J. R. Statist. Soc. B (2001)
63, *Part* 3, *pp.* 465–514

A penalized likelihood approach to image warping

C. A. Glasbey

Biomathematics and Statistics Scotland, Edinburgh, UK

and K. V. Mardia

University of Leeds, UK

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, February 14th, 2001*, Professor P. J. Diggle *in the Chair*]

Summary. A warping is a function that deforms images by mapping between image domains. The choice of function is formulated statistically as maximum penalized likelihood, where the likelihood measures the similarity between images after warping and the penalty is a measure of distortion of a warping. The paper addresses two issues simultaneously, of how to choose the warping function and how to assess the alignment. A new, Fourier–von Mises image model is identified, with phase differences between Fourier-transformed images having von Mises distributions. Also, new, null set distortion criteria are proposed, with each criterion uniquely minimized by a particular set of polynomial functions. A conjugate gradient algorithm is used to estimate the warping function, which is numerically approximated by a piecewise bilinear function. The method is motivated by, and used to solve, three applied problems: to register a remotely sensed image with a map, to align microscope images obtained by using different optics and to discriminate between species of fish from photographic images.

Keywords: Bijective transformation; Conjugate gradients; Cross-covariance; Digital microscopy; Distortion criteria; Fast Fourier transform; Fish species discrimination; Phase correlation; Polynomial transformation; Registration; Similarity transformation; Synthetic aperture radar; Thin plate splines; von Mises distribution

1. Introduction

Image analysis, the extraction of information from pictures, is a broad interdisciplinary field with many challenging problems to which statistical methods are applicable (for overviews, see Mardia (1994) and Glasbey and Horgan (1995)). One such topic is *image warping*, a function that deforms images by mapping between image domains. Warping is a fundamental stage in many applications of image analysis, whether to register an image with a map or template, or to align multiple images. It dates back over a century, to Galton (1878), who used analogue methods to construct average faces of criminals and mental patients from photographs. Since then, the subject has had a large and diverse literature. The images to be aligned may be different specimens to be compared to characterize population variation, or the same specimen at different times to be interpolated between ('morphed') or complementary sources of information to be fused. Alternatively, they may be either successive two-dimensional sections or stereoscopic pairs, from which a three-dimensional scene is to be reconstructed. In some applications, different types of deformation or even

Address for correspondence: C. A. Glasbey, Biomathematics and Statistics Scotland, James Clerk Maxwell Building, King's Buildings, Mayfield Road, Edinburgh, EH9 3JZ, UK.
E-mail: chris@bioss.sari.ac.uk

 1369–7412/01/63465

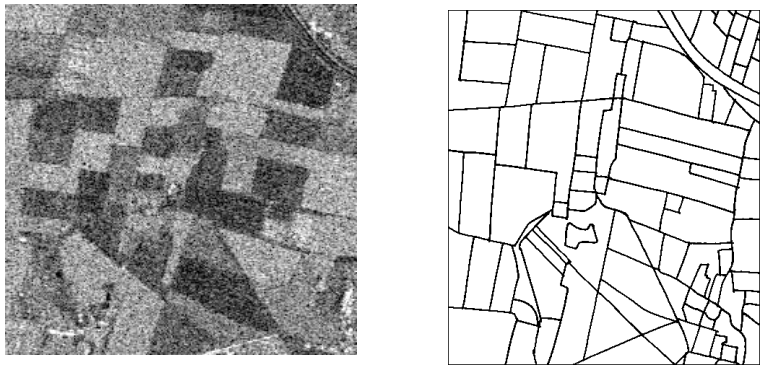
discontinuities may be permissible in parts of images. The transformation may be constrained to be one to one, i.e. bijective, or folding may be acceptable. Also, it may or may not be appropriate for the boundaries of one image domain to map to the boundaries of the other domain. The accuracy required of the alignment is another issue: if a radiologist is to make a visual assessment of two medical images then a precise alignment may be unnecessary, whereas in remote sensing, where quantitative use is to be made of images, subpixel registration may be critical.

To illustrate, we consider three applied problems, the data for which are obtainable from

`http://www.blackwellpublishers.co.uk/rss/`

(a) Fig. 1(a) shows a remotely sensed synthetic aperture radar (SAR) image of an area near Feltwell, England. SAR is an active remote sensing system: microwave radiation is beamed down to the earth's surface from a plane or satellite, a sensor detects the reflected signal and from this an image is constructed. Before any practical use can be made of such an image, it needs to be registered with a map, such as the digitized map of field boundaries in Fig. 1(b) (problem 1). Registration of remotely sensed images, including SAR, is often performed manually (see, for example, Vornberger and Bindschadler (1992) and Dobson *et al.* (1996)). Li *et al.* (1995) reviewed automatic methods, distinguishing between area- and feature-based methods. To locate features, Caves *et al.* (1992) used linear filters, whereas Kher and Mitra (1993) used morphological methods. Registration of SAR can also simplify the task of segmenting the images into homogeneous regions (Glasbey, 1997).

(b) Fig. 2 shows a sample of algae imaged using three light microscope modalities: brightfield, differential interference contrast and phase contrast. Brightfield microscopy reveals the optical attenuation of the specimen, whereas differential interference contrast microscopy responds to the refractive properties of the specimen and phase contrast microscopy shows diffractive properties. By fusing the images, these sources of complementary information can be combined (Modrusan *et al.*, 1994; Ried *et al.*, 1992). However, this requires a translation to be applied to the images to compensate for changes in image alignment resulting from imperfect centring of the different lens



(a) (b)

Fig. 1. Area to the north of the village of Feltwell in East Anglia: (a) an aerial SAR image, 250×250 pixels in size (3 km $\times$ 3 km); (b) digital line drawing of field, road and other boundaries for approximately the same region

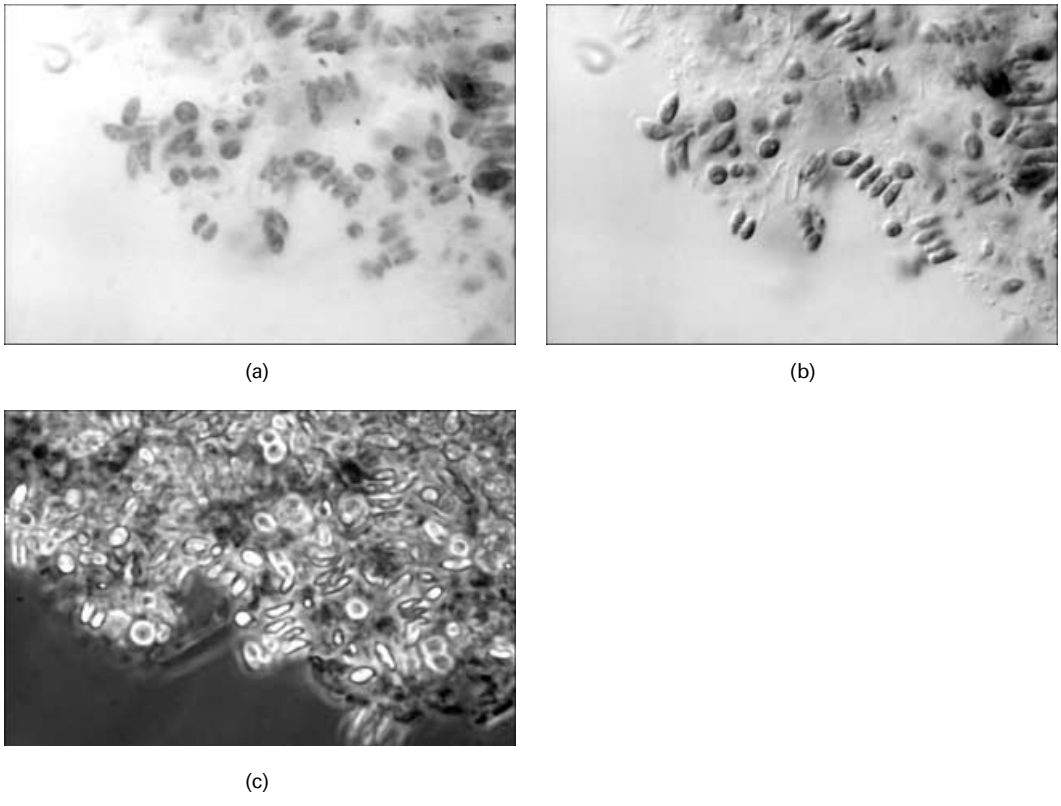


Fig. 2. Sample of algae imaged using three light microscope modalities: (a) brightfield; (b) differential interference contrast; (c) phase contrast (the images are 512×768 pixels in size)

systems (problem 2). Galbraith and Farkas (1993) described two methods for aligning images, involving either the imaging of a rectangular grid or the manual identification of control points.

(c) Fig. 3 shows photographic images, obtained under controlled conditions, of two species of fish (haddock and whiting) that we wish to discriminate (problem 3). These are part of a larger data set consisting of images of 10 haddocks and 10 whitings. Strachan *et al.* (1990) analysed images of seven species and found these two species to be the most difficult to distinguish. One way of comparing images is by warping them to align with each other. These images have already been aligned globally, and our concern is with local alignment. The study of fish shape is a subject with a long history. Comparisons have typically been restricted to the fish outlines and a few other features, but simple measures such as length-to-width ratios are not sufficient in this application. Thompson (1917) used a mapping which superimposed an outline of one fish on another as a way of comparing shapes. Bookstein (1991) developed this further, whereas Strachan *et al.* (1990) used summary statistics derived from outlines to discriminate between seven species of fish and Mokhtarian (1995) used curvature scale space to recognize marine animals.

Our proposal is a statistical formulation of image warping, using a penalized likelihood approach. However, we first summarize a large number of alternative approaches,

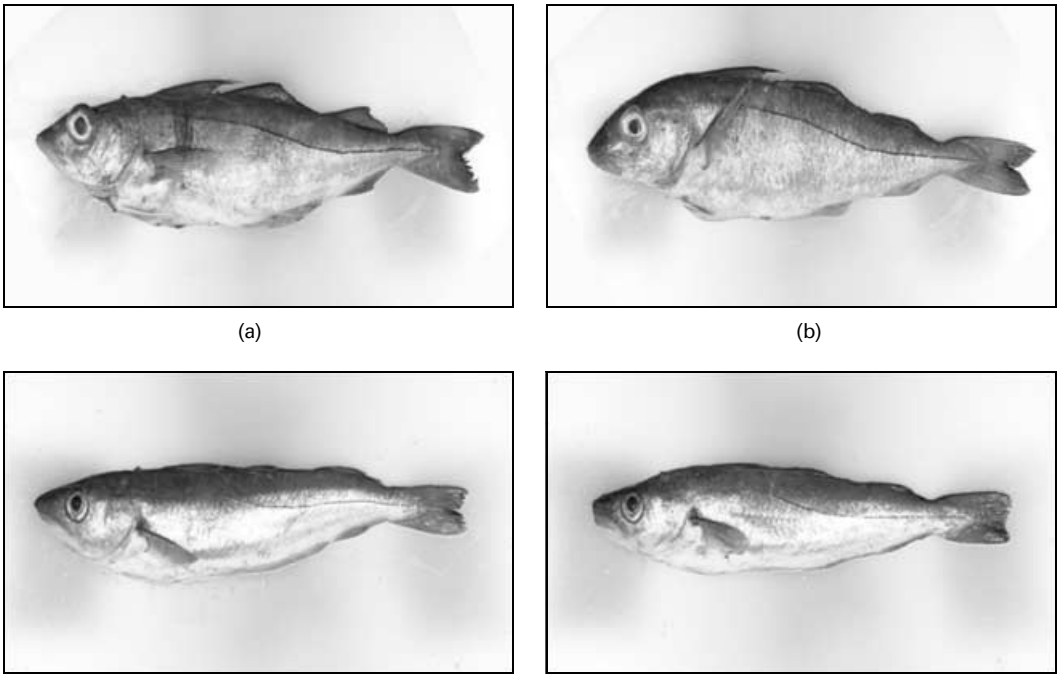


Fig. 3. Images of two species of fish, photographed on a light table: (a) haddock 1; (b) haddock 2; (c) whiting 1; (d) whiting 2 (the images are 300×500 pixels in size)

predominantly in the computer vision and engineering literatures. There are recent reviews of image warping in general (Glasbey and Mardia, 1998; Goshtasby and Le Moigne, 1999), of medical applications and computational anatomy (Grenander and Miller, 1998; Maintz and Viergever, 1998; Singh *et al*., 1998), and brain imaging specifically (Toga, 1999; Cao and Worsley, 1999), of comparisons of faces (Hallinan *et al*., 1999) and of templates and shape analysis (McInerney and Terzopoulos, 1996; Dryden and Mardia, 1998; Loncaric, 1998). For the special case of one-dimensional curve registration, see Ramsay and Li (1998). Measures of similarity to assess the quality of image alignment have included mean-square differences and correlation between pixels in images (see, for example, Rosenfeld and Kak (1982), section 9.4), phase correlation (Kuglin and Hines, 1975), coincidence of landmark points (Cross and Hancock, 1998; Hill *et al*., 2000) or edges in images (Bajcsy and Kovacic, 1989; Moshfeghi, 1991), mutual information (Meyer *et al*., 1996; Viola and Wells, 1997; Rangarajan *et al*., 1999; Studholme *et al*., 1999) and distance metrics (Baddeley and Molchanov, 1998; Kaijser, 1998). Measures used to ensure that the warping is not too severe have been motivated by thin plate splines (Bookstein, 1991), elastic deformations (Burr, 1981; Younes, 1999), optical or fluid flow (Barron *et al*., 1994; Christensen *et al*., 1996; Joshi and Miller, 2000), diffusion (Amit *et al*., 1991), numerical regularizers (Thompson *et al*., 1991), Hopfield neural networks (Cote and Tatnall, 1997) and Bayesian prior distributions (Carstensen, 1996; Gee, 1999).

Our proposal builds on much of this earlier work but is distinctive. It is also our intention to give image warping more exposure to a statistical audience, which we think it needs. In Section 2 we formulate image warping as a penalized likelihood problem, incorporating new classes of both similarity measures and distortion penalties. We restrict attention to

two-dimensional images, though the theory extends in a straightforward manner to three and higher dimensions. Then, in Section 3 we apply the method to solve the three problems above. Finally, in Section 4 we discuss the results.

2. Method

Suppose that we have a single image Y that we wish to align with another, given, image μ, sometimes referred to in the computer vision literature as a grey scale template. We propose to do so by estimating the warping function f: $\Re^2 \mapsto \Re^2$ to maximize a *penalized likelihood* functional P consisting of two components:

$$P(Y|\mu, f, \xi, \mathcal{C}, \lambda) = L(Y|\mu, f, \xi) - \lambda D(f, \mathcal{C}). \quad (1)$$

Here, L is the log-likelihood for Y, which depends on the warping function f and parameters ξ. The log-likelihood operates as a measure of similarity between μ and the warped version of Y. The second component, D, is a non-negative *measure of distortion* of f, chosen to be 0 if and only if $f \in \mathcal{C}$, a *null set* of functions, and λ is a non-negative constant that determines the relative weighting between L and D. We use the term 'distortion' in preference to the commonly used term 'roughness', because we sometimes wish to penalize warpings that would not be considered rough in the general sense of that word. As f is infinite dimensional, in the absence of a measure of distortion, the problem would be ill conditioned. Penalized likelihoods have appeared in the statistical literature in many other contexts and may be justified in several ways, including as regularizers and in Bayesian formulations (see, for example, Green and Silverman (1994) and Green (1999)). We could place a probabilistic interpretation on f, via $D(f, \mathcal{C})$ (see, for example, Grenander and Miller (1998)) but we prefer to leave it ambiguous.

If we have two images, then it may be natural to use one as the grey scale template μ. For example, in problem 1 we take the SAR image to be Y, which we align with the digital map, taken to be μ. However, if we have two images that we wish to treat interchangeably, or K (> 2) images, $Y^{(1)}, \ldots, Y^{(K)}$, then μ takes on the role of a consensus image that we also need to estimate. We generalize equation (1) to

$$P^{(K)}(Y^{(1)}, \ldots, Y^{(K)}|\mu, f^{(1)}, \ldots, f^{(K)}, \xi^{(1)}, \ldots, \xi^{(K)}, \mathcal{C}, \lambda) = \sum_k P(Y^{(k)}|\mu, f^{(k)}, \xi^{(k)}, \mathcal{C}, \lambda) \quad (2)$$

where $f^{(k)}$ denotes the warping function from $Y^{(k)}$ to μ, and we maximize $P^{(K)}$ also with respect to μ, which is an array of location parameters. Note, however, that it is not always possible to estimate μ and ξ simultaneously, a topic to which we shall return in Section 3.2.

We consider specific forms for L in Section 2.1 and for D in Section 2.2; then we describe an algorithm to estimate f in Section 2.3. There are many ways to choose λ (see, for example, Thompson *et al.* (1991)). We illustrate some specific strategies in the applications, namely cross-validation (Section 3.1), prior knowledge (Section 3.2) and discriminatory power (Section 3.3).

2.1. Fourier–von Mises image model

We consider two log-likelihoods; the first is based on a Gaussian model for Y after warping, but our main, novel proposal is for a *Fourier–von Mises image model*. First we need further notation.

Image μ is a real function, either on a discrete domain, $X = \{1, \ldots, n_1\} \times \{1, \ldots, n_2\}$, so

that μ: $X \mapsto \Re$, or on a continuous domain, so that μ: $(0, n_1) \times (0, n_2) \mapsto \Re$, according to which is more convenient (although values of μ are typically only known or estimated on the discrete domain). In this section the domain is taken to be discrete. We use μ_x to denote the pixel value at location $x = (x_1, x_2)$. Image Y is similarly specified and has a possibly different size, $n' = (n_1', n_2')$. We define Y_f to be the warped version of Y under f. It is an array of size $n = (n_1, n_2)$ specified by

$$(Y_f)_x \equiv Y_{f(x)} \qquad \forall x \in X,$$

and so its pixel value at x is defined to be the value of Y at location $f(x)$. Typically, $(f_1(x), f_2(x))$ are not integers, so $Y_{f(x)}$ is obtained by interpolation, and, if $f(x)$ lies outside the domain of Y, $Y_{f(x)}$ is defined to be a constant: either 0 or a mean pixel value. We use bilinear interpolation, though it would be possible to use alternatives such as splines or kernels.

Consider a simple Gaussian model for Y, conditional on f, of the form

$$Y_{f(x)} \sim N(\mu_x, \sigma^2) \qquad \forall x \in X, \tag{3}$$

with Y_x for other values of x specified deterministically, by bilinear interpolation, for example. We regard Y as a single entity rather than as an array of individual observations, as in the philosophy in functional data analysis of Ramsay and Silverman (1997), pages 37–38. The log-likelihood of Y, to within additive and scaling constants, is

$$L^*(Y|\mu, f) = -\sum_x (Y_{f(x)} - \mu_x)^2, \tag{4}$$

where, throughout the paper, the x-summation is over X. This Gaussian model may be reasonable for problem 3, the fish images in Section 1, but not for all applications. For example, in problem 2, algal cells appear dark in Fig. 2(a), whereas in Fig. 2(b) one side of each cell appears dark whereas the other side appears light, and in Fig. 2(c) cells look different again. Therefore, it would be more appropriate to model the relationship between the edges of cells in the two images, rather than the image intensities directly. By constructing an image model in the Fourier domain, we can be flexible in allowing either intensities or edges to be related, provided that the edges can be extracted by using linear filters, as we show below.

The Fourier representation of Y_f is

$$Y_{f(x)} = \frac{1}{\sqrt{(n_1 n_2)}} \sum_\omega A_\omega^{(Y_f)} \cos(\theta_\omega^{(Y_f)} + 2\pi\omega^{\mathrm{T}} x) \qquad \forall x \in X, \tag{5}$$

where $A^{(Y_f)}$ and $\theta^{(Y_f)}$ are respectively the arrays of amplitudes and phases of the Fourier transform of Y_f. Throughout the paper, the ω-summation is over Ω, the set of frequencies $\omega = (j_1/n_1, j_2/n_2)$ for $j_i = -\frac{1}{2}n_i, (-\frac{1}{2}n_i + 1), \ldots, -1, 0, 1, \ldots, (\frac{1}{2}n_i - 2), (\frac{1}{2}n_i - 1)$ if n_i is even or $j_i = -\frac{1}{2}(n_i - 1), \ldots, \frac{1}{2}(n_i - 1)$ if n_i is odd. Similarly, we define $A^{(\mu)}$ and $\theta^{(\mu)}$ to be the arrays of amplitudes and phases of the Fourier transform of μ. Note that the arrays have a rotational symmetry, as $A_\omega^{(Y_f)} = A_{-\omega}^{(Y_f)}$ and $\theta_\omega^{(Y_f)} = -\theta_{-\omega}^{(Y_f)}$. Also, any linear filter applied to μ can be interpreted and computed simply as a rescaling of each element in $A^{(\mu)}$ and the addition of a constant to $\theta^{(\mu)}$. The arrays can be computed efficiently by using fast Fourier transforms, and we taper the image boundaries by using a cosine bell, to remove artificial image discontinuities produced by wraparound of the image domain. For an introductory background to Fourier analysis of images, see, for example, Glasbey and Horgan (1995), chapter 3, pages 60–70.

We now specify our Fourier–von Mises image model for Y, conditional on f. The Fourier phases, $\theta^{(Y_f)}$, are independently von Mises distributed, conditional on $A^{(Y_f)}$, as follows:

$$(\theta_\omega^{(Y_f)} | A^{(Y_f)}) \sim M\{\theta_\omega^{(\mu)}, \kappa_\omega(\xi)\} \qquad \forall \omega \in \Omega, \tag{6}$$

where the concentration κ (> 0) is given by

$$\kappa_\omega(\xi) = \exp\{\xi_0 + \xi_1|\omega| + \xi_2|\omega|^2 + \xi_3 \log(A_\omega^{(\mu)}) + \xi_4 \log(A_\omega^{(Y_f)})\}. \tag{7}$$

Here, $|\omega|$ denotes the modulus of ω, a non-direction frequency, and κ is a log-linear function of $|\omega|$, $|\omega|^2$, $A^{(\mu)}$ and $A^{(Y_f)}$, with parameters ξ. The Fourier amplitudes $A^{(Y_f)}$ are regarded as an array of constants rather than random variables, and we use equation (5) to define Y_x for all values of x, not just for $f(x) \in X$. Therefore, the log-likelihood for Y, to within an additive constant including terms in $A^{(Y_f)}$ for the Jacobian of the transformation, is

$$L(Y|\mu, f, \xi) = \sum_\omega \kappa_\omega(\xi) \cos(\theta_\omega^{(Y_f)} - \theta_\omega^{(\mu)}) - \sum_\omega \log[I_0\{\kappa_\omega(\xi)\}], \tag{8}$$

where I_0 is the normalizing term for the von Mises distribution, a modified Bessel function of the first kind and of order 0 (see, for example, Mardia and Jupp (1999)). We have the following four theoretical and empirical motivations for choosing this model.

