Application of the Fisher-Rao Metric to Structure Detection

Stephen J. Maybank

28 July 2005

School of Computer Science and Information Systems, Birkbeck College, Malet Street, London, WC1E 7HX, UK

(sjmaybank@dcs.bbk.ac.uk)

Abstract - Certain structure detection problems can be solved by sampling a parameter space for the different structures at a finite number of points and checking each point to see if the corresponding structure has a sufficient number of inlying measurements. The measurement space is a Riemannian manifold and the measurements relevant to a given structure are near to or on a submanifold which constitutes the structure. The probability density function for the errors in the measurements is described using a generalisation of the Gaussian density to Riemannian manifolds. The conditional probability density function for the measurements yields the Fisher information which defines a metric, known as the Fisher-Rao metric, on the parameter space. The main result is a derivation of an asymptotic approximation to the Fisher-Rao metric, under the assumption that the measurement noise is small. Using this approximation to the Fisher-Rao metric, the parameter space is sampled, such that each point of the parameter space is near to at least one sample point, to within the level of accuracy allowed by the measurement errors. The probability of a false detection of a structure is estimated. The feasibility of this ap-

proach to structure detection is tested experimentally using the example of line detection in digital images.

Index Terms-asymptotic approximation, false detection, Fisher-Rao metric, heat equation, Hough transform, line detection, parameter space, Riemannian manifold.

1 Introduction

In many practical cases of structure detection the measurements take values in a d-dimensional manifold D, the set of all structures is parameterised by the points θ in an n-dimensional parameter manifold T and the measurements exactly compatible with a given point θ in T form a submanifold $M(\theta)$ of D. For example, if the structures to be detected are lines in two dimensional images, then D is a subset of the image, T is a two dimensional parameter space for lines and the $M(\theta)$ are lines in D [22]. If the structures are epipolar transforms between pairs of images, then D is the Cartesian product of two images, T is a seven dimensional parameter space for the epipolar transforms and the $M(\theta)$ are certain quadric hypersurfaces in D [13].

The key quantity required for an analysis of the structure detection problem is the probability density function $p(x|\theta)$ for the measurement x conditional on the presence of the structure θ. In forming $p(x|\theta)$ it is assumed that x is caused by or arises from the structure θ. The formation of $p(x|\theta)$ is divided into two stages. In the first stage it is assumed that x arises from a single unknown error free measurement $\tilde{x}$ which is located on the structure θ. The probability density function for x conditional on $\tilde{x}$ is $p(x|\tilde{x})$. In the second stage it is assumed that $\tilde{x}$ has a known density $p(\tilde{x}|\theta)$ on $M(\theta)$. The density

$p(x|\theta)$ is obtained by integrating $p(x|\bar{x})$ over $M(\theta)$ with weight $p(x|\bar{x})$. Further details are given in Section 3.1.

The conditional density $p(x|\theta)$ gives rise to a Riemannian metric on T known in statistics as the Fisher-Rao metric [1,10,18,27]. The Fisher-Rao metric, J, has two key roles. Firstly, it defines a level of resolution on T, namely the level or scale at which pairs of points of T can be distinguished. If two points $\theta(1)$, $\theta(2)$ of T are close together under J, then the corresponding structures are difficult to distinguish given only a single measurement. The second key role of J is that it defines the volume [6,11] $V(T, J)$ of T. This volume measures the difficulty of detecting the structures parameterised by T. If $V(T, J)$ is small, then detection is easy because there is only a small number of distinguishable structures that might be detected in D. Conversely, if $V(T, J)$ is large, then detection is difficult because there is a large number of distinguishable structures that might be detected in D. This interpretation of $V(T, J)$ is one of the most important aspects of this work. It provides a quantitative measure of the difficulty of solving the associated detection problem.

The above properties of J suggest an algorithm for structure detection based on sampling T at a discrete set of points $\theta(i)$, $1 \leq i \leq n_s$. The $\theta(i)$ are chosen such that each point θ of T is near to at least one of the $\theta(i)$. If $V(T, J)$ is finite, then only a finite number of sample points is needed. Each $\theta(i)$ is checked in turn to see if the number of measurements near to the submanifold $M(\theta(i))$ of D exceeds a given threshold $r(N, \theta(i))$ where N is the total number of measurements. If the threshold is exceeded, then the structure $\theta(i)$ is detected. The number, n_s, of sample points is of order $O(V(T, J))$. The threshold $r(N, \theta(i))$ can be calculated under the assumption that the outliers for $\theta(i)$ are

distributed uniformly in D.

The two main advantages of this sample based approach are firstly that it is possible to detect structures in the presence of very large numbers of outliers and secondly any structure that is present in the image is detected. The accuracy with which a structure is located depends on the density of the sample points $\theta(i)$. If the density is high, then the structure is located accurately but the run time of the algorithm is long. In practice, a relatively low density of sample points is used. Once a structure is detected, the measurements inlying to it can be identified and the position of the structure estimated more accurately. The problem of finding the best position estimate, given a set of measurements known to be inliers, is outside the scope of this paper.

The main result of this paper is that $J(\theta)$ has an asymptotic approximation $K(\theta)$ which is accurate if the noise level is low. The approximation, $K(\theta)$, is useful because in all except the simplest cases the expression for $J(\theta)$, as given below in (1), is intractable. In many cases of practical interest, including line detection [22], and the detection of projective transformations of the line [23], the components of $K(\theta)$ are given by closed form expressions and it is possible to design and use statistically sound detection algorithms based on $K(\theta)$.

Related work on the Fisher-Rao metric and structure detection is described in Section 2. The probabilistic model for structure detection is described in Section 3 and the asymptotic approximation to J is obtained. Two special cases of the Fisher-Rao metric are described in Section 4. The sample based algorithm for structure detection is described in Section 5 and thresholds for structure detection are obtained. Experimental results are reported in Sections 6 and 7. Some concluding remarks and suggestions for future work

are made in Section 8. Mathematical results relating to Section 3 are given in Appendix A.

2 Related Work

Rissanen [28,29] shows that T can be sampled by a discrete set of points $\theta(i)$, $1 \leq i \leq O(V(T,J))$ such that any point of T is indistinguishable from at least one $\theta(i)$, given the measurements. In effect T can be replaced by a finite discrete approximation with only a negligible effect on the ability to detect structures. In [24], Myung *et al.* point out the importance of $V(T,J)$ as a measure of the complexity of structure detection, or to use their terminology, as a measure of model complexity. Jeffreys [16] and Balasubramanian [2,3] show that the correct prior density on T is the normalized canonical measure obtained from J. Under this prior density, the probability of a subset B of T is $V(B,J)/V(T,J)$. The prior probability depends on the number of distinguishable densities $x \mapsto p(x|\theta)$ as θ ranges over B. If B contains many densities which are easily distinguishable from each other given the measurements, then B is assigned a high prior probability. An important observation is that J and the prior density $B \mapsto V(B,J)/V(T,J)$ transform under reparameterisations of T in such a way that the probability assigned to any subset of T remains unchanged. If these probabilities were to change under reparameterisation of T, then the choice of parameterisation of T would depend on the statistical properties of the measurements. The invariance of the probabilities ensures that any mathematically convenient parameterisation of T can be chosen.

There are two well established structure detection methods based on discrete approximations to T. These methods are RANSAC and the Hough transform. To describe

RANSAC [9], let S be the set of measurements and let k be the minimum number of measurements sufficient to determine a structure uniquely in the noise free case. For example, if the structures are lines, then $k = 2$. A set of k measurements $S(k) = \{x(i(1)), \ldots, x(i(k))\}$ is chosen from S at random, and T is sampled at the point $\hat{\theta}(S(k))$. If $\hat{\theta}(S(k))$ has a sufficient number of inlying measurements, then a structure is detected. Sets $S(k)$ are chosen repeatedly and randomly, in order to ensure that any structure with a large number of inliers has a high probability of being detected.

