Principal Warps: Thin-Plate Splines and the Decomposition of Deformations

FRED L. BOOKSTEIN

Abstract—One conventional tool for interpolating surfaces over scattered data, the *thin-plate spline*, has an elegant algebra expressing the dependence of the physical bending energy of a thin metal plate on point constraints. For interpolation of a surface over a fixed set of nodes in the plane, the bending energy is a quadratic form in the heights assigned to the surface. The spline is the superposition of eigenvectors of the bending energy matrix, of successively larger physical scales, over a tilted flat plane having no bending energy at all.

When these splines are paired, one representing the x-coordinate of another form and the other the y-coordinate, they aid greatly in the modeling of biological shape change as *deformation*. In this context, the pair becomes an interpolation map from R^2 to R^2 relating two sets of *landmark points*. The spline maps decompose, in the same way as the spline surfaces, into a linear part (an affine transformation) together with the superposition of *principal warps*, which are geometrically independent, affine-free deformations of progressively smaller geometrical scales. The warps decompose an empirical deformation into orthogonal features more or less as a conventional orthogonal functional analysis decomposes the single scene. This paper demonstrates the decomposition of deformations by principal warps, extends the method to deal with curving edges between landmarks, relates this formalism to other applications of splines current in computer vision, and indicates how they might aid in the extraction of features for analysis, comparison, and diagnosis of biological and medical images.

Index Terms—Affine transformations, biharmonic equation, biomedical image analysis, deformation, principal warps, quadratic variation, shape, thin-plate splines, warping.

I. The Thin-Plate Spline as an Interpolant

A. The Function $U(r)$

THIS paper proposes an algebraic approach to the description of deformations specified by finitely many point-correspondences in an irregular spacing. At the root of the analysis is the special function sketched in Fig. 1. This is the surface

$$z(x, y) = -U(r) = -r^2 \log r^2,$$

where r is the distance $\sqrt{x^2 + y^2}$ from the Cartesian origin. The minus sign is for ease of reading the form of this surface: in this pose, it appears to be a slightly dented but otherwise convex surface viewed from above. The surface incorporates the point (0, 0, 0), as marked by the X in the figure. Also, the function is zero along the indicated circle, where $r = 1$. The maximum of the surface is achieved all along a circle of radius $1/\sqrt{e} \sim 0.607$ concentric with the circle of radius 1 that is drawn.

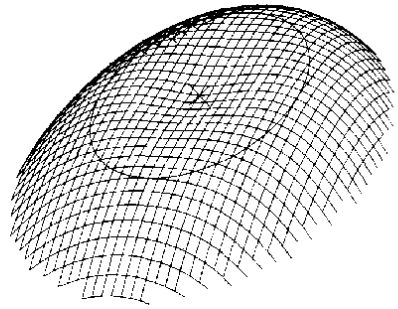


Fig. 1. Fundamental solution of the biharmonic equation: a circular fragment of the surface $z(x, y) = -r^2 \log r^2$ viewed from above. The X is at (0, 0, 0); the remaining zeros of the function are on the circle of radius 1 drawn.

The function $U(r)$ satisfies the equation

$$\Delta^2 U = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^2 U \propto \delta_{(0,0)}.$$

The right-hand side of this expression is proportional to the "generalized function" $\delta_{(0,0)}$ zero everywhere except at the origin but having an integral equal to 1. That is, U is a so-called *fundamental solution* of the *biharmonic equation* $\Delta^2 U = 0$, the equation for the shape of a thin steel plate lofted as a function $z(x, y)$ above the (x, y)—plane. This basis function is the natural generalization to two dimensions of the function $|x|^3$ that underlies the familiar one-dimensional cubic spline.

B. Bounded Linear Combinations of Terms $U(r)$

Fig. 2 is a mathematical model of a thin steel plate which should be imagined as extending to infinity in all directions. Passing through the plate is a rigid armature in the form of a square of side $\sqrt{2}$, drawn in perspective view as the rhombus at the center of the figure. The steel plate is tacked (fixed in position) some distance above two diagonally opposite corners of the square, and the same distance below the other two corners of the square. In the figure, this tacking is indicated by the X's, which are to be taken as lying exactly upon the steel sheet but also as rigidly welded, via their "stalks," to the corresponding corners of the underlying square.