Firstly, for particular choices of ξ, the log-likelihood simplifies to commonly used measures of similarity between images. If $\xi = (\xi_0, 0, 0, 1, 1)$ then $\kappa \propto A^{(\mu)}A^{(Y_f)}$ and L can be re-expressed as

$$L\{Y|\mu, f, (\xi_0, 0, 0, 1, 1)\} = \sum_x \mu_x Y_{f(x)} - \sum_\omega \log\{I_0(A_\omega^{(\mu)}A_\omega^{(Y_f)})\}. \tag{9}$$

The first term is the cross-covariance or cross-product between μ and Y, which is closely related to L^*, given by equation (4). If $\xi = (\xi_0, 0, 0, 0, 0)$ then κ is a constant and we obtain the phase correlation measure (Kuglin and Hines, 1975), the cross-covariance between the images after the application of a high pass filter which results in the filtered images having flat spectra (Glasbey and Horgan (1995), Fig. 3.6b, page 65). In general, L can be re-expressed, and interpreted, as the cross-covariance between Y_f and a filtered version of μ, which we denote by $\mu^{<\xi>}$,

$$L(Y|\mu, f, \xi) = \sum_x \mu_x^{<\xi>} Y_{f(x)} - \sum_\omega \log[I_0\{\kappa_\omega(\xi)\}], \tag{10}$$

where

$$\mu_x^{<\xi>} = \frac{1}{\sqrt{(n_1 n_2)}} \sum_\omega \frac{\kappa_\omega(\xi)}{A_\omega^{(Y_f)}} \cos(\theta_\omega^{(\mu)} + 2\pi\omega^{\mathrm{T}}x) \qquad \forall x \in X. \tag{11}$$

Array $\mu^{<\xi>}$ is a filtered version of μ, obtained by modifying the amplitudes in the Fourier transform and then back-transforming. It is also a function of f but we suppress this dependence for reasons discussed in the optimization algorithm in Section 2.3. Alternatively, we could have applied the filter to Y_f, or shared its effects between both μ and Y_f, but we shall make use of equation (10) in Section 2.3. Typically the effect of the filter will be to enhance edges in images. Thus, we have combined intensity matching and edge matching in one measure, unlike, for example, Hallinan *et al.* (1999), who treated them separately. Note that our approach is different from those of others, who have used local Fourier methods, such as Gabor filters (Lades *et al.*, 1993) and frequency varying chirp-like filters (Bonmassar and Schwartz, 1997; Tabernero *et al.*, 1999) to extract landmarks from images.

Secondly, we specify $A^{(Y_f)}$ to be an array of constants because most information about the warping f is contained in $\theta^{(Y_f)}$ rather than in $A^{(Y_f)}$, and also because in general it is difficult to construct a realistic stochastic model for $A^{(Y_f)}$. In particular, when f is simply a translation function, as is appropriate for problem 2, the microscopy images in Section 1, all the information is in $\theta^{(Y_f)}$. This follows because there is a simple relationship between the Fourier transforms of Y and Y_f, given by

$$A_\omega^{(Y_f)} = A_\omega^{(Y)}, \quad \theta_\omega^{(Y_f)} = \theta_\omega^{(Y)} + 2\pi\omega^{\mathrm{T}}\alpha \qquad \forall\omega \in \Omega, \quad \text{where } f_i = \alpha_i + x_i \mod(n_i), \quad i = 1, 2, \tag{12}$$

for constants α_1 and α_2, provided that we allow modulo n wraparound in the translation.

Thirdly, a von Mises distribution is a natural choice for an angular variable such as $\theta^{(Y_f)}$, as it is, in many ways, the circular equivalent of the Gaussian distribution. It can also be derived from the Gaussian model (3), because then

$$A_\omega^{(Y_f)} \cos(\theta_\omega^{(Y_f)}) \sim N\{A_\omega^{(\mu)} \cos(\theta_\omega^{(\mu)}), \sigma^2\}, \qquad A_\omega^{(Y_f)} \sin(\theta_\omega^{(Y_f)}) \sim N\{A_\omega^{(\mu)} \sin(\theta_\omega^{(\mu)}), \sigma^2\}, \qquad \forall\omega \in \Omega.$$

These are all independently distributed terms, except for the symmetry constraints already mentioned, which we shall ignore as they simply introduce a scaling factor of $\frac{1}{2}$ into the final log-likelihood. The joint probability density over all frequencies ω, including the Jacobian of the transformation to $(A^{(Y_f)}, \theta^{(Y_f)})$, is

$$p(A^{(Y_f)}, \theta^{(Y_f)}) \propto \prod_\omega A_\omega^{(Y_f)} \exp\left[-\frac{1}{2\sigma^2}\{A_\omega^{(Y_f)^2} + A_\omega^{(\mu)^2} - 2A_\omega^{(\mu)} A_\omega^{(Y_f)} \cos(\theta_\omega^{(Y_f)} - \theta_\omega^{(\mu)})\} \right].$$

We see that the distribution for $\theta^{(Y_f)}$, conditional on $A^{(Y_f)}$, is independent von Mises with concentration $\kappa = A^{(\mu)}A^{(Y_f)}/\sigma^2$.

Fourthly, rather than restricting κ to this form, we generalize to the log-linear model given in equation (7). Fisher and Lee (1992) also modelled the concentration of circular data by using log-linear models. This choice ensures positivity and is isotropic, and some experimentation indicated that a quadratic function in $|\omega|$ is sufficiently flexible to model the observed patterns of κ in our examples. Terms in $A^{(\mu)}$ and $A^{(Y_f)}$ are included, both because we would expect phases to be less susceptible to sampling variability when amplitudes are large and because it leads to some standard models as special cases, as already discussed. In applications where there is near collinearity in the explanatory variables there will be some redundancy in this model and a lack of identifiability in ξ, but this should not affect the estimation of f. Other forms of κ have been considered, particularly in the one-dimensional case of signal processing. Hamon and Hannan (1974) showed that the optimal choice is $\kappa_\omega = c_\omega^2/(1 - c_\omega^2)$, where c^2 is the coherence between two series, which they estimated nonparametrically. See also Hannan and Thomson (1988) and, on the subject to subpixel alignment, Berman *et al.* (1994), to which we shall return in Section 3.2.

2.2. Null set distortion criteria

We formulate distortion criteria D that are uniquely minimized by particular null sets of functions $\mathcal{C}$. To motivate this approach, consider problem 3 introduced in Section 1, the discrimination of fish species. Fish are not rigid bodies, so shape comparisons should allow for small distortions. Fig. 4 shows an abstracted version of this problem, with the object in Fig. 4(b) having the same shape as the triangle in Fig. 4(a) except for a small non-linear deformation. Figs 4(c) and 4(d) show a second, differently shaped triangle and a non-linearly

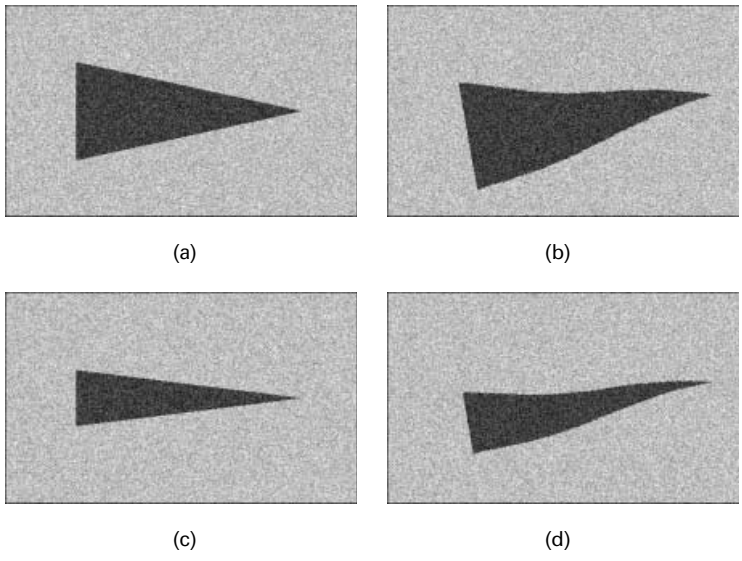


Fig. 4. Test images: (a) and (b) are triangles of the same shape, except for a smooth deformation of (a) into (b), and similarly for (c) and (d)

deformed version. (Gaussian white noise has been added to all four images, for reasons that will become apparent in Section 3.3.) We need a distortion criterion that penalizes warpings that align Figs 4(a) and 4(c) more than those that align Figs 4(a) and 4(b). As far as we are aware, no existing distortion criterion is tailored to this problem. For example, the thin plate spline distortion criterion (see D_{B_2} in equation (16) below) does not penalize affine transformations such as that which would warp Fig. 4(a) to align exactly with Fig. 4(c). To penalize such an affine transformation, we construct $D(f, \mathcal{C})$, taking for $\mathcal{C}$ the set of Euclidean similarity transformations (see equation (20) below). By making λ arbitrarily large in equation (1), the function that maximizes P will be a similarity transformation and, as λ is reduced, warpings are obtained which are nonparametric departures of increasing magnitude from this transformation.

Let $D_B(f)$ be a functional, called the *base distortion criterion*, such that

$$\begin{aligned} D_B(f) &\geqslant 0, \\ D_B(0) &= 0. \end{aligned} \tag{13}$$

We define the *null set distortion criterion* to be

$$D(f, \mathcal{C}) = \min_{g \in \mathcal{C}} \{D_B(f - g)\}. \tag{14}$$

Therefore, for $f \in \mathcal{C}$, $D(f, \mathcal{C}) = 0$, and by an appropriate choice of D_B we seek to ensure that

$$D(f, \mathcal{C}) > 0 \qquad \text{for } f \notin \mathcal{C}.$$

If, as will usually be the case, $0 \in \mathcal{C}$, then $D(f, \mathcal{C}) \leqslant D_B(f)$, and a necessary condition for D_B is that $\{f\colon D_B(f) = 0\} \subseteq \mathcal{C}$. Silverman (1982), pages 116–117, was the first to look for criteria that were 0 if and only if f was in a specified set of functions: for imaging applications, see Arad *et al.* (1994) and Hallinan *et al.* (1999), chapter 4. Our idea is qualitatively similar, but different in that we construct the functional which annihilates a specific set of functions.

This enables us to construct many criteria, each appropriate to a specific set of imaging applications.

For the null set $\mathcal{C}$, we consider subsets of pth-order polynomial transformations:

$$g_i = \alpha_i + \sum_{j_1=1}^{2} \alpha_{ij_1} x_{j_1} + \sum_{j_1 j_2} \alpha_{ij_1 j_2} x_{j_1} x_{j_2} + \ldots + \sum_{j_1,\ldots,j_p} \alpha_{ij_1\ldots j_p} x_{j_1} \ldots x_{j_p} \qquad \text{for } i = 1, 2. \quad (15)$$

Low order polynomials occur repeatedly in image warping applications, including third and higher orders in registration of remotely sensed images (see Glasbey and Mardia (1998) for specific references). For the base distortion criterion D_B, we use functionals of partial derivatives, such as the following first and second partial derivatives:

$$D_{B_1}(f) = \sum_{i=1}^{2} \sum_{j=1}^{2} \int_{\square} \left(\frac{\partial f_i}{\partial x_j} \right)^2 \mathrm{d}x,$$
$$D_{B_2}(f) = \sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=1}^{2} \int_{\square} \left(\frac{\partial^2 f_i}{\partial x_j \, \partial x_k} \right)^2 \mathrm{d}x, \quad (16)$$

integrated over domain $\square = (0, n_1) \times (0, n_2)$, where $\mathrm{d}x$ denotes $\mathrm{d}x_1 \, \mathrm{d}x_2$. (In this section it is convenient to treat image domains as continuous.) Both these functionals satisfy condition (13) and have been proposed many times: D_{B_1} is referred to as the Gaussian prior (Hallinan *et al.* (1999), page 93) and D_{B_2} is the bending energy of a pair of thin plate splines in a finite window (see Green and Silverman (1994), pages 150–155). If the domains of integration were $\Re^2$, both functionals would be translationally and rotationally invariant, particular cases of the functionals considered by Wahba (1990). By specifying distortion using first partial derivatives, warpings are produced which are similar to the deformations of elastic membranes and can have discontinuous second derivatives. For a detailed treatment of such penalties from a general viewpoint see Blake and Zisserman (1987). As with snakes, which are linear templates that deform smoothly to align with features in images (Kass *et al.*, 1988), first-order derivatives can be regarded as tension constraints and second-order derivatives as rigidity constraints.

By combining equations (14)–(16), many null set distortion criteria are produced, most of which are new, and add to the range of first- and second-derivative functionals used by others. For example, if we choose for $\mathcal{C}$ the set of bilinear transformations

$$\mathcal{B} = \{g: g_i = \alpha_i + \alpha_{i1} x_1 + \alpha_{i2} x_2 + \alpha_{i12} x_1 x_2, \quad i = 1, 2\}$$

and use D_{B_2}, then

$$D(f, \mathcal{B}) = \min_{\alpha_{112}, \alpha_{212}} \left\{ \sum_{i,j,k} \int_{\square} \left(\frac{\partial^2 f_i}{\partial x_j \, \partial x_k} - \alpha_{ijk} \right)^2 \mathrm{d}x \right\},$$

where $\alpha_{i11} = \alpha_{i22} = 0$. It can be seen that $D(f, \mathcal{B}) = 0$ if and only if $f \in \mathcal{B}$. The minimizing values of α_{i12} are

$$\tilde{\alpha}_{i12} = \frac{1}{n_1 n_2} \int_{\square} \frac{\partial^2 f_i}{\partial x_1 \, \partial x_2} \, \mathrm{d}x \qquad i = 1, 2,$$

producing

$$D(f, \mathcal{B}) = D_{B_2}(f) - 2 n_1 n_2 (\tilde{\alpha}_{112}^2 + \tilde{\alpha}_{212}^2). \quad (17)$$

If instead we were to choose the set of affine transformations

$$\mathcal{A} = \{g\colon g_i = \alpha_i + \alpha_{i1}x_1 + \alpha_{i2}x_2, \quad i = 1, 2\}, \tag{18}$$

then, as is well known, $\{f\colon D_{B_2}(f) = 0\} = \mathcal{A}$, and therefore

$$D(f, \mathcal{A}) = \min_{g \in A}\{D_{B_2}(f - g)\} = D_{B_2}(f). \tag{19}$$

If, for the null set $\mathcal{C}$, we wish to consider a subset of $\mathcal{A}$, such as a translation, a translation in combination with either a rotation or scaling, or the Euclidean similarity transformation

$$\mathcal{S} = \{g\colon g_1 = \alpha_1 + \alpha_{11}x_1 + \alpha_{12}x_2, \quad g_2 = \alpha_2 - \alpha_{12}x_1 + \alpha_{11}x_2\}, \tag{20}$$

then we cannot use D_{B_2} alone as the base distortion criterion, because $\{f\colon D_{B_2}(f) = 0\} \nsubseteq \mathcal{S}$. Instead we use D_{B_1}. Set $\mathcal{S}$ is important since shapes are defined to be invariant under these transformations (see, for example, Dryden and Mardia (1998)). We have

$$D(f, \mathcal{S}) = \min_{\alpha_{11}, \alpha_{12}} \left\{ \sum_{i,j} \int_{\square} \left(\frac{\partial f_i}{\partial x_j} - \alpha_{ij} \right)^2 \mathrm{d}x \right\},$$

where $\alpha_{21} = -\alpha_{12}$ and $\alpha_{22} = \alpha_{11}$. Again, $D(f, \mathcal{S}) = 0$ if and only if $f \in \mathcal{S}$, so we have an appropriate distortion criterion. The minimizing values of α_{11} and α_{12} are

$$\tilde{\alpha}_{11} = \frac{1}{2n_1n_2} \int_{\square} \left(\frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2} \right) \mathrm{d}x,$$

$$\tilde{\alpha}_{12} = \frac{1}{2n_1n_2} \int_{\square} \left(\frac{\partial f_1}{\partial x_2} - \frac{\partial f_2}{\partial x_1} \right) \mathrm{d}x,$$

producing

$$D(f, \mathcal{S}) = D_{B_1}(f) - 2n_1n_2(\tilde{\alpha}_{11}^2 + \tilde{\alpha}_{12}^2). \tag{21}$$

To illustrate, if $g \in \mathcal{A}$, given by equation (18), then

$$D_{B_1}(g) = n_1n_2(\alpha_{11}^2 + \alpha_{12}^2 + \alpha_{21}^2 + \alpha_{22}^2),$$

$$\tilde{\alpha}_{11} = \tfrac{1}{2}(\alpha_{11} + \alpha_{22}),$$

$$\tilde{\alpha}_{12} = \tfrac{1}{2}(\alpha_{12} - \alpha_{21})$$

and

$$D(g, \mathcal{S}) = \frac{n_1n_2}{2}\{(\alpha_{11} - \alpha_{22})^2 + (\alpha_{12} + \alpha_{21})^2\},$$

which is 0 if and only if $\alpha_{11} = \alpha_{22}$ and $\alpha_{12} = -\alpha_{21}$, the constraints for $g \in \mathcal{S}$, given by equation (20).

For other subsets of $\mathcal{A}$, we can similarly derive $D(f, \mathcal{C})$ based on equation (14) using D_{B_1}. We could also add a term involving D_{B_2} or higher order derivatives to $D(f, \mathcal{C})$, to constrain f to have a continuous first derivative, while still retaining the property that $D(f, \mathcal{C})$ is 0 if and only if $f \in \mathcal{C}$. Therefore, it is important to note that, although our null set distortion criterion $D(f, \mathcal{C})$ is uniquely minimized by $f \in \mathcal{C}$, $D(f, \mathcal{C})$ is not itself unique: there are many alternative distortion criteria with the same property.

2.3. Optimization algorithm

We first consider maximizing P, given by equation (1), and then the multi-image problem of maximizing $P^{(K)}$, given by equation (2). The maximization of P, with respect to f and parameters ξ, has, in general, no known analytic solution for the functionals which we have considered in Section 2.1 and 2.2. Therefore, we must resort to numerical methods.

Numerically, we approximate f by specifying its values at a $(q_1 + 1) \times (q_2 + 1)$ lattice of points:

$$f\left(\frac{k_1 n_1}{q_1}, \frac{k_2 n_2}{q_2}\right) = \beta_k \qquad k_1 = 0, \ldots, q_1, \quad k_2 = 0, \ldots, q_2, \tag{22}$$

involving an array of parameters, β, and interpolate $f(x)$ elsewhere by using the piecewise bilinear transformation

$$f(x) = \beta_k + \beta_k^{+0}\left(\frac{x_1 q_1}{n_1} - k_1\right) + \beta_k^{0+}\left(\frac{x_2 q_2}{n_2} - k_2\right) + \beta_k^{++}\left(\frac{x_1 q_1}{n_1} - k_1\right)\left(\frac{x_2 q_2}{n_2} - k_2\right). \tag{23}$$

Here

$$k_i = \text{int}\left[\frac{x_i q_i}{n_i}\right] \qquad i = 1, 2,$$

with int$[z]$ used to denote the integer part of z, and

$$\begin{aligned}
\beta_k^{+0} &= \beta_{k+(1,0)} - \beta_k, \\
\beta_k^{0+} &= \beta_{k+(0,1)} - \beta_k, \\
\beta_k^{++} &= \beta_{k+(1,1)} - \beta_{k+(1,0)} - \beta_{k+(0,1)} + \beta_k.
\end{aligned}$$

Alternatively we could have interpolated using B-splines (Rueckert *et al.*, 1999). Fig. 5 illustrates the case when $q_1 = q_2 = 3$.

For a piecewise bilinear transformation, it is straightforward to evaluate first-derivative terms in our null set distortion criteria. For example,

$$\int_{\square} \frac{\partial f}{\partial x_1}\, \mathrm{d}x = \frac{n_2}{q_2} \sum_{k_1=0}^{q_1-1} \sum_{k_2=0}^{q_2-1} (\beta_k^{+0} + \tfrac{1}{2}\beta_k^{++}),$$

$$\int_{\square} \left(\frac{\partial f}{\partial x_1}\right)^2 \mathrm{d}x = \frac{n_2}{q_2} \sum_{k_1=0}^{q_1-1} \sum_{k_2=0}^{q_2-1} \{(\beta_k^{+0})^2 + \beta_k^{+0}\beta_k^{++} + \tfrac{1}{3}(\beta_k^{++})^2\},$$

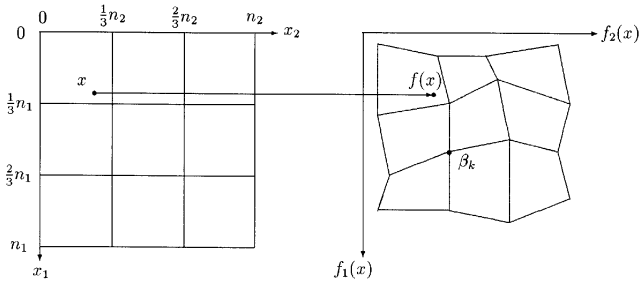


Fig. 5. Illustration of a piecewise bilinear approximation to f for a 3×3 grid

and other terms in $D(f, \mathcal{S})$, given by equation (21), can be similarly computed. If D involves derivatives of higher order, these can only be approximated. For example, $D(f, \mathcal{A})$ given by equation (19), can be approximated as

$$D(f, \mathcal{A}) \approx \frac{q_1^3 n_2}{n_1^3 q_2} \sum_{k=(1,0)}^{q-(1,0)} (\beta_k^{+0} - \beta_{k-(1,0)}^{+0})^2 + 2\,\frac{q_1 q_2}{n_1 n_2} \sum_{k=(0,0)}^{q-(1,1)} (\beta_k^{++})^2 + \frac{n_1 q_2^3}{q_1 n_2^3} \sum_{k=(0,1)}^{q-(0,1)} (\beta_k^{0+} - \beta_{k-(0,1)}^{0+})^2,$$

using an abbreviated summation notation. Incidentally, this particular distortion criterion can be interpreted as a negative log-density of a Gaussian Markov random field on the $(q_1 + 1) \times (q_2 + 1)$ lattice. In all cases, $D(f, \mathcal{C})$ is a positive definite quadratic form in β.

We use a conjugate gradient method to maximize P (see, for example, Press (1994)). This is a general optimization algorithm which requires only first partial derivatives. At each iteration, the search direction is that of steepest ascent, modified by the previous search direction, in such a way that, if the function were an n-dimensional quadratic, it would be optimized in n steps. This algorithm is well suited to our problem, as first partial derivatives can be obtained relatively simply, as

$$\frac{\partial P}{\partial \beta} = \frac{\partial}{\partial \beta}\left(\sum_x \mu_x^{<\xi>} Y_{f(x)} - \sum_\omega \log[I_0\{\kappa_\omega(\xi)\}] - \lambda\, D(f, \mathcal{C})\right) \approx \sum_x \mu_x^{<\xi>} \frac{\partial Y_{f(x)}}{\partial \beta} - \lambda \frac{\partial D(f, \mathcal{C})}{\partial \beta}, \tag{24}$$

using the formulation of the Fourier–von Mises log-likelihood L, given by equation (10), and

$$\begin{aligned} \frac{\partial P}{\partial \xi} &= \frac{\partial}{\partial \xi}\left(\sum_\omega \kappa_\omega(\xi)\, \cos(\theta_\omega^{(Y_f)} - \theta_\omega^{(\mu)}) - \sum_\omega \log[I_0\{\kappa_\omega(\xi)\}]\right) \\ &= \sum_\omega \frac{\partial \kappa_\omega(\xi)}{\partial \xi} \cos(\theta_\omega^{(Y_f)} - \theta_\omega^{(\mu)}) - \sum_\omega \frac{\partial(\log[I_0\{\kappa_\omega(\xi)\}])}{\partial \xi}, \end{aligned} \tag{25}$$

using the formulation of L given by equation (8). Derivatives are computed by using difference methods, taking advantage of changes in β_k only affecting a subset of terms in Y_f and D, and we achieve substantial gains in speed by ignoring the second-order dependence of $\mu^{<\xi>}$ and $\kappa(\xi)$ on β. Various strategies can be adopted to guard against becoming trapped in local suboptima. These include a multiresolution approach, where q is increased as iterations proceed, and permitting greater distortion by decreasing λ as iterations proceed. For the examples in Section 3, the algorithm typically took 30 min of central processor unit time on a single processor of a SUN Enterprise 450 computer using Fortran 77. However, parallelization would considerably reduce this time.