In the Hough transform [15,19] T is divided into a finite set of regions called buckets or accumulator cells. Each bucket B has an associated integer valued counter $a(B)$ which is initialized at 0. Each measurement $x(i)$ is associated with a set $C(i)$ consisting of those $\theta \in T$ exactly compatible with $x(i)$. If $C(i)$ meets a bucket B, then the counter $a(B)$ is incremented by 1. After all the measurements have been examined, structures are detected corresponding to those buckets B for which $a(B)$ is sufficiently large.

In the literature the Hough transform has many variations depending on *i)* the way in which T is divided into cells; *ii)* the criteria for deciding if a structure θ is present; and *iii)* techniques for reducing the computational cost. For example, in [4, 20] a hierarchical division of the parameter space is used to speed up line detection. The parameter space is enclosed in a single box. If the box is proved not to contain any line with sufficient support from the measurements, then the box is discarded, otherwise it is subdivided and the search for lines continued recursively. A similar approach is developed in [26] and applied to circle detection.

The Hough transform is extended to more complicated shapes in the Generalised Hough Transform (GHT). A plane object is described by a set of points on the boundary

of the object. At each boundary point the direction of the normal to the boundary is recorded. In [31] a real time object recognition system based on the GHT is described. Recognition is speeded up by constructing multiresolution pyramids for the object and for the image. The search is begun at low resolutions and extended to higher resolutions.

Olson [25] describes an algorithm for detecting curves belonging to parameterised families for which the number, k, of parameters is large. The algorithm is a combination of RANSAC and the Hough transform. A set of j measurements is chosen at random and the search for curves supported by the measurements is restricted to those curves which pass exactly through the j measurements. This restriction fixes the values of j parameters. The remaining $k - j$ parameters are found using a Hough transform.

In [30] Rucklidge describes a sampling algorithm to find an affine transformation from a geometrical model to an image. The model consists of a finite set of points in known relative positions. The space of affine transformations is rasterised, i.e. sampled at a set of points defined by the vertices of a discrete lattice, and a divide and conquer strategy is used to search the vertices of the lattice for the affine transformation which most nearly matches the model points with feature points in the image. The lattice spacing is chosen such that the affine transformations associated with neighbouring lattice vertices differ on the image by no more than the width of one pixel.

The main contribution of this paper is a proof that the Fisher-Rao metric on the parameter space has an asymptotic approximation which is accurate if the noise level is small. The volume of the parameter space under this approximating metric is a fundamental measure of the difficulty of detecting the particular structure under consideration.

3 Probabilistic Model

The information contained in the differential structures of the manifolds D, T is only sufficient for a superficial analysis of the structure detection problem. In order to carry out a deeper analysis, it is necessary to include enough information about the measurement errors to determine the conditional probability density $p(x|\theta)$ for a measurement x given a structure θ.

A suitable definition of $p(x|\theta)$ is given in Section 3.1. The Fisher-Rao metric J is defined in Section 3.2 and the asymptotic approximations to $p(x|\theta)$ and J are obtained in Sections 3.3 and 3.4 respectively. The differential entropy of $p(x|\theta)$ is estimated in Section 3.5. Some mathematical results relevant to Sections 3.3, 3.4 and 3.5 are described in Appendix A.

3.1 Definition of the conditional density

The definition of $p(x|\theta)$ is a generalisation of one given by Werman and Keren [33] for the special case $D = \mathbb{R}^d$. It is assumed that D, T are differentiable manifolds with $\dim(D) = d$ and $\dim(T) = n$. The manifold T is parameterised by a vector θ containing n coordinates. It is assumed that D is compact and that each set $M(\theta)$ of measurements exactly compatible with θ is a compact submanifold of D with $\dim(M(\theta)) = m$ for all θ in T. The manifold D is given a Riemannian metric g with the associated canonical measure $d\mu$ [6,11]. The metric g usually arises naturally from the fact that D is a measurement space. For example, if D is an image, then g is the Euclidean metric defined on D by the pixel coordinates.

It is assumed that each measurement x is derived from an underlying noise free mea-

surement $\tilde{x}$ and that the conditional density $p(x|\tilde{x})$ for x given $\tilde{x}$ is the result of a heat flow or diffusion on D [5]. The heat flow begins at time 0 as a delta function concentration of heat at $\tilde{x}$ and lasts for a time t, giving a probability density function $p(x|\tilde{x}) = p_t(x|\tilde{x})$. If g is Euclidean in a neighbourhood of $\tilde{x}$, then $p_t(x|\tilde{x})$ is closely approximated by the Gaussian density with expected value $\tilde{x}$ and covariance $2tI$ where I is the $d \times d$ identity matrix.

Each submanifold $M(\theta)$ is given a probability measure dh which specifies the distribution of $\tilde{x}$ on $M(\theta)$. The simplest, default choice is to make dh equal to a scaled version of the measure induced on $M(\theta)$ by g. The scaling of dh is chosen such that the total volume of $M(\theta)$ under dh is 1.

The density $p(x|\theta) = p_t(x|\theta)$ is obtained by integrating the contributions $p_t(x|\tilde{x})$ as $\tilde{x}$ varies over $M(\theta)$, or equivalently, by solving the heat equation on D with the condition that at time 0 the heat is distributed on $M(\theta)$ in accordance with the probability measure dh [5]. As the time increases away from 0 towards t, the heat flows from $M(\theta)$ into the rest of D. If t is small, then the density $p_t(x|\theta)$ is concentrated in a neighbourhood of $M(\theta)$.

3.2 Definition of the Fisher-Rao metric

The Fisher-Rao metric [1,7] is given on T by the $n \times n$ symmetric matrix $J(\theta)$ defined at each point θ of T by

$$J_{ij}(\theta) = -\int_D \left(\partial^2_{\theta^i\theta^j} \ln p(x|\theta)\right) p(x|\theta)\, d\mu(x), \qquad 1 \le i, j \le n, \tag{1}$$

where θ^i, θ^j are components of the vector θ, and $\partial^2_{\theta^i\theta^j}$ is the differential operator defined such that

$$\partial^2_{\theta^i\theta^j} \ln p(x|\theta) = \frac{\partial^2}{\partial\theta^i \partial\theta^j} \ln p(x|\theta), \qquad 1 \leq i, j \leq n.$$

The superscript notation θ^i for the components of θ is usual in the context of Riemannian geometry [6,11]. The matrix $J(\theta)$ is the Fisher information at θ for a single measurement x. The Fisher information for k measurements sampled independently from D according to the density $p(x|\theta)$ is $kJ(\theta)$. The matrix $J(\theta)$ is used rather than $kJ(\theta)$ because it is not known a priori which subsets of measurements are inliers to a single structure.

If θ, θ' are nearby points in T then the squared distance between θ, θ' is

$$(\theta' - \theta)^\top J(\theta)(\theta' - \theta) + O_3(\theta' - \theta), \tag{2}$$

where $O_3(\theta' - \theta)$ consists of terms which are third order or higher in the components of $\theta' - \theta$. Another measure of the distance between θ', θ is the Kullback-Leibler distance [7,21], $D(\theta||\theta')$, defined by

$$D(\theta||\theta') = \int_D \ln(p(x|\theta)/p(x|\theta'))p(x|\theta)\, d\mu(x). \tag{3}$$

The log likelihood ratio $\ln(p(x|\theta)/p(x|\theta'))$ in the integrand of (3) is the key quantity needed for choosing between θ and θ' as 'explanations' for a given measurement x. The Kullback-Leibler distance is related to the squared distance (2) by

$$D(\theta||\theta') = \frac{1}{2}(\theta' - \theta)^\top J(\theta)(\theta' - \theta) + O_3(\theta' - \theta). \tag{4}$$

Let $\tau(\theta)\, d\theta$ be the canonical measure [6,11] on T associated with the Fisher-Rao metric,

$$\tau(\theta)\, d\theta = |\det(J(\theta))|^{1/2}\, d\theta.$$

The probability density $V(T,J)^{-1}\tau(\theta)\,d\theta$ is Jeffrey's prior density on T [16]. The volume $V(B,J)$ of any subset B of T under the canonical measure is independent of the choice of parameterisation of T. It is defined by

$$V(B,J) = \int_B \tau(\theta)\,d\theta. \tag{5}$$

If $V(B,J)$ is small, then the densities $p(x|\theta)$, $\theta \in B$ are similar. If x is a measurement chosen independently of B, then there is a relatively low probability that $p(x|\theta)$ will be large for some $\theta \in B$. Conversely, if $V(B,J)$ is large, then the densities $p(x|\theta)$ are dissimilar as θ varies over B and there is an increased probability that $p(x|\theta)$ will be large for some θ in B.

