Elsevier Editorial System(tm) for Neurocomputing
Manuscript Draft

Manuscript Number:

Title: Salient regions for image matching

Article Type: Special Issue: Image Feature Detection

Keywords: discrete cosine transform; elliptically symmetric distributions; image statistics; Kullback-Leibler divergence; salient image regions; stereo matching.

Corresponding Author: Professor Stephen J Maybank, PhD

Corresponding Author's Institution: Birkbeck College

First Author: Stephen J Maybank, PhD

Order of Authors: Stephen J Maybank, PhD

Abstract: A probabilistic definition of saliency is given in a form suitable for applications to image matching. In order to make this definition, the values of the pixels in pairs of matching regions are modeled using an elliptically symmetric distribution (ESD). The values of the pixels in background pairs of regions are also modeled using an ESD. If a region is given in one image, then the conditional probability density function for the matching region in a second image of the same scene can be calculated. The saliency of the given region is defined to be the Kullback-Leibler divergence between this conditional pdf and a background conditional pdf. Experiments carried out using images in the Middlebury stereo database show that if the salience of a given image region is high, then there are few background regions with a better match to the given region than the true matching region.

Salient Regions for Image Matching

S.J. Maybank
1st December 2011

Department of Computer Science and Information Systems, Birkbeck College, Malet Street, London, WC1E 7HX, UK.
sjmaybank@dcs.bbk.ac.uk

Abstract. A probabilistic definition of saliency is given in a form suitable for applications to image matching. In order to make this definition, the values of the pixels in pairs of matching regions are modeled using an elliptically symmetric distribution (ESD). The values of the pixels in background pairs of regions are also modeled using an ESD. If a region is given in one image, then the conditional probability density function for the matching region in a second image of the same scene can be calculated. The saliency of the given region is defined to be the Kullback-Leibler divergence between this conditional pdf and a background conditional pdf. Experiments carried out using images in the Middlebury stereo database show that if the salience of a given image region is high, then there are few background regions with a better match to the given region than the true matching region.

Keywords: discrete cosine transform, elliptically symmetric distributions, image statistics, Kullback-Leibler divergence, salient image regions, stereo matching.

1 Introduction

The matching of regions between two images is an important task in computer vision, with applications to stereo vision, structure from motion, image registration, object recognition and content based image retrieval (Brown et al. 2011). A widely used strategy for finding matching regions is to first establish matches between pairs of salient regions and then extend the matching to the remaining image regions. Salient regions are also important in psychophysical studies of visual attention. For example, experiments reported by Itti et al. (1998) show that viewers often examine the salient regions of an image first. Computer based experiments using biologically plausible networks show that the use of saliency can improve recognition rates (Han and Vasconcelos 2010).

There are many different definitions of saliency in the image processing literature. They can be divided into three types, as follows.

1. Local extrema: an image region is salient if it differs in some marked way from the regions that surround it (Tuytelaars and Mikolajczyk 2008; Itti et al. 1998).

2. Predefined features: an image region is salient because of the nature of the values of the pixels within it. For example the region might contain an edge or a corner or it might be highly textured (Schmid et al. 2000).

3. Application oriented: an image region is salient if it is well suited to a particular task, for example image matching or image interpretation (Mudge et al. 1987; Walker et al. 1998; Gao and Vasconcelos 2004).

The definition of saliency adopted in this paper is of the third type. The application is the matching of image regions between two images of the same scene. In heuristic terms, a region in one image is said to be salient if there is a high probability of identifying the correct matching region in the set of candidate matching regions. It follows from this definition that the saliency of a region depends on the properties of the image as a whole. For example, if the image contains very few low contrast regions then a low contrast region would be salient, because a correct match to a low contrast region could be found easily. If, as is usually the case, the image contains a large number of low contrast regions, then a low contrast region would not be salient because there are likely to be many good candidates for the correct matching region. Methods from probability theory are used to make the above heuristic definition of saliency quantitative. Once this definition is in place, it is not necessary to rely on intuition to decide, for example, whether corners or edges or highly textured regions are salient. Instead, the saliency of any given region such as a corner or an edge can be calculated.

The probabilistic methods used to define saliency are briefly described. For further details, see Sections 2 and 3 below. Let $R(1)$ and $R(2)$ be regions in separate images and let $v(1)$, $v(2)$ be feature vectors obtained from $R(1)$ and $R(2)$ respectively. Let H be the hypothesis that $R(1)$ and $R(2)$ are a pair of matching regions and let B be the hypothesis that $R(1)$ and $R(2)$ are a background pair of regions which do not necessarily match. Let $p(v(2)|v(1), H)$ be the probability density function (pdf) for $v(2)$ conditional on $v(1)$ and the hypothesis H. Let $p(v(2)|v(1), B)$ be defined similarly for the hypothesis B. If $R(1)$ is given, then the matching region can be found with a low probability of error provided the two conditional pdfs, $p(v(2)|v(1), H)$ and $p(v(2)|v(1), B)$ are very different from each other. Conversely, if the two pdfs are similar, then it is hard to find the correct matching region to $R(1)$. The difference between the two pdfs is measured using the Kullback-Leibler divergence of $p(v(2)|v(1), B)$ from $p(v(2)|v(1), H)$ (Cover and Thomas 1991). The value of this divergence is by definition the saliency of $R(1)$.

Let v be the vector obtained by concatenating $v(1)$ and $v(2)$. In the implementation of the above definition of saliency, the distributions of v given H and v given B are assumed to be elliptically symmetric (Chmielewski 1981), with probability density functions $p(v|H)$ and $p(v|B)$ respectively. These two pdfs are estimated using two images for which the pairs of corresponding regions are known. The advantages of the elliptically symmetric distributions (ESDs) are firstly that they depend directly on the most important statistical properties of v, namely the expected value and the covariance, and secondly, that they are computationally tractable. In particular, if $v(1)$ is given then the conditional pdfs $p(v(2)|v(1), H)$ and $p(v(2)|v(1), B)$ are also elliptically symmetric, and they can be calculated numerically from $p(v|H)$ and $p(v|B)$ respectively. The symmetries of the covariance matrix of v ensure that the calculation of the Kullback-Leibler divergence between $p(v(2)|v(1), H)$ and $p(v(2)|v(1), B)$ can be reduced to the numerical evaluation of a two dimensional integral.

In order to obtain the ESDs for v it is necessary to estimate the covariance matrix of v up to a scale factor. This is done in two stages. In the first stage, a covariance matrix is obtained for the pixel values in an image region, using a Gaussian model for these values. The Gaussian model is a two dimensional generalisation of the one dimensional model described by Clarke (1981). The eigenvectors of the covariance matrix are listed in decreasing order of the associated eigenvalues, and the first k_1 eigenvectors are used

to project a vector of the pixel values to the vector $v(1)$ in $\mathbb{R}^{k_1}$. The vector $v(2)$ is obtained in a similar way. In the second stage, a covariance matrix for v is estimated using the covariance matrices for $v(1)$, $v(2)$ and a single parameter which measures the extent to which $v(1)$ and $v(2)$ are likely to differ, under the hypothesis that $v(1)$ and $v(2)$ are obtained from a pair of matching image regions. This latter parameter is estimated empirically using a sequence of pairs of correctly matched image regions.