In applications where it is important, bijectivity can be ensured. Necessary and sufficient conditions for the piecewise bilinear transformation to be bijective are that the transformed boundary does not self-intersect, and that each quadrilateral, specified by the ordered set of four vertices β_k, $\beta_{k+(1,0)}$, $\beta_{k+(1,1)}$ and $\beta_{k+(0,1)}$, is convex, with the vertices ordered anticlockwise. Convexity ensures that the bilinear interpolant is bijective within quadrilaterals (a result that is best seen geometrically by plotting the deformations of lines parallel to the axes), and the anticlockwise constraint prevents the mapping from folding at the divisions between quadrilaterals. Computationally, convexity can be ensured by checking that the two diagonals intersect inside the quadrilateral.

To maximize $P^{(K)}$, given by equation (2), we alternate between estimating the consensus image μ and aligning each individual image with it, in a way analogous to generalized Procrustes analysis. Fluete and Lavallée (1998) used a similar method to align shape outlines

and Ramsay and Li (1998) to align curves. If we use the log-likelihood L^* given by equation (4), then

$$P^{(K)} = -\sum_{k=1}^{K} \left\{ \sum_x (Y^{(k)}_{f^{(k)}(x)} - \mu_x)^2 + \lambda D(f^{(k)}, \mathcal{C}) \right\} \tag{26}$$

is maximized with respect to μ simply by averaging at each pixel:

$$\hat{\mu}_x = \frac{1}{K} \sum_k Y^{(k)}_{f^{(k)}(x)}. \tag{27}$$

Alternatively, if we use the Fourier–von Mises log-likelihood L given by equation (8), then we can only estimate the phases of the Fourier transform of μ, denoted ζ, by

$$\hat{\zeta}_\omega = \tan^{-1} \left\{ \sum_k \kappa_\omega(\xi^{(k)}) \sin\left(\theta_\omega^{(Y_f^{(k)})}\right) \Big/ \sum_k \kappa_\omega(\xi^{(k)}) \cos\left(\theta_\omega^{(Y_f^{(k)})}\right) \right\}. \tag{28}$$

In either case, given $\hat{\mu}$, each warping $f^{(k)}$ and $\xi^{(k)}$ can be estimated by maximizing the component functional $P^{(k)}$, using the conjugate gradients algorithm already discussed. Because $P^{(K)}$ increases at each iteration and is bounded above, the algorithm is guaranteed to converge. However, $P^{(K)}$ need not have a unique maximum, and the solution can depend on the initial choice of an average image. In general, it will not be adequate to start by averaging all the unwarped images. We consider one solution for problem 3, in Section 3.3.

3. Applications

We now apply the methodology developed in Section 2 to the three practical problems introduced in Section 1.

3.1. Problem 1: synthetic aperture radar registration

For problem 1, we wish to align the SAR image, Fig. 1(a), with the digital map, Fig. 1(b), which it is natural to take as μ. The appropriate transformation is a projection,

$$\begin{aligned} f_1 = \alpha_1 + \gamma[&x_1\{-\cos(\phi_1)\sin(\phi_2)\sin(\phi_3) + \cos(\phi_2)\cos(\phi_3)\} \\ &+ x_2(-\cos(\phi_1)\cos(\phi_2)\sin(\phi_3) - \sin(\phi_2)\cos(\phi_3)\} + h(x)\sin(\phi_1)\sin(\phi_3)], \\ f_2 = \alpha_2 + \gamma[&x_1\{\cos(\phi_1)\sin(\phi_2)\cos(\phi_3) + \cos(\phi_2)\sin(\phi_3)\} \\ &+ x_2\{\cos(\phi_1)\cos(\phi_2)\cos(\phi_3) - \sin(\phi_2)\sin(\phi_3)\} - h(x)\sin(\phi_1)\cos(\phi_3)], \end{aligned} \tag{29}$$

as shown in Fig. 6. In addition to the elevation function h: $\Re^2 \mapsto \Re$, there are six unknown parameters, the translation and scale parameters (α, γ) and the Euler angles ϕ. The first Euler angle, ϕ_1, is shown in Fig. 6, and the other two relate to the orientations of the two sets of axes. If the ground were planar the transformation would be affine, as given by equation (18), but with a different parameterization. We penalize non-linear functions h, using the thin plate spline distortion criterion $D(h, \mathcal{A})$, a one-dimensional version of that in equation (19).

We use the Fourier–von Mises image model (6). However, for Y, we use an edge-filtered version of the SAR image, as shown in Fig. 7(a). Details are given in Glasbey (1997). It is not possible to subsume this filter in the Fourier–von Mises image model, because linear filters are incapable of transforming Fig. 1(a) to an image that looks like Fig. 1(b). However, our

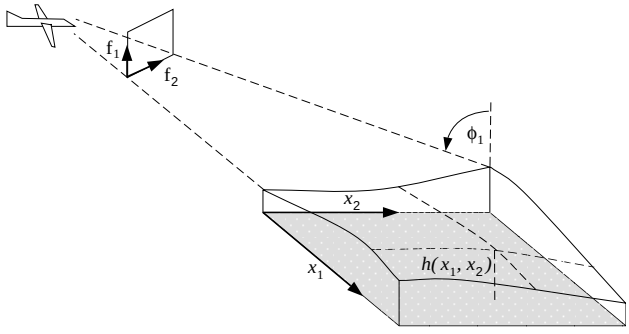


Fig. 6. Illustration of the projective transform from the map to the SAR image in Fig. 1

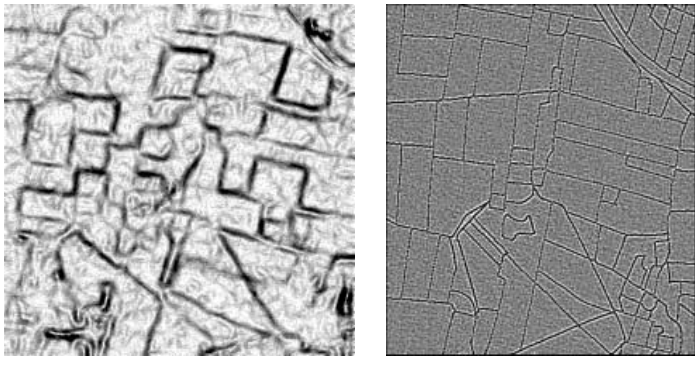


Fig. 7. Filtered SAR image and map, with larger values displayed as darker shades of grey: (a) edge-filtered SAR image; (b) filtered map, given by $\mu^{<\xi>}$

model does allow the fine-tuning, by linear filters, of this empirically chosen edge filter, to optimize its performance for the warping task.

The algorithm of Section 2.3 was used to maximize P with respect to f, given by equation (29), and parameters ξ. For simplicity, in this application we set $\xi_3 = \xi_4$. Experimentation with different values of q showed $q_1 = q_2 = 16$ to be sufficiently large to approximate f. The warping transformation was constrained to be bijective, although this need not be the case in this type of application: the presence of hills could lead to occlusions which would need to be taken into account also in the image model. A range of values of λ was used, as is common practice (see, for example, Silverman (1986)). Table 1 summarizes the results. It can be seen that, as λ decreases, P, L and $\lambda D(h, \mathcal{A})$ all increase, except for some tailing off in $\lambda D(h, \mathcal{A})$ for the smallest values of λ. We used a cross-validatory approach to choose λ, by estimating f with a 50×50 block of pixels in the 250×250 array Y set to a constant mean value, then evaluating the covariance between the complete images. This was repeated for each of the nine blocks with pixel locations x_1 and x_2 in the range 51–100, 101–150 or 151–200. We chose this size of block, rather than individual pixels, because adjacent pixels are likely to be correlated, and to reduce the computational effort. The final column of Table 1 gives the results, from which it can be seen from the value in bold that $\lambda = 100$ appears to be best.

Table 1. Effect of varying λ on criteria for aligning the SAR image with the map

λ	P	L	$\lambda D(h, \mathcal{A})$	*Cross-validated covariance*
100000	285	286	1	34.5
30000	290	293	4	34.9
10000	302	312	10	35.6
3000	321	335	14	36.2
1000	343	363	20	37.4
300	367	406	38	37.2
100	415	513	97	**40.4**
30	513	683	170	34.0
10	599	817	216	31.6
3	869	1084	215	21.9
1	1086	1244	158	31.1

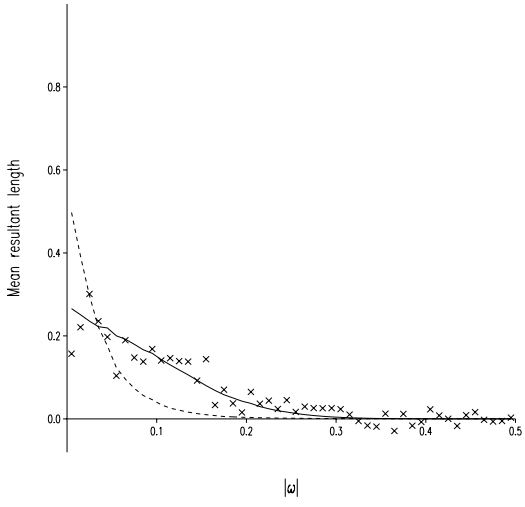


Fig. 8. 'Mean resultant length' for the SAR edge-filtered image and map, averaged over all orientations, plotted against $|\omega|$: $\times$, sample values, obtained by using equation (30); ———, expected values from the full model, obtained by using equation (31); - - - - -, expected values when $\xi = (\xi_0, 0, 0, 1, 1)$

For $\lambda = 100$, we obtained $\hat{\xi} = (-2.9, 15, -61, 0.45, 0.45)$. The sample 'mean resultant length' was calculated for a range of values of the non-directional frequency $|\omega|$,

$$\frac{1}{N(\Lambda_{|\omega|})} \sum_{\nu \in \Lambda_{|\omega|}} \cos(\theta_\nu^{(Y_f)} - \theta_\nu^{(\mu)}), \qquad \text{where } \Lambda_{|\omega|} = \{\nu: \, ||\nu| - |\omega|| < 0.05\} \tag{30}$$

and $N(\Lambda_{|\omega|})$ denotes the number of elements in set $\Lambda_{|\omega|}$. According to the von Mises model, the mean resultant length has expectation

$$\frac{1}{N(\Lambda_{|\omega|})} \sum_{\nu \in \Lambda_{|\omega|}} \frac{I_1(\kappa_\nu)}{I_0(\kappa_\nu)}, \tag{31}$$

where I_0 and I_1 are Bessel functions. Fig. 8 shows these sample and expected values plotted

Fig. 9. Superposition of the aligned SAR image with the map

against $|\omega|$, from which we see that the agreement is excellent. For comparison, the expected values for the best-fitting model of the form $\xi = (\xi_0, 0, 0, 1, 1)$, with $\hat{\xi}_0 = -4.6$, is also shown. The value of $\hat{\xi}_3 = \hat{\xi}_4 = 0.45$ in the full model suggests that the best choice of model leads to a measure of similarity which is a half-way compromise between covariance and phase correlation, i.e. a *band pass filter*. In contrast, Koch and Snowdon (1994) advocated the use of a low pass filter in an application involving the alignment of X-ray images. The filtered map, denoted $\mu^{<\xi>}$, as defined in equation (11), is displayed in Fig. 7(b).

Fig. 9 shows the SAR image registered with the digital map, obtained by applying the estimated warping to the original SAR image. The alignment can be seen to be very good and automatically yields an almost complete segmentation of the image into homogeneous regions.

3.2. Problem 2: multimodal microscopy

For problem 2, we know *a priori* that a translation

$$f = \alpha + x \tag{32}$$

is sufficient to align any pair of microscope images. Therefore, formally, we choose the null set distortion criterion D to be uniquely minimized by translations, using the method of Section 2.2, but we also take $\lambda \to \infty$. In practice, we simply use the parametric transformation.

By combining equations (12) and (8), the Fourier–von Mises log-likelihood can be re-expressed as

$$L(Y|\mu, f, \xi) = \sum_{\omega} \kappa_\omega(\xi) \cos(\theta_\omega^{(Y)} - \theta_\omega^{(\mu)} + 2\pi\omega^{\mathrm{T}}\alpha) - \sum_{\omega} \log[I_0\{\kappa_\omega(\xi)\}], \tag{33}$$

provided that we allow modulo n wraparound in the translation. L can be evaluated simultaneously for all integer values of α by a single fast Fourier transform. For problem 2, an alignment to the nearest pixel is sufficiently accurate and is probably all that is achievable. The model is a two-dimensional variant of the 'barber's pole' proposed by Gould (1969). In some applications it is possible to estimate α to subpixel accuracy, especially if adjustments to take account of aliasing, proposed by Berman *et al.* (1994), are also included.

We wish to align all three microscopy images under translation simultaneously. However, it turns out that we cannot maximize $P^{(K)}$ with respect to parameters in both the concentration function and the consensus image. This is similar to the Neyman–Scott problem (see, for example, Stuart *et al.* (1999), pages 80–81). So, instead we propose to maximize a pseudo-log-likelihood:

$$P^* = \sum_{k<l} L(Y^{(l)} | Y^{(k)}, f^{(k,l)}, \xi^{(k,l)}), \tag{34}$$

with respect to $\alpha^{(k,l)}$, which specifies $f^{(k,l)}$ as given in equation (32), and $\xi^{(k,l)}$, subject to constraints

$$f^{(k,m)} = f^{(k,l)} \circ f^{(l,m)} \qquad \forall k < l < m, \tag{35}$$

where $\circ$ denotes a composite of functions. This construction eliminates the consensus image μ. In general, these constraints are difficult to enforce, but for parametric transformations they take simple forms. In particular, for translations

$$\alpha_i^{(1,3)} = \alpha_i^{(1,2)} + \alpha_i^{(2,3)} \mod(n_i), \qquad i = 1, 2. \tag{36}$$

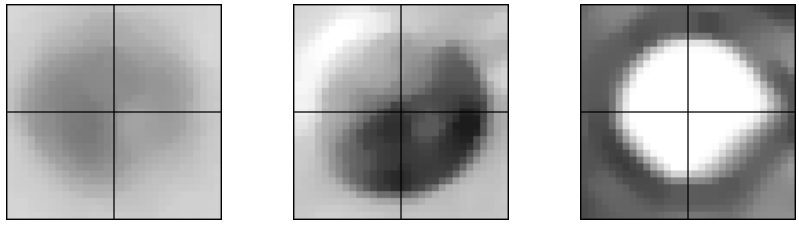
Also, when f is a translation, Y and μ are interchangeable in L, given by equation (33), so we need only to consider all unordered pairs in P^*.

We use the conjugate gradient method described in Section 2.3, to maximize P^* with respect to $\xi^{(1,2)}$, $\xi^{(1,3)}$ and $\xi^{(2,3)}$, but for each value of ξ we conduct a grid search to estimate α. An exhaustive search would have to consider $n_1^2 n_2^2$ possibilities. Therefore, we approximate by a local optimum, by only searching values around $\arg\max_{\alpha^{(1,2)}}(L^{(1,2)})$ and $\arg\max_{\alpha^{(2,3)}}(L^{(2,3)})$. Similarly, we consider values around each of the other two pairs of maxima. Table 2 gives the results, which agree with those reported in Glasbey and Martin (1996), using an *ad hoc* similarity criterion. So, for example, we estimate that Fig. 2(b) needs to be shifted down by three rows and shifted right by six columns to align with Fig. 2(a). Fig. 10 shows a single algal cell, after alignment, in the three microscope modalities. A cross-wire has been superimposed to aid the comparison in alignments, which can be seen to be very good. The individual pixels can be discerned at this magnification, and it can be appreciated that even a shift as small as three rows and six columns has a marked effect.

We compare our method of alignment with two alternatives: the covariance and phase correlation criteria. Each of 70 subimages of Fig. 2(a) (128 × 192 pixels in size) was aligned

Table 2. Parameter estimates to align the microscope images

k	l	$\hat{\alpha}_1^{(k,l)}$	$\hat{\alpha}_2^{(k,l)}$	$\hat{\xi}_0^{(k,l)}$	$\hat{\xi}_1^{(k,l)}$	$\hat{\xi}_2^{(k,l)}$	$\hat{\xi}_3^{(k,l)}$	$\hat{\xi}_4^{(k,l)}$
1	2	3	6	−6	43	−118	0.91	0.28
1	3	28	170	−20	616	−14000	0.72	1.53
2	3	25	164	−17	502	−11900	0.65	1.33



(a) (b) (c)

Fig. 10. Single algal cell, after alignment of three microscope images, and with a cross-wire superimposed: (a) brightfield; (b) differential interference contrast; (c) phase contrast

Table 3. Summary of results for estimating translation parameters to align 70 different 128×192 subimages of Fig. 2(a) with subimages of Fig. 2(b) shifted by 20 rows and 20 columns, using three similarity criteria

Similarity criterion	*Mean*		*Standard deviation*	
	α_1	α_2	α_1	α_2
Covariance	17.5	22.7	8.0	4.3
Phase correlation	21.2	23.6	11.4	23.4
Fourier–von Mises log-likelihood L	22.8	25.8	0.7	0.7

with a subimage of Fig. 2(b) *shifted* by 20 rows and 20 columns. For the new criterion and covariance and phase correlation criteria, the means and standard deviations of the 70 estimates of α were evaluated, as given in Table 3. We see that L produces by far the most consistent results, with standard deviations of less than one pixel. Also, although the subimages contain far less information than the full images, the estimated translation agrees well with the earlier results, which with the additional translation of (20, 20) should now be (23, 26).

3.3. Problem 3: fish species discrimination

For problem 3, it is natural to use the null set distortion criterion based on the Euclidean similarity transformation $D(f, \mathcal{S})$, given by equation (21), since this is then shape invariant. To assess our procedure, we first apply the method to the synthetic example of triangles in Fig. 4.

We propose to use the Gaussian image model with log-likelihood L^* given by equation (4). The algorithm of Section 2.3 was used to align each image in Fig. 4 with every other image, 12 ordered pairs in total, for each of a range of values of λ. Experimentation with different values of q showed $q_1 = q_2 = 64$ to be sufficiently large to approximate f and, again, the bijective constraint was used. The results are summarized in Table 4, by the average values of the criterion P for within- and between-shape comparisons. Within-shape comparisons are defined to be those between Figs 4(a) and 4(b), and between Figs 4(c) and 4(d), of which there are four. The remaining eight ordered pairs are regarded as between-shape comparisons. In both cases, P increases with λ, because the warping is progressively less constrained to be smooth and can therefore achieve a greater agreement in pixel values between images. The criterion is smaller when two images of different shape are aligned than with two of the same shape, for all values of λ. Standard deviations of values of P are also

Table 4. Effect of varying λ on discrimination between the two shapes of triangle in Fig. 4

λ	*Values of* $P = L^* - \lambda D(f, \mathcal{S})$		
	Mean†		*Studentized difference in means*
	Within shape	*Between shape*	
1	−2090 (15.6)	−2662 (159.0)	3.6
0.3	−1972 (5.7)	−2397 (64.6)	**6.6**
0.1	−1865 (9.1)	−2142 (55.4)	4.9
0.03	−1743 (8.8)	−1869 (26.4)	4.5
0.01	−1633 (8.9)	−1694 (10.6)	4.4
0.003	−1546 (3.2)	−1567 (7.1)	2.7
0.001	−1498 (3.2)	−1510 (3.9)	2.3

†Standard deviations are given in parentheses.

Table 5. Maximized values of penalized likelihood for pairwise comparisons of images in Fig. 4, using two distortion criteria

Image	*Values of* $P = L^* - 0.3\, D(f, \mathcal{S})$ *for the following images:*				*Values of* $P = L^* - 10^6\, D_{B_2}$ *for the following images:*			
	(a)	*(b)*	*(c)*	*(d)*	*(a)*	*(b)*	*(c)*	*(d)*
(a)	—	**−1975**	−2309	−2350	—	−1906	**−1854**	−1908
(b)	**−1979**	—	−2348	−2357	−1903	—	−1912	−1910
(c)	−2475	−2470	—	**−1967**	**−1853**	−1894	—	−1893
(d)	−2409	−2458	**−1968**	—	−1911	−1911	−1890	—

given in Table 4, from which the Studentized difference between the means of the two groups can be computed, as the difference in the means divided by the square root of the sum of the two variances. However, note that it is inappropriate to perform t-tests as samples are not independently distributed. The distances show that the two shapes are well discriminated, with the best choice shown in bold for $\lambda = 0.3$. Thus we conclude that our method provides satisfactory answers for this simplified problem.

Table 5 compares the results that we obtained using $D(f, \mathcal{S})$ with what we would have obtained if we had instead used the thin plate spline distortion criterion D_{B_2}, having selected an appropriate value of $\lambda = 10^6$. In both cases, the largest values are shown in bold. We see that, using D_{B_2}, Figs 4(a) and 4(c) are assessed as being most similar, which is as we would expect, as an affine transformation is sufficient to transform one triangle to the other and this is not penalized by D_{B_2}. It is clear that this distortion criterion will not enable us to discriminate between the two shapes of triangle. Similar inadequate results will be produced by using any distortion criterion other than $D(f, \mathcal{S})$.

The same algorithm was then used to align all pairs of images of fish in Fig. 3 for each of a range of values of λ. Optimized values of P are given in Table 6, summarized as before. The two species are well discriminated, with marginally the best choice of λ being 0.01. Fig. 11 illustrates the warping for this optimal choice of λ, for the alignment of a haddock with another haddock, and with a whiting. Figs 11(a) and 11(b) show grids of the two estimated warps. The deformations in Fig. 11(a) are less severe than in Fig. 11(b), and the distortion is less when the two haddocks are aligned than when a haddock and a whiting are aligned. Fig. 11(c) shows how haddock 1 (Fig. 3(a)) is warped to look like haddock 2 (Fig. 3(b)), and

Table 6. Effect of varying λ on discrimination between haddock and whiting, using the four images in Fig. 3

λ	*Values of* $P = L^* - \lambda D(f, \mathcal{S})$		
	Mean†		*Studentized difference in means*
	Within species	*Between species*	
1	−643 (152)	−1064 (175)	1.8
0.3	−538 (163)	−952 (114)	2.1
0.1	−383 (138)	−784 (49)	2.7
0.03	−255 (92)	−586 (22)	3.5
0.01	−182 (60)	−433 (15)	**4.1**
0.003	−130 (40)	−298 (15)	3.9
0.001	−99 (29)	−218 (16)	3.6

†Standard deviations are given in parentheses.