Manuscript received July 17, 1987; revised August 2, 1988. Recommended for acceptance by W. E. L. Grimson. This work was supported in part by the National Institutes of Health under Grant GM-37251.

The author is with the Center for Human Growth and Development, University of Michigan, Ann Arbor, MI 48109.

IEEE Log Number 8927505.

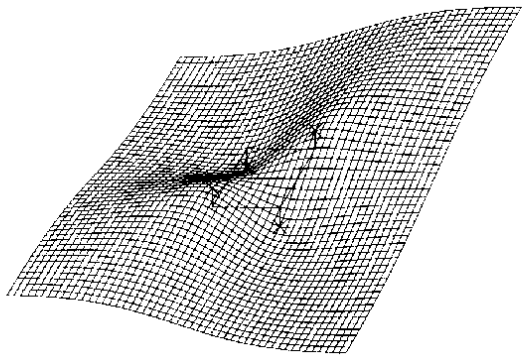


Fig. 2. Part of an infinite thin metal plate constrained to lie at some distance above a ground plane at points $(0, \pm 1)$ and the same distance below it at points $(\pm 1, 0)$. The rhombus at the center represents the rigid square armature enforcing these constraints by fixing the positions of the X's, which all lie on the surface above the corners of the square. Far from the armature, the height of the plate approaches a multiple of the cosine of the double central angle.

The surface in Fig. 2 corresponds to a multiple of the function

$$\begin{aligned} z(x, y) &= U\left(\sqrt{x^2 + [y - 1]^2}\right) \\ &\quad - U\left(\sqrt{[x + 1]^2 + y^2}\right) \\ &\quad + U\left(\sqrt{x^2 + [y + 1]^2}\right) \\ &\quad - U\left(\sqrt{[x - 1]^2 + y^2}\right) \\ &= \sum_{k=1}^{4} (-1)^k U\left(\left|(x, y) - D_k\right|\right) \end{aligned}$$

where the D_k are the corners $(1, 0)$, $(0, 1)$, $(-1, 0)$, $(0, -1)$ of the square. The functions $U(r)$ are taken with coefficients $+1$ for the ends of one diagonal, -1 for the ends of the other. It can be shown that this function $z(x, y)$ is, indeed, the solution of the biharmonic equation $\Delta^2 z = 0$ consistent with the tacking of a previously flat infinite plate to points alternately above and below the corners of the square as shown. The physical steel takes this form, as long as the displacements are small, because the function $z(x, y)$ is the configuration of lowest physical bending energy consistent with the given constraints. For a thin plate subjected to only slight bending, the bending energy at a point is proportional to the quantity $(\partial^2 z/\partial x^2)^2 + 2(\partial^2 z/\partial x \partial y)^2 + (\partial^2 z/\partial y^2)^2$ at that point, and $z(x, y) = \Sigma(-1)^k U(|(x, y) - D_k|)$ minimizes

$$\iint_{R^2} \left(\left(\frac{\partial^2 z}{\partial x^2}\right)^2 + 2\left(\frac{\partial^2 z}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 z}{\partial y^2}\right)^2 \right) dx\, dy$$

over the class of all functions z taking the values $(-1)^k$ at D_k, as drawn. As a physical model this idealization incorporates several assumptions, such as zero energy cost for in-plane deformations and the absence of gravity, which do not concern us here.

As one travels far away from the origin, this plate is asymptotically flat and level in all directions. In Fig. 2, for instance, the corner of the plate facing the viewer in the diagram has apparently become nearly level somewhat underneath the level of the constraint at the nearest corner of the armature, and likewise the other three corners. Algebraically, we have, by adding and subtracting $2U(\sqrt{x^2 + y^2 + 1})$ in the definition of $z(x, y)$,

$$\begin{aligned} z(x, y) &= \Sigma(-1)^k U\left(\left|(x, y) - D_k\right|\right) \\ &= V\left(x^2 + [y + 1]^2\right) - 2V(x^2 + y^2 + 1) \\ &\quad + V\left(x^2 + [y - 1]^2\right) \\ &\quad - \left\{V\left([x + 1]^2 + y^2\right) - 2V(x^2 + y^2 + 1)\right. \\ &\quad \left. + V\left([x - 1]^2 + y^2\right)\right\} \end{aligned}$$