3.3 Asymptotic approximation to the conditional density

The density $p(x|\theta)$, as defined in Section 3.1, is a solution to the heat equation on D with the initial condition given by the measurement dh on $M(\theta)$. There is a well developed mathematical theory for the heat equation and its solutions [5]. The results of most interest here are that the heat equation has a fundamental solution, known as the heat kernel, and the heat kernel has an asymptotic expansion valid for small times $t > 0$. The expansion is stated as (37) in Appendix A. A proof can be found in [5].

It is assumed that $M(\theta)$ is a smooth compact submanifold of D. Let $\mathrm{dist}_g(x,x')$ be the geodesic distance between any two points x, x' of D. It is assumed that the initial probability measure dh on $M(\theta)$ has the form $dh = f(y)\,dy$ for a choice of coordinates y on $M(\theta)$. The notation $\mathrm{dist}_g(x,y)$ is also used for the distance between the point x of D and the point y of $M(\theta)$. In this latter use of the notation $\mathrm{dist}_g(x,y)$ the coordinates of x, y are specified in different coordinate systems. The coordinate system for x is defined

on D and the coordinate system for y is defined on $M(\theta)$.

It follows from (37), Appendix A that $p(x|\theta)$ has an asymptotic approximation,

$$p(x|\theta) = \frac{1}{(4\pi t)^{d/2}} \int_{M(\theta)} \exp\left(-\frac{\mathrm{dist}_g(x,y)^2}{4t}\right) \rho(\mathrm{dist}_g(x,y)) \left(u_0(x,y) + O(t)\right) f(y)\, dy,$$

$$0 < t. \tag{6}$$

The functions ρ and u_0 are defined in Appendix A. It is apparent from (6) that $2t$ is analogous to the variance σ^2 of a Gaussian density.

The approximation (6) to $p(x|\theta)$ is simplified by applying a result from the asymptotic analysis of multi-dimensional integrals given in Chapter IX of [35]. Let $x \mapsto w(x,\theta)$ be the function defined on D by

$$w(x,\theta) = \inf\{\mathrm{dist}_g(x,y) \mid y \in M(\theta)\}. \tag{7}$$

It is assumed that $M(\theta) \subset U(\theta)$, where $U(\theta)$ is an open set in D such that *i)* $p(x|\theta)$ is negligible for $x \notin U(\theta)$; and *ii)* if $x \in U(\theta)$, then there is a unique point $x(y) \in M(\theta)$ such that $w(x,\theta) = \mathrm{dist}_g(x, x(y))$. On taking a chart on $M(\theta)$, it can be assumed that $y \in \mathbb{R}^m$. These choices of coordinates are closely related to the Fermi coordinates, as described in [6]. Let $A(x,\theta)$ be the $m \times m$ matrix defined by

$$A_{ij}(x,\theta) = \partial^2_{y^i y^j} \mathrm{dist}_g(x,y)^2 \Big|_{y=y(x)}, \qquad 1 \le i,j \le m. \tag{8}$$

The matrix $A(x,\theta)$ is non-singular provided x is sufficiently close to $M(\theta)$, i.e. $w(x,\theta) < \delta$ for some $\delta > 0$. The assumption that $M(\theta)$ is compact ensures that a suitable δ exists for each θ. It is assumed that *i)* δ can be chosen independently of θ, *ii)* $w(x,\theta) > \delta$ if $x \notin U(\theta)$ and *iii)* the value of δ is significantly larger than $\sigma = (2t)^{1/2}$. In practice the lower bound $\delta > 3\sigma$ is sufficient. In order to ensure the existence of a suitable δ it may

be necessary to remove from T those structures θ for which $M(\theta)$ contains regions of very high curvature.

It follows from (6), (8) and the Laplace approximation in $\mathbb{R}^m$ [35] that

$$p(x|\theta) = \frac{2^{m/2} f(y(x)) \rho(\mathrm{dist}_g(x, y(x))) \left(u_0(x, y(x)) + O(t)\right)}{(4\pi t)^{(d-m)/2} (\det(A(x,\theta)))^{1/2}} \exp\left(-\frac{w(x,\theta)^2}{4t}\right). \tag{9}$$

It follows from (9) that

$$\ln p(x|\theta) = -\frac{w(x,\theta)^2}{4t} - \frac{1}{2}(d-m)\ln t + O(t^0). \tag{10}$$

Equation (10) is related to a result of Varadhan [5,32].

3.4 Asymptotic approximation to the Fisher information

It follows from (1), (10) and (36) that

$$\begin{aligned} J_{ij}(\theta) &= \frac{1}{4t}\int_D \left(\partial^2_{\theta^i\theta^j} w(x,\theta)^2\right) p(x|\theta)\, d\mu(x) + O(t^0), \\ &= \frac{1}{4t}\int_D \int_D \left(\partial^2_{\theta i\theta^j} w(x,\theta)^2\right) k(x,y,t) dh(y) d\mu(x) + O(t^0), \quad 1 \le i,j \le n, \end{aligned} \tag{11}$$

where k is the heat kernel on D. On exchanging the order of integration in (11) and applying (38), it follows that

$$\begin{aligned} J_{ij}(\theta) &= \frac{1}{4t}\int_D \int_D \left(\partial^2_{\theta^i\theta^j} w(x,\theta)^2\right) k(x,y,t) d\mu(x) dh(y) + O(t^0), \\ &= \frac{1}{4t}\int_D \left(\partial^2_{\theta^i\theta^j} w(x,\theta)^2\right)_{x=x(y)} dh(y) + O(t^0), \\ &= \frac{1}{4t}\int_{M(\theta)} \left(\partial^2_{\theta^i\theta^j} w(x,\theta)^2\right)_{x=x(y)} f(y)\, dy + O(t^0). \end{aligned} \tag{12}$$

Let $K(\theta)$ be the leading order term in the asymptotic expansion of $J(\theta)$ with respect to t. It follows from (12) that

$$K_{ij}(\theta) = \frac{1}{4t}\int_{M(\theta)} \left(\partial^2_{\theta^i\theta^j} w(x,\theta)^2\right)_{x=x(y)} f(y)\, dy, \qquad 1 \le i,j \le n. \tag{13}$$

Let N be the number of measurements. The matrix $N\,K(\theta)$ is closely related to the matrix $\bar{\boldsymbol{M}}_\infty$ defined by Kanatani in Section 14.4 of [17]. Kanatani identifies the inverse of $\bar{\boldsymbol{M}}_\infty$ as an asymptotic approximation, as $N \to \infty$, to the Cramer-Rao lower bound for the covariance of an unbiased estimate of θ.

The expression (13) for $K(\theta)$ can be simplified if $M(\theta)$ has codimension 1 in D, $m = d - 1$. Let $M(\theta)$ be fixed and let $M(\theta')$ be variable but with θ' close to θ. Let x be a point of D near to $M(\theta')$ and let y be a point on $M(\theta)$ near to x. Let $\eta(x, \theta') = \pm 1$ be defined such that

$$(x, \theta') \mapsto \tilde{w}(x, \theta') \equiv \eta(x, \theta') w(x, \theta'), \qquad x \in D,$$

is a C^2 function of x. Let $v(x, \theta)$ be the vector with components defined by

$$v_i(x, \theta) = \partial_{\theta'^i} \tilde{w}(x, \theta')|_{\theta'=\theta}\,, \qquad 1 \le i \le n. \tag{14}$$

The Taylor series expansion of $\tilde{w}(x, \theta')$ about θ is

$$\tilde{w}(x, \theta') = \tilde{w}(x, \theta) + \sum_{i=1}^{n} v_i(x, \theta)(\theta'^i - \theta^i) + O_2(\theta' - \theta), \tag{15}$$

where $O_2(\theta - \theta')$ consists of terms of second order or higher in the components of $\theta - \theta'$. It follows from (15) that

$$\begin{aligned} w(x, \theta')^2 &= \tilde{w}(x, \theta')^2 \\ &= \tilde{w}(x, \theta)^2 + \left(\sum_{i=1}^{n} v_i(\theta'^i - \theta^i) \right)^2 + \tilde{w}(x, \theta) O_2(\theta' - \theta) + O_3(\theta' - \theta), \end{aligned}$$

thus

$$\partial^2_{\theta'^i \theta'^j} w(x, \theta')^2 \Big|_{x=x(y), \theta'=\theta} = 2 v_i(x(y), \theta) v_j(x(y), \theta), \qquad 1 \le i, j \le n. \tag{16}$$

The simplified expression for $K(\theta)$ follows from (13), (16),

$$K(\theta) = \frac{1}{2t} \int_{M(\theta)} v(x(y), \theta) \otimes v(x(y), \theta)\, f(y)\, dy. \tag{17}$$

It is apparent from (17) that the components of $K(\theta)$ are large if small changes in θ produce large changes in the position of $M(\theta)$.