Fig. 11(e) shows the pixel-by-pixel difference between the two images after alignment. In comparison, Fig. 11(d) shows how haddock 1 is warped to look like whiting 1 (Fig. 3(c)), and Fig. 11(f) shows the pixel-by-pixel difference between the two images. The sum of squared differences is greater than for the within-species comparison.

We now consider the analysis of the larger data set, consisting of images of 10 haddocks and 10 whitings. For each species, we used eight images to characterize the population average and variation, by maximizing $P^{(K)}$, given by equation (26). Two images of each species, chosen at random, were then available to validate the method.

An *ad hoc* procedure to overcome some of the numerical problems in estimating the template $\hat{\mu}$ for each species is as follows. To obtain an initial estimate of $\hat{\mu}$, we warped image 2 to image 1 and formed a composite image:

$$\hat{\mu}^{(1,2)}_{\{x+f(x)\}/2} = \frac{Y^{(1)}_x + Y^{(2)}_{f(x)}}{2} \qquad \forall x \in X. \tag{37}$$

Here, we have taken the average of the pixel at location x in image 1 and the pixel at location $f(x)$ in image 2, and assigned it to the pixel at location $\{x + f(x)\}/2$ in the composite image. Unassigned pixels in $\hat{\mu}^{(1,2)}$ were given the same value as their nearest neighbour. We similarly formed the average of images 3 and 4 and then averaged image (1, 2) with image (3, 4) to obtain image ((1, 2), (3, 4)), and so on until finally

$$\hat{\mu} = \hat{\mu}^{(((1,2),(3,4)),((5,6),(7,8)))}.$$

We then warped the eight original images to $\hat{\mu}$ to maximize $P^{(K)}$, re-estimated μ using equation (27) and repeated until convergence. This procedure treats all eight images equivalently and could be modified to handle other sizes of training set. For both haddock and whiting, values stabilized within a couple of iterations. Fig. 12 shows the two average fish.

Finally, each of the 20 images in turn was warped to the average haddock, by maximizing P. Fig. 13 shows the maximized values of P plotted against the corresponding values when the images were instead warped to the average whiting. The maximized values of the penalized likelihoods are measures of similarity of individual images from the two species. We can see two clusters of points, and the two species are clearly distinguishable. The 10 haddocks, including the two not used previously, are far more similar to the average haddock than to the average whiting, and a similar pattern occurs with the whiting. However, we see

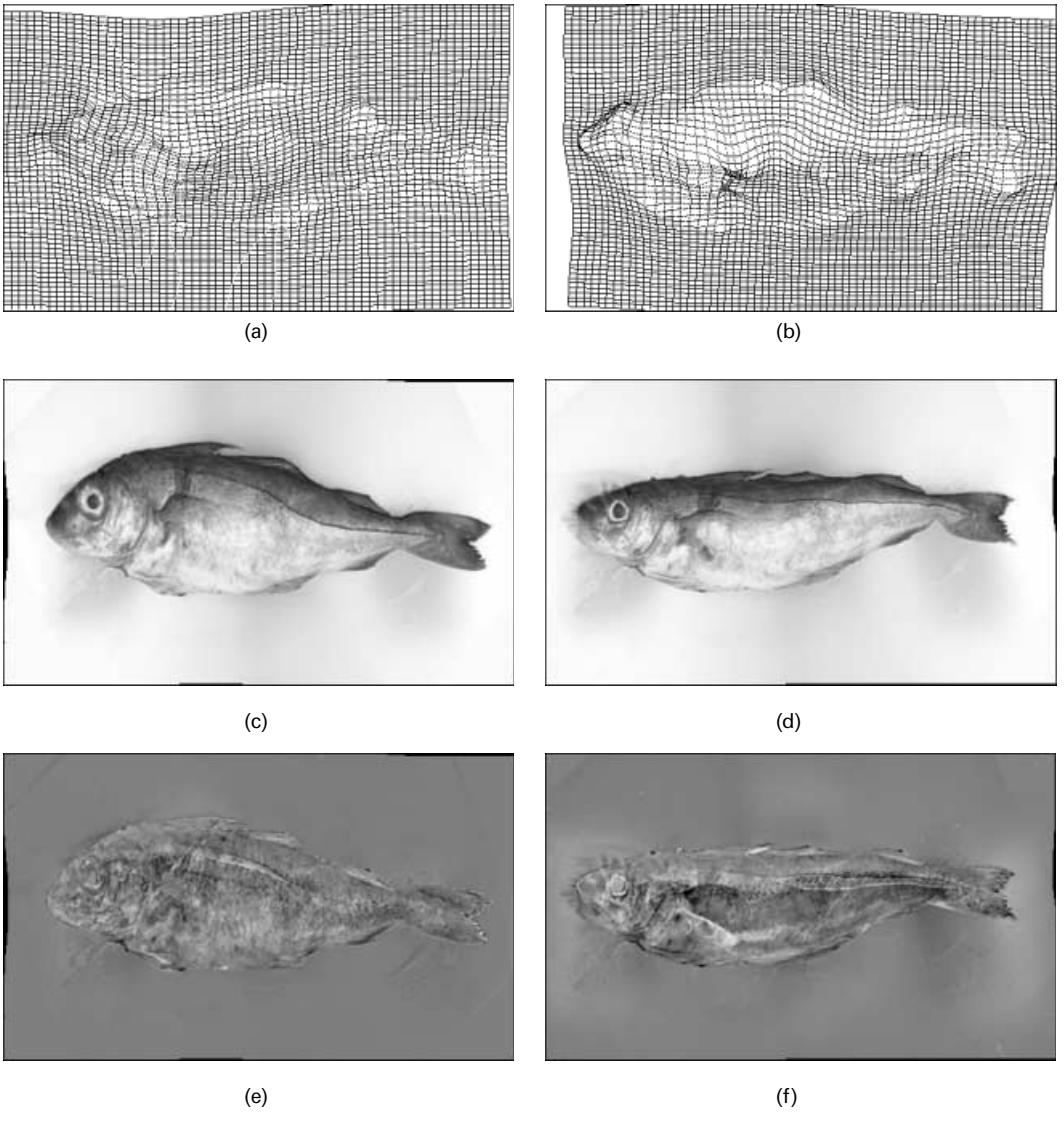
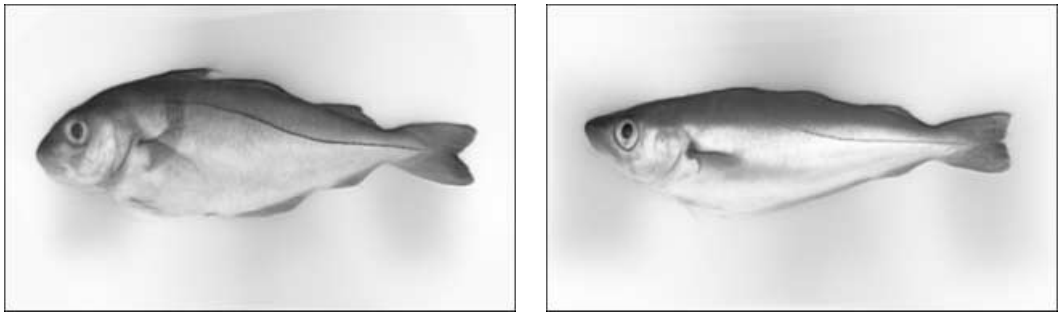


Fig. 11. Result of warping haddock 1, with $\lambda = 0.01$: grid of deformations for alignment with (a) haddock 2 and (b) whiting 1; warped image of haddock 1, to align with (c) haddock 2 and (d) whiting 1; pixel-by-pixel difference (zero values are displayed as mid-grey) between a warped image of haddock 1 and (e) haddock 2 and (f) whiting 1

that the circled points lie on the extremes of the two clusters, nearer to the other cluster, indicating some slight overfitting in the model.

By making full use of the grey level information (including texture) in the photographic images of fish, we have improved on the discriminating power of Strachan *et al*. (1990) and Glasbey *et al*. (1995). We could develop the model further, and apply principal components analysis, both to the grids of warpings β, given by equation (22), and to the differences between pixel values in the aligned images, as Lanitis *et al*. (1995) did with images of faces. These could be used to replace L^* by differences between principal component scores,



(a) (b)

Fig. 12. Average images of two species of fish, obtained by averaging eight images: (a) haddock; (b) whiting

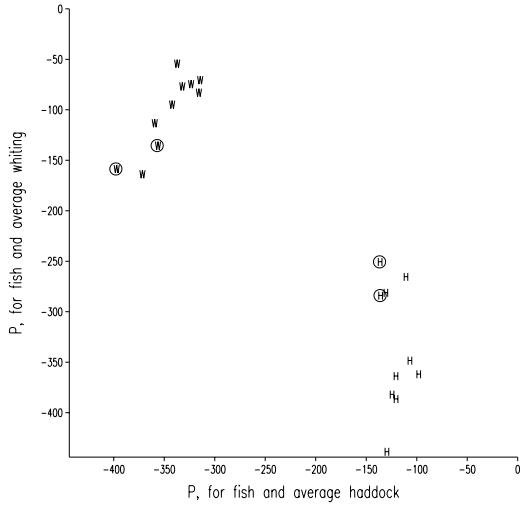


Fig. 13. Plot of *P* between individual images and the average haddock and the average whiting: H, haddock; W, whiting; ○, images not used in obtaining the averages

thereby weighting pixel values according to how variable they are in different parts of an image (Cootes *et al.*, 1998). Further, Moghaddam *et al.* (1996) modelled variation in facial expression, both within and between individuals, and Duta *et al.* (1999) formed clusters of similar shapes and modelled intracluster variation. Rao (2000), pages 580–583, has discussed an unbiased and a consistent estimator of the template under certain conditions when the errors are coloured and the warps are elements of a similarity group.

4. Discussion

The subject of warping has been reviewed comprehensively in Glasbey and Mardia (1998) and elsewhere, as cited in Section 1. Therefore, here we shall focus on the techniques developed in this paper. We have shown that image warping can be formulated statistically, as

maximum penalized likelihood, and this has allowed us to understand and generalize existing methods. Our approach has produced good results for the three applied problems. We do not claim that ours are the only methods that are capable of producing such results, but there would seem to be little opportunity for alternative methods to improve on them. Also, it is clear from Table 3 that the Fourier–von Mises log-likelihood outperforms simpler measures of image similarity, and from Table 5 that the thin plate spline distortion criterion is inappropriate for making similarity shape comparisons, though also see point (b) below. Also, although we have focused on point estimation, as we have formulated image warping in a statistical framework we could also obtain measures of precision of estimators. For example, we could base inferences on multiple samples drawn from the Fourier–von Mises image model.

We now deal with the three main ingredients of the paper: the Fourier–von Mises image model, the null set distortion criterion and the algorithm.

(a) The Fourier–von Mises image model offers a flexible approach to modelling the relationship between images, which will work for general lighting conditions because of its Fourier basis. We showed the model to be plausible and used it in both problem 1 and problem 2. However, the general form is inappropriate to solving problem 3, where we are concerned with discrimination between fish species, and a simple Gaussian model is what is required. Other grey level metrics, such as the Kantorovich distance (Kaijser, 1998) are more computationally expensive. Our image model is elegant in that it combines intensity matching and edge matching in one measure. In the machine vision literature, the two terms for intensity matching and edge matching are treated separately (see, for example, Hallinan *et al.* (1999), page 79). We believe that our procedure has some advantages since it removes the necessity of estimating the weights required to combine the two terms.

(b) The null set distortion criteria furnish us with a rich class. In problem 1 we have taken the roughness penalty from thin plate splines as our distortion criterion, whereas in problem 2 the distortion criterion is used only implicitly to constrain the warp to be in the null set of translations ($\lambda \to \infty$). In problem 3 the distortion criterion is a shape invariant criterion. Thus, our formulation allows us to select a criterion that is appropriate to the application within our general null class. Our shape invariant criterion uses only first derivatives, so there can be degenerate solutions if only a limited number of points such as landmarks are used (e.g. Green and Silverman (1994), page 159). However, in our case there is always a unique solution for finite q since D is quadratic in β. Note that, for landmark-based methods, there is an explicit expression for kriging warps including thin plate splines (see, for example, Mardia and Hainsworth (1993) and Kent and Mardia (1994)). However, we need to add an extra penalty term to the thin plate spline criterion if we want to penalize affine transformations that are not shape preserving.

(c) Our algorithm has some similarities with finite difference methods, though our use of a piecewise bilinear transformation eliminates numerical integration in calculating distortion criteria only involving first derivatives. The conditions that we have imposed on the piecewise bilinear transformation lead to local as well as global bijectivity. Bijectivity is important if we wish to warp a standard co-ordinate system to the image. However, in the SAR example, the projective transformation may not be bijective or continuous owing to occlusion. The method extends in a straightforward manner to three and higher dimensions, although the computational cost will be high. Also,

stochastic methods such as simulated annealing and Markov chain Monte Carlo methods could be implemented, which would also give more information on the posterior distributions of $\hat{\xi}$ etc., but again the method may prove expensive. The issue of global registration has not arisen because, for example, in problem 3, the fish were placed in a prespecified orientation with a fixed camera position. If this were not so, we could resort to any of a number of global registration methods, such as the matching of low order moments (Wong and Hall, 1978; Yang and Cohen, 1999). Alternatively, an additional penalty term could be added to P (see, for example, Mardia *et al.* (1997)). We have taken different q for problems 1–3. Its selection depends on the size of the images, and the overall accuracy required. A wavelet-based distortion criterion in turn is another approach (see, for example, Downie *et al.* (1996)). Also, whether one should use compositional warps at different resolutions or additive warps is another issue.

There remain many challenging problems in image analysis, to which statistical methods are applicable, both in general and in particular in image warping. This paper follows earlier ground breaking papers on image analysis by Besag (1986) and Grenander and Miller (1994). We hope that our paper similarly stimulates further work in this area.

Acknowledgements

Our profound thanks go to John Kent for his generous comments on earlier drafts and, in particular, the suggestion to use equation (14) to formalize the method of construction of distortion criteria. Our thanks are also due to Ulf Grenander, Anil Jain, the Royal Statistical Society's Research Section Committee and the referees for their comments, and to Kevin de Souza for help with preparing the manuscript. We also thank Nick Martin and Norval Strachan for permission to use the microscopy and fish images respectively. The first author's work was supported by funds from the Scottish Executive Rural Affairs Department, and the second author was in receipt of a grant from the Engineering and Physical Sciences Research Council.

References

Amit, Y., Grenander, U. and Piccioni, M. (1991) Structural image restoration through deformable templates. *J. Am. Statist. Ass.*, **86**, 376–387.

Arad, N., Dyn, N., Reisfeld, D. and Yeshurun, Y. (1994) Image warping by radial basis functions: applications to facial expressions. *Graph. Models Image Process.*, **56**, 161–172.

Baddeley, A. and Molchanov, I. (1998) Averaging of random sets based on their distance functions. *J. Math. Imgng Vis.*, **8**, 79–92.

Bajcsy, R. and Kovacic, S. (1989) Multiresolution elastic matching. *Comput. Vis. Graph. Image Process.*, **46**, 1–21.

Barron, J. L., Fleet, D. J. and Beauchemin, S. S. (1994) Performance of optical flow techniques. *Int. J. Comput. Vis.*, **12**, 43–77.

Berman, M., Bischof, L. M., Davies, S. J., Green, A. A. and Craig, M. (1994) Estimating band-to-band mis-registrations in aliased imagery. *Graph. Models Image Process.*, **56**, 479–493.

Besag, J. (1986) On the statistical analysis of dirty pictures (with discussion). *J. R. Statist. Soc.* B, **48**, 259–302.

Blake, A. and Zisserman, A. (1987) *Visual Reconstruction*. Cambridge: Massachusetts Institute of Technology Press.

Bonmassar, G. and Schwartz, E. L. (1997) Space-variant Fourier analysis: the exponential chirp transform. *IEEE Trans. Pattn Anal. Mach. Intell.*, **19**, 1080–1089.

Bookstein, F. L. (1991) *Morphometric Tools for Landmark Data*. Cambridge: Cambridge University Press.

Burr, D. J. (1981) A dynamic model for image registration. *Comput. Graph. Image Process.*, **15**, 102–112.

Cao, J. and Worsley, K. J. (1999) The detection of local shape changes via the geometry of Hotelling's T^2 fields. *Ann. Statist.*, **27**, 925–942.

Carstensen, J. M. (1996) An active lattice model in a Bayesian framework. *Comput. Vis. Image Understndng*, **63**, 380–387.

Caves, R. G., Harley, P. J. and Quegan, S. (1992) Matching map features to synthetic aperture radar (SAR) images using template matching. *IEEE Trans. Geosci. Remote Sensng*, **30**, 680–685.
Christensen, G. E., Rabbitt, R. D. and Miller, M. I. (1996) Deformable templates using large deformation kinetics. *IEEE Trans. Image Process.*, **5**, 1435–1447.
Cootes, T. F., Edwards, G. J. and Taylor, C. J. (1998) Active appearance models. In *Proc. Eur. Conf. Computer Vision* (eds H. Burkhardt and B. Neumann), vol. 2, pp. 484–498. Berlin: Springer.
Cote, S. and Tatnall, A. R. L. (1997) The Hopfield neural network as a tool for feature tracking and recognition from satellite sensor images. *Int. J. Remote Sensng*, **18**, 871–885.
Cross, A. D. J. and Hancock, E. R. (1998) Graph matching with a dual-step EM algorithm. *IEEE Trans. Pattn Anal. Mach. Intell.*, **20**, 1236–1253.
Dobson, M. C., Pierce, L. E. and Ulaby, F. T. (1996) Knowledge-based land-cover classification using ERS-1/JERS-1 SAR composites. *IEEE Trans. Geosci. Remote Sensng*, **34**, 83–99.
Downie, T. R., Shepstone, L. and Silverman, B. W. (1996) A wavelet based approach to deformable templates. In *Proc. Leeds A. Statistics Research Wrkshp* (eds K. V. Mardia, C. A. Gill and I. L. Dryden), pp. 163–169. Leeds: Leeds University Press.
Dryden, I. L. and Mardia, K. V. (1998) *Statistical Shape Analysis*. Chichester: Wiley.
Duta, N., Jain, A. K. and Dubuisson-Jolly, M.-P. (1999) Learning 2D shape models. In *Proc. IEEE Computer Science Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 8–14. Piscataway: Institute of Electrical and Electronics Engineers.
Fisher, N. I. and Lee, A. J. (1992) Regression models for an angular response. *Biometrics*, **48**, 665–677.
Fluete, M. and Lavallée, S. (1998) Building a complete surface model from sparse data using statistical shape models: applications to computer assisted knee surgery. *Med. Image Comput. Comput. Assist. Interven.*, **22**, 878–887.
Galbraith, W. and Farkas, D. L. (1993) Remapping disparate images for coincidence. *J. Microsc.*, **172**, 163–176.
Galton, F. (1878) Composite portraits. *J. Anth. Inst. Gr. Br. Ire.*, **8**, 132–142.
Gee, J. C. (1999) On matching brain volumes. *Pattn Recogn.*, **32**, 99–111.
Glasbey, C. A. (1997) SAR image registration and segmentation using an estimated DEM. *Lect. Notes Comput. Sci.*, **1223**, 507–520.
Glasbey, C. A. and Horgan, G. W. (1995) *Image Analysis for the Biological Sciences*. Chichester: Wiley.
Glasbey, C. A., Horgan, G. W., Gibson, G. J. and Hitchcock, D. (1995) Fish shape analysis using landmarks. *Biometr. J.*, **37**, 481–495.
Glasbey, C. A. and Mardia, K. V. (1998) A review of image warping methods. *J. Appl. Statist.*, **25**, 155–171.
Glasbey, C. A. and Martin, N. J. (1996) Multimodality microscopy by digital image processing. *J. Microsc.*, **181**, 225–237.
Goshtasby, A. A. and Le Moigne, J. (eds) (1999) Image registration: special issue. *Pattn Recogn*, **32**, 1–149.
Gould, A. L. (1969) A regression model for angular variates. *Biometrics*, **25**, 683–700.
Green, P. J. (1999) Penalized likelihood. In *Encyclopedia of Statistical Sciences*, update vol. 3, pp. 578–586. Chichester: Wiley.
Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.
Grenander, U. and Miller, M. I. (1994) Representations of knowledge in complex systems (with discussion). *J. R. Statist. Soc.* B, **46**, 549–603.
———(1998) Computational anatomy: an emerging discipline. *Q. Appl. Math.*, **56**, 617–694.
Hallinan, P. W., Gordon, G. G., Yuille, A. L., Giblin, P. and Mumford, D. (1999) *Two- and Three-dimensional Patterns of the Face*. Natick: Peters.
Hamon, B. V. and Hannan, E. J. (1974) Spectral estimation of time delay for dispersive and non-dispersive systems. *Appl. Statist.*, **23**, 134–142.
Hannan, E. J. and Thomson, P. J. (1988) Time delay estimation. *J. Time Ser. Anal.*, **9**, 21–33.
Hill, A., Taylor, C. J. and Brett, A. D. (2000) A framework for automatic landmark identification using a new method of nonrigid correspondence. *IEEE Trans. Pattn Anal. Mach. Intell.*, **22**, 241–251.
Joshi, S. C. and Miller, M. I. (2000) Landmark matching via large deformation diffeomorphisms. *IEEE Trans. Pattn Anal. Mach. Intell.*, **9**, 1357–1370.
Kaijser, T. (1998) Computing the Kantorovich distance for images. *J. Math. Imgng Vis.*, **9**, 173–191.
Kass, M., Witkin, A. and Terzopoulos, D. (1988) Snakes: active contour models. *Int. J. Comput. Vis.*, **1**, 321–331.
Kent, J. T. and Mardia, K. V. (1994) The link between kriging and thin plate splines In *Probability, Statistics and Optimization: a Tribute to Peter Whittle* (ed. F. P. Kelly), pp. 325–339. New York: Wiley.
Kher, A. and Mitra, S. (1993) Registration of noisy SAR imagery using morphological feature extractor and 2-D cepstrum. *Appl. Dig. Im. Process.*, **15**, 281–291.
Koch, I. and Snowdon, S. (1994) Image registration by smoothed Fourier methods: an application to medical X-ray images. *Technical Report 94-4*. Department of Statistics, University of Newcastle, Callaghan.
Kuglin, C. D. and Hines, D. C. (1975) The phase correlation image alignment method. In *Proc. IEEE 1975 Int. Conf. Cybernetics and Society*, pp. 163–165. Piscataway: Institute of Electrical and Electronics Engineers.
Lades, M., Vorbruggen, J., Buhmann, J., Lange, J., Vandermalsburg, C., Wurtz, R. P. and Konen, W. (1993) Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Comput.*, **42**, 300–311.