The Gaussian model for the pixel values in a pair of matching image regions is described in Section 2. The elliptically symmetric distributions for image regions are described in Section 3. An experimental investigation of saliency is described in Sections 4 and 5. Some concluding remarks are made in Section 6. All the experiments were carried out using Mathematica, version 5.

1.1 Related work

Salient image regions or image points are often referred to as interest points or as local features. In some applications interest points are found using one criterion and then classified as salient or otherwise using a second criterion. For example, see Lowe (1999) and Schmid et al. (2000). In view of these differences in terminology, it is convenient in this section to use the terms "salient point" and "interest point" interchangeably. A comprehensive review of the vast literature on interest point detectors can be found in Tuytelaars and Mikolajczyk (2008).

1.1.1 Type 1 saliency: local extrema

The use of local extrema to define saliency is widespread. Only a few representative examples are given here. Tuytelaars and Mikolajczyk (2008) use the term local feature rather than salient point, and define a local feature as "an image pattern which differs from its immediate neighborhood". They note in their conclusion that there is at present no theory to specify which features should be extracted from an image in any particular application. Salient points in the form of corners can be extracted from an image using the local maxima of the Harris-Stephens corner and edge detector (Harris and Stephens 1988). In Lowe (1999) salient points are found using the local maxima and minima obtained after filtering the image with a difference of Gaussians. Kadir and Brady (2001) detect salient points at different scales using the local maxima of a measure of the entropy of the pixel values in image patches.

1.1.2 Type 2 saliency: predefined features

In Schmid et al. (2000) an interest point is by definition the centre of a region in which the contour lines defined by the pixel values are curved rather than straight. They note that corners, the ends of branches and highly textured regions are all examples of interest points. Edge elements are excluded. Schmid et al. (2000) survey the literature on interest point detectors prior to 2000. A general discussion of interest point detection and further references can be found in the book by Forsyth and Ponce (2003).

Saal et al. (2006) identify salient points using the properties of the structure tensor obtained from the image gradients in the neighbourhoods of the points. Schmid et al. (2000) classify an interest point as salient if the values of the pixels in a region centred on

the point have a high information content. The information content is measured using a Gaussian model for vectors of differential invariants obtained from regions centred on the interest points. Experiments show that the highest information content is obtained using the interest points found by an improved version of the Harris-Stephens detector (Harris and Stephens 1988).

1.1.3 Type 3 saliency: application oriented

Mudge et al. (1987) define features using pairs of segments in the boundary of a flat object. The saliency of a feature is inversely proportional to the number of times that the feature appears in a training set. The saliencies of the features are used in an algorithm to detect individual objects in a set of objects thrown together in a bin. Walker et al. (1998) define salient features as "those which have a low probability of being misclassified with any other feature". Each feature has associated with it a pdf for a corresponding feature. A given feature is salient if a comparison of the associated pdf with the pdfs associated with the other features shows that the probability of misclassifying the given feature is low. The pdfs in question are mixtures of Gaussians. Walker et al. (1998) test their method on images of faces.

Gao and Vasconcelos (2004) define an attribute of an object to be salient if it is useful for object recognition, in that it distinguishes the object from all other objects of interest. This qualitative definition is made quantitative using the Kullback-Leibler divergence between a pdf for a feature, conditional on class membership, and a background pdf.

1.1.4 Probabilistic models for feature vectors

Clarke (1981) shows that the discrete cosine transform (DCT) of a one dimensional image, consisting of a row or column of m pixels, can be obtained from a Gaussian model for the pixel values. The eigenvectors of the covariance matrix for the vectors of pixel values form a basis of $\mathbb{R}^m$ and DCT base is obtained in the limit as the correlations of the values of adjacent pixels tends to 1. A detailed proof of this result can be found in Hoggar (2006). Uenohara and Kanade (1998) derive a similar result for the vectors of pixel values obtained from sets of rotated images. Hafed and Levine (2001) compare the performance of two algorithms for face recognition, one based on the Karhunen-Loève transform and one based on the DCT. They show that both algorithms yield similar results, even though their Karhunen-Loève transform is adaptive to the data but the DCT is not.

2 Feature Vectors

Let m be a relatively small strictly positive integer and let I be an image. An image region is, by definition, an $m \times m$ square of pixels in I. A probabilistic model for the pixel values in pairs of matching image regions is obtained in this section. The pixel values in an $m \times m$ image region R are summarised by a vector in a feature space $\mathbb{R}^{k_1}$, where k_1 is an integer such that $1 \leq k_1 \leq m^2$. The pixel values in pairs $(R(1), R(2))$ of matching image regions are summarised by a vector in a feature space $\mathbb{R}^{2k_1}$.

The feature vectors for $m \times m$ image regions are described in Section 2.1 and the feature vectors for pairs of matching regions are described in Section 2.2.

2.1 Feature vector for an image region

Let R be an $m \times m$ image region and let R_{ij} for $1 \leq i, j \leq m$ be the pixel values in R. The pixel values are concatenated to yield a vector in $\mathbb{R}^{m^2}$. This feature vector is simplified in the usual way by projecting it to a lower dimensional subspace $\mathbb{R}^{k_1}$. Suppose that each R_{ij} is a zero mean Gaussian random variable. Let $\|.\|$ be the Euclidean norm and let κ be a real number such that $1-\kappa$ is small and positive. The covariance of R_{i_1,j_1} and R_{i_2,j_2} is defined by

$$\mathrm{Cov}(R_{i_1j_1}, R_{i_2j_2}) = \kappa^{\|(i_1,j_1)-(i_2,j_2)\|}, \quad 1 \leq i_1, j_1, i_2, j_2 \leq m. \tag{1}$$

Define the $m \times m$ matrix Z of Gaussian random variables by

$$Z_{ij} = R_{ij} - m^{-2} \sum_{i_1,j_1=1}^{m} R_{i_1j_1}, \quad 1 \leq i, j \leq m, \tag{2}$$

and let z be the vector in $\mathbb{R}^{m^2}$ obtained by flattening Z. More precisely, the rows of Z are concatenated and the resulting row vector is transposed to give the column vector z. Let Λ be the covariance matrix of z. The eigenvectors $e_1, \ldots, e_{m^2}$ of Λ are ordered such that the corresponding eigenvalues, λ_i, form a decreasing sequence. It is noted that $\lambda_{m^2} = 0$ and that every entry of e_{m^2} is equal to m^{-2}. Let k_1 be an integer chosen such that $1 \leq k_1 \leq m^2$ and let $M(k_1)$ be the $k_1 \times m^2$ matrix such that the ith row of $M(k_1)$ is $e_i^\top$ for $1 \leq i \leq k_1$. The values of the pixels in R are summarised by the feature vector $M(k_1)z$. Clarke (1981) obtains explicit expressions for the eigenvectors in a similar Gaussian model in which the image consists of a single row of pixels. See also Hoggar (2006).