3.5 Differential entropy

An approximation to the differential entropy $H(\theta)$ of $p(x|\theta)$ is obtained. The differential entropy is the usual entropy, as defined for a continuous probability density function. The definition of $H(\theta)$ is

$$H(\theta) = -\int_D (\ln p(x|\theta)) p(x|\theta)\, d\mu(x). \tag{18}$$

It follows from (10) and (18) that

$$H(\theta) = \frac{1}{4t}\int_D w(x,\theta)^2 p(x|\theta)\, d\mu(x) + \frac{1}{2}(d-m)\int_D (\ln t) p(x|\theta)\, d\mu(x) + O(t^0). \tag{19}$$

It follows from (36),(38) in Appendix A that

$$\begin{aligned}
\frac{1}{4t}\int_D w(x,\theta)^2 p(x|\theta)\, d\mu(x) &= \frac{1}{4t}\int_D \int_D w(x,\theta)^2 k(x,z,t)\, dh(z) d\mu(x), \\
&= \frac{1}{4t}\int_D \left(\int_D w(x,\theta)^2 k(x,z,t)\, d\mu(x)\right) dh(z), \\
&= \frac{1}{4t}\int_D \Big(w(z,\theta)^2 + O(t)\Big)\, dh(z), \\
&= \frac{1}{4t}\int_{M(\theta)} O(t)\, dh, \\
&= O(t^0),
\end{aligned}$$

thus

$$H(\theta) = \frac{1}{2}(d-m)\ln t + O(t^0). \tag{20}$$

It is shown in Section 4.2 below that under certain conditions (20) leads to a simple expression for the trace of $K(\theta)$.

4 Special Cases

In certain very special cases, partial information about J and K can be obtained quickly. Two such cases are described in Sections 4.1 and 4.2 below.

4.1 Symmetry

Suppose that D has a coordinate system in which the Riemannian metric g is independent of one of the components x^k of x, and suppose that θ has a component, without loss of generality also labelled k, such that

$$\partial_{x^k} p(x|\theta) = -\partial_{\theta^k} p(x|\theta). \tag{21}$$

If x^k is an angular coordinate taking values in the circle, or if $p(x|\theta)$ and $\partial_{x^k} p(x|\theta)$ are negligibly small outside an open set on which the coordinates x are defined, then $J(\theta)$ is independent of θ^k. In the following proof the arguments of $p(x|\theta)$ are omitted in order to simplify the notation.

$$\begin{aligned}
\partial_{\theta^k} J_{ij}(\theta) &= -\int_D \partial_{\theta^k} \left(\left(\partial^2_{\theta^i \theta^j} \ln(p) \right) p \right) \, d\mu, \\
&= \int_D \partial_{x^k} \left(\left(\partial^2_{\theta^i \theta^j} \ln(p) \right) p \right) \, \det(g(x))^{1/2} \, dx, \\
&= \int_D \partial_{x^k} \left(\left(\partial^2_{\theta^i \theta^j} \ln(p) \right) p \, \det(g(x))^{1/2} \right) \, dx, \\
&= 0, \qquad\qquad 1 \leq i, j \leq n.
\end{aligned} \tag{22}$$

The matrix $K(\theta)$ is an asymptotic approximation to $J(\theta)$, thus $\partial_{\theta^k} K_{ij}(\theta) = 0$, $1 \leq i, j \leq n$.

Examples of (21) arise when D has a family of isometries which act in a simple way on the submanifolds $\theta \mapsto M(\theta)$. For example, suppose that D is the open unit disk and the $M(\theta)$ are lines in D. Cartesian coordinates are chosen in D with origin at the centre

of D. The set of lines in D is parameterised by $\theta = (\rho, \alpha)$, where $\rho(\cos(\alpha), \sin(\alpha))$ is the point on the line $M(\theta)$ closest to the origin. The manifold D is not compact, but the boundary effects are negligible for those lines (ρ, α) with $\rho \leq 1 - O(\sqrt{t})$. The length of the line $M(\theta)$ contained in D is $2(1-\rho^2)^{1/2}$ and the conditional density $p(x|\theta)$ is

$$p(x|\theta) = \frac{1}{(16\pi t)^{1/2}(1-\rho^2)^{1/2}} \exp\left(-w(x,\theta)^2/(4t)\right) + O(t).$$

The isometries of D are the rotations about the centre of D. A rotation through an angle β transforms the line $M(\rho, \alpha)$ to the line $M(\rho, \alpha + \beta)$. If polar coordinates (r, ϕ) are chosen in D, then $w(x, \theta) = (r\cos(\phi - \alpha) - \rho)^2$ and $\partial_\phi p(x|\theta) = -\partial_\alpha p(x|\theta)$. It follows from (22) that $K(\theta)$ is independent of α. An expression for $K(\theta)$ is stated below as (34), and derived in [22].

4.2 Trace of the Fisher-Rao metric on a torus

In this subsection the notation $p_t(x|\theta)$ is used rather than $p(x|\theta)$ because the main result, stated as (24) below, involves the differential operator ∂_t. Let D be the d-dimensional torus with g equal to a flat metric, let x^i, $1 \leq i \leq d$ be a set of coordinates on D, let θ be a d-dimensional vector of parameters such that

$$\partial_{x^i} p_t(x|\theta) = -\partial_{\theta^i} p_t(x|\theta), \qquad 1 \leq i \leq d, \tag{23}$$

and let $H(\theta)$ be the differential entropy of $p_t(x|\theta)$, as defined by (18). Then a version of de Bruijn's identity [7] holds,

$$\partial_t H(\theta) = \text{trace}(J(\theta)). \tag{24}$$

To prove (24), note first that $p_t(x|\theta)$ satisfies the heat equation on D,

$$\partial_t p_t(x|\theta) = \sum_{i=1}^{d} \partial^2_{x^i x^i} p_t(x|\theta).$$

Equation (24) is obtained by the following steps,

$$\begin{aligned}
\partial_t H(\theta) &= -\int_D (\ln p_t)\partial_t p_t \, dx, \\
&= -\int_D (\ln p_t) \left(\sum_{i=1}^{d} \partial^2_{x^i x^i} p_t \right) dx, \\
&= -\int_D \left(\sum_{i=1}^{d} \partial^2_{\theta^i \theta^i} \ln p_t \right) p_t \, dx, \\
&= \text{trace}(J(\theta)). \qquad (25)
\end{aligned}$$

It follows from (20) and (24) that $\text{trace}(J) = (d-m)/2t + O(t^0)$, thus

$$\text{trace}(K(\theta)) = \frac{1}{2t}(d-m). \qquad (26)$$

5 Algorithm for Structure Detection

If the volume $V(T, J)$ is small, then it is possible to search efficiently for structures θ by sampling T at a finite set of points $\theta(i)$, $1 \leq i \leq n_s$ and checking each $\theta(i)$ to see if it is supported by a sufficient number of measurements. As noted in Section 1, algorithms of this type can detect structures $M(\theta(i))$ even in the presence of large numbers of outliers. The outliers could be measurements arising from other structures $M(\theta(j))$ or they could be random in that they are not associated with any recognizable structure in the image.