Lanitis, A., Taylor, C. J. and Cootes, T. F. (1995) Automatic face identification system using flexible appearance models. *Im. Vis. Comput*., **3**, 393–401.
Li, H., Manjunath, B. S. and Mitra, S. K. (1995) A contour-based approach to multisensor image registration. *IEEE Trans. Im. Process*., **4**, 320–334.
Loncaric, S. (1998) A survey of shape analysis techniques. *Pattn Recogn*, **31**, 983–1001.
Maintz, J. B. A. and Viergever, M. A. (1998) A survey of medical image registration. *Med. Im. Anal*., **2**, 1–36.
Mardia, K. V. (ed.) (1994) *Statistics and Images*, vol. 2. Abingdon: Carfax.
Mardia, K. V. and Hainsworth, T. J. (1993) Image warping and Bayesian reconstruction with grey-level templates. *J. Appl. Statist*., **20**, 257–280.
Mardia, K. V. and Jupp, P. E. (1999) *Directional Statistics*. Chichester: Wiley.
Mardia, K. V., McCulloch, C., Dryden, I. L. and Johnson, V. (1997) Automatic scale-space method of landmark detection. In *Proc. Leeds A. Statistics Research Wrkshp* (eds K. V. Mardia, C. A. Gill and R. G. Aykroyd). Leeds: Leeds University Press.
McInerney, T. and Terzopoulos, D. (1996) Deformable models in medical image analysis: a survey. *Med. Im. Anal*., **1**, 91–108.
Meyer, C. R., Boes, J. L., Kim, B., Bland, P. H., Zasadny, K. R., Kison, P. V., Koral, K., Frey, K. A. and Wahl, R. L. (1996) Demonstration of accuracy and clinical versatility of mutual information for automatic multimodality image fusion using affine and thin-plate spline warped geometric deformations. *Med. Im. Anal*., **1**, 195–206.
Modrusan, Z., Reiser, R., Feldmann, R. A., Fischer, R. L. and Haughn, G. W. (1994) Homeotic transformation of ovules into carpel-like structures in *Arabidopsis*. *Plant Cell*, **6**, 333–349.
Moghaddam, B., Nastar, C. and Pentland, A. (1996) A Bayesian similarity measure for direct image matching. *Int. Conf. Pattern Recognition, Vienna*.
Mokhtarian, F. (1995) Silhouette-based isolated object recognition through curvature scale space. *IEEE Trans. Pattn Anal. Mach. Intell*., **17**, 539–544.
Moshfeghi, M. (1991) Elastic matching of multimodality medical images. *Comput. Vis. Graph. Im. Process*., **53**, 271–282.
Press, W. H. (ed.) (1994) *Numerical Recipes in Fortran: the Art of Scientific Computing*, 2nd edn. Cambridge: Cambridge University Press.
Ramsay, J. O. and Li, X. (1998) Curve registration. *J. R. Statist. Soc.* B, **60**, 351–363.
Ramsay, J. O. and Silverman, B. W. (1997) *Functional Data Analysis*. New York: Springer.
Rangarajan, A., Chui, H. and Duncan, J. S. (1999) Rigid point feature registration using mutual information. *Med. Im. Anal*., **3**, 425–440.
Rao, M. M. (2000) *Stochastic Processes: Inference Theory*. Dordrecht: Kluwer.
Ried, T., Baldini, A., Rand, T. C. and Ward, D. C. (1992) Simultaneous visualisation of seven different DNA probes by *in situ* hybridisation using combinatorial fluorescence and digital imaging microscopy. *Proc. Natn. Acad. Sci. USA*, **89**, 1388–1392.
Rosenfeld, A. and Kak, A. C. (1982) *Digital Picture Processing*, 2nd edn. San Diego: Academic Press.
Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L. G., Leach, M. O. and Hawkes, D. J. (1999) Nonrigid registration using free-form deformations: applications to breast MR images. *IEEE Trans. Med. Imgng*, **18**, 712–721.
Silverman, B. W. (1982) On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist*., **10**, 795–810.
———(1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
Singh, A., Goldgof, D. and Terzopoulos, D. (eds) (1998) *Deformable Models in Medical Image Analysis*. Los Alamitos: Institute of Electrical and Electronic Engineers Computing Society.
Strachan, N. J., Nesvadba, P. P. and Allen, A. R. (1990) Fish species recognition by shape analysis of images. *Pattn Recogn*., **23**, 539–544.
Stuart, A., Ord, J. K. and Arnold, S. (1999) *Kendall's Advanced Theory of Statistics*, vol. 2A, *Classical Inference and the Linear Model*. London: Arnold.
Studholme, C., Hill, D. L. G. and Hawkes, D. J. (1999) An overlap invariant entropy measure of 3D medical image alignment. *Pattn Recogn*., **32**, 71–86.
Tabernero, A., Portilla, J. and Navarro, R. (1999) Duality of log-polar image representations in the space and spatial-frequency domains. *IEEE Trans. Signal Process*., **47**, 2469–2479.
Thompson, A. M., Brown, J. C., Kay, J. W. and Titterington, D. M. (1991) A study of methods of choosing the smoothing parameter in image restoration by regularization. *IEEE Trans. Pattn Anal. Mach. Intell*., **13**, 326–339.
Thompson, D. A. W. (1917) *On Growth and Form*. Cambridge: Cambridge University Press.
Toga, A. W. (ed.) (1999) *Brain Warping*. San Diego: Academic Press.
Viola, P. and Wells III, W. M. (1997) Alignment by maximization of mutual information. *Int. J. Comput. Vis*., **24**, 137–154.
Vornberger, P. L. and Bindschadler, R. A. (1992) Multispectral analysis of ice sheets using co-registered SAR and TM imagery. *Int. J. Remote Sensng*, **13**, 637–645.

Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.
Wong, R. Y. and Hall, E. L. (1978) Scene matching with invariant moments. *Comput. Graph. Im. Process.*, **8**, 16–24.
Yang, Z. W. and Cohen, F. S. (1999) Cross-weighted moments and affine invariants for image registration and matching. *IEEE Trans. Pattn Anal. Mach. Intell.*, **21**, 804–814.
Younes, L. (1999) Optimal matching between shapes via elastic deformations. *Im. Vis. Comput.*, **17**, 381–389.

Discussion on the paper by Glasbey and Mardia

Ilya Molchanov (*University of Glasgow*)

The authors consider an important problem that has many applications in image analysis and encompasses several approaches that have already been developed in this and related areas. I would like to revisit several issues related to this paper, sometimes rephrasing them in a more abstract language.

A general formulation of warping problems

Warping ideas appear in probability theory and statistics under many different names. A general formulation of such problems involves the minimization of a functional

$$m(x, y) = \min_{\lambda \in \Lambda} \{\rho(x, \lambda y) + D(\lambda)\}$$

over a family Λ of transformations defined on a space X, where $x, y \in X$ and ρ is a symmetry on X, i.e. $\rho(x, y) = \rho(y, x)$ and $\rho(x, y) = 0$ implies that $x = y$. One of the first examples of such functionals appears in the definition of the Skorohod topology (Skorohod, 1956) that is widely used to formulate limit theorems for stochastic processes. Then x and y are cadlag functions on [0, 1] (i.e. continuous from the right and having limits from the left), ρ is the uniform metric and λ determines the change of variable using a monotonic bijective transformation of [0, 1]. The smallest value of $m(x, y)$ is then called the Skorohod distance between x and y. Its counterpart in statistics concerns the alignment of curves and dynamic time warping; see Wang and Gasser (1999). In studies of shapes and figures, $D(\lambda)$ vanishes if $\lambda \in \Lambda$ and is infinite otherwise, so $m(x, y)$ is obtained by minimizing $\rho(x, \lambda y)$ over $\lambda \in \Lambda$. Furthermore, $m(x, y)$ with $\rho(x, y)$ interpreted as the likelihood and $D(\lambda)$ as the penalization or prior is one of the key ideas in smoothing and Bayesian statistics. It also is widely used as a regularization technique in numerical methods for ill-posed problems. In image analysis some effort has been put into constructing geodesics (or interpolation) that provide the 'shortest' warping transformation between binary images (Serra, 1998) and smoothing of image sequences (Friel, 1999).

Image dissimilarity measures

The likelihood term in the penalized likelihood functional provides an example of a dissimilarity measure for grey scale images. In fact this paper is not concerned with modelling images and the main application of the Fourier–von Mises model advocated by the authors is not for modelling but for defining a loss function that may also be called an image dissimilarity measure or image metric. It is generally recognized that conventional distances (e.g. L^2 or root mean square) do not perform well for grey scale images. For binary images, a family of useful image metrics was proposed by Baddeley (1992). However, his idea cannot be easily extended to grey scale images at a reasonable computational cost; see Wilson *et al.* (1997), Friel and Molchanov (1998) and Kaijser (1998) for further discussions of grey scale image metrics. The Fourier–von Mises image model offers a sufficiently flexible and computationally efficient approach to defining grey scale image metrics. It would be interesting to investigate its performance for typical examples that involve assessing distances between grey scale images from the above-mentioned references.

Warping as preprocessing for averaging

While calculating averages of fish images, the authors used warping to align the images before averaging. In this case the target image μ that is used to warp individual images is unknown and the approach is to warp images 'close together'. The averaging is performed for the post-warped images and therefore must match warping. In other words, the average of several warped images $y_1, \ldots, y_n$ should be defined as an image μ that minimizes $\Sigma\, \rho(\mu, y_i)^2$, where ρ is the same loss functional that was used to determine the optimal warpings, i.e. the Fourier–von Mises loss function in the context of the current paper. In application to binary images, this idea was pursued by Stoyan and Molchanov (1997).

Consistency of estimators
The warping functions are estimated as minimizers of the loss functional. In most applications the warped images may be subject to misregistration. This calls for an application of stochastic optimization methods to deduce that the estimators suggested by the authors are consistent. Quite similar ideas were developed by Wang and Gasser (1999) whose techniques may be applied to assess the order of bias when the warping transforms are being obtained using a sample of images produced by kernel smoothing from observed data. It is also essential to ensure that the gradient algorithms that are used by the authors do not become trapped at one of the local minima.

Distortion measures
The null set distortion criterion that is used by the authors can be equivalently formulated as the 'distance' between the transformation (warping) λ and the set Λ that consists of 'neutral' (or null) transformations.

In my opinion, the approach pursued by the authors looks promising and worthy of further exploration. The examples presented are convincing and the algorithms look computationally efficient. I congratulate the authors on a stimulating paper and have great pleasure in proposing the vote of thanks.

C. Jennison (*University of Bath*)
Dr Glasbey and Professor Mardia have offered us a stimulating paper. They have combined models for image data, a general method of controlling the warping function which can allow certain transformations without penalty, and an impressive computational algorithm. The three illustrative examples are well motivated and each introduces its own special complexities.

The technique of penalized likelihood that is employed by the authors has a curious status. Its connection with Bayesian methods is well known (e.g. Green and Silverman (1994), page 51). The penalized likelihood of the observed image Y obtained from the penalized log-likelihood in equation (1) is

$$l(Y|\mu, f, \xi)\ \exp\{-\lambda\, D(f, \mathcal{C})\}$$

where $l(Y|\mu, f, \xi)$ is the likelihood of Y given μ, f and ξ. In a Bayesian interpretation this formula is a multiple of the posterior density of f and ξ, given the observed Y, for the case where f has prior density proportional to $\exp\{-\lambda\, D(f, \mathcal{C})\}$ and ξ an improper flat prior. The posterior distribution of f provides a basis for inference; one could also place a prior on λ and estimate this from the data along with everything else. Despite the current popularity of Bayesian methods in statistical image analysis, the authors do not advocate such an interpretation of their work: perhaps they do not regard $\exp\{-\lambda\, D(f, \mathcal{C})\}$ as a reasonable prior for f. However, if the penalty term is viewed simply as an *ad hoc* way of regularizing maximum likelihood estimation in an otherwise ill-conditioned problem, one is left with a rather *ad hoc* method for obtaining a point estimate of f and no simple way of quantifying uncertainty in this estimate.

In their first example, the authors choose an edge-filtered version of the original synthetic aperture radar data as their 'image' Y and assume that $Y|\mu$ follows the Fourier–von-Mises distribution. This is a pragmatic and, evidently, very effective choice. Despite the high noise level, direct processing of such synthetic aperture radar data is possible: Hurn and Jennison (1995) presented a multiresolution algorithm for fitting a Markov random field image model of the type proposed by Geman and Reynolds (1992) which encourages sharp edges in the image. I wonder whether stochastic image models might also be incorporated in the authors' methods; for example, in the third example, one could consider a Markov random field model under which grey levels vary slowly with occasional sharp discontinuities to specify spatial properties of μ, the average haddock or whiting. What do the authors think about

(a) the feasibility of incorporating this extra ingredient in their algorithms and
(b) when this might lead to significant improvements in results?

I have two technical points. The first concerns the summation in the data log-likelihood (4) which is over points x in the μ_x-lattice. Corresponding values $Y_{f(x)}$ are formed as weighted combinations of values observed at the lattice points of the image Y but no account appears to be taken of the correlations between the variables that are thus created. In any case, since the Y are the observed data it would be natural to define the data log-likelihood as a sum over observations Y at lattice points of the observed image and then to use bilinear interpolation of f^{-1}, the inverse of the warping function, to give a μ-value for each Y.

My second point of detail pertains to the calculation of $D(f, \mathcal{C})$ in the optimization algorithm of Section 2.3. The function f is specified through its values at a set of grid points with remaining values of f obtained by interpolation. Piecing together the interpolating functions on each square of the grid gives a function f with discontinuities in high order derivatives at the boundaries of these squares. Either this form of interpolation is too crude when the penalty term involves such high order derivatives, or contributions are needed from the behaviour at the edges of grid squares as well as from integrals within each square. Sibson and Thomson (1981) used piecewise quadratic functions with derivatives matched at the 'seams' to tackle a related problem.

Turning to a more general perspective, the authors are a prime example of statisticians participating in a field where others, including in this case electronic engineers and computer scientists, have already created a battery of effective techniques. We may ask what a statistician brings to such a field, and the methods presented by the authors without a Bayesian setting are of particular interest as one cannot just fall back on the ability of statistical methods to provide a measure of confidence in their results. I believe that we do bring a new perspective, drawing on a different and often complementary heritage. The detailed study of the philosophy and methodology of statistical inference helps in judging what is achievable in new application areas—and the pitfalls that may await. We should certainly not be timid: there are plenty of researchers from other disciplines exploring fields one might have regarded as the rightful domain of statisticians!

The authors have tackled difficult problems with a substantive statistical component. They have been inventive in pursuing these problems to real, effective solutions, demonstrating the value of their statistical approach. It is a pleasure to congratulate them on their achievements and to second the vote of thanks.

The vote of thanks was passed by acclamation.

Bjarne K. Ersbøll (*Technical University of Denmark, Lyngby*)

I congratulate the authors on their extremely neat unification of image warping techniques. Apart from Carstensen (1996) which—as mentioned—presents a Bayesian type of approach to image warping, we have at my department experience from both correlation-based (matching two-dimensional electrophoretic gels; Conradsen and Pedersen (1992)), landmark-based (matching human mandibles; Andresen *et al.* (2000)) and local phase-based warping (matching stereo image pairs, following Granlund and Knutsson (1995)). A further implication of matching and warping is the field of optical flow. Larsen *et al.* (1998) adopted an approach with some resemblance to the methodology mentioned in the paper in order to interpolate and extrapolate image sequences. Your paper ingeniously seems to have combined all these techniques. Furthermore, the Fourier–von Mises idea has a nice appeal to it.

Using these conceptually different techniques on real applications we found that it is usually necessary to operate on multiple resolutions with a stepwise coarse to fine refinement of the warping to obtain reliable results. The aspect of a multiresolution approach is only briefly mentioned in the paper, however, so I would be extremely interested in the opinion and experiences of the authors on the matter.

Below are three examples which raise some questions about and might challenge your technique.

(a) When matching fundus images from the same patient over time an important problem is the occurrence or disappearance of features. This could probably be modelled as an occlusion or folding as it is termed in the paper. Furthermore, the structures to be matched—here blood vessels—change over time. Occlusion occurs for many types of images; how does the methodology proposed perform here? Which assumptions are necessary?

(b) A spot image and an orthophoto have extremely different resolutions ($20\,\mathrm{m} \times 20\,\mathrm{m}$ *versus* $62.5\,\mathrm{cm} \times 62.5\,\mathrm{cm}$ pixels), making the matching problem far from trivial. The questions are the same as before.

(c) Finally, many image data are in colour or even multispectral. Consider matching two Landsat images, one taken during the summer where the altitude of the sun is higher and the vegetation is drier than on the one taken during the winter. Does the methodology generalize to include image cues such as colour or multispectral information, or do we have to make do with a suitable projection onto grey scales?

Edwin Hancock and Richard Wilson (*University of York*)

Mardia and Glasbey have presented an important paper which we believe contains ideas which will also

prove influential on researchers in the fields of computer vision and image analysis. They address the pervasive problem of image alignment. They offer a way of gauging data closeness in the Fourier domain by using the von Mises distribution and for regularizing the alignment parameters by using the null set distortion criterion.

For the past decade, we and our co-workers have been studying the problems of image alignment and correspondence matching. Our contributions have been threefold. First, we have shown how to use the expectation–maximization (EM) algorithm for spatial domain alignment under similarity, affine and perspective transformations (Cross and Hancock, 1998); the resulting algorithms have been applied both to synthetic aperture radar images and to the more complicated images delivered by millimetre radars (Moss and Hancock, 1997). Our second contribution has been to develop probabilistic methods for graph matching, which can be used to find correspondences between point or line features in different images (Wilson and Hancock, 1997; Myers *et al.*, 2000). When working with point or line features, as is frequently the case in high level vision, then the alignment and correspondence problems have a chicken-and-egg relationship to one another. Our third contribution has been to develop a dual step EM algorithm in which the recovery of alignment parameters is constrained by the pattern of correspondence matches residing on a relational graph which represents the arrangement of image features (Cross and Hancock, 1998).

We can see the scope for important synergies between our work and that of Mardia and Glasbey. From our experiences, we have three suggestions for the authors. First, and as alluded to by Professor Molchanov, the Fourier domain alignment process could be usefully integrated into an EM algorithm which iterates between estimating distortion parameters in the M-step and computing *a posteriori* matching probabilities in the E-step. Second, it might prove fruitful to develop a localized feature representation along the lines of the Von der Maalsburg bunch graph and to organize these features by using relational graphs. Thirdly, we would be interested in whether there might be advantages in using alternatives to the thin plate spline to generate the null set distortion criterion.

Graham Horgan (*Biomathematics and Statistics Scotland, Aberdeen*)

The importance of the work presented by Glasbey and Mardia derives from the generality with which it

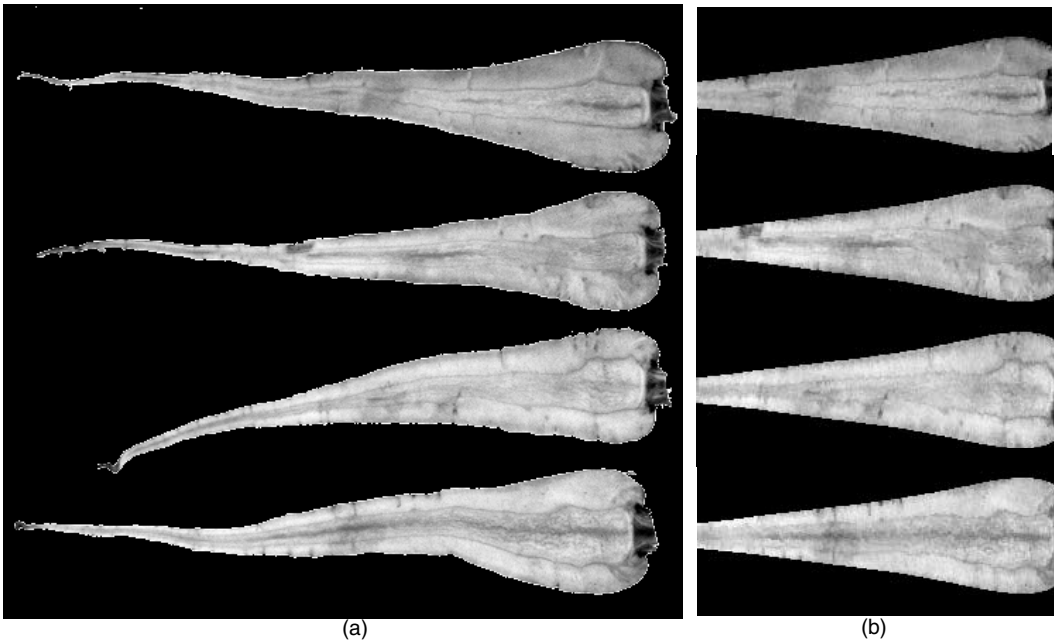


Fig. 14. (a) Longitudinal slices of four parsnips and (b) the parsnips warped to a common outline by a piecewise affine transformation

may be applied to many diverse applications. In two of the examples used (synthetic aperture radar and microscopy), the interest is the warping itself. The effectiveness of the method is demonstrated in Figs 9 and 10. In the former we can see precise fitting over much of the image, where it is possible, combined with the rigidity of the warping at the bottom right-hand side.

The third application (fish) is of particular interest, in that warping fish images to match each other is not the central motivation, which is rather the discrimination of different species. This gives the work in the paper a great potential breadth of application. Much research effort has been expended on the recognition of individuals from images of faces, for example (Craw *et al*., 1999). Intermediate between the recognition of species and of individuals is the horticultural application of cultivar discrimination (Horgan *et al*., 2001). Implicit in much work on these topics is the idea that differences between images are split into two types—differences in outline shape and differences in grey levels within the outline. After the former have been described, warping is carried out to remove outline shape differences and to make a separate study of the latter (Fig. 14). The authors' method would handle both these aspects of appearance simultaneously.

One significant advantage of splitting the discrimination into two stages is that the background, and any variation in it, within or between images, becomes irrelevant. This may be important in some applications. It can also be useful to study the components of variation of both types separately. The question which naturally arises, then, is whether the authors' methods can be adapted to preserve these benefits, perhaps by some preprocessing of the images, or in some other way.

John T. Kent (*University of Leeds*)

The paper contains several fascinating statistical proposals for image warping. I would like to draw attention to some related work by two of my recent doctoral students. The first result is by Debbie Godwin. Let R_0 and R_1 denote two simply connected compact regions in $\mathbb{R}^2$ with smooth boundaries parameterized by functions $\varphi_j: S_1 \to \mathbb{R}^2$, $j = 1, 2$, where S_1 denotes the unit circle. Consider the interpolation problem of finding the 'smoothest' possible deformation $f: R_0 \to R_1$, such that f is required to match the parameterized boundaries, $f \circ \varphi_0 = \varphi_1$, and where smoothness is defined in terms of minimizing a penalty. Two possible penalties are given by equations (16) and (21) in the paper, which in the current context take the form

$$\begin{aligned} D_1(f) &= \int_{R_0} \sum_{i,j=1}^{2} (\partial f_i/\partial x_j)^2 \, \mathrm{d}x \\ &= 0 \qquad \text{for } f \text{ constant,} \\ D_2(f) &= D_1(f) - \text{correction terms} \\ &= 0 \qquad \text{for } f \text{ a similarity transformation.} \end{aligned}$$

Godwin (2000) proved the surprising result that the optimal deformations under the penalties $D_1(f)$ and $D_2(f)$ are the *same*, i.e. the introduction of correction terms is irrelevant to the optimal fit, though of course the value of the optimal penalty is different in the two cases. As a consequence of this result, one can ask whether the introduction of correction terms makes much difference to the fitted deformations in the examples of this paper.

A contrasting conclusion was reached by Gary Walker, who was jointly supervised by me, Ian Dryden and Chris Glasbey. In a one-dimensional 'smoothing spline' version of the problem, again based on first derivatives, Walker (2000) found that the introduction of a correction term made a substantial and useful difference. In his setting the desired deformation is close to linear. But, without the introduction of the correction term, the fitted deformation tends to flatten out beyond the range of the data since only the constant function lies in the null space of the penalty. With the introduction of the correction term, linear functions also have a zero penalty, and the fitted deformation is much closer to a linear function.