2.2 Feature vector for pairs of matching image regions

Let $R(1)$ and $R(2)$ be matching $m \times m$ image regions, let $Z(1)$, $Z(2)$ be the corresponding $m \times m$ matrices of Gaussian random variables defined as in (2) and let $z(i)$ be the random vector in $\mathbb{R}^{m^2}$ obtained by flattening the $Z(i)$ for $i = 1, 2$. Let κ_1, κ_2 be numbers such that $0 < \kappa_1, \kappa_2 < 1$ and such that $1-\kappa_2$ is small. The pixel values in $R(1)$, $R(2)$ are given by Gaussian random variables with zero mean values and with covariances defined by

$$\begin{aligned} \mathrm{Cov}(R_{i_1j_1}(h), R_{i_2j_2}(h)) &= \kappa_2^{\|(i_1,j_1)-(i_2,j_2)\|}, \quad 1 \leq i_1, j_1, i_2, j_2 \leq m, h = 1, 2, \\ \mathrm{Cov}(R_{i_1j_1}(1), R_{i_2j_2}(2)) &= \kappa_1\kappa_2^{\|(i_1,j_1)-(i_2,j_2)\|}, \quad 1 \leq i_1, j_1, i_2, j_2 \leq m. \end{aligned} \tag{3}$$

The effect of κ_1 in (3) is to decrease the covariances between the pixel values in $R(1)$ and the pixel values in $R(2)$.

Let z be the vector obtained by concatenating $z(1)$ and $z(2)$, let $v(i) = M(k_1)z(i)$ for $i = 1, 2$ and let v be the vector obtained by concatenating $v(1)$ and $v(2)$. Let $\tilde{C}$ be the covariance matrix of z and let C be the covariance matrix of v. It follows that

$$C = \begin{pmatrix} M(k_1) & 0 \\ 0 & M(k_1) \end{pmatrix} \tilde{C} \begin{pmatrix} M(k_1)^\top & 0 \\ 0 & M(k_1)^\top \end{pmatrix}. \tag{4}$$

The value of κ_1 is estimated by comparing C with an empirical covariance matrix obtained from a training set of matching regions. Further information is given in Section 4.1. The

covariance matrix C_B for background pairs of regions is constructed in a similar way to C, but using instead pairs $R(1)$, $R(2)$ of image regions which do not necessarily match. The subscript B refers to the background.

The covariance matrices C, C_B have the form

$$C = \begin{pmatrix} A & \mu A \\ \mu A & A \end{pmatrix}, \quad C_B = \begin{pmatrix} \mu_{B1} A & \mu_{B2} A \\ \mu_{B2} A & \mu_{B1} A \end{pmatrix}, \tag{5}$$

where A is a diagonal $k_1 \times k_1$ matrix and μ, μ_{B1}, μ_{B2} are real numbers. When C_B is estimated empirically, μ_{B2} is negligibly small.

3 Probabilistic model for feature vectors

In this section the pdf for an elliptically symmetric distribution (ESD) is defined and it is shown how the conditional pdfs required by the definition of saliency in Section 1 can be calculated. Other applications of ESDs to image processing are described by Verdoolaege and Scheunders (2011) and Kwitt et al. (2009).

3.1 The elliptically symmetric distribution

An elliptically symmetric distribution (ESD) is by definition the affine transform of a spherically symmetric distribution. A spherically symmetric distribution is one which is invariant under all orthogonal transformations of the Euclidean space on which it is defined (Chmielewski 1981). If an ESD has a probability density function then this pdf is constant on each member of a one parameter family of ellipsoids. From now on, only ESDs with pdfs are considered. The pdf for an ESD E in $\mathbb{R}^k$ is written as $p(x|E)$.

Let a be a vector in $\mathbb{R}^k$, let C be a symmetric, strictly positive $k \times k$ matrix and let f be a function, $f : \mathbb{R} \to \mathbb{R}$. The pdf $p(x|E)$ for an elliptically symmetric distribution E is written in the following form,

$$p(x|E) = \det(C)^{-1/2} f\left(-\frac{1}{2}(x-a)^\top C^{-1}(x-a)\right), \quad x \in \mathbb{R}^k. \tag{6}$$

It is convenient to use the notation $E = \{a, C, f\}$. The right hand side of (6) is integrable over $\mathbb{R}^k$, such that the integral takes the value 1. It can be shown that if the pdf $p(x|E)$ in (6) has an expected value and a covariance, then the expected value is a and the covariance is proportional to C.

On applying the transformation $y = C^{-1/2}(x-a)$ to (6), it follows that

$$p(y|E) = f\left(-\|y\|^2/2\right), \quad y \in \mathbb{R}^k, \tag{7}$$

where $\|.\|$ is the Euclidean norm in $\mathbb{R}^k$. Let $V(k-1)$ be the volume of the $k-1$ dimensional hypersphere in $\mathbb{R}^k$ with unit radius. It follows from (7) that

$$V(k-1) \int_0^\infty f\left(-r^2/2\right) r^{k-1}\, dr = 1. \tag{8}$$

3.2 Conditional probability density function

Let $2k_1 = k$ and let x be a vector in $\mathbb{R}^k$ such that $x = (x(1), x(2))$, with $x(1)$, $x(2)$ in $\mathbb{R}^{k_1}$. (The condition that the vectors $x(1)$ and $x(2)$ have the same dimension is not essential, but it simplifies the notation.) Let $p(x|E)$ be the pdf for x, such that $E = \{a, C, f\}$. It is shown that the conditional distribution $p(x(2)|x(1), E)$ is an ESD.

The expression

$$(x-a)^\top C^{-1}(x-a)$$

is quadratic in $x(2)$, thus it can be written in the form

$$(x(2)-b)^\top Q^{-1}(x(2)-b) + s^2 \tag{9}$$

where b is a vector in $\mathbb{R}^{k_1}$, Q is a symmetric strictly positive $k_1 \times k_1$ matrix and s is a real number. The variables b, s in (9) depend on $x(1)$, but this does not cause any difficulty, because $x(1)$ is fixed in this calculation.

It follows from (9) and the definition of conditional pdfs that

$$p(x(2)|x(1), E) =$$

$$\alpha(\det(Q))^{-1/2} f\left(-\frac{1}{2}\left((x(2)-b)^\top Q^{-1}(x(2)-b) + s^2\right)\right), \quad x(2) \in \mathbb{R}^{k_1}. \tag{10}$$

where α is a scale factor. The right hand side of (10) is reduced to the standard form for an ESD by first finding a real valued function g such that

$$g(-r^2/2) = f(-(r^2+s^2)/2), \quad r \geq 0, \tag{11}$$

and then choosing the scale factor α such that

$$p(x(2)|x(1), E) = \alpha(\det(Q))^{-1/2} g\left(-(1/2)(x(2)-b)^\top Q^{-1}(x(2)-b)\right), \quad x(2) \in \mathbb{R}^{k_1}$$

is correctly normalised. The scale factor α is removed from the notation by redefining g.