The number $n(T, J)$ of distinguishable structures is estimated in Section 5.1. Methods for defining inliers and for reducing the probability of false detections are described in Sections 5.2 and 5.3. The algorithm itself is described in Section 5.4.

5.1 The number of sample points

Let γ be a strictly positive constant and define $B(\theta) \subseteq T$ by

$$B(\theta) = \left\{ \theta' \mid \theta' \in T \text{ and } (\theta' - \theta)^\top J(\theta)(\theta' - \theta) \leq \gamma \right\}. \tag{27}$$

If θ' is in $B(\theta)$ and if γ is sufficiently small, then the submanifolds $M(\theta)$, $M(\theta')$ of D are so close together that a measurement x sampled from $p(x|\theta)$ is with a high probability also a plausible sample from $p(x|\theta')$. The vector θ is chosen as a single representative for all the vectors θ' in $B(\theta)$. The sets $B(\theta)$ are similar to the buckets used in the Hough Transform [15,19].

Let $b(n)$ be the volume of the unit ball in the Euclidean space $\mathbb{R}^n$. It can be shown that the volume $V(B(\theta), J)$ of $B(\theta)$ under J, as defined by (5), satisfies $V(B(\theta), J) \approx \gamma^{n/2} b(n)$. In particular, $V(B(\theta), J)$ is independent of θ to leading order. The number, $n(T, J)$, of distinguishable structures in T is defined by

$$n(T, J) = \frac{V(T, J)}{\gamma^{n/2} b(n)}.$$

The quantity $n(T, J)$ is independent of the choice of parameterisation of T. If $n(T, J)$ is small, then it is easy to detect the presence of any structures θ in D by checking each one of $n_s = O(n(T, J))$ sample points in turn. Note that $n_s > n(T, J)$, because the sets $B(\theta)$ in any covering of T overlap if $n \geq 2$.

If γ is too small, then the sample will contain points $\theta(i)$, $\theta(j)$ for which $M(\theta(i))$, $M(\theta(j))$ are so close together in D that they cannot be distinguished using only a single measurement. Either $\theta(i)$ or $\theta(j)$ could be removed from the set of sample points without affecting the performance of a structure detection algorithm based on the sample. If γ is too large, then measurements far from $M(\theta)$ will be mistakenly classified as inliers

to $M(\theta)$. If θ' is on the boundary of $B(\theta)$, then $D(\theta||\theta')$, or equivalently, the weighted average of $\ln(p(x|\theta)/p(x|\theta'))$ over $x \in D$, is approximately $\gamma/2$. This suggests that a suitable value for γ is $\gamma \approx 1$.

5.2 Inliers and false detections

A measurement is defined to be an inlier to θ if it is contained in the subset $A(\theta)$ of D defined by

$$A(\theta) = \cup\{M(\theta') \mid \theta' \in B(\theta)\}.$$

Any measurements not contained in $A(\theta)$ are outliers for θ although they may be inliers for other structures θ'. The structure θ is detected if the number of measurements in $A(\theta)$ exceeds a threshold $r(N, \theta)$, where N is the total number of measurements.

If the threshold $r(N, \theta)$ is too small, then false detections occur: the structure θ is not in the image, but by chance $A(\theta)$ contains $r(N, \theta)$ or more measurements and θ is detected. If $r(N, \theta)$ is too big, then missed detections occur: θ is not detected, even though it is present, because $A(\theta)$ does not contain enough measurements. The exact value for $r(N, \theta)$ depends on the application, but a convenient default is to make $r(N, \theta)$ large enough to ensure that the false detection rate is small when the measurements are random in that they are sampled uniformly and independently from D.

The probability, $p(\theta)$, that a measurement sampled uniformly on D is contained in $A(\theta)$ is

$$p(\theta) = V(A(\theta), g)/V(D, g).$$

Let S be a set of N measurements sampled independently and uniformly from D. Let $E(\theta, r)$ be the event that r or more of the measurements in S are contained in $A(\theta)$. The

probability $\text{Prob}(E(\theta, r))$ of $E(\theta, r)$ is given by

$$\text{Prob}(E(\theta, r)) = \sum_{i=r}^{N} \binom{N}{i} p(\theta)^i (1 - p(\theta))^{N-i}.$$

The values of the $r(N, \theta)$ are chosen such that the $\text{Prob}(E(\theta, r(N, \theta)))$ are constant, to a first approximation. In effect, each $B(\theta)$ is regarded as a single structure and the different $B(\theta)$ contribute equally to the false detection rate. This choice of the $r(N, \theta)$ is in line with the observation that to first approximation the Jeffrey's prior density $\theta \mapsto V(T, J)^{-1}\tau(\theta)\, d\theta$ assigns the same probability to each set $B(\theta)$.

Let e_f be a user specified false detection rate. The term 'rate' is used rather than 'probability' because e_f might be greater than 1. The common value e_s of the $\text{Prob}(E(\theta, r(N, \theta))$ for θ in T is defined by

$$e_s = e_f / n(T, J). \tag{28}$$

The thresholds $r(N, \theta)$ are defined by

$$r(N, \theta) = \min\{r \mid \text{Prob}(E(\theta, r)) \leq e_s\}. \tag{29}$$

Numerical experiments, described in Section 6.2, strongly suggest that if the $r(N, \theta)$ are defined by (29), then the false alarm rate over the whole of T is approximately equal to e_f.

The probabilities $\text{Prob}(E(\theta, r))$ and hence the values of $r(N, \theta)$ defined by (29) can be estimated using the Poisson approximation to the Binomial distribution [8]. The parameter of the Poisson approximation is $\lambda(\theta) = Np(\theta)$ and

$$\text{Prob}(E(\theta, r)) \approx \exp(-\lambda(\theta)) \sum_{i=r}^{N} \lambda(\theta)^i / i!. \tag{30}$$

The threshold $r_p(N, \theta)$ estimated using (30) is often equal to $r(N, \theta)$ for the ranges of

parameter values relevant to structure detection, especially if $p(\theta)$ is small. For example, if $p(\theta) = 0.01$, $e_s = 0.5 \times 10^{-4}$ and $N = 450$, then $r_p(N, \theta) = r(N, \theta) = 16$.

5.3 Volume of $A(\theta)$

An estimate of $V(A(\theta), g)$ is obtained for the case in which $M(\theta)$ has codimension 1 in D. The strategy is to fix a point $x(y)$ on $M(\theta)$, take the normal to $M(\theta)$ at $x(y)$ and then find a point x' on the normal as far from $M(\theta)$ as possible subject to the condition that x' is in $A(\theta)$. Equivalently, x' is in $M(\theta')$ for some θ' on the boundary of $B(\theta)$. The volume, $V(A(\theta), g)$, is estimated by integrating $y \mapsto 2\text{dist}_g(x', y)$ over $M(\theta)$.

Let ξ be the distance from $x(y)$ to x'. It follows that $w(x', \theta) = \xi$. Let $\tilde{w}(x, \theta')$, $v(x, \theta)$ be as defined in Section 3.4. A Taylor expansion of $w(x', \theta')$ yields

$$\tilde{w}(x', \theta') = \tilde{w}(x', \theta) + \sum_{i=1}^{n} v_i(x', \theta)(\theta'^i - \theta^i) + O_2(\theta' - \theta), \tag{31}$$

where $O_2(\theta' - \theta)$ consists of terms second order or higher in the components of $\theta' - \theta$. It follows from (31) and the condition $\tilde{w}(x', \theta') = 0$ that

$$\xi = -v(x', \theta).(\theta' - \theta) + O_2(\theta' - \theta) = -v(x, \theta).(\theta' - \theta) + O_2(\theta' - \theta). \tag{32}$$

The second order terms in (32) are ignored and the absolute value of ξ is maximized subject to the condition that θ' is in the boundary of $B(\theta)$, or equivalently,

$$(\theta' - \theta)^\top J(\theta)(\theta' - \theta) = \gamma. \tag{33}$$

Let ξ_m be the value of ξ for which $|\xi|$ is a maximum. It follows from (32) and (33) that

$$\xi_m(x(y)) = -\gamma^{1/2}(v(x(y), \theta)^\top J(\theta)^{-1} v(x(y), \theta))^{1/2}.$$

Let $d\mu_{M(\theta)}$ be the measure induced on $M(\theta)$ by g. The volume $V(A(\theta), g)$ is estimated by

$$\begin{aligned} V(A(\theta), g) &= 2\int_{M(\theta)} |\xi_m(x(y))|\, d\mu_{M(\theta)}, \\ &= 2\gamma^{1/2} \int_{M(\theta)} (v(x(y), \theta)^\top J(\theta)^{-1} v(x(y), \theta))^{1/2}\, d\mu_{M(\theta)}. \end{aligned}$$

5.4 Algorithm

The discussion in Sections 5.1-5.3 above leads to the following algorithm for detecting structures. Note that $|S|$ is the number of elements in the finite set S.