John Ashburner (*Institute of Neurology, London*)

A feature of this paper that I like is that it ensures consistency by using consensus images. Internal consistency is an often overlooked factor in image warping. For example, warping one image to match another does not necessarily produce the inverse of the deformation obtained by warping the images the other way around. This is partly because derivatives of the log-likelihood function with respect to the warping function's parameters often depend on the gradient of only one of the images.

Another reason may be 'asymmetries' in the measure of distortion that is used to penalize improbable warps. These arise when a particular deformation is not deemed as probable as its inverse. Similarly, individually warping together many pairs of images may not produce deformations that are consistent with each other. The use of a consensus image maintains consistency between all the warps, as the mapping between any pair of registered images can be obtained by combining a forward and an inverse (bijective) warp.

The paper describes linearly regularizing distortions in addition to those that can be represented by a rigid body transformation and isotropic zoom. By including these transformations within the null set of functions, the estimated shapes should not be influenced by object size and pose. One complication that may occur is if one (very supple) fish is bent in the middle by 90°. This fish only has a different shape in the middle, whereas its head and tail are both normal shapes. There may be enough evidence in the images to bend the outline of another fish to match it. However, the term that penalizes distortions may still have unwanted effects, as the null set distortion criterion does not model the different rigid body transformation between the head and tail.

My own thoughts are that measures of distortion should be used that are completely rotationally invariant. This can be achieved by using singular value decompositions of the Jacobian matrices at each point of the deformation, which effectively decompose each matrix into a rotation, a set of orthogonal zooms and another rotation (Ashburner *et al.*, 1999). A rotationally invariant measure of distortion can then be derived from the zooms. A penalty function that is identical for both a forward transform and its inverse can then be constructed by assuming that the logarithms of the zooms are normally distributed. This also means that logarithms of areas and volumes would both also be normally distributed.

Ian L. Dryden (*University of Nottingham*)

My comments centre on the new methods of this interesting paper.

Independence

The independence model (3) is given as motivation for the von Mises model, leading to $\xi = (-\log(\sigma^2), 0, 0, 1, 1)$. The authors actually use a much more general model with $\kappa_\omega(\xi)$ given by equation (7). How do these von Mises distributions relate to possible joint distributions for the grey levels $Y_{f(x)}$ and their joint correlation structure?

Consider an $n_1 \times n_2$ image to have toroidal boundaries and let $Y_{f(x)}$ be a Gaussian random field on the discrete image with mean μ_x and covariance matrix Σ which is block circulant with circulant blocks. If W_n is the usual $n \times n$ discrete Fourier transform matrix,

$$(W_{n_1} \otimes W_{n_2})^* \Sigma (W_{n_1} \otimes W_{n_2}) = \mathrm{diag}(\sigma^2/\lambda_\omega),$$

say. Hence, under this model the displayed equation after equation (12) has σ^2 replaced by σ^2/λ_ω, and then equation (6) is replaced by

$$\kappa_\omega(\xi) = \exp\{-\log(\sigma^2) + \log(\lambda_\omega) + \log(A_\omega^{(\mu)}) + \log(A_\omega^{(Y_f)})\}.$$

There is some overlap with the model derived from this family of Gaussian random fields and the class considered by the authors. A simple practical model (not in the authors' class) is a homogeneous Gaussian Markov random field for $Y_{f(x)} - \mu_x$, where

$$\lambda_\omega = 1 - \sum_{h \in N} \beta_h \cos(2\pi h^{\mathrm{T}} \omega),$$

where N is a finite symmetric neighbourhood of the origin and $\beta_h = \beta_{-h}$. It would be interesting to see how these models compare with, for example, Fig. 8.

Null set criterion

How should one choose the null set criterion in practical applications? Typically there are two types of matching situations:

(a) different views and/or modalities of the same geometrical object (e.g. applications 1 and 2);
(b) images of different objects (e.g. application 3).

In (a) the geometrical variability of the object itself is zero so one ought to aim to match exactly. In (b)

there are geometrical differences in the objects as well as possibly differences in the modalities or views. Partitioning this variability is not so straightforward in general.

The authors do have a method for choosing λ based on discrimination in application 3, but how do we interpret the value? Are the fish populations significantly different in shape and size or is it just texture information that is helping with discrimination?

Variability of estimators

No method for obtaining standard errors or credible regions is given. The probabilistic interpretation on f via $D(f, \mathcal{C})$ would be helpful, and posterior credibility intervals could be obtained by Markov chain Monte Carlo methods of course.

Higher dimensions

Although in principle warping is 'straightforward' in higher dimensions, there is much greater complexity when considering certain aspects, e.g. shape theory (see Kendall *et al.* (1999)).

Homogeneity and robustness

Different weightings to parts of the image are often useful (see the end of Section 3) but may make a substantial difference to the solution. Thus, a robust solution may be called for (for example see Dryden and Walker (1999) for a matching example using S-estimators).

M. Petrou (*University of Surrey, Guildford*)

This is an impressive piece of work that brings together two major topics: image registration and invariant feature construction.

An interesting problem is the identification of functionals which annul the effect of the functions of a particular group of transformations (Kadyrov and Petrou, 2000).

We have developed what we call the 'trace transform' which computes functionals along tracing lines of the image, to map the image onto the line parameter space. Further functionals are chosen applied to the parametric representation of the image so that they yield a single number that is invariant to the group of transforms that we have chosen.

The philosophy behind such an approach is that the first functional is computed along the curves that are left unchanged by the group of transformations. For linear and affine distortions these curves are just straight lines. For more complicated transformations, however, the unaffected curves are much more complicated and it is difficult to apply the theory in such cases. This is the point where we may have to abandon the approach based on the group of transformations and use an approach based on a set of locally applied operators that cause deformations.

We have developed an image registration algorithm for the elastic registration of three-dimensional images based on this idea.

In the optimization approach we invoke at random operators that are applied locally and deform one image to match the reference image. This way we do not restrict ourselves to a particular group of distortions, but to a particular set of deformations appropriate for the images of the application that we are interested in, and which may be different in different places of the image. Such is, for example, the case of a medical image where a tumour is growing locally.

The following contributions were received in writing after the meeting.

José M. Angulo (*University of Granada*)

I would like first to thank the Research Section for this opportunity to contribute to the discussion of this interesting paper, as well as to congratulate the authors for their significant contribution with this work. One of the most challenging problems in image warping consists of the proper definition of the concept of distortion and its formal treatment in applications. In this regard, the authors propose a methodological approach based on penalizing the likelihood associated with the warping in terms of distortion.

From my perspective, the innovative contributions in the paper, regarding both the methodological approach and the technical solutions given to its implementation, provide a significant basis for future research, such as a consideration of possible alternatives. In this respect, I shall focus my comments on certain aspects related to *scales* on the basis of the methodology proposed.

First, D_B as defined in criterion (13) and then used in the definition of $D(f, \mathcal{C})$ in equation (14) can be viewed as an 'absolute displacement-based' distortion criterion. Formally and also partly conceptually, an alternative would be to formulate, say $D^*(f, \mathcal{C}^*)$ as the minimum of $D_B^*(f \circ h)$, for $h \in \mathcal{C}^*$, with $D_B^*(\cdot)$

now being a non-negative functional null at the identity. This would also require a slightly different consideration for class $\mathcal{C}^*$ instead of $\mathcal{C}$. In the case of bijective deformations, the simple relationship $f - g = (f \circ g^{-1} - Id) \circ g$ suggests that, whereas in $D(f, \mathcal{C})$ the function f is compared with each function g on the (absolute) scale of the domain of f, in $D^*(f, \mathcal{C}^*)$ the comparison of f with each function $h = g^{-1}$ would be performed on the (relative) scale of the domain of g^{-1}, i.e. 'removing' g from f in a compositional sense rather than by subtraction. For f in the null set of functions, we have a distortion of 0 in both cases. For other f, we would measure distortion as the 'departure' in terms of different scales. Such a difference can be related to the primary question of defining a concept for *distortion*.

Second, equation (1) involves a mixture of quantities, the log-likelihood L and the measure of distortion D, which are defined, in principle, by scales of a different nature. Under the same idea of *penalizing the likelihood in terms of distortion*, we might then think of considering different algebraic alternatives from this construction of the objective functional, and the problem remains to study the properties and a proper justification for each specific formulation, as well as comparative performances.

Mark Berman (*Commonwealth Scientific and Industrial Research Organisation, Sydney*)

I congratulate the authors on their interesting paper, which neatly combines aspects of Fourier theory, circular statistics and regularization to address the general problem of image warping. The paper generalizes a number of simpler image matching procedures and elegantly combines intensity matching and edge matching in one measure. In addition, the methodology provides performance measures for assessing the quality of a warp.

However, the methodology is mathematically complex and computationally intensive, and is only likely to be taken up by the computer vision community (and optimized by them, especially for speed purposes) if the value of the methodology can be demonstrated in an application of wide interest to that community and their commercial partners. One possible area is in face recognition, to which the authors briefly allude in their discussion of the fish discrimination problem. However, in the fish images the backgrounds appear reasonably homogeneous. This is unlikely to be the case in many practical face recognition problems. How would the authors' methodology work with variable backgrounds, or would the fish or faces need to be segmented out first?

My main technical comment is about the use of cross-validation in the synthetic aperture radar example and in image analysis problems generally. Replacing a block of pixels by a constant (rather than omitting them altogether) does not seem a very natural approach. The authors apparently need to do this because the discrete Fourier transform does not normally cope with 'missing values', although this can be dealt with in some linear problems (Berman, 1994). In image analysis problems where more classical cross-validation can be performed, it is not necessarily the best thing to do. For instance, in the image segmentation context, Lee (2000) demonstrated the superiority of a minimum description length-based approach over a cross-validation procedure suggested by Bose and O'Sullivan (1997).

Finally, I have a query about the choice of q_1 and q_2 in the optimization algorithm. Are these parameters chosen 'by eye' or in some more objective way? This would be an issue for those interested in a fully automated solution.

Rodney Coleman (*Imperial College of Science, Technology and Medicine, London*)

There is much to admire in the methodology described in this paper, and I wish to offer suggestions for areas in which it might be further developed.

In particular, the Fourier–von Mises image model hints at the possibility of fashioning functionals that might be used as image signatures that can be compared for similarities, but which are not dependent on the viewer's 'expert knowledge' being applied directly to the actual images. With this in mind, the haddock *versus* whiting example (Fig. 3) appears to be a poor illustration, since expert knowledge is a natural starting-point for discriminating between the two species and requires no technology. Thus, a chicken sexer can identify by eye the sex of day old chicks at a rate of one every second.

On a second point, I feel that attention might also be turned to the potential of new technological advances that could make warping for image registration unnecessary. By way of illustration, with conventional radiography, the wide variation in object thickness means that, even with automatic exposure devices, optimal exposure over the entire field is impossible. In the mid-1980s scanning equalization devices which rely on a sophisticated feed-back system to modulate the exposure of the X-ray beam were developed and are in use, though not yet in everyday clinical operation (Hansell *et al*., 1991).

N. Duta and A. K. Jain (*Michigan State University, East Lansing*)

We would like to congratulate Dr Glasbey and Professor Mardia for proposing an interesting solution to the practical and difficult problem of warping bidimensional signals. Warping-based approaches to object recognition and identity verification in digital images have been successfully applied in several domains. However, most of the applications have been limited to one-dimensional signals. We believe that the methods introduced in this paper will increase the applicability and effectiveness of warping-based pattern recognition systems.

We have recently been investigating the object identification and localization problem based on two-dimensional shape and one-dimensional grey level pattern warping. The biometrics (hand-shape-based) security system (Jain and Duta, 1999) and the medical diagnosis (cardiac ventricle localization–segmentation) system (Duta *et al.*, 1999) which we outline in Fig. 15 complement the applications reported in the paper. We followed the same methodology as the authors propose for object identification. Two object patterns (e.g. the hand shapes in Fig. 15(a)) are warped onto each other (Fig. 15(b)). Subsequently, a warping distance measuring the non-linear distortion (after the similarity transformations have been factored out) between the two patterns is computed. A threshold is applied to this distance for deciding whether the two patterns belong to the same class of objects. The distributions of the distances between objects belonging to the same (left-hand curve) and different (right-hand curve) classes are shown in Fig. 15(c). The method's potential in separating object classes is quite high: a 95% correct acceptance rate corresponds to 1% false acceptance rate. A different type of image warping application is feature or object localization based on registration to a template (atlas or map similar to the first application described in the paper). Multiple detections produced by a classifier on the cardiac image in Fig. 15(d) must be combined to obtain one accurate position of the ventricle (the white circle in Fig. 15(g)). The intersection points between the medial axis of the ventricle and the four main directions can be estimated by warping the corresponding signal profiles (the thin curves in Figs 15(e) and 15(f)) to an average profile (the thick curve). In this way, the position of some salient points along the ventricle (marked by asterisks in Fig. 15(e)) can be estimated after alignment of each profile to the template (Fig. 15(f)).

J. K. Ghosh and C. A. Murthy (*Indian Statistical Institute, Calcutta*)

This is a very interesting paper with many novel ideas and three similar but not identical problems. We focus on one of them to motivate a Bayesian approach, which is applicable to all three examples. We also sketch a computing strategy. Much fine tuning would be needed to make it work.

Let $Y^{(i,j)}$ be the ith image from the jth species, μ_j some average for the jth species and $f^{(i,j)}$ the restoring functions. The log-likelihood

$$L(Y^{(i,j)},\ i, j = 1, 2|\mu_j, f^{(i,j)},\ i, j = 1, 2)$$

is assumed Gaussian and the log-prior for μs and fs is of the form

$$p(\mu) - \lambda\left\{\sum_{i,j} D(f^{(i,j)}, C)\right\} - \log\{C(\lambda)\}$$

where $C(\lambda)$ is a normalizing constant. To reduce the calculations, we drop p and replace μ by an empirical Bayes estimate μ_j along the lines of equation (37) of the paper and using the (posterior) mode of $f = \{f^{(i,j)}, i, j = 1, 2\}$.

For any fixed λ, the evaluation of $C(\lambda)$ is sensitive to a precise specification of the function class being searched. Using the notion of a weak version of Laplace integration and data-dependent priors, we suggest using the class of functions actually searched in the course of maximization with respect to f for fixed λ. Interpreting

$$P = L - \lambda\left\{\sum D(f^{(i,j)}, C)\right\} - \log\{C(\lambda)\} \tag{38}$$

as the logarithm of likelihood and prior for fs, we can calculate the posterior of the fs, the posterior mode of f and a measure of deviation

$$S_W = \sum_{i,j,x,f} (Y^{ij} f^{(i,j)} - \mu_{jx})^2 p(f|Y_N) \tag{39}$$

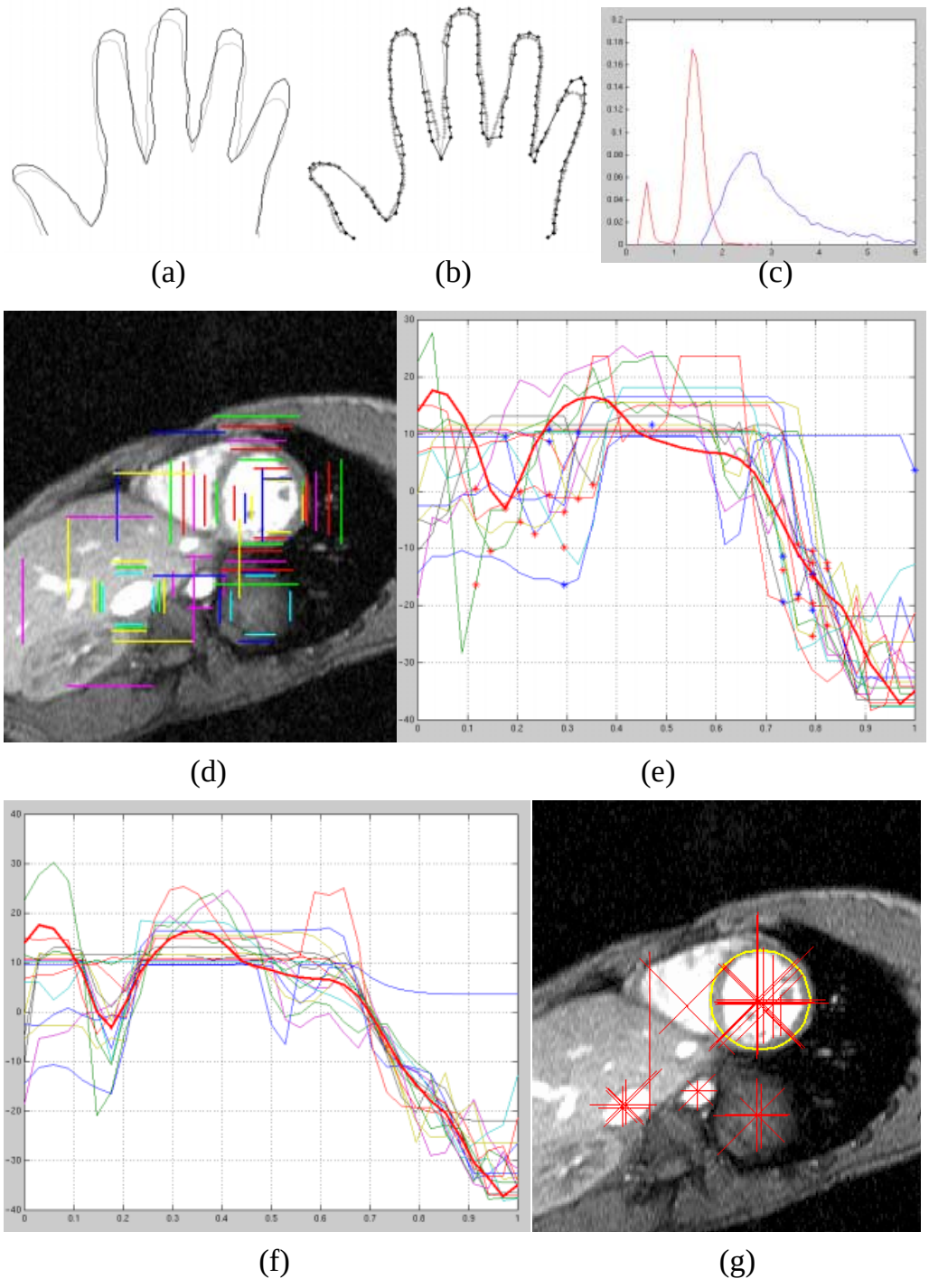


Fig. 15. Practical applications of warping-based object identification: (a)–(c) identification of identity based on hand shapes; (d)–(g) localization of a cardiac ventricle

where μ depends on f as indicated earlier. To choose λ we may maximize a Studentized mean as in Glasbey and Mardia's paper but a natural criterion is S_B/S_W, where

$$S_B = \sum_x (\mu_{1x} - \mu_{2x})^2.$$

A more natural method is to maximize expression (38) with respect to λ. The presence of $C(\lambda)$ prevents equation (38) from being monotone.

Finally, we have a couple of questions. If two images are different because the object has moved a hand or face in space, what would be the two-dimensional function class for the images? An answer would throw light on deformable template matching. Indeed, a discussion of this problem from the point of view of this paper would be very interesting. Secondly, how sensitive are the final images to variations in λ? Also, would the inclusion of a variance term in the Gaussian likelihood make the problem ill posed?

A. Gray (*University of Strathclyde, Glasgow*)

This paper offers a sophisticated statistical solution to the image warping problem, and the penalized likelihood framework that is used is neat and familiar.

An appealing feature of the method is that it is not necessary to supply control points, manually or otherwise, and the proposed Fourier–von Mises image match criterion can use edge information as well as intensity information. However, it is still necessary to specify a suitable set of transformations, in the form of the null set C, rather than these being suggested by the data (as would be ideal) and also to specify an appropriate base distortion criterion to find the distortion penalty function $D(f, C)$. Although the likelihood is given a recommended form, this also involves unknown parameters to be estimated at the same time as the warping function, and the optimization appears very time consuming.

The results shown are impressive compared with those of more standard approaches. Nevertheless the method is complex to implement and to compute, and an appropriate value of λ still requires to be chosen or else a range of different values used, as well as needing to experiment with the fineness of the grid on which the warping function f is approximated. Given the time that is needed for the optimization, achieving good results will therefore be slow.

As it is described the approach appears very flexible and powerful; however, in all three examples some *ad hoc* method or approximation appears to be necessary to make the approach work. Therefore the methodology is not necessarily quite as usable as its presentation suggests. Despite the superior results, its extra complexity may mean that in practice simpler, quicker, methods will be preferred to this new approach.

John Gustafsson and Mats Rudemo (*Chalmers University of Technology, Gothenburg*)

We would like to discuss another application of the approach that is presented in the paper: matching of two-dimensional electrophoresis gel images. Two-dimensional gel electrophoresis is a method for the simultaneous separation of thousands of proteins from a complex protein solution on the basis of their isoelectric point and molecular weight. It is currently one of the major methods in proteomic research. One crucial step in two-dimensional gel analysis is to match spots in different gel images that correspond to the same protein. This matching step seems to be a bottle-neck in the gel analysis. It still requires extensive and time-consuming manual interference, although several semi-automatic techniques exist (Voss and Haberl, 2000).

The statistical formulation in the paper by Glasbey and Mardia provides a general framework for the formulation of an automatic warping method to find an image alignment that can aid the matching of spot patterns. We are currently investigating a combination of two warping methods. First, we have formulated a simple physicochemical model of what might be one main cause of spatial distortion of the spot pattern: current leakage. Basically the model is a set of coupled partial differential equations including Laplace's equation for the electric potential. We apply the model to a global warping of each gel image to correct for the effect of current leakage. Thereafter, we use the penalized likelihood approach to align the images locally with piecewise bilinear transformations to handle distortions that cannot be explained by current leakage. In this application the simple Gaussian image model is a natural choice, and for the null set in the distortion criterion we use the set of affine transformations.

Given an undistorted reference gel it might be possible to unify these two steps by choosing as the null set the functions that satisfy the partial differential equations in the physicochemical model. One might also add a probabilistic interpretation of the transformation by introducing a stochastic Poisson

equation for the electric potential. Further, in this application we wish to conserve the spot contents during the warping process. Compared with the paper it is therefore natural to use a slightly different definition of the warped version of Y under f:

$$(Y_f)_x \equiv \left|\frac{\partial f}{\partial x}\right| Y_{f(x)} \qquad \forall x \in X,$$

where the area scale introduced is the Jacobian determinant of the warping transformation.

David Hogg (*University of Leeds*)
I would like to start by congratulating the authors on this stimulating paper.

On reading the paper, I found myself looking for ways in which the intriguing method proposed by Glasbey and Mardia could be adapted to improve a different kind of approach to certain kinds of warping problem.

For situations in which the aim is to locate a familiar object within a two-dimensional image (as for the fish example), a powerful approach has been to model the variation of shape found in a representative data set, using for example a Gaussian density or mixture of such densities. This contrasts with generic models based on smoothness criteria or likelihood functions, even though the parameters of such models could in principle be estimated from a data set. Blake extended such an approach to permit simple transformations such as an affine mapping. In his recent book, he demonstrates that such an extension can also be made for class-specific models of the variation in shape. It is interesting to speculate whether the null set distortion criterion could be similarly chosen, thereby integrating learnt models of variation with a set of distortions that arise from the physics of the imaging situation—specifically, a projectivity for roughly planar objects or an affine approximation to this.