3.3 Saliency

Let $p(v(2)|v(1), H)$ and $p(v(2)|v(1), B)$ be the conditional pdfs defined as described in Section 1, and let $R(1)$ be the image region from which $v(1)$ is obtained. As noted in Section 1, the saliency $\zeta(v(1), H, B)$ of $R(1)$ is defined to be the Kullback-Leibler divergence between $p(v(2)|v(1), H)$ and $p(v(2)|v(1), B)$,

$$\zeta(v(1), H, B) = \int_{\mathbb{R}^{k_1}} p(v(2)|v(1), H) \ln(p(v(2)|v(1), H)/p(v(2)|v(1), B))\, dv(2). \tag{12}$$

It is shown that the right hand side of (12) can be reduced to a two dimensional integral.

Let the ESDs for $p(v(2)|v(1), H)$ and $p(v(2)|v(1), B)$ be respectively,

$$\begin{aligned} E &= \{a, D, f\}, \\ E_B &= \{b, D_B, f_B\}. \end{aligned}$$

The matrices D, D_B are equal up to a scale factor, because the respective scaled covariance matrices C, C_B for $p(v(1), v(2)|H)$ and $p(v(1), v(2)|B)$ have the form (5). In addition, D

and D_B are diagonal, because of the choice of basis for the feature vectors $v(2)$, however, this latter fact is not required in the following calculation. Let λ be defined such that

$$D_B^{-1} = \lambda^2 D^{-1}. \tag{13}$$

and define the vectors y, u by

$$y = D^{-1/2}(v(2) - a), \tag{14}$$
$$u = -D^{-1/2}(a - b). \tag{15}$$

It follows from (12), (13), (14) and (15) that

$$\zeta(v(1), H, B) = -k_1 \ln(\lambda) + \int_{\mathbb{R}^{k_1}} f\left(-\|y\|^2/2\right) \ln\left(\frac{f(-\|y\|^2/2)}{f_B(-\lambda^2\|y-u\|^2/2)}\right) dy. \tag{16}$$

Let $r = \|y\|$, let $\hat{u}$ be the unit vector in the direction of u and let ω be a unit vector perpendicular to u. The range of ω is a sphere in $\mathbb{R}^{k_1}$ of dimension $k_1 - 2$. Coordinates are chosen in $\mathbb{R}^{k_1}$ such that the variable y in (16) is given by

$$y = r\sin(\phi)\omega + r\cos(\phi)\hat{u}, \quad 0 \le r < \infty, 0 \le \phi \le \pi.$$

It follows that

$$\|y - u\|^2 = r^2 - 2r\|u\|\cos(\phi) + \|u\|^2.$$

Let $\xi(\|u\|, r, \phi)$ be the function defined by

$$\xi(\|u\|, r, \phi) = (r^2 - 2r\|u\|\cos(\phi) + \|u\|^2)^{1/2}.$$

On transforming to the new coordinate system r, ω, ϕ, and integrating over ω, it follows from (16) that

$$\zeta(v(1), H, B) = -k_1 \ln(\lambda) + V(k_1 - 2)\iota(\|u\|)$$

where $\iota(\|u\|)$ is the following integral,

$$\iota(\|u\|) = \int_0^{\pi}\int_0^{\infty} f\left(-r^2/2\right) \ln\left(\frac{f(-r^2/2)}{f_B(-\lambda^2\xi(\|u\|, r, \phi)^2/2)}\right) r^{k_1-1}\sin^{k_1-2}(\phi)\, dr d\phi. \tag{17}$$

The two dimensional integral $\iota(\|u\|)$ can be evaluated numerically. One method is to convert to Cartesian coordinates $x_1 = r\cos(\phi)$, $x_2 = r\sin(\phi)$, sample the x_1, x_2 plane uniformly at the vertices of a square grid, convert the continuous distributions to discrete distributions and then use the Kullback-Leibler divergence for discrete distributions.

4 Empirical Distributions for Pairs of Regions

Pairs of images with known ground truth matchings were obtained from the Middlebury stereo database (Scharstein and Szeliski 2002). The images are known to be corrected for radial distortion and rectified such that corresponding pixels are of the form

$$(i, j) \leftrightarrow (i, j - d)$$

where $d > 0$ is the disparity. The estimation of the joint pdfs $p(v|H)$ and $p(v|B)$ is described in Section 4.1. The estimation of the conditional pdfs $p(v(2)|v(1), H)$ and $p(v(2)|v(1)|B)$ is described in Section 4.2.

4.1 Estimation of joint densities

The $k \times k$ covariance matrix C for v, under the hypothesis H, is defined in (4). It contains a single adjustable parameter κ_1. Let C_E be an empirical estimate of $C \equiv C(\kappa_1)$ obtained from a training set of pairs of matching image regions. The discrepancy between C, C_E is measured using the Kullback-Leibler divergence $D(C\|C_E)$ between Gaussian distributions with expected value 0 and respective covariance matrices C, C_E. It is only necessary to determine C up to scale, thus let d be the function defined on pairs of covariance matrices by

$$d(C, C_E) = \min_{\lambda>0} D(\lambda C, C_E).$$

The parameter κ_1 is estimated by

$$\kappa_1 = \text{argmin}\{\kappa \mapsto d(C(\kappa), C_E)\}.$$

The parameter κ_{B1} for the covariance matrix C_B for v under the hypothesis B is estimated in a similar way. In applications to stereo matching the background pairs of feature vectors $v(1)$, $v(2)$ are obtained from pairs $R(1)$, $R(2)$ of image regions centred on corresponding epipolar lines.

The covariance matrices C and C_B were estimated using the images view1.png and view2.png in the directory

http://vision.middlebury.edu/stereo/data/scenes2006/FullSize/Aloe/Illum3/Exp2

in the Middlebury stereo database. The image view1.png was reduced to a grey scale image Aloe1 by averaging the R, G, B values at each pixel. Similarly, view5.png was reduced to a grey scale image Aloe5. The images Aloe1, Aloe5 are of size 1110×1282 pixels^2. Each image was smoothed with an approximation to a Gaussian filter using a 3×3 mask. The images are shown in Fig. 1.

Graphs of the function

$$k_1 \mapsto d(C(m, k_1, \kappa_1), C_E) \tag{18}$$

are shown in Fig. 2a for $m = 3, 5, 7, 9$. Similarly, graphs of the function

$$k_1 \mapsto d(C_B(m, k_1, \kappa_{B1}), C_{BE}) \tag{19}$$

are shown in Fig 2b, also for $m = 3, 5, 7, 9$. It can be seen that (18) and (19) are increasing functions of k_1 for fixed m. The graphs for $m = 5, 7, 9$ are similar and show a steady increase as k_1 increases. The graphs for $m = 3$ are more erratic. These graphs are affected by the Gaussian smoothing of the images. For example, if smoothing is carried out using a mask of size 5×5 that approximates to a Gaussian filter mask, then the graphs for $m = 3$ and $m = 5$ become erratic.