Algorithm 1

Input: values of t, γ, e_f and a set S of measurements.

Output: A set of structures.

1. Choose a finite set $\theta(i)$, $1 \leq i \leq n_s$ of samples in T such that each point θ of T is in at least one of the sets $B(\theta(i))$.

2. Find the least thresholds $r(|S|, \theta(i))$, $1 \leq i \leq n_s$ sufficient to ensure that the probability of falsely detecting $\theta(i)$ is no greater that $e_f/n(T, J)$ when $|S|$ measurements are distributed independently and uniformly in D.

3. Form the set L of those $\theta(i)$ for which $|S \cap A(\theta(i))| \geq r(|S|, \theta(i))$, $1 \leq i \leq n_s$.

4. Output L.

5. Stop.

Step 1 can be carried out off-line if the values of t, γ are known. Similarly, Step 2 can

be carried out off-line if the $r(|S|, \theta(i))$, $1 \leq i \leq n_s$ can be calculated for the ranges of values of $|S|$ and e_f that arise in practice.

When Algorithm 1 is applied to the detection of lines in digital images the set L found in Step 3 is too large. The reason is that the measurements tend to clump together. A single clump of measurements can strongly support the presence of any one of a large number of lines, all of which pass through the clump. For this reason Algorithm 1 is modified. The modification involves editing L to remove those lines which are not required to 'explain' the measurements. Further details are given in Section 7.

6 Experimental Investigation of the Thresholds for Line Detection

The software for the experiments described in this Section and the next was written in Mathematica [34].

6.1 The Fisher-Rao metric for line detection

The theory developed in Sections 3 and 5 was applied to line detection in digital images. The measurement space D is the unit disk centred at the origin of Cartesian coordinates. The family of lines in D is parameterised by $\theta = (\rho, \alpha)$ where $\rho(\cos(\alpha), \sin(\alpha))$ is the point on the line θ closest to the origin. The parameter space is $T = [0, 1) \times [0, 2\pi)$ and the approximation $K(\theta)$ to $J(\theta)$ is [22]

$$K(\theta) = \frac{1}{2t} \begin{pmatrix} 1 & 0 \\ 0 & 3^{-1}(1 - \rho^2) \end{pmatrix}. \tag{34}$$

The approximation (34) to $J(\theta)$ fails if ρ is near to 1, or equivalently if only a small part of the line θ is included in D.

A short calculation yields

$$\begin{aligned} V(T,K) &= \pi^2/(4\sqrt{3}\,t), \\ n(T,K) &= \pi/(4\sqrt{3}\,\gamma t), \\ p(\theta) &= \pi^{-1}(8\gamma t(1-\rho^2))^{1/2}\left(2+\sinh^{-1}\left(\sqrt{3}\right)/\sqrt{3}\right). \end{aligned}$$

Let s be an arc length parameter on $M(\theta)$, chosen such that $s=0$ corresponds to the point $\rho(\cos(\alpha),\sin(\alpha))$. Let $\xi_m(s)$ be the distance from the point with arc length parameter s to the boundary of $A(\theta)$. It can be shown that to leading order

$$\xi_m(s) = \sqrt{2\gamma t}\left(1+3s^2(1-\rho^2)^{-1}\right)^{1/2}, \qquad -(1-\rho^2)^{1/2} \le s \le (1-\rho^2)^{1/2}.$$

The probability $p(\theta)$ is required to calculate the thresholds $r(N,\theta)$ and $\xi_m(s)$ is required to find those measurements which are in $A(\theta)$. A measurement x is in $A(\theta)$, at least to a first approximation, if $w(x,\theta) \le \xi_m(x(s))$ where $x(s)$ is the projection of x in a direction parallel to θ in D.

It is straightforward to choose the sample points $\theta(i)$, $1 \le i \le n_s$ because $K(\theta)$ is diagonal and $K_{11}(\theta)$ is constant. The quantity $\Delta\rho$ is defined by $\Delta\rho = 2\sqrt{\gamma t}$ and the ρ-axis is sampled at the points $\rho(i) = \Delta\rho/2 + i\Delta\rho$, $0 \le i \le \lfloor \Delta\rho^{-1} - 1/2 \rfloor$. To sample in the α direction, the quantities $\Delta\alpha(i)$ are defined by

$$\Delta\alpha(i) = \left(\frac{12\gamma t}{1-\rho(i)^2}\right)^{1/2}, \qquad 1 \le i \le \lfloor \Delta\rho^{-1} - 1/2 \rfloor,$$

and samples $(\rho(i), \Delta\alpha(i)/2 + j\Delta\alpha(i))$ chosen for $0 \le j \le \lfloor 2\pi\Delta\alpha(i)^{-1} - 1/2 \rfloor$.

6.2 Thresholds for line detection

Experiments were carried out to assess the prediction (29) for the threshold $r(N, \theta)$ for a false detection. The results are shown in Table 1. The lower part of the left hand column shows the values of the number N of random measurements, the top row shows the values of e_f, $e_f = 0.004, 0.02, 0.1, 0.5, 2.5$. Each row in the main body of Table 1 shows the number of false detections summed over the 50 trials with $\gamma = 1$, $t = 2 \times 10^{-4}$. In each trial all the contributions to the relevant row are included. The N measurements were sampled independently and uniformly from D. To give an example, if $N = 50$ and $e_f = 2.5$, then the average number of false detections per trial is estimated at 71/50.

e_f:	0.004	0.02	0.1	0.5	2.5
$N = 50$	0	0	1	12	71
$N = 150$	0	0	1	12	104
$N = 250$	0	0	6	26	125
$N = 350$	0	2	11	40	139

Table 1. Numbers of false detections.

An examination of Table 1 suggests that the number of false detections does not have a strong dependence on N. The key parameter controlling the number of false detections is e_f. This suggestion is supported by the graphs shown in Figure 1.

The entries in Table 1 suggest a simple model for estimating the number of false detections: assume that the $n(T, K)$ distinct structures in T contribute independently to the total number of false detections and that each structure has a probability of false detection $e_s = e_f/n(T, K)$. The expected number of false detections over 50 trials is

e_f:	0.004	0.02	0.1	0.5	2.5
$N = 50$	-0.45	2.00	0.89	-0.40	-2.60
$N = 150$	1.79	0.00	-0.89	-0.20	-0.35
$N = 250$	-0.45	-1.00	-1.79	-1.40	-1.16
$N = 350$	-0.45	-1.00	-1.34	-1.80	-1.43

Table 2. Normalised entries from Table 1.

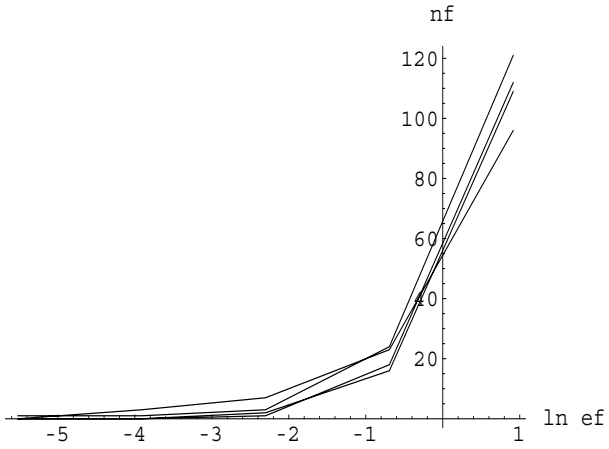

Figure 1: Numbers of false detections as a function of $\ln e_f$ for $N = 50, 150, 250, 350$.