In addition to class-specific models of deformation, the variations in intensity within a deformed image (i.e. the intensity array after alignment) can also be modelled in a similar way. A problem here has been to deal with lighting variation under different imaging conditions. The Fourier–von Mises image model and the way in which it is used by Glasbey and Mardia may provide a way for factoring out these sources of variation before modelling. Related approaches based on wavelets have been used in these situations with some success.

Inge Koch (*University of Newcastle, Callaghan*)
The paper makes an important contribution to the development of registration methods and represents a serious attempt at making registration and imaging methods acceptable to the statistics literature. Registration methods evolved from the statistical concept of cross-correlation in the early 1970s and have been applied in science and computing areas since then, but they have not enjoyed the same popularity in the statistical community.

The idea of the warping function extends many of the classical registration methods that are currently in use in engineering and imaging applications which are restricted to translation and rotation between the images. The warping function, and its estimation by penalized likelihood, is one of the strong points of this paper, as it allows—at least in theory—a large class of functions to be treated simultaneously. If the relationship between the images is simple, such as a translation, then the new approach may not lead to an improvement, since Fourier-based registration methods in particular perform extremely well and can be implemented very efficiently. For more complex transformations between images the new method shows its strength through its flexible definition of the warping function and the penalty term.

The examples are clear and indicate the scope of the new approach. However, by being very general, the method also shows a certain weakness, as it is not clear how to apply it in cases which do not fit into the framework given in the three examples. There is no strategy of how to choose the parameters ξ for the concentration κ of the phase factors or which distortion criteria should be applied. Indeed, the motivation for the choice of the parameter ξ is not very convincing, apart from the fact that it includes special cases such as cross-correlation and simple Gaussian models. It is not even clear why particular values of ξ would lead to useful or good measures of similarity. To become of more general interest, the choices for the parameter ξ, the concentration κ, the distortion and the class of functions f used in the maximization of the likelihood will need to be further addressed.

In some sense, the likelihood approach based on warping works too well: as shown in example 3, it is possible to specify the parameters and the warping to such an extent that haddock and whiting can

easily be distinguished. This is good, but how do we know when a haddock is no longer a haddock? This question is of interest in registration problems where we need to monitor whether change has occurred over time, for example. By choosing the parameters and the warping function well, the change is not likely to be detectable, since the warping function will attempt to obtain the best transformation between the original image and the new image or object.

Alf Linney (*University College London*)
In work which involves understanding changes in the human body brought about by such factors as growth, surgery, injury, illness and treatment or diet it is often necessary to compare sequential images, or to compare images to a reference or to calculate averages to understand group and individual patterns of change. For many years *ad hoc* methods have been devised, which although conveying some appreciation of changes in shape and size have not been statistically robust and have not been based on any methodology which would allow the derivation of probability distributions.

This paper provides a statistically robust method which meets the requirements for two-dimensional medical image comparisons for clinical monitoring and audit, and provides methods of warping which may be used as a preprocess to averaging or as a basis for comparisons both within a class or with a set of independent standards. The penalties generated in accordance with the rules associated with the degree of warp established here allow for probabilistic classification which should prove useful both in the differential diagnosis of individuals and in understanding the strength of relationships between group averages. The latter is likely to find application in the understanding of the genetically determined inheritance of body and facial shapes.

The fact that the algorithms developed in this paper may optimize the match of both edges and intensities within an image makes them particularly useful for dealing with images produced by computerized tomography scanners and magnetic resonance imaging systems, which are two of the most used imaging modalities in current medical imaging. This directly arises from the fact that the grey level intensities in these images relate directly to the physical properties of the body matter and reproducible scanner settings, unlike the case of photographic images where illumination is often less controlled.

At last it does appear that a robust statistical method has been developed for warping medical images, and I look forward very much to see the results of its application in the numerous areas in which it has significant relevance.

J. O. Ramsay (*McGill University, Montreal*) **and T. O. Ramsay** (*Statistics Canada, Ottawa*)
Dr Glasbey and Professor Mardia have used the two-dimensional Fourier transform (FT), with its clean separation of phase and amplitude parameters, to define a registration routine which permits the choice of a remarkably large and useful class of optimization criteria. Their idea of defining null sets within transformation spaces seems to us to be fundamental. We hope that the general availability of the FT will result in a wide range of registration applications taking advantage of their work.

Almost all images contain landmarks, usually in the form of points, open and closed curves; and a useful registration might want to make use of this extra information. A closed boundary outside which there is no information of interest is an especially important landmark; and, indeed, much of the interesting information in the interior may occur close to that boundary. In registering a functional magnetic resonance image (typically 1.5×10^6 voxels), for example, it is often known in advance that the interesting event is restricted to a small area within 4 mm of the surface of the cerebral cortex.

The need for a hybrid of spatial landmark- and intensity-based registration within a complicated boundary such as that of a fish or the coastline of the British Isles raises the question of what basis to use in the representation of an image and its deformation. Tensor product bases such as the two-dimensional FT or the commonly used spline bases have trouble dealing accurately with boundaries, both regular and irregular, and are not tied to landmarks. Also, too many basis functions may be allocated to regions where little or no fitting power is needed.

We are betting, instead, on finite element methods. Both of us have studied two-dimensional smoothing and registration problems (Ramsay, T., 2000; Ramsay, J., 2000; Malfait *et al.*, 2000; Ramsay, 1999) represented by triangular meshes readily constructed by available software tools such as MATLAB (MathWorks, 1995). Basis systems defined on these meshes, such as the Lagrange polynomial bases, permit any level of localized accuracy in representing either intensity or spatial structure. Smoothing and registration problems can be represented by systems of partial differential

equations, and these systems can be solved in only a few minutes on personal computers using standard sparse matrix algorithms. For example, the registration of 20 faces to a template by Ramsay (1999), using 3481 basis functions, took 51 s on a 280 MHz processor.

The basis system that one chooses to represent a problem imposes severe limitations on all that follows, and getting that choice right, even in one-dimensional problems, seems essential. A good choice should separate those parameters carrying amplitude or intensity information from those carrying phase or spatial information. The FT perhaps does not go sufficiently far, at least for some applications, and the co-ordinate-free finite element approach seems to offer a wider range of two-dimensional and three-dimensional image analysis options.

M. M. Rao (*University of California, Riverside*)

Dr Glasbey and Professor Mardia highlighted several important questions of practical interest in this paper on image and shape analysis. A typical problem here is one of nonparametric estimation of an image or an unknown function describing the (usually deformed) shape based on some (several) non-independent observations. An analogous question on template estimation, in computational anatomy, was raised by Professor Ulf Grenander a couple of years ago (as noted in Rao (2000)). The observations there come from a carefully constructed mapping, resulting in deformed images which are diffeomorphisms of a three-dimensional compact object (e.g. a brain), and it is desired to estimate it consistently by using a large set of observations, similar to the problems discussed here. One can consider a modification of the method of Kampé de Fériet and Frenkiel (1962) that depends on a certain averaging process. The diffeomorphism group D here is too large (it is not locally compact) and for this procedure we need the availability of an invariant integral. In Rao (2000), example IX.4.7, I outlined a method involving locally compact subgroups of D with the idea of extending it for a larger class using a projective limit process. The details of the latter have not yet been completely worked out. Another related problem based on the researches of Grenander and his associates has also been sketched in complement IX.6.4 of Rao (2000), giving a lower bound of the risk function of estimators. Some technical problems remain for a successful implementation of these ideas. It is noted that, generally, in shape analysis the problems to be solved are non-elementary as amply illustrated by Kendall *et al.* (1999).

Thus the required technical results lag far behind the applications. The authors have outlined some methods that illuminate the underlying theoretical set-up. They should be commended for bringing out the importance and practicality of the problems and for suggesting certain procedures that may be used now until the necessary theoretical basis has been developed, possibly in the near future.

Giovanni Sebastiani (*Istituto per le Applicazioni del Calcolo "M. Picone", Rome*)

I shall present here some general considerations about the null set distortion criteria adopted in this work. Given a non-negative regularization functional D with kernel $\ker(D)$ and such that $D(0) = 0$, a new non-negative functional $D_{\mathcal{C}}$ is built by minimizing the action of D on the difference between the functional argument and any element of the set $\mathcal{C}$. Any set $\mathcal{C} \supseteq \ker(D)$ closed with respect to addition can be chosen. The new functional is strongly related to D and will exploit its main features. Furthermore, since $\ker(D_{\mathcal{C}}) = \mathcal{C}$ the new functional is minimized by the elements of $\mathcal{C}$. A suitable choice of $\mathcal{C}$ which is meaningful for the problem under study may allow us to take advantage of this property. The procedure is also applicable to the sum of functionals D_i satisfying $\cap_i \ker(D_i) \subseteq \mathcal{C}$. The resulting functional will be minimized by the elements of $\mathcal{C}$.

In the paper it is shown for a particular D that $D_{\mathcal{C}} = D$ when $\mathcal{C} = \ker(D)$. From the text it seems that this always happens if we choose $\mathcal{C} = \ker(D)$. For the particular functional D chosen

$$D(f) = \sum_{i,j,k=1}^{2} D_1\{D_{jk}(f_i)\},$$

where $D_1(u) = \int_{\square} u^2 \mathrm{d}x$ and $D_{jk} = \partial^2/\partial x_j \partial x_k$ the result follows because $\mathcal{A} = \ker(D) = \mathcal{W} \times \mathcal{W}$, where

$$\mathcal{W} = \bigcap_{j,k=1}^{2} \ker(D_{jk}).$$

In fact, we have

$$D_{\mathcal{A}}(f) = \min_{g\in\mathcal{A}}\{D(f-g)\} = \min_{g\in\mathcal{A}}\left[\sum_{i,j,k=1}^{2} D_1\{D_{jk}(f_i) - D_{jk}(g_i)\}\right] = \min_{g\in\mathcal{A}}\left[\sum_{i,j,k=1}^{2} D_1\{D_{jk}(f_i)\}\right] = D(f),$$

because $g_i \in \mathcal{W}$. The result is not true in general for any functional D when we choose $\mathcal{C} = \ker(D)$. In fact, if we choose

$$D(f) = \sum_{i=1}^{2} \int_{\square} \left(\frac{\partial f_i}{\partial x_1}\right)^2 \left(\frac{\partial f_i}{\partial x_2}\right)^2 \mathrm{d}x$$

and we evaluate its action on the point $\tilde{f} = (x_1 + x_2,\ x_1 + x_2)$ we have $D(\tilde{f}) = 2n_1n_2$. The point $\tilde{g} = (x_2/2,\ x_2/2)$ belongs to $\mathcal{A} = \ker(D)$. Since we have $D_{\mathcal{A}}(\tilde{f}) \leqslant D(\tilde{f} - \tilde{g}) = n_1n_2/2 < D(\tilde{f})$, we cannot have $D_{\mathcal{A}} = D$.

As also stated in the paper, given D and $\mathcal{C} \supseteq \ker(D)$, different functionals can be built that exploit the main features of D and are minimized by the elements of $\mathcal{C}$. As an alternative to the $D_{\mathcal{C}}$ proposed, we can consider f as parameterized by $f = g + d$ with $g \in \mathcal{C}$ and define $D_{\mathcal{C}}(f) = D(d)$. If $f \in \ker(D_{\mathcal{C}})$ we have that $d \in \ker(D)$. Since $\ker(D) \subseteq \mathcal{C}$ it follows that $d \in \mathcal{C}$. Now, $f = g + d$ will belong to $\mathcal{C}$ because $g,\ d \in \mathcal{C}$ and $\mathcal{C}$ is closed with respect to addition. Therefore, $\ker(D_{\mathcal{C}}) \subseteq \mathcal{C}$. Let us assume now that $\mathcal{C}$ is also closed with respect to subtraction, as is the case for the set $\mathcal{S}$ of the paper. Given $f \in \mathcal{C}$, we can choose any $d \in \ker(D)$ and write $f = (f - d) + d$. Now, d will belong to $\mathcal{C}$ since $d \in \ker(D)$ and $\ker(D) \subseteq \mathcal{C}$. Since we assumed $\mathcal{C}$ to be closed with respect to subtraction, $f - d$ also belongs to $\mathcal{C}$. We have therefore written f as the sum of an element of $\mathcal{C}$ and an element of $\ker(D)$, so that $f \in \ker(D_{\mathcal{C}})$. This means that $\ker(D_{\mathcal{C}}) \supseteq \mathcal{C}$. Combining the two results we have that $\ker(D_{\mathcal{C}}) = \mathcal{C}$. I have no reason to prefer one of these two choices for $D_{\mathcal{C}}$ to the other apart from the larger computational complexity of the $D_{\mathcal{C}}$ proposed in the paper (which is given as the sum of D plus other terms). Further considerations can be taken into account, like the stability of the minimizers of $D_{\mathcal{C}}$.

Kevin de Souza (*University of Leeds*)

I wish to congratulate the authors on an enjoyable and stimulating paper. I would also like to draw attention to some unpublished joint work with J. T. Kent and K. V. Mardia. In the current paper deformations are parameterized at a single scale by specifying locations for a set of vertices on a fine rectangular grid. Further, a penalized likelihood objective function is maximized through an iterative algorithm in which the vertices of a fitted deformation are successively updated. In our work we found it fruitful to represent a deformation in terms of a composition of deformations on a hierarchical arrangement of triangular grids. Such a construction needs to be done carefully, but it allows a simple specification of large scale changes in the deformation through the adjustment of vertices at the coarser scales of the grid.

As in the current paper, this representation of a deformation can be incorporated into a penalized likelihood framework in which an objective function is iteratively improved. We used stochastic updates based on the Markov chain Monte Carlo algorithm rather than deterministic updates to avoid becoming trapped in local optima. The advantage of the hierarchical approach is that it allows faster movement through the space of possible deformations, thus allowing faster and more reliable optimization. Some limited experimentation has demonstrated the value of this approach, but further work is needed to understand its properties more fully.

Changming Sun and Michael Buckley (*Commonwealth Scientific and Industrial Research Organisation, Sydney*)

We congratulate the authors on an interesting and significant piece of work. We have one suggestion for potential application, as well as some specific comments.

A common aim, particularly in industrial quality assessment, is the measurement of sample characteristics in digital images. Two of many examples are the measurement of the distance between two features—e.g. between the rear of a gill and the start of the tail in fish—and the measurement of average (or 'typical') colour within a particular region—e.g. the body of a fish, excluding the head, tail and fins.

This is a difficult problem, especially if the features or regions concerned are difficult to characterize objectively—e.g. the 'start of the tail'. Image warping provides an interesting solution to such problems via a standard or 'average' object—in the examples above, an 'average fish'. After such a standard fish

has been obtained, points or regions of interest can be manually specified once and for all on the standard fish. Thereafter, when other fish of the same species are warped to match the standard, the feature points or regions in the new fish are immediately available, and measurements can be taken.

Some further specific points and questions are as follows.

(a) What would happen with these algorithms if, for example, one of the fish had its mouth open? Is this treated as local or global distortion?
(b) In Section 2.1, the authors claim that most of the information about the warping is contained in the phase rather than in the amplitude. This is certainly true in the case of translation. However, what evidence is there that this is true in more general warping?
(c) The authors mention that a multiresolution approach can be adopted to guard against becoming trapped in local suboptima. To make the multiresolution approach work, it has been found that image smoothing is required—more smoothing at lower resolutions (Moulin, 2000).
(d) The authors assert that parallelization would considerably reduce the computational cost of their algorithm, but it is not clear at which stage of the algorithm parallel processing could be applied.
(e) We believe that the transformation (29) used in the synthetic aperture radar example is not correctly called a 'projection'. Projection transformations contain a denominator term; see, for example, section 14.1 of Haralick and Shapiro (1993).
(f) These days colour images are very common. It would be interesting if the authors could comment on how the Fourier–von Mises criterion could be modified to apply effectively to colour image data.

D. M. Titterington (*University of Glasgow*)

The paper has presented an innovative approach to the oft-visited problem of image registration, the key new ideas being the Fourier–von Mises image model and the null set distortion criterion, which together create the quantity whose minimizer provides the solution to the problem. The results are impressive and the method seems to have considerable flexibility, although maybe the penalized likelihood interpretation is more difficult to justify than in other implementations of that paradigm; I mean here that the log-likelihood term in previous applications usually comes from a noise model that is arguably more plausible as a physical model than is the model underlying, say, expression (10). This is usually the case in the so-called Bayesian approaches to image analysis described famously in this journal by Besag (1986), but in that scenario one can in turn be sceptical about the contextual realism of the prior that underlies the corresponding penalty function. Maybe therefore the importance of the method is to be judged on a purely empirical basis, in which case the illustrations in the paper speak well for the approach. Of course, it will be important to make sure that the method competes effectively with the many existing methods for image registration, and I apologize for my laziness in not contributing to such empirical comparisons in this discussion. If the new approach were to hold sway, then I wonder whether either or both of the two new ideas can be transferred to other regularization contexts; I would be glad to hear the authors' thoughts about this. Secondary issues of importance in the method include the ubiquitous problem of choosing the regularization constant and the thorny problem of possible multiple local optima. Again, I feel that the former issue is less requiring of theoretical investigation here than it is in contexts such as spline smoothing or density estimation, but I do have concerns about the question of multiple optima, and I would be grateful for any further reassurance from the authors in the light of their practical experiences.

A. Trubuil (*Institut National de la Recherche Agronomique, Jouy-en-Josas*)

Dr Glasbey and Professor Mardia have given us an interesting and most useful paper. It collects for us many references on image warping and presents in a new and elegant way an effective approach to the non-rigid multimodal registration problem. Considering Fourier decomposition of the images, they build the likelihood on the phase variables. This has some drawbacks but also many advantages as explained in the paper. In brief, they use at most five parameters for the likelihood model and look for a non-rigid deformation in a space whose dimension can be chosen by the user. They can also control the deformation to forbid folds.

If we consider medical image registration of three-dimensional ultrasound images with magnetic resonance images, there is often an imbalance between parameters devoted to the likelihood and parameters devoted to deformation: with many parameters for the likelihood term and only few parameters for the deformation (e.g. rigid registration). Hence Roche *et al.* (2000) propose comparing

grey levels obtained from ultrasound with estimated values from a function of the vector of grey level and gradient of the magnetic resonance. This leads me to a question: could medical image multimodal registration benefit from the parsimony of parameters associated with the phase model proposed by the authors and estimate non-rigid deformation?

Incidentally, normalization using the variance of grey levels inside the template domain is a technique used in correlation criteria to avoid a match between only a few voxels of the test domain and the template domain. How is such a spurious solution eliminated using the criteria proposed by the authors?

A technique considered useful in medical image registration is the so-called partial volume interpolation (Maes *et al.*, 1997), which consists in interpolating between voxels at the criteria level. Hence, if x_k denotes a voxel in the test image domain and $f(x_k)$ the position in the template domain, instead of considering intensity interpolation

$$\left[I(x_k) - \Phi\left\{\sum_l w_{kl}\, J(y_l)\right\}\right]^2,$$

where $J(y_l)$ are the intensities of voxels in the neighbourhood of $f(x_k)$ and w_{kl} are weights dependent on the location of $f(x_k)$ with respect to voxels of this neighbourhood, we consider

$$\sum_l w_{kl}[I(x_k) - \Phi\{J(y_l)\}]^2.$$

It may be interesting to think about this technique also for the phase-related criteria.

With regard to applications, the approach presented by the authors is very interesting and could be considered in medical image registration and three-dimensional microscopy, unless computation time is a limitation.

K. J. Worsley (*McGill University, Montreal*)

There has been considerable interest in image warping in the brain mapping literature in recent years. The problem here is to align or register three-dimensional magnetic resonance images of the human brain to an atlas standard. The main reason for doing this is to compare regions of the brain 'activated' by a task such as a visual or cognitive stimulus measured by three-dimensional functional magnetic resonance imaging (FMRI) across different subjects (Lange and Zeger, 1997). Once the MRI images are registered to an atlas standard, the FMRI images can be deformed in the same way and then averaged to increase the signal-to-noise ratio. Furthermore, regions of high signal can be identified on the brain atlas (Collins *et al.*, 1995). The problem is made difficult by the fact that brain anatomy is never quite the same: sometimes the auditory cortex consists of two 'folds' or gyrii instead of one, so no reasonable warping can ever achieve a perfect match. On top of this, three-dimensional data are much more difficult to warp than two-dimensional data; typical data sets consist of a million voxels (three-dimensional pixels).

There have been two main approaches to this problem. The first is a 'brute force' approach in which blurred images are registered by penalized intensity matching; then the amount of blurring is gradually reduced until the desired resolution is achieved (Collins *et al.*, 1995). The second method is to parameterize the warping by a set of basis functions, usually cosine transform bases (Ashburner and Friston, 1999). Expanding the matching criterion as a linear function of the unknown coefficients results in a linear model that can be fitted by ridge regression. The maximum feasible number of basis functions is 8 per dimension, which, together with their products, gives $8^3 \times 3 = 1536$ unknown coefficients to be estimated, requiring the inversion of a 1536×1536 matrix. Thus the resolution is lower than that of the first method, but it is much faster, taking minutes rather than hours.

Finally, the warpings themselves can be used for the statistical analysis of shape. They are modelled as a trivariate Gaussian random field, and a three-dimensional field of Hotelling's T^2-statistics is used to detect localized shape differences between say groups of subjects (Cao and Worsley, 1999). Recent advances in the geometry of random fields have enabled us to set a threshold for Hotelling's T^2-field that can control the probability of detecting false positive shape changes in regions where no change has taken place to say 0.05.

Keming Yu (*University of Plymouth*)

The paper does a fine job in explaining the important topic of image warping to a larger statistical audience. The whole model can be expressed by means of equation (1) of the paper. I would like to

mention *two* alternative approaches, both of which may be applied to the *Fourier–von Mises image model* adopted by the authors.

Full Bayesian approach

As the paper mentions, the penalized log-likelihood approach may be justified by a Bayesian formulation. Moreover, the penalized log-likelihood approach corresponds to the mode of a posterior density defined by means of a *partially improper* prior related to a Brownian motion or Wiener process. Estimating parameter λ is very important in this approach. However, if we adopt a full Bayesian approach with prior related to the distortion criteria to estimate the warping function f by the posterior mean, we can effectively integrate out λ. No matter how complicated the likelihood function, an advanced Markov chain Monte Carlo method or Gibbs sampling can be used to do the full Bayesian approach in spite of possibly heavy computation.

Kernel smoothing

The choice of warp is a compromise between a two-dimensional smooth distortion and one which achieves a good match. This paper reminds me about a challenging problem: how to introduce kernel smoothing techniques for image warping or how to smooth the noise out of image data with kernels while achieving a good match at the same time. Why would the authors prefer their penalized log-likelihood approach over such an approach? Clearly, the local average with image warping is important in the presence of local distortions, so that a kernel smoothing method has a potential application here. Although standard kernel smoothing techniques such as Nadaraya–Watson-type estimation may blur some unsmooth features such as edges, spikes and jumps, the *kernel-weighted log-likelihood* should be applicable in image warping. For this, the proposed *Fourier–von Mises model* in the paper is just equivalent to identifying a new log-likelihood function which is suitable for some image warping problems such as problem 2 in the paper, whereas a simple *Gaussian* likelihood is reasonable for problem 3 of the paper.