Experiments show that the expected values of v under the hypotheses H or B are negligibly small. For this reason, the expected values are set to zero. Let $E = \{0, C, f\}$ specify the ESD $p(v|H)$ and let $k = 2k_1$. The function f is estimated as follows. On making the substitution $y = C^{-1/2}v$, the ESD $p(v|H)$ reduces to

$$p(y|H) = f(-\|y\|^2)/2, \quad y \in \mathbb{R}^k.$$

Let $\tilde{f}$ be the pdf defined in $\mathbb{R}$ for $\|y\|$, and let $V(k-1)$ be the volume of the unit $k-1$ dimensional sphere. It follows that

$$\tilde{f}(r) = f(-r^2/2)r^{k-1}V(k-1), \quad r \geq 0. \tag{20}$$

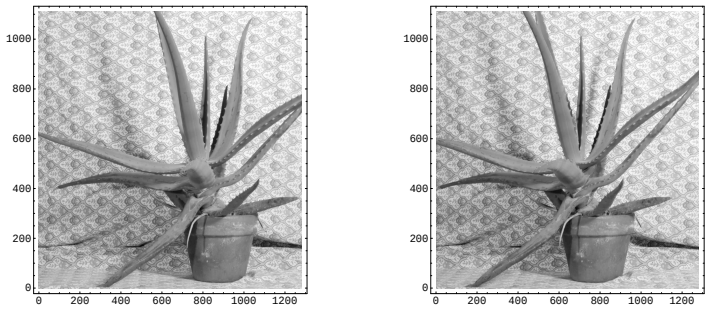
The pdf $\tilde{f}$ is estimated using the values of $\|C^{-1/2}v\|$ obtained from a set of sample values of v. The function f is then obtained from (20). In order to represent $\tilde{f}$ numerically, points $r_1, r_2, \ldots$ are chosen in the interval $r \geq 0$ such that $r_1 = 0$ and $r_i < r_{i+1}$ for $i \geq 1$. The function $\tilde{f}$ is given on the interval $[r_1, r_2]$ by

$$r \mapsto ar^{k-1} + br^k, \quad r_1 \leq r \leq r_2$$

where a, b are constants, and it is given on each of the remaining intervals by

$$r \mapsto m_i r + c_i, \quad i \geq 2.$$

This numerical representation of $\tilde{f}$ is not required to be continuous. The ESD $E_B = \{0, C_B, f_B\}$ for $p(v|B)$ is estimated in a similar way.



(a) (b)

Figure 1: Gray scale images (a) Aloe1; (b) Aloe5.

4.2 Estimation of conditional densities

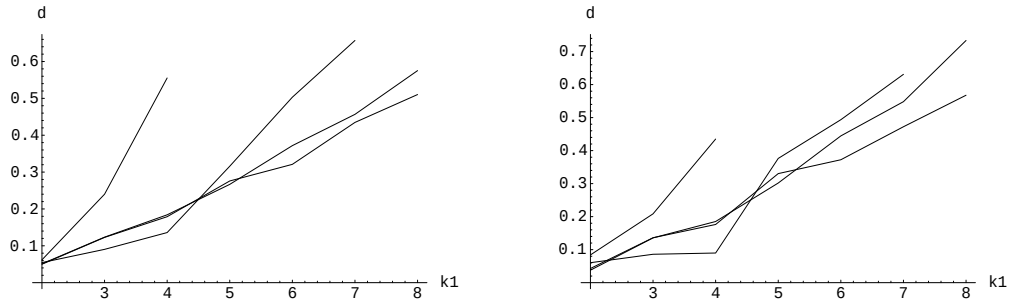
Let $\{0, C, f\}$ specify the ESD $p(v(1), v(2)|H)$ and let $\{a, D, g\}$ specify $p(v(2)|v(1), H)$. The parameters a, D are obtained as described in Section 3.2. The function $\tilde{g}$ corresponding to g is

$$\tilde{g}(r) = \frac{V(k_1 - 1)r^{k_1 - 1}\tilde{f}((r^2 + s^2)^{1/2})}{V(k-1)(r^2 + s^2)^{(k-1)/2}}, \quad r \geq 0. \tag{21}$$

It is convenient to approximate $\tilde{g}$ by a function of the same form as $\tilde{f}$ in Section 4.1. The interval $r \geq 0$ is divided into subintervals using the same points r_i as those used to specify $\tilde{f}$. The function $\tilde{g}$ is approximated in $[r_1, r_2)$ by

$$r \mapsto a_1 r^{k_1 - 1} + b_1 r^{k_1}, \quad r_1 \leq r < r_2,$$

where a_1, b_1 are constants, and it is approximated on the remaining intervals (r_i, r_{i+1}), $i \geq 2$ by straight line segments. The parameters of the ESD $p(v(2)|v(1), B)$ are obtained in a similar way.



(a) (b)

Figure 2: (a) Graphs of $d(C(m, k_1, \kappa_1), C_E)$ as a function of k_1 for $m = 3, 5, 7, 9$. The respective end points are (4,0.55), (7,0.65), (8,0.51) and $(8, 0.57)$. (b) Graphs of $d(C_B(m, k_1, \kappa_{B1}), C_{BE})$ as a function of k_1 for $m = 3, 5, 7, 9$. The respective end points are (4,0.43), (7,0.63), (8,0.56), (8,0.73).

5 Experimental Investigation of Saliency

Examples of the saliencies of natural 5×5 image regions are shown in Section 5.1, and the saliencies of some computer generated image regions are shown in Section 5.2. An application of saliency to stereo matching is described in Section 5.3. Experiments with stereo matching show that if an image region has a high saliency, then the number of candidate false matches is reduced.

5.1 Saliencies of regions in natural images

Fig. 3a shows an image region $\tilde{I}$ of size 101×101, taken from Aloe1. Fig. 3b shows the corresponding image $\tilde{S}$ of saliencies for $m = 5$ and $k_1 = 4$. The value $\tilde{S}_{ij}$ of the pixel (i, j) in $\tilde{S}$ is the salience of the $m \times m$ region in Aloe1 and centred at the pixel corresponding to the pixel (i, j) in $\tilde{I}$.

Let s_1, s_2 be thresholds such that 10% of the values $\tilde{S}_{ij}$ are at or below s_1 and 10% of the values $\tilde{S}_{ij}$ are at or above s_2. The pixels in $\tilde{I}$ associated with saliencies equal to or greater than s_2 are shown in peak white in Fig. 3c. Similarly, the pixels in $\tilde{I}$ associated with saliencies equal to or less than s_1 are shown in peak white in Fig. 3d. It is apparent from Fig. 3 that the low contrast regions in $\tilde{I}$ tend to have low saliencies, while regions containing well defined edges tend to have higher saliencies. However, there is no systematic variation of saliency with local contrast. See Walker et al. (1998) for alternative maps of saliency.

The histogram of the saliencies in $\tilde{S}$ is shown in Fig. 4. It is apparent that there is a relatively large number of regions with a high saliency. In fact half of all the pixels in $\tilde{S}$ have a saliency of 4.94 or greater.