for $V(s) = U(\sqrt{s})$. In this we recognize two copies of the familiar approximation to the second derivative of a function,

$$d^2 f/ds^2 \sim \frac{f(s + h) - 2f(s) + f(s - h)}{h^2}$$

for h equal to $2y$ in the top line, $2x$ in the bottom line. Then

$$z(x, y) \sim (2y)^2 \frac{d^2 V(s)}{ds^2} - (2x)^2 \frac{d^2 V(s)}{ds^2}$$

pertaining to the function $V(s) = s \log s$ evaluated at $s = x^2 + y^2 + 1$. Thus $z(x, y)$ reduces to $4(y^2 - x^2)/x^2 + y^2 + 1$ together with terms that drop to zero as $1/r$ or faster. Except for the term $+1$ in the denominator, this value is just -4 times the cosine of the double angle $2 \tan^{-1}(y/x)$. Thus, a long way from $(0, 0)$ our metal sheet takes the form of a very slowly rising and falling circuit of the armature, *of bounded variation*: 4 units above the armature at points far out along one diagonal, 4 units below the armature at points far out along the other.

C. Displacements in the Coordinate Plane: The Thin-Plate Spline as an Interpolant

In Fig. 2, the displacement of the thin plate lies in a direction orthogonal to the lie of the plate itself. This orthogonality is not necessary. (Of course, we are no longer modeling a physical plate: that applied only for bending of small extent normal to the coordinate plane of x and y.) We may imagine the displacements $z(x, y)$ to be applied directly to one or both of the coordinates x or y of the plate with which we started.

Thus we may interpret the scheme of Fig. 2 as the *interpolation function* shown in Fig. 3. The four points begin in the form of a square; then one diagonal is displaced with respect to the other diagonal until there results the form of a kite, right. Over the square on the left there is superimposed a grid of points so that we can visualize the effect of this transformation on the elements of area surrounding the X's. The x-coordinate is transferred from

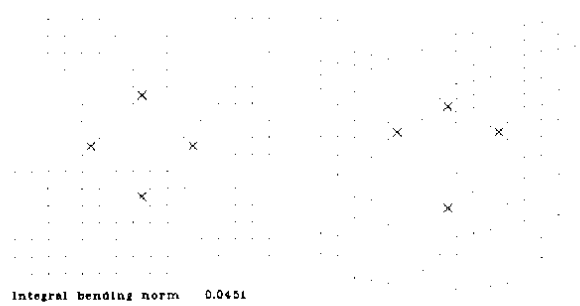


Fig. 3. A multiple of the same function added to the y-coordinate of points of a square grid rather than lofted as a z-coordinate. There results an interpolation function between a square and a kite. Another interpretation of this same landmark reconfiguration is relative translation of the two diagonals of the square.

left to right without change, while the y-coordinate is altered by the value $z(x, y)$ which was the z-coordinate of the metal sheet in Fig. 2. We thus arrive at the mapping function

$$(x, y) \rightarrow (x', y') = (x, y + z(x, y))$$

where $z(x, y)$ is the same function $\Sigma(-1)^k U(|(x, y) - D_k|)$ we have been viewing in three dimensions. (The limitation of the z-adjustment to the y-coordinate is only for purposes of this particular example, and will be lifted presently.)

In this manner the thin-plate spline we have been examining can be used to solve a two-dimensional interpolation problem, the computation of a map $\boldsymbol{R}^2 \rightarrow \boldsymbol{R}^2$ from arbitrary data. In this special case, it represents the mapping consistent with the assigned correspondence of X's and adapted to their reconfiguration in the manner which *uniquely* minimizes a certain sort of "bending energy," namely, the (linearized) energy which would have been required had the landmark displacements in question been normal to the plane of the figure rather than within that plane.

