered background results for the motorbike experiments.

There may be several other relevant factors. First, both the DoG and MSHar methods blur the image, hence causing a greater degree of similarity between objects and background. Second, in most images the objects of interest tend to be in focus while backgrounds are out of focus and hence blurred. Blurred regions tend to exhibit slowly varying statistics which result in a relatively low entropy and inter-scale saliency in the saliency detector. Third, the DoG and MSHar methods define saliency with respect to specific properties of the local surface geometry. In contrast, the saliency detector uses a much broader definition.

5 Discussion and Future Work

In this paper we have presented a new region detector which is comparable to the state of the art [19,20] in terms of co-variance with viewpoint. We have also demonstrated that it has superior performance on two further criteria: robustness to image perturbations, and repeatability under intra-class variability. The new detector extends the original method of Kadir and Brady to affine invariance; we have developed a properly anti-aliased implementation and a fast optimisation based on a local search.

We have also proposed a new methodology to test detectors under intra-class variations and background perturbations. Performance under this extended criterion is important for many applications, for example part detectors for object recognition.

The intra-class experiments demonstrate that defining saliency in the manner of the saliency detector is, on average, a better search heuristic than the other region detectors tested on at least the three data sets used here.

It is interesting to consider how the design of feature detectors affects performance. Many global effects, such as viewpoint, scale or illumination variations can be modelled mathematically and as such can be tackled directly provided the detector also lends itself to such analysis. Compared to the diffusion-based scale-spaces, relatively little is currently known about the properties of spaces generated by statistical methods such as that described here. Further investigation of its properties seems an appealing line of future work.

We plan to compare the saliency detector to other region detection approaches which are not based on filter response extrema such as [17,22]

Acknowledgements. We would like to acknowledge Krystian Mikolajczyk for supplying the MSHar and embedded (David Lowe) DoG feature detector code. Thanks to Mark Everingham and Josef Sivic for many useful discussions and suggestions. This work was funded by EC project CogViSys.

References

1. S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *Proc. European Conf. Computer Vision*, pages 113–130, 2002.
2. A. Baumberg. Reliable feature matching across widely separated views. In *Proc. Computer Vision Pattern Recognition*, pages 774–781, 2000.
3. E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *Proc. European Conf. Computer Vision*, pages 109–124, 2002.
4. R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. Computer Vision Pattern Recognition*, pages II: 264–271, 2003.
5. J.A. Foley and A. Van Dam. *Fundamentals of Interactive Computer Graphics*. Addison-Wesley, 1982.
6. L.D. Griffin. Scale-imprecision space. *Image and Vision Computing*, 15:369–398, 1997.

7. C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. Alvey Vision Conf.*, pages 189–192, 1988. Manchester.
8. T. Kadir. *Scale, Saliency and Scene Description*. PhD thesis, University of Oxford, 2002.
9. T. Kadir, D. Boukerroui, and J.M. Brady. An analysis of the scale saliency algorithm. Technical Report OUEL No: 2264/03, University of Oxford, 2003.
10. T. Kadir and J.M. Brady. Scale, saliency and image description. *Intl. J. of Computer Vision*, 45(2):83–105, 2001.
11. J.J. Kœnderink and A.J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, 63:291–297, 1987.
12. J.J. Kœnderink and A.J. van Doorn. The structure of locally orderless images. *Intl. J. of Computer Vision*, 31(2/3):159–168, 1999.
13. S. Kok-Wiles, M. Brady, and R. Highnam. Comparing mammogram pairs for the detection of lesions. In *Proc. Intl. Workshop on Digital Mammography*, pages 103–110, 1998.
14. T. Lindeberg. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *Intl. J. of Computer Vision*, 11(3):283–318, 1993.
15. T. Lindeberg and B.M. ter Haar Romeny. Linear scale-space: I. basic theory, II. early visual operations. In B.M. ter Haar Romeny, editor, *Geometry-Driven Diffusion*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994.
16. D.G. Lowe. Object recognition from local scale-invariant features. In *Proc. Intl. Conf. on Computer Vision*, pages 1150–1157, 1999.
17. J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. British Machine Vision Conf.*, pages 384–393, 2002.
18. K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proc. Intl. Conf. on Computer Vision*, 2001.
19. K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. European Conf. Computer Vision*, 2002.
20. F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In *Proc. European Conf. Computer Vision*, pages 414–431, 2002.
21. C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(5):530–535, 1997.
22. T. Tuytelaars and L. Van Gool. Wide baseline stereo based on local, affinely invariant regions. In *Proc. British Machine Vision Conf.*, pages 412–422, 2000.
23. M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proc. European Conf. Computer Vision*, June 2000.