5.2 Saliency of computer generated image regions

Fig. 5a shows graphs of saliency as a function of grey level contrast for an $m \times m$ image region that has a uniform light 3×3 square at the centre, against a uniform dark

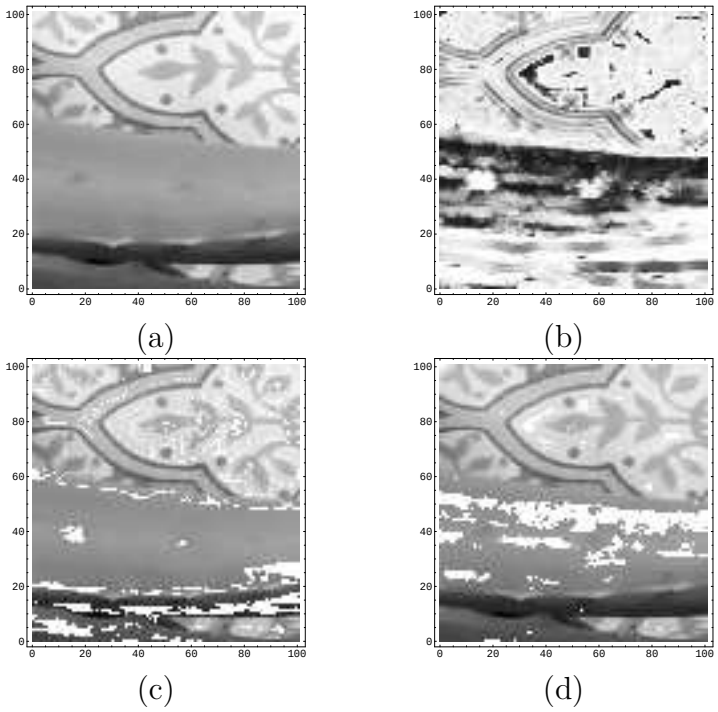


Figure 3: (a) Image region $\tilde{I}$ in Aloe1, of size 101×101; (b) saliencies of regions in Aloe1 with centres corresponding to pixels in $\tilde{I}$, for $m = 5$, $k_1 = 4$; (c) centres of regions with high salience marked in peak white; (d) centres of regions with low salience marked in peak white.

background. The three graphs are obtained using $(m, k_1) = (9, 4)$, $(m, k_1) = (9, 8)$ and $(m, k_1) = (5, 4)$. It is apparent that the three graphs are similar and that the saliency is not a monotonic function of the contrast. Fig 5b shows graphs of saliency as a function of grey level contrast for an image consisting of horizontal edge separating two regions, such that each region has a uniform grey level. Fig 5c shows graphs of saliency as a function of grey level contrast for an image consisting of an edge at 45°, separating two regions, such that each region has a uniform grey level. The values of m, k_1 for the graphs in Figs 5b and 5c are the same as in Fig. 5a. The graphs for the two types of edges are very similar.

The graphs in Fig. 5 suggest that higher saliencies can be achieved by increasing k_1, and that increases in m with k_1 fixed have less effect.

5.3 Stereo matching

A stereo matching criterion is defined using the ESDs $p(v(2)|v(1), H)$ and $p(v(2)|v(1), B)$. This criterion is compared experimentally with a well known stereo matching criterion based on the sum of absolute differences (SAD). See Scharstein and Szeliski (2002). The comparison shows that if the dimension k_1 of $v(2)$ is sufficiently large, then the accuracy of stereo matching using the ESDs is similar to the accuracy of stereo matching using SAD. This result confirms that the ESDs $p(v|H)$ and $p(v|B)$, from which $p(v(2)|v(1), H)$ and $p(v(2)|v(1), B)$ are obtained, contain a large amount of useful information about the $m \times m$ image regions.

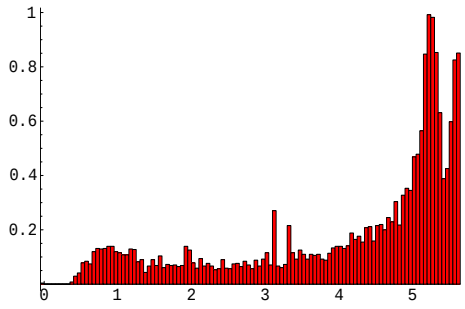
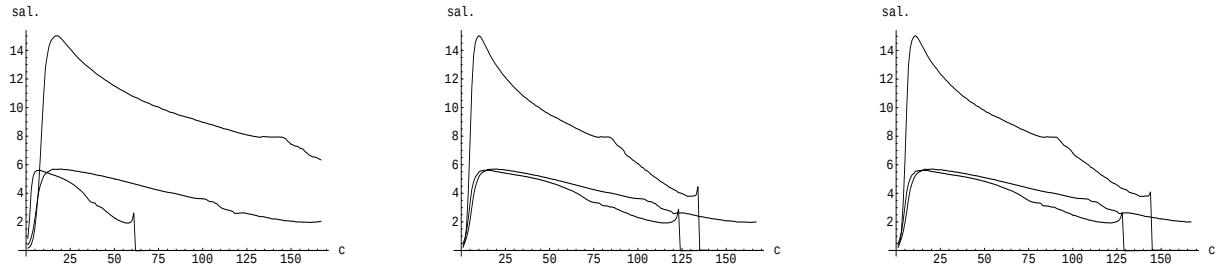


Figure 4: Histogram of the saliencies in Fig. 3b.



(a) (b) (c)

Figure 5: Saliencies of computer generated images as functions of grey level contrast. In each of (a),(b) and (c), $(m, k_1) = (9, 8), (9, 4), (5, 4)$ in order from the top down. (a) Uniform 3×3 square against a uniform background; (b) horizontal edge; (c) 45° edge.

5.3.1 Criteria for stereo matching

Let $I(1)$ and $I(2)$ be a stereo pair of images for which the ground truth stereo matches are known. Let $R(1)$ be an $m \times m$ region in $I(1)$ and let $l(2)$ be the epipolar line in $I(2)$ such that $l(2)$ corresponds to the central pixel of $R(1)$. Let $R(2, i)$ for $1 \leq i \leq N$ be a list of $m \times m$ regions in $I(2)$ such that the central pixel of each region $R(2, i)$ is on $l(2)$. Each region $R(2, i)$ is a candidate match to $R(1)$. Let $z(1)$ and $z(2, i)$ be the vectors in $\mathbb{R}^{m^2}$ obtained from $R(1)$ and $R(2, i)$ respectively, as described in Section 2.1, and let $v(1)$, $v(2, i)$ be the vectors in $\mathbb{R}^{k_1}$ obtained from $z(1)$, $z(2, i)$ respectively, as described in Section 2.2. Let the integer i^*_{esd} be defined using the log likelihood, as follows,

$$i^*_{esd} = \text{argmax}_{1 \leq i \leq N}\, i \mapsto \ln\left(p(v(2, i)|v(1), H)/p(v(2, i)|v(1), B)\right). \tag{22}$$

The region $R(2, i^*_{esd})$ is chosen as the matching region to $R(1)$. See Han and Vasconcelos (2010). The accuracy of the stereo matching is measured by the ratio of the number of correct matches to the total number of matches obtained using (22), as $R(1)$ varies in $I(1)$. Let this ratio be $r_{esd}(m, k_1)$.