$50e_f$. The variance over 50 trials is $50n(T, K)e_s(1 - e_s) = 50e_f(1 - e_s)$. These predictions are tested in Table 2. If m is an entry in a column of Table 1 headed by e_f, then the corresponding entry in Table 2 is

$$(m - 50e_f)/(50e_f(1 - e_s))^{1/2}.$$

The entries in Table 2 are O(1) in magnitude. This suggests that the model for the number of false detections is a useful guide for estimating thresholds, at least for the values of N in the range 50 to 350.

7 Line Detection

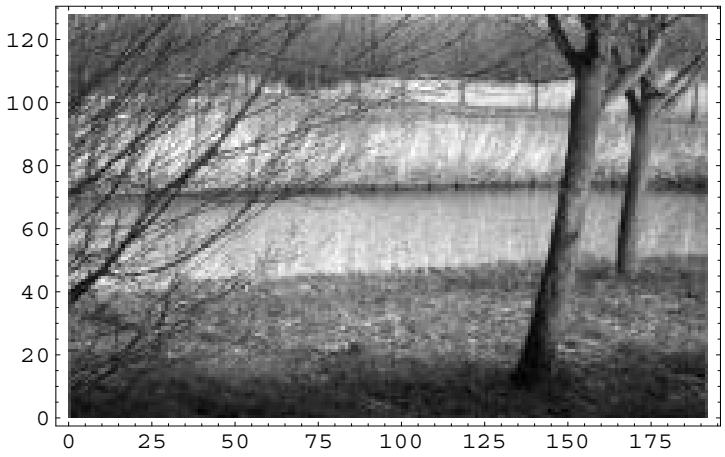

Figure 2: Image imk00038 in the Groningen natural image data base

Algorithm 1 was applied to the detection of lines in the grey level image shown in Figure 2. This image is imk00038 in the Groningen natural image data base [14] available at http://hlab.phys.rug.nl/l1_200/index.html. The image is of size 128×192, and it was taken by a Kodak DCS420 camera.

The centre of img00038 is $c = (64, 96)$. Let D' be the disk centred at c and with radius $r = 64$. The measurements were obtained by applying Sobel operator [12] to D' and choosing the N pixels with the largest responses. Each measurement x' in D' was mapped to D by $x' \mapsto x \equiv (x' - c)/r$. The following parameter values were used: $\gamma = 1$, $t = 0.5r^{-2}$, $N = 1000$, $e_f = 0.1$.

As noted in Section 5.4, the set L of lines produced by Step 3 of Algorithm 1 is too large. The difficulty is overcome by editing L in order to obtain a subset L_e such that

i) the inliers to each θ in L_e are distributed along θ with no gap greater than a given fraction δ of the length of θ;

ii) each θ in L_e has a sufficient number of inliers which are not shared with any other line in L_e.

To achieve *(i)*, let $\theta = (\rho, \alpha)$ be a line in L, let $\nu(\theta)$ be a unit vector parallel to θ in D and define the set $R(\theta)$ by

$$R(\theta) = \{2^{-1} + 2^{-1}(1-\rho^2)^{-1/2} x.\nu(\theta) \mid x \in A(\theta) \cap S\} \cup \{0, 1\}.$$

Let L' be the subset of L consisting of all lines θ in L such that the length of the largest gap between consecutive elements of $R(\theta)$ is less than or equal to δ. In these experiments, $\delta = 0.25$.

To achieve *(ii)*, the lines $\theta(1)$, $\theta(2)$, ... in L' are first ordered such that

$$|S \cap A(\theta(1))| \geq |S \cap A(\theta(2))| \geq \ldots.$$

The lines $\theta(i)$ are checked in order. Suppose that $\theta(i)$, $1 \leq i \leq k$ have been checked, and let L_e be the edited list of lines at this stage. Let $S(k+1)$ be the set defined by

$$S(k+1) = S \cap A(\theta(k+1)) \setminus \cup\{S \cap A(\theta) \mid \theta \text{ in } L_e\}.$$

The line $\theta(k+1)$ is added to L_e if and only if $|S(k+1)| \geq r(N, \theta(k+1))$.

The number of lines detected in Step 3 of Algorithm 1 is 452. The number of lines obtained after restricting the maximum size of the gap between the measurements is 160, and the number of lines obtained after restricting the inliers is 8. The detected lines are shown in Figure 3, superimposed on the measurements. The same detected lines are shown in Figure 4 superimposed on the original image. The circle in Figures 3 and 4 is the boundary of D'.

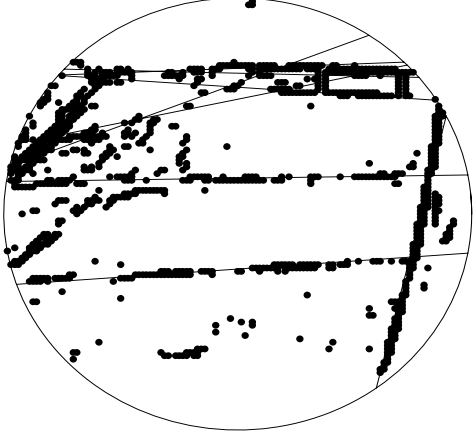

Figure 3: Lines detected in imk00038, superimposed on the measurements.

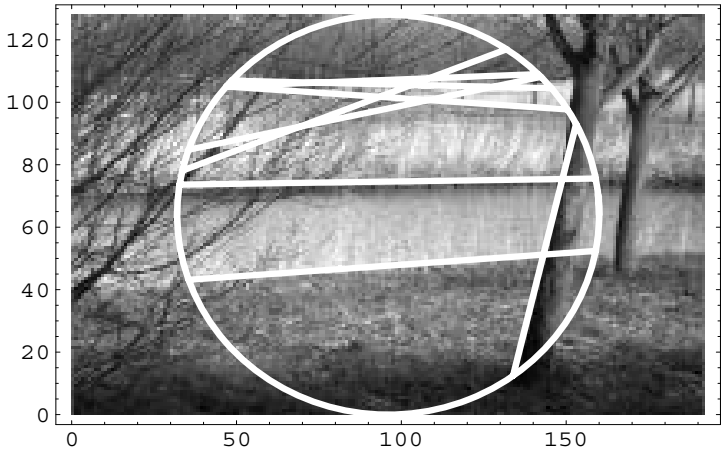

Figure 4: Lines detected in imk00038, superimposed on the original image.

8 Conclusion

A general probabilistic model for structure detection has been described. In the model the probability density function for the measurement errors is defined using a diffusion or heat flow from an unknown true measurement located on the structure. In the low noise case known results on solutions to the heat equation are applied to obtain asymptotic approximations to the probability density function for a measurement and to the Fisher-Rao metric on the parameter space. These results are the basis of an algorithm, Algorithm

1, for structure detection based on approximating the parameter space by a finite set of points.

The theoretical results are applied to the special case of line detection in digital images. A numerical study suggests that the false detection rate can be modelled by assuming that the false detections arise independently from $n(T, J)$ different sources, and that the probability that a single source gives rise to a false detection is $e_f/n(T, J)$ where e_f is a user specified estimate of the false detection rate. In this model the expected number of false detections is e_f and the the variance in the number of false detections is $e_f(1 - e_f/n(T, J))$.

The experiments with natural digital images show that a modification of Algorithm 1 is required to take account of the fact that the measurements tend to clump together. Once this modification is made, the lines in the images can be detected reliably.

There are two different ways in which the theory might be developed. The first way is to apply the results obtained so far to specific structure detection problems, especially those for which the asymptotic approximation to the Fisher-Rao metric takes a simple form. The second way is to construct more realistic models for the measurement process, in order to make better use of the information in the image grey levels.

Acknowledgement. I thank the referees for providing the references [4,20,26,31].