The splined mapping is defined everywhere in the plane of the picture, and is differentiable everywhere in the picture; it is a diffeomorphism as long as it does not fold. (Folds do not arise in the applications intended here, dealing with realistic biological data.) In this it contrasts with other familiar global maps, notably the projection and the bilinear mapping [4], [12], both of which are singular along entire lines in all nontrivial (inhomogeneous) applications. In comparison to interpolants produced by inverse distance-weighting methods (cf. [13]), the spline is flat at infinity, but nevertheless does not reduce to a constant there. The asymptotic cosine of two cycles is clearly visible in Fig. 3 as the vertical shifting of the corners of the grid alternately upward and downward.

We know that if physical steel sheets are merely *tilted*, changed from level to oblique, they need not bend: in tilting, energy does work against gravity, not against elasticity. To maintain the analogy between displacements of the plate normal to its plane and displacements of points in their plane, transformations of the landmarks which can be assigned the same effect as "tilting" should have no "bending energy." If we also allow the general increase or decrease of geometric scale (as by rerolling the physical plate), the class of these transformations is coextensive with the *affine transformations* or homogeneous shears, those which leave parallel lines parallel. Because the plate of Fig. 2 has no linear part at infinity, its energetics are independent of any net shearing of the configuration (tilting of the plate), such as a change from square to rectangle, before applying the alternating displacements. This separation of nonlinear part from linear by behavior at infinity will be pursued further in the next section. Lifting and tilting aside, to bend the plate requires energy; and the sharper the bending, the greater the second derivatives of the surface $z(x, y)$ and the greater the energy required. A pattern of "tacks" differing by a given height requires much less energy to install when they are far apart than when they are close together.

This observation hints at a very useful application of the splines to the problem of localizing the information content of plane deformations. In the next section I review the algebra of these splines for more general configurations of points than squares, and then indicate how to interpret the eigenvectors of a bending-energy matrix as serving for the analogous decomposition in a certain large class of plane deformations. These eigenvectors will become the *principal warps* of a given point-configuration; they represent features of deformation at distinct geometrical scales.

Because the pictures which I measure are primarily biomedical, I will refer to the data points X specifying these deformations, analogous to the points where the physical plate is tacked, as *landmarks*. To the biologist, landmarks are points in one form for which objectively meaningful and reproducible biological counterparts exist in all the other forms of a data set [9]. The landmarks of one form are said to be *homologous* to the corresponding landmarks in other forms. In other applications these same points may be called registration points, fiducials, and the like.

D. Algebra of the Thin-Plate Spline for Arbitrary Sets of Landmarks

One can imagine the steel plate of Fig. 2 to be fixed in position arbitrarily high or low above the base plane at any combination of points, not just the corners of a square as shown. Subject to whatever constraints are posed, the plate will still adopt the position of least net bending energy, and the description of its form will still be a linear combination of terms $r^2 \log r^2$ (Fig. 1), fundamental solutions of the biharmonic equation, centered at each point where information (here, height) is specified. The exploitation of the thin-plate equation to provide interpolatory splines in this way seems to have been originated by Duchon [11]—"le principe des plaques minces"—and was later formalized by Meinguet [15]–[17] in a very general mathematical setting (see also [20]).

The present application of thin-plate splines should be

distinguished from earlier applications to computer vision of *real* surfaces, the problem corresponding to Fig. 2 rather than Fig. 3. For instance, Terzopoulos' [19] approach to stereopsis realizes the sampled surface as that which minimizes a certain energy functional, incorporating bending energy as one of its terms, over a finite region. In this setting, no analytic solution is available; Terzopoulos offers a hierarchical finite-element method incorporating a local iterative algorithm. In a context of deformation, the splines have appeared previously in a study of "signal matching" by [21]. They consider the general case of continuous "signals" in one or more dimensions. The algorithm they propose minimizes, again, a generalization of the bending-energy used here. Neither of these approaches seems to support any equivalent of the very finite spectrum of principal warps to be introduced presently.

The remainder of this section comprises a terse overview of the algebraic crux of the thin-plate method. Let $P_1 = (x_1, y_1)$, $P_2 = (x_2, y_2)$, $\cdots$, $P_n = (x_n, y_n)$ be n points in the ordinary Euclidean plane according to any convenient Cartesian coordinate system. We are concerned with functions f taking specified values at the points P_i; should certain pairs or triples of P's be closely adjacent, the effect is that of specifying derivatives of f as well as values. Write $r_{ij} = |P_i - P_j|$ for the distance between points i and j.