Let the integer i^*_{sad} be defined by

$$i^*_{sad} = \text{argmin}_{1 \leq i \leq N}\, i \mapsto \sum_{j=1}^{m^2} |z_j(1) - z_j(2, i)|, \tag{23}$$

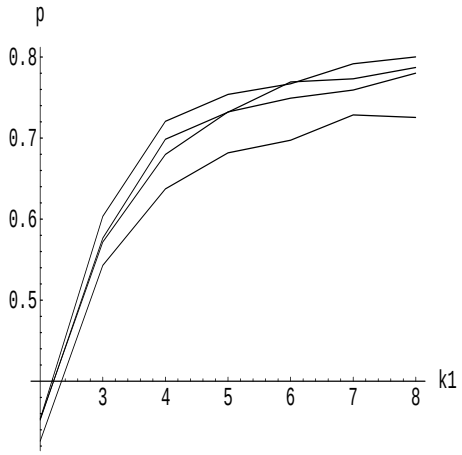


Figure 6: Graphs of the functions $k_1 \mapsto r_{esd}(m, k_1)$ for $m = 3, 5, 7, 9$.

where $|.|$ is absolute value. The region $R(2, i^*_{sad})$ is chosen as the matching region to $R(1)$. Let $r_{sad}(m)$ be the ratio of the number of correct matches to the total number of matches, as $R(1)$ varies in $I(1)$.

Epipolar lines were chosen from Aloe1 with a space of 50 pixels between consecutive lines, and pixels were sampled from each line in Aloe1 with a space of 20 pixels between consecutive pixels. The maximum allowed disparity was 211 pixels. The total number of pixels chosen from Aloe1 was approximately 1400. The values of $r_{esd}(m, k_1)$ and $r_{sad}(m)$ are shown in Table 1. Graphs of the functions $k_1 \mapsto r_{esd}(m, k_1)$ are shown in Fig. 6.

m	$k_1 = 2$	$k_1 = 3$	$k_1 = 4$	$k_1 = 5$	$k_1 = 6$	$k_1 = 7$	$k_1 = 8$	$r_{sad}(m)$
3	0.32	0.54	0.63	0.68	0.69	0.72	0.72	0.67
5	0.35	0.57	0.69	0.73	0.76	0.77	0.78	0.79
7	0.35	0.60	0.72	0.75	0.76	0.79	0.80	0.81
9	0.35	0.57	0.67	0.73	0.74	0.75	0.77	0.80

Table 1. Values of the functions $k_1 \mapsto r_{esd}(m, k_1)$ and values of $r_{sad}(m)$ for $m = 3, 5, 7, 9$.

Fig. 7 shows six scatter plots. The left hand column contains scatter plots for pairs (s, l) in which s is the saliency of a region in Aloe1, and l is the log likelihood ratio obtained from the feature vector $v(1)$ of the region and the feature vector $v(2)$ of the true matching region in Aloe5,

$$l = \ln(p(v(2)|v(1), H)/p(v(2)|v(1), B)). \tag{24}$$

The right hand column contains scatter plots for pairs $(s, n_1/n_2)$, where s is the saliency of a region in Aloe1, n_1 is the number of background regions which are candidate matches and for which the log likelihood ratio is greater than or equal to the log likelihood ratio (24) for the true matching region and n_2 is the total number of background regions which are candidate matches. The scatter plots in the first two rows contain 1287 points and the scatter plots in the last row contain 1296 points. In the first row $m = 5$, $k_1 = 4$, in the second row $m = 5$, $k_1 = 8$ and in the third row $m = 7$, $k_1 = 4$.

It is apparent from the scatter plots that regions in Aloe1 with large saliencies tend to have large values for the log likelihood ratio (24). In addition, if a region has a large

saliency then in most cases there is only a small proportion of candidate matching regions which yield log likelihood ratios larger than or equal to the log likelihood ratio for the correct matching region.

6 Conclusion

A new definition of the saliency of image regions has been given. This definition is based on a probabilistic model for the pixel values in image regions. It is appropriate for the task of matching regions between images. Experimental results show that a significant proportion of image regions have relatively high saliencies. This suggests that in searching for matching image regions there is little to be gained by focusing on regions with very high saliencies. Very low contrast image regions tend to have low saliencies, however, experiments with computer generated images containing squares or edges suggest that there is no simple relation between saliency and local contrast.

The effectiveness of the new definition of saliency was tested using stereo matching. A stereo matching algorithm based on the probabilistic model for pixel values and a log likelihood ratio test for finding matching regions achieved a similar performance to a stereo matching algorithm in which matching regions were found using the sum of the absolute differences of pixel values. Experiments showed that a region with a high saliency is easier to match than a region with low saliency. This is because there are relatively few background regions with a better match to the high saliency region than the true matching region.

References

1. Brown, M., Gang Hua & Winder, S. (2011) Discriminative learning of local image descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, No. 1, pp. 43-57.

2. Chmielewski, M. A. (1981) Elliptically symmetric distributions: a review and bibliography. *International Statistical Review*, vol. 49, no. 1, pp. 67-74.

3. Clarke, R.J. (1981) On the relation between the Karhunen-Loève and cosine transforms. *IEE Proceedings, Part F: Communications, Radar and Signal Processing*, vol. 128, pp. 359-360.

4. Cover, T.M. & Thomas, J. A. (1991) *Elements of Information Theory.* John Wiley and Sons.

5. Forsyth, D. & Ponce, J. (2003) *Computer Vision: a modern approach.* Prentice Hall.

6. Gao, D. & Vasconcelos, N. (2004) Discriminant saliency for visual recognition from cluttered scenes. *Proc. Neural Information Processing Systems (NIPS)*, Vancouver, Canada, pp. 481-488.

7. Hafed, Z.M. & Levine, M.D. (2001) Face recognition using the discrete cosine transform. *International Journal of Computer Vision*, vol. 43, No. 3, pp. 167-188.

8. Han, S. & Vasconcelos, N. (2010) Biologically plausible saliency mechanisms improve feedforward object recognition. *Vision Research*, vol. 50, pp. 2295-2307.

9. Harris, C. & Stephens, M. (1988) A combined corner and edge detector. In *Alvey Vision Conference*, pp. 147-151.

10. Hoggar, S.G. (2006) *Mathematics of Digital Images: creation, compression, restoration, recognition.* CUP.

11. Itti, L., Koch, C. & Niebur, E. (1998) A model of saliency-based attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-259.

12. Kadir, T. & Brady, M. (2001) Saliency, scale and image description. *International Journal of Computer Vision*, vol. 45, No. 2, pp. 83-105.

13. Kwitt, R., Meerwald, P. & Uhl, A. (2009) Color-image watermarking using multivariate power-exponential distribution. *International Conference on Image Processing*, pp. 4245-4248.