A Appendix

The aim in this appendix is to describe the asymptotic expansion of the fundamental solution to the heat equation on the measurement space D equipped with the Riemannian

metric g. Let $\triangle_x$ be the Laplace-Beltrami operator on D. The sign convention for $\triangle_x$ is chosen such that if $D = \mathbb{R}^d$ under the Euclidean metric, then $\triangle_x f$ is the negative of the Laplacian of f,

$$\triangle_x f = -\sum_{i=1}^{d} \partial^2_{x^i x^i} f.$$

Let f be a continuous real value function defined on $D \times (0, \infty)$ such that $x \mapsto f(x,t)$ is C^2 for all t in $(0,\infty)$ and $t \mapsto f(x,t)$ is C^1 for all x in D. The heat equation on D is $\triangle_x f = \partial_t f$. A fundamental solution to the heat equation on D is a continuous function $(x,y,t) \mapsto k(x,y,t)$ defined on $D \times D \times (0,\infty)$ such that k is C^2 with respect to x, C^1 with respect to t, k satisfies the heat equation $\triangle_x k = \partial_t k$, $t > 0$, $x \in D$ and if f is any bounded, continuous real function on D, then

$$\lim_{t \to 0+} \int_D k(x,y,t) f(x)\, d\mu(x) = f(y), \qquad y \in D. \tag{35}$$

The solution $p(x|\theta)$ to the heat equation with initial conditions given by the measure dh on $M(\theta)$ is obtained from the heat kernel k by

$$p(x|\theta) = \int_D k(x,y,t)\, dh \tag{36}$$

Let $\mathrm{inj}(D)$ be the injectivity radius of D, as defined in [5]. It is assumed that $\mathrm{inj}(D) > 0$. Let $\rho : [0,\infty) \to [0,1]$ be a C^∞ function such that $\rho(r) = 1$ for all r in $[0, \mathrm{inj}(D)/4]$, and $\rho(r) = 0$ for all r in $[\mathrm{inj}(D)/2, \infty)$. Let the set W be defined by

$$W = \{(x,y) \mid (x,y) \in D \times D \text{ and } \mathrm{dist}_g(x,y) < \mathrm{inj}(D)\},$$

In [5] functions u_0, u_1, ... are constructed with the following properties: each u_i is a real valued C^∞ function on W and $k(x,y,t)$ has the asymptotic expansion

$$k(x,y,t) = (4\pi t)^{-d/2} \exp\left(-\frac{\mathrm{dist}_g(x,y)^2}{4t}\right) \left(\rho(\mathrm{dist}_g(x,y)) \sum_{j=0}^{k} t^j u_j(x,y) + O(t^{k+1})\right),$$

$$t > 0, k > d/2 + 2. \tag{37}$$

It follows from (35) and (37) that if f is a bounded continuous real valued function defined on D, then

$$\int_D k(x, y, t) f(x)\, d\mu(x) = f(y) + O(t), \qquad y \in D. \tag{38}$$

References

[1] S.-I. Amari, *Differential-Geometrical Methods in Statistics.* Lecture Notes in Computer Science, vol. 28. Springer, 1985.

[2] V. Balasubramanian, "A geometric formulation of Occam's razor for inference of parametric distributions". Report No. PUPT-1588, Dept. of Physics, Princeton University, http://www.arxiv.org/list/nlin.AO/9601, 1996.

[3] V. Balasubramanian, "Statistical inference, Occam's razor and statistical mechanics on the space of probability distributions", *Neural Computation*, vol. 9, pp. 349-368, 1997.

[4] T.M. Breuel, "Finding lines under bounded error". *Pattern Recognition*, vol. 29,No. 1, pp. 167-178, 1996.

[5] I. Chavel, *Eigenvalues in Riemannian Geometry.* Academic Press Inc., 1984.

[6] I. Chavel, *Riemannian Geometry: a modern introduction.* Cambridge Tracts in Mathematics, vol. 108, CUP, 1996.

[7] T.M. Cover and J.A. Thomas, *Elements of Information Theory.* John Wiley and Sons, 1991.

[8] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. 1. Third edition, revised printing, Wiley, 1968.

[9] M.A. Fischler and R.C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". *Communications of the ACM*, pp. 381-395, 1981.

[10] R.A. Fisher, "On the mathematical foundations of theoretical statistics". *Phil. Trans. R. Soc. Lond. Series A*, Vol. 222 A, pp. 309-368, 1922.

[11] S. Gallot, D. Hulin and J. LaFontaine, *Riemannian Geometry.* 2nd edition, Universitext, Springer, 1990.

[12] R.C. Gonzalez and RE Woods, *Digital Image Processing.* Second edition, Prentice Hall, 2002.

[13] R.Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision.* CUP, 2003.

[14] J.H. van Hateren and A. van der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex". *Proc. R. Soc. Lond. Series B*, Vol. 265, pp. 359-366, 1998.

[15] J. Illingworth and J. Kittler, "A survey of the Hough transform". *Computer Vision, Graphics, and Image Processing*, Vol. 44, pp. 87-116, 1988.

[16] H. Jeffreys, *Theory of Probability.* Clarendon Press, Oxford, 1998.

[17] K. Kanatani, *Statistical Computation for Geometric Optimization: theory and practice.* Elsevier, 1996.

[18] S. Kotz and N.L. Johnson (eds.), *Breakthroughs in Statistics. Vol. 1. foundations and basic theory.* Springer Series in Statistics, Springer-Verlag, 1992.

[19] V.F. Leavers, *Shape Detection in Computer Vision Using the Hough Transform.* Springer Verlag, 1992.

[20] H. Li, M.A. Lavin and R.J. Le Master, "Fast Hough transform: a hierarchical approach. *Computer Vision, Graphics, and Image Processing.* vol. 36, pp. 139-161, 1986.

[21] S.J. Maybank, "Fisher information and model selection for projective transformations of the line". *Proceedings of the Royal Society of London*, Series A, vol. 459, pp. 1-21, 2003.

[22] S.J. Maybank, "Detection of image structures using Fisher information and the Rao metric". *IEEE Trans. Pattern Anal. and Machine Intell.*, Vol. 26, No. 12, pp. 1579-1589, 2004.

[23] S.J. Maybank, "The Fisher-Rao metric for projective transformations of the line". *International Journal of Computer Vision*, vol. 63, No. 3, pp. 191-206, 2005.

[24] I.J. Myung, V. Balasubramanian and M.A. Pitt, "Counting probability distributions: differential geometry and model selection". *Proc. Natl Acad. Sci.*, vol. 97, pp. 11170-11175, 2000.

[25] C.F. Olson, "Constrained Hough transforms for curve detection". *Computer Vision and Image Understanding*, Vol. 73, No. 3, pp. 329-345, 1999.

[26] C.F. Olson, "Locating geometric primitives by pruning the parameter space". *Pattern Recognition*, vol. 34, pp. 1247-1256, 2001.

[27] C.R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters". *Bull. Calcutta Math. Soc.*, vol. 37, pp. 81-91, 1945.

[28] J. Rissanen, "Fisher information and stochastic complexity". *IEEE Trans. Inf. Theory*, vol. 42, pp. 40-47, 1996.

[29] J. Rissanen, "Complexity and information in data". In A. Greven, G. Keller and G.

Warnecke (eds.) *Entropy*. Princeton Series in Applied Mathematics, pp. 299-312, 2003.

[30] W.J. Rucklidge, "Efficiently locating objects using the Hausdorff distance". *Int. Journal Computer Vision*, vol. 24, pp. 251-270, 1997.

[31] M. Ulrich, C. Steger and A. Baumgartner, "Real-time object recognition using a modified generalized Hough transform. *Pattern Recognition*, vol. 36, pp. 2557-2570, 2003.

[32] S.R.S. Varadhan, "On the behavior of the fundamental solution of the heat equation with variable coefficients". *Communications on Pure and Applied Mathematics*, vol. 20, pp. 431-455, 1967.

[33] M. Werman and D. Keren, "A novel Bayesian method for fitting parametric and non-parametric models to noisy data". *Proc. Computer Vision and Pattern Recognition*, Fort Collins, Colorado, June 1999, vol. 2, pp. 552-558, 1999.

[34] S. Wolfram, *The Mathematica Book*. 5th Edition, Wolfram Media, Inc., 2003.

[35] R. Wong, *Asymptotic Approximations of Integrals*. SIAM, 2001.