The **authors** replied later, in writing, as follows.

We are pleased by the number and diversity of the contributions to the discussion, which come from both statistical and computer vision communities and cover a spectrum from theoretical issues to additional applications. We address the topics raised in the order in which they appear in the paper.

Penalized likelihood approach

Several discussants suggest a Bayesian formulation (Jennison, Dryden, Ghosh and Murthy, and Yu). This would facilitate a quantification of uncertainty in the estimators and possibly simplify the choice of λ, though at the price of greater computational effort. In the paper we focused on point estimation and were concerned about computing time, so we did no more than to point out the opportunities for a Bayesian approach. However, we are not opposed to it: in our applications $\exp\{-\lambda D(f, \mathcal{C})\}$ looks to be a reasonable measure of prior belief in f, and it would be relatively straightforward to embed our ideas within current Bayesian methodology. In reply to Professor Jennison's question, it should also be possible to incorporate a stochastic model for μ, which would be beneficial in the fish application if there were fewer or noisier images.

Both Professor Molchanov and Professor Rao raise issues regarding the theoretical underpinning of image warping and image averaging, and Professor Hancock provides useful links to work in computer vision. We hope that our paper encourages our more theoretical colleagues to study these important problems further. We also thank Professor Angulo for his suggestion to combine L and D non-additively, and Dr Yu for the idea of using kernel smoothing.

With regard to the choice of λ, we agree with Dr Berman that cross-validation is not ideal, though it seems to give reasonable results in the synthetic aperture radar application. His suggested alternative, minimum description length coding, would be considerably more complicated to implement, we think. In response to Professor Ghosh and Professor Murthy, our results are insensitive to the choice of λ to within an order of magnitude, according to Tables 1, 4 and 6.

Fourier–von Mises image model

We thank Professor Dryden for generalizing our result on the distribution of $\theta^{(Y_f)}$, conditional on $A^{(Y_f)}$, when Y_f is a Gaussian random field. As regards the reverse problem, of deriving the distribution of Y_f from that of $\theta^{(Y_f)}$, this is unlikely to have a tractable form. Also, it would be necessary to specify a joint distribution for $\theta^{(Y_f)}$ and $A^{(Y_f)}$ in order for Y_f to have full degrees of freedom.

Dr Ersbøll asks how our method copes with images at very different resolutions. In principle there is no problem, as either one image is interpolated or the other is subsampled. In practice, this is probably best done in the Fourier domain. In reply to Dr Sun and Dr Buckley, our experience is that phases carry most information about images because sinusoids at different frequencies need to combine to produce edges. Empirical evidence is given in Glasbey and Horgan (1995), Figs 3.6(c) and 3.6(d): when the phases of one image are combined with the amplitudes of another image, the result looks much more similar to the first one. Dr Koch questions the choice of ξ. Although the functional form of κ in equation (7) is somewhat arbitrary, we would expect the maximum likelihood estimator $\hat{\xi}$ to give the best matching criterion, as borne out in Table 3. We agree with Dr Trubuil, that the Fourier–von Mises model could provide a parsimonious representation of differences between medical images.

In reply to Professor Jennison's point, we sum over x rather than $f(x)$ in equations (4) and (10), because we can then obtain analytic expressions for the average images, given by equations (27) and (28), and because some aspects of the algorithm are simplified. We note that Ramsay and Li (1998) did likewise, and we have a similar philosophy to them, of regarding Y as a single entity rather than as an array of individual observations. Such a strategy is natural for functional data analysis, and the choice of summand is then somewhat arbitrary.

We are interested in Professor Molchanov's suggestion to explore the use of the Fourier–von Mises likelihood as the basis for a grey scale image metric. We agree with Professor Hancock and Professor Ramsay and Dr Ramsay that the inclusion of local features in the likelihood criterion will be effective in some applications. However, we do not wholly share Dr Coleman's view on expert knowledge. Humans are almost always better than computers at image analysis, but that expertise can be exceedingly difficult to encode in computer algorithms in general and it can be better to develop automatic methods independently.

Null set distortion criteria

We thank Professor Hancock and Dr Sebastiani for the suggestion to use other base distortion criteria than D_{B_1} and D_{B_2} to derive alternative $D(f, \mathcal{C})$ with the same null set property but with other features inherited from D_B. Further, Professor Angulo and Dr Sebastiani propose alternatives to equation (14) for constructing $D(f, \mathcal{C})$ from D_B. Also, we agree with Dr Sebastiani's point and had not meant to imply that $D_C = D_B$ whenever $\mathcal{C} = \ker(D_B)$. Dr Ashburner and Professor Petrou propose yet other null set distortion criteria. In particular, criterion

$$\frac{1+|J|}{2}\{\log(s_1)^2 + \log(s_2)^2\},$$

where J is the Jacobian of an affine transformation and s_1 and s_2 are its singular values (Ashburner *et al.*, 1999), is minimized per unit area by the family of translations and rotations

$$\mathcal{R} = \{g\colon g_1 = \alpha_1 + x_1\cos(\theta) + x_2\sin(\theta),\ \ g_2 = \alpha_2 - x_1\sin(\theta) + x_2\cos(\theta)\}.$$

In comparison, our method yields

$$D(f, \mathcal{R}) = D_{B_1}(f) + 2n_1n_2 - 4n_1n_2\sqrt{(\tilde{\alpha}_{11}^2 + \tilde{\alpha}_{12}^2)},$$

using the same notation as in equation (21), and is also rotationally invariant. If $g \in \mathcal{A}$, an affine transformation given by equation (18),

$$D(g, \mathcal{R}) = n_1n_2[\alpha_{11}^2 + \alpha_{12}^2 + \alpha_{21}^2 + \alpha_{22}^2 + 2 - 2\sqrt{\{(\alpha_{11} + \alpha_{22})^2 + (\alpha_{12} - \alpha_{21})^2\}}],$$

which can be re-expressed as

$$D(g, \mathcal{R}) = n_1n_2\{s_1^2 + s_2^2 + 2 - 2\sqrt{(s_1^2 + s_2^2 + 2\,|J|)}\}.$$

Thus our criterion also depends on $|J|$ and a symmetric function of s_1 and s_2, and has the benefit of having been derived under a unified approach.

We are impressed by Mr Gustafsson and Professor Rudemo's work with electrophoresis gels, and we welcome the combining of our methods with realistic physicochemical models in specific applications. We only partly support Dr Gray's wish that $\mathcal{C}$ be obtainable empirically from data. In particular, in our three applications we have prior knowledge which, in our opinion, should be given overriding consideration. However, this could be an interesting area for further work, developing on Professor

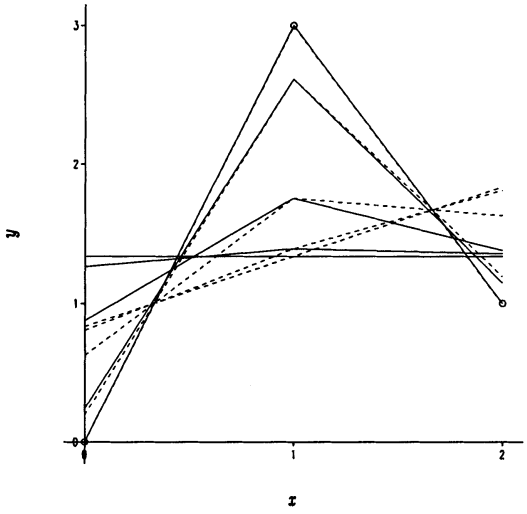


Fig. 16. Illustration of nonparametric regression: ○, data; ———, $\hat{f}$ obtained by minimizing expression (40); - - - - -, $\hat{f}$ obtained by minimizing expression (41) (in both cases for a range of values of $\lambda \rightarrow 0$, 0.1, 1, 10, ∞, from top to bottom)

Hogg's ideas for using training data and addressing also Dr Koch's concerns regarding how to choose an appropriate distortion criterion in new applications.

It would seem that the contrast in results between the two students in using D_{B_1} and $D(f, \mathcal{S})$, to which Professor Kent refers, is that one case involves interpolation and the other smoothing and extrapolation. Our approach also involves smoothing, and we find that the two distortion criteria give different results. Consider a simplified one-dimensional case, where we estimate f by either minimizing

$$\sum_{x=0}^{n} \{Y_x - f(x)\}^2 + \lambda \int_0^n \left(\frac{\partial f}{\partial x}\right)^2 \mathrm{d}x, \tag{40}$$

using the one-dimensional version of D_{B_1}, or

$$\sum_{x=0}^{n} \{Y_x - f(x)\}^2 + \lambda \left\{\int_0^n \left(\frac{\partial f}{\partial x}\right)^2 \mathrm{d}x - \frac{1}{n}\left(\int_0^n \frac{\partial f}{\partial x}\,\mathrm{d}x\right)^2\right\}, \tag{41}$$

using the one-dimensional version of $D(f, \mathcal{S})$. In particular, expression (40) is the formulation of a linear spline. In both cases, the optimal f is piecewise linear, with $\widehat{f(x)}$ obtained by standard methods. Fig. 16 illustrates the two sets of results for $Y^T = (0, 3, 1)$ for a range of values of λ. We see that, except when $\lambda \rightarrow 0$ and $\widehat{f(x)} \rightarrow Y_x$, the two sets of results are distinct. As $\lambda \rightarrow \infty$, $\hat{f} \rightarrow \bar{Y}$ in case (40) and the best-fitting straight line in case (41). We thank Professor Kent for raising this interesting point.

We apologize to Professor Jennison, for not making it clear that our numerical approximation for $D(f, \mathcal{A})$ includes contributions from the edges of grid squares, so that, for example

$$\int_{\square} \frac{\partial^2 f}{\partial x_1^2}\mathrm{d}x \approx \frac{q_1^3 n_2}{n_1^3 q_2} \sum_{k=(1,0)}^{q-(1,0)} (\beta_k^{+0} - \beta_{k-(1,0)}^{+0})^2.$$

Optimization algorithm

Professor Hancock advocates the use of EM-type optimization algorithms. At an earlier stage in the work we used a form of alternating algorithm (Glasbey and Mardia, 1998), switching between estimating β and ξ, before opting for the more elegant conjugate gradient algorithm. In fact, the former algorithm was computationally faster, a topic of concern to some discussants (Berman and Gray),

whereas others suggest Markov chain Monte Carlo and other stochastic methods which would be even more computationally intensive. We are interested in Professor Ramsay and Dr Ramsay's use of finite element methods and Dr de Souza's use of a composition of functions. In reply to Dr Berman's question, we used a stopping rule to choose q, halving the grid size until it made no difference to the estimated warp. Pathological solutions, where an image maps to a single point on μ, can arise in certain formulations, as Dr Trubuil points out, but we did not encounter them.

Dr Sun and Dr Buckley raise several computational points. We agree that it makes sense to apply a smoothing filter to images when working at a coarser resolution, to reduce the risk of being trapped in a local optimum, a concern also expressed by Professor Titterington. This we did, and we were remiss in not mentioning it in Section 2.3. As regards the potential for parallelization: because effects on P of changes in β are locally computable, terms in $\partial P/\partial\beta$ can be computed simultaneously.

Applications

Several discussants question how the algorithm will perform if there are differences between images, such as movement, occlusion, growth or physiological differences (Ashburner, Ersbøll, Ghosh and Murthy, Sun and Buckley, and Worsley). The term λD is set by the user to constrain the warp so that a distinction is made between minor differences, which are accommodated by the warp, and major ones, which are not. In particular cases, where certain differences are to be expected, tailor-made methods may perform better. For example, there are specific algorithms for interpreting stereoscopic pairs (e.g. Weng *et al.* (1993)), where occlusions occur frequently and can be a source of additional information. We agree with Dr Coleman that it is desirable, wherever possible, to capture images in a way that avoids warping, but objects will always differ inherently in spite of imaging technology, such as fish even of the same species. Professor Dryden makes an interesting distinction between different images of the same object and images of different objects. However, we think that the same warping methodology is appropriate in both cases, though the interpretation may be different.

Another question raised is how to handle applications where only subsets of images are informative such as where backgrounds vary (Berman) or all information is within a boundary (Horgan and Ramsay and Ramsay). If differences in the background are other than minimal, then image warping, which treats the whole image as equally informative, is not the appropriate methodology. Instead, it would be more appropriate to isolate the regions of interest, perhaps by matching templates to certain features. In answer to Dr Horgan's and Professor Dryden's questions, we have not been concerned with the relative contributions of differences in boundaries and textures to fish discrimination, as we see it as a strength of our approach that all differences are synthesized into a single criterion.

Dr Ersbøll and Dr Sun and Dr Buckley ask about extensions to the methodology to colour and multispectral images. This should be possible, the main complication we expect being the necessity to use a multivariate von Mises distribution, as it would probably be unrealistic to assume that phase differences from different variates at a common frequency were independent. Image warping in three dimensions is common practice (see, for example, the contributions of Professor Petrou and Professor Worsley), and we see this as technically straightforward though computationally intensive. However, it is possible that extra complexities may arise in applications involving shape.

We thank Dr Sun and Dr Buckley for pointing out that equation (29) is not a *perspective* projection. Rather, it is a *parallel* projection, which is an asymptotic approximation requiring three fewer parameters. In response to Professor Titterington, we are not aware of extensions of our ideas to other regularization problems, though it may be possible to include null set distortion criteria in image deconvolution.

Finally, we thank the discussants for bringing to our attention a diverse range of other applications: radar (Hancock), fundus images of the retina (Ersbøll), parsnips (Horgan), hand outlines and X-ray computer tomography cardiac images (Duta and Jain), electrophoretograms (Gustafsson and Rudemo) and magnetic resonance imaging (Worsley and Linney). Much remains to be done, both theoretically and in applications, in this challenging area.

References in the discussion

Andresen, P. R., Bookstein, F. L., Conradsen, K., Ersbøll, B. K., Marsh, J. L. and Kreiborg, S. (2000) Surface-bounded growth modeling applied to human mandibles. *IEEE Trans. Med. Imagng*, **19**, 1053–1063.

Ashburner, J., Andersson, J. and Friston, K. J. (1999) High-dimensional nonlinear image registration using symmetric priors. *NeuroImage*, **9**, 619–628.

Ashburner, J. and Friston, K. J. (1999) Nonlinear spatial normalisation using basis functions. *Hum. Brain Map.*, **7**, 254–266.
Baddeley, A. J. (1992) Errors in binary images and an L^p version of the Hausdorff metric. *Nieuw Arch. Wisk.*, **10**, 157–183.
Berman, M. (1994) Automated smoothing of image and other regularly spaced data. *IEEE Trans. Pattn Anal. Mach. Intell.*, **16**, 460–468.
Besag, J. (1986) On the statistical analysis of dirty pictures (with discussion). *J. R. Statist. Soc.* B, **48**, 259–302.
Bose, S. and O'Sullivan, F. (1997) A region-based image segmentation method for multi-channel data. *J. Am. Statist. Ass.*, **92**, 92–106.
Cao, J. and Worsley, K. J. (1999) The detection of local shape changes via the geometry of Hotelling's T^2 fields. *Ann. Statist.*, **27**, 925–942.
Carstensen, J. M. (1996) An active lattice model in a Bayesian framework. *Comput. Vis. Image Understndng*, **63**, 380–387.
Collins, D. L., Holmes, C. J., Peters, T. M. and Evans, A. C. (1995) Automatic 3-D model-based neuroanatomical segmentation. *Hum. Brain Map.*, **3**, 190–208.
Conradsen, K. and Pedersen, J. (1992) Analysis of two-dimensional electrophoretic gels. *Biometrics*, **48**, 1273–1287.
Craw, I., Costen, N., Kato, T. and Akamatsu, S. (1999) How should we represent faces for automatic recognition? *IEEE Trans. Pattn Anal. Mach. Intell.*, **21**, 725–736.
Cross, A. D. J. and Hancock, E. R. (1998) Graph matching with a dual step EM algorithm. *IEEE Trans. Pattn Anal. Mach. Intell.*, **20**, 1236–1253.
Dryden, I. L. and Walker, G. (1999) Highly resistant regression and object matching. *Biometrics*, **55**, 820–825.
Duta, N., Jain, A. K. and Jolly, M. P. (1999) Learning-based object detection in cardiac MR images. In *Proc. Int. Conf. Computer Vision, Corfu*, pp. 1210–1216. Institute of Electrical and Electronics Engineers Computer Society Press.
Friel, N. (1999) Application of random sets to image analysis. *PhD Thesis*. University of Glasgow, Glasgow.
Friel, N. and Molchanov, I. S. (1998) Distances between grey-scale images. In *Mathematical Morphology and Its Applications to Image and Signal Processing* (eds H. J. A. M. Heijmans and J. B. T. M. Roerdink), pp. 283–290. Dordrecht: Kluwer.
Geman, D. and Reynolds, G. (1992) Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattn Anal. Mach. Intell.*, **14**, 367–383.
Glasbey, C. A. and Horgan, G. W. (1995) *Image Analysis for the Biological Sciences*. Chichester: Wiley.
Glasbey, C. A. and Mardia, K. V. (1998) Statistical aspects of image warping. *Technical Report STAT98/15*. Department of Statistics, University of Leeds, Leeds.
Godwin, D. J. (2000) Deformations in shape analysis. *PhD Thesis*. University of Leeds, Leeds.
Granlund, G. H. and Knutsson, H. (1995) *Signal Processing for Computer Vision*. Dordrecht: Kluwer.
Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.
Hansell, D. M., Coleman, R., Du Bois, R. M., Carr, D. H., Goodman, L. R., Kerr, I. H., Pearson, M. C. and Rubens, M. B. (1991) Advanced Multiple Beam Equalisation Radiography (AMBER) in the detection of diffuse lung disease. *Clin. Radiol.*, **44**, 227–231.
Haralick, R. and Shapiro, L. (1993) *Computer and Robot Vision*, vol. II. Reading: Addison-Wesley.
Horgan, G. W., Talbot, M. and Davey, J. (2001) Use of statistical image analysis to discriminate carrot cultivars. *Comput. Electr. Agric.*, **31**, 191–199.
Hurn, M. A. and Jennison, C. (1995) A study of simulated annealing and a revised cascade algorithm for image reconstruction. *Statist. Comput.*, **5**, 175–190.
Jain, A. K. and Duta, N. (1999) Deformable matching of hand shapes for user verification. In *Proc. Int. Conf. Image Processing, Kobe*. Institute of Electrical and Electronics Engineers Computer Society Press.
Kadyrov, A. and Petrou, M. (2000) The trace transform and its applications. *IEEE Trans. Pattn Anal. Mach. Intell.*, to be published.
Kaijser, T. (1998) Computing the Kantorovich distance for images. *J. Math. Imagng Vis.*, **9**, 173–191.
Kampé de Fériet, J. and Frenkiel, F. N. (1962) Correlations and spectra for nonstationary random functions. *Math. Comput.*, **16**, 1–21.
Kendall, D. G., Barden, D., Carne, T. K. and Le, H. (1999) *Shape and Shape Theory*. Chichester: Wiley.
Kovalev, V. A. and Petrou, M. (1998) Non-rigid volume registration of medical images. *J. Comput. Inform. Technol.*, **6**, 181–190.
Lange, N. and Zeger, S. L. (1997) Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging (with discussion). *Appl. Statist.*, **46**, 1–29.
Larsen, R., Conradsen, K. and Ersbøll, B. K. (1998) Estimation of dense image flow fields in fluids. *IEEE Trans. Geosci. Remote Sensng*, **36**, 256–264.
Lee, T. C. M. (2000) A minimum description length-based image segmentation procedure, and its comparison with a cross-validation-based segmentation procedure. *J. Am. Statist. Ass.*, **95**, 259–270.
Maes, F., Collington, A., Vermeulen, D., Marchal, G. and Suetens, P. (1997) Multimodality image registration by maximization of mutual information. *IEEE Trans. Med. Imagng*, **16**, 187–198.

Malfait, N., Ramsay, J. O. and Froda, S. (2000) The historical functional linear model. McGill University, Montreal.
MathWorks (1995) *Partial Differential Equation Toolbox User's Guide*. Natick: MathWorks.
Moss, S. and Hancock, E. R. (1997) Registering incomplete radar images with the EM algorithm. *Image Vis. Comput.*, **15**, 637–648.
Moulin, P. (2000) Multiscale image decomposition and wavelets. In *Handbook of Image and Video Processing* (ed. A. Bovick), ch. 4.2, pp. 289–300. New York: Academic Press.
Myers, R., Wilson, R. C. and Hancock, E. R. (2000) Bayesian graph edit distance. *IEEE Trans. Pattn Anal. Mach. Intell.*, **22**, 628–635.
Ramsay, J. O. (2000) Differential equation models for statistical functions. *Can. J. Statist.*, **28**, 225–240.
Ramsay, J. O. and Li, X. (1998) Curve registration. *J. R. Statist. Soc.* B, **60**, 351–363.
Ramsay, T. O. (1999) A bivariate finite element smoothing spline applied to image registration. *Doctoral Dissertation*. Queen's University, Kingston.
——— (2000) Spline smoothing over difficult regions. Submitted to *J. R. Statist. Soc.* B.
Rao, M. M. (2000) *Stochastic Processes: Inference Theory*. Dordrecht: Kluwer.
Roche, A., Pennec, X., Rudolph, M., Auer, D. P., Malandain, G., Ourselin, S., Auer, L. M. and Ayache, N. (2000) Generalized correlation ratio for rigid registration of 3D ultrasound with MR images. In *Proc. 3rd Int. Conf. Medical Robotics, Imaging and Computer Assisted Surgery, Pittsburgh, Oct. 11th–14th* (eds A. DiGioia and S. Delp).
Serra, J. (1998) Hausdorff distances and interpolations. In *Mathematical Morphology and Its Applications to Image and Signal Processing* (eds H. J. A. M. Heijmans and J. B. T. M. Roerdink), pp. 107–114. Dordrecht: Kluwer.
Sibson, R. and Thomson, G. D. (1981) A seamed quadratic element for contouring. *Comput. J.*, **24**, 378–382.
Skorohod, A. V. (1956) Limit theorems for stochastic processes. *Theory Probab. Applic.*, **1**, 261–290.
Stoyan, D. and Molchanov, I. S. (1997) Set-valued means of random particles. *J. Math. Imgng Vis.*, **7**, 111–121.
Voss, T. and Haberl, P. (2000) Observations on the reproducibility and matching efficiency of two-dimensional electrophoresis gels: consequences for comprehensive data analysis. *Electrophoresis*, **21**, 3345–3350.
Walker, G. (2000) Robust, non-parametric and automatic methods for matching spatial point patterns. *PhD Thesis*. University of Leeds, Leeds.
Wang, K. and Gasser, T. (1999) Synchronising sample curves nonparametrically. *Ann. Statist.*, **27**, 439–460.
Weng, J., Ahuja, N. and Huang, T. S. (1993) Optimal motion and structure estimation. *IEEE Trans. Pattn Anal. Mach. Intell.*, **15**, 864–884.
Wilson, D. L., Baddeley, A. J. and Owens, R. A. (1997) A new metric for grey-scale image comparison. *Int. J. Comput. Vis.*, **24**, 5–17.
Wilson, R. C. and Hancock, E. R. (1997) Structural matching by discrete relaxation. *IEEE Trans. Pattn Anal. Mach. Intell.*, **19**, 634–648.