14. Lowe, D.G. (1999) Object recognition from local scale invariant features. *International Conference on Computer Vision*, Corfu, Greece.

15. Mudge, T.N., Turney, J.L. & Volta, R.A. (1987) Automatic generation of salient features for the recognition of partially occluded parts. *Robotica*, vol. 5, pp. 117-127.

16. Saal,H., Nortmann, N., Krüger, N. & König, P. (2006) Salient image regions as a guide for useful visual features. *Proc. IEEE Advances in Cybernetic Systems (AICS)*, Sheffield, UK.

17. Scharstein, D. & Szeliski, R. (2002) A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, vol. 47, No. 1, pp. 7-42.

18. Schmid, C., Mohr, R. & Bauckhage, C. (2000) Evaluation of interest point detectors. *International Journal of Computer Vision*, vol. 37, No. 2, pp. 151-172.

19. Tuytelaars, T. & Mikolajczyk, K. (2008) Local invariant feature detectors: a survey. *Foundations and Trends in Computer Graphics and Vision*, vol. 3, No. 3, pp. 177-280.

20. Uenohara, M. & Kanade, T. (1998) Optimal approximation of uniformly rotated images: relationship between Karhunen-Loève expansion and discrete cosine transform. *IEEE Transactions on Image Processing*, vol. 7, pp. 116-119.

21. Verdoolaege, G. & Scheunders, P. (2011) Geodesics on the manifold of multivariate generalized Gaussian distributions with an application to multicomponent texture discrimination. *International Journal of Computer Vision*, vol. 95, No. 3, pp. 265-286.

22. Walker, K.N., Cootes, T.F. & Taylor, C.J. (1998) Locating salient object features. *Proc. British Machine Vision Conf.*, vol. 2, pp. 557-566.

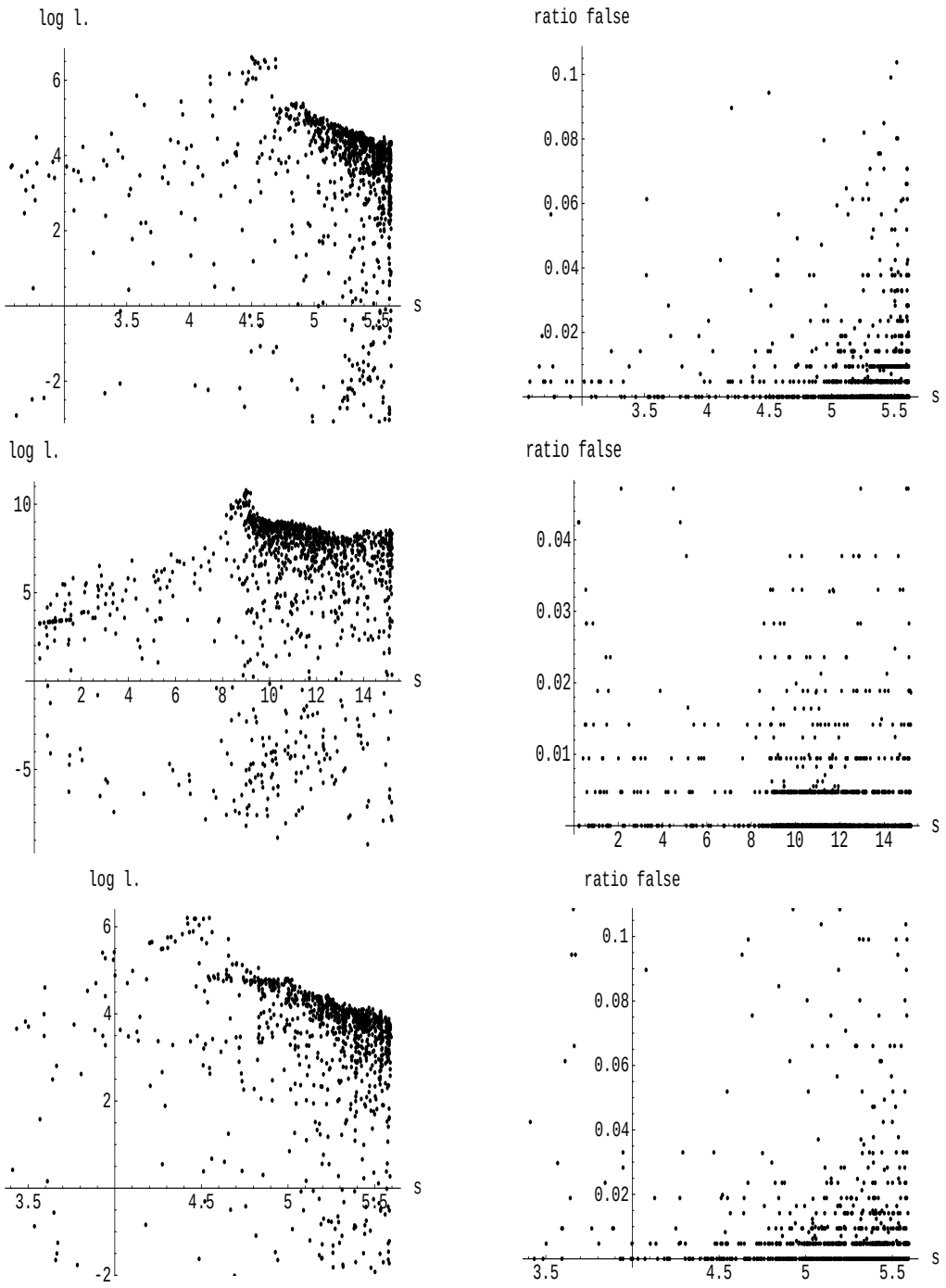


Figure 7: First column: scatter plots of the saliency s and the log likelihood ratio obtained using the true matching region. Second column: scatter plots of s and the proportion of background regions that yield a log likelihood ratio greater than or equal to the log likelihood ratio for the true match. First row: $m = 5$, $k_1 = 4$; second row: $m = 5$, $k_1 = 8$; third row: $m = 7$, $k_1 = 4$.

Steve Maybank received the BA degree in mathematics from King's College, Cambridge in 1976 and the PhD degree in computer science from Birkbeck College, University of London in 1988. He joined the Pattern Recognition Group at Marconi Command and Control Systems, Frimley in 1980 and moved to the GEC Hirst Research Centre, Wembley in 1989. From 1993-1995 he was a Royal Society/ Engineering and Physical Sciences Reasearch Council (EPSRC) Industrial Fellow in the Department of Engineering Science at the University of Oxford. In 1995 he joined the University of Reading as a lecturer in the Department of Computer Science. In 2004 he became a professor in the School of Computer Science and Information Systems, Birkbeck College. His research interests include the geometry of multiple images, camera calibration, visual surveillance, information geometry and the applications of statistics to computer vision. He was awarded the Koederink Prize at ECCV 2008 and is a Fellow of the IEEE.

***Photo of the author(s)**

Click here to download high resolution image