Define matrices

$$K = \begin{bmatrix} 0 & U(r_{12}) & \cdots & U(r_{1n}) \\ U(r_{21}) & 0 & \cdots & U(r_{2n}) \\ \cdots & \cdots & \cdots & \cdots \\ U(r_{n1}) & U(r_{n2}) & \cdots & 0 \end{bmatrix}, \; n \times n;$$

$$P = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \cdots & \cdots & \cdots \\ 1 & x_n & y_n \end{bmatrix}, \; 3 \times n;$$

and

$$L = \left[\begin{array}{c|c} K & P \\ \hline P^T & O \end{array}\right], \; (n+3) \times (n+3),$$

where ^{T} is the matrix transpose operator and O is a 3×3 matrix of zeros.

Let $V = (v_1, \cdots, v_n)$ be any n-vector, and write $Y = (V|0\ 0\ 0)^T$, a column vector of length $n + 3$. Define the vector $W = (w_1, \cdots, w_n)$ and the coefficients a_1, a_x, a_y by the equation

$$L^{-1}Y = (W|a_1 \quad a_x \quad a_y)^T.$$

Use the elements of $L^{-1}Y$ to define a function $f(x, y)$ everywhere in the plane:

$$f(x, y) = a_1 + a_x x + a_y y + \sum_{i=1}^{n} w_i U(|P_i - (x, y)|).$$

The role of the last three rows of L is to guarantee that the coefficients w_i sum to zero and that their crossproducts with the x- and y-coordinates of the points P_i are likewise zero. The function f is divided into two parts: a sum of functions $U(r)$ which can be shown to be bounded and asymptotically flat, just like the example in Fig. 2, and an affine part representing the behavior of f at infinity.

Then the following three propositions hold:

1) $f(x_i, y_i) = v_i$, all i. (This is just a restatement of the equations represented by the first n rows of L, those not involved in regularizing the function at infinity.)

2) The function f minimizes the nonnegative quantity

$$I_f = \int\int_{R^2} \left(\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x \, \partial y} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right) dx \, dy$$

over the class of such interpolants. Call this the "integral quadratic variation" or the "integral bending norm."

3) The value of I_f is proportional to

$$WKW^T = V(L_n^{-t} K L_n^{-1}) V^T,$$

where L_n^{-1} is the upper left $n \times n$ subblock of L^{-1}. This integral is zero only when all the components of W are zero: in this case, the computed spline is $f(x, y) = a_1 + a_x x + a_y y$, a flat surface.

In the present application we take the points (x_i, y_i) to be landmarks and V to be the $n \times 2$ matrix

$$V = \begin{bmatrix} x_1' & x_2' & \cdots & x_n' \\ y_1' & y_2' & \cdots & y_n' \end{bmatrix}$$

where each (x_i', y_i') is the landmark homologous to (x_i, y_i) in another copy of $\boldsymbol{R}^2$. The application of L^{-1} to the first column of V^T specifies the coefficients of 1, x, y, and the U's for $f_x(x, y)$, the x-coordinate of the image of (x, y); the application of L^{-1} to the second column of V^T does the same for the y-coordinate $f_y(x, y)$.

The resulting function $f(x, y) = [f_x(x, y), f_y(x, y)]$ is now vector-valued: it maps each point (x_i, y_i) to its homolog (x_i', y_i') and is least bent (according to the measure I_f, integral quadratic variation over all $\boldsymbol{R}^2$, computed separately for real and imaginary parts of f and summed) of all such functions. These vector-valued functions $f(x, y)$ are the *thin-plate spline mappings* of this paper. If the pairing of points between the sets is in accordance with biological homology, the function f models the comparison of biological forms as a *deformation*, as suggested by D'Arcy Thompson in 1917. For a review of the history of this idea in quantitative biology, see [3].

The whole procedure is invariant under translation or rotation of either set of landmarks. Invariance under (x, y)-rotation follows from the fact that the minimand, bending energy, is a scalar. As for (x', y')-rotation, because $L^{-1}(Y_x \cos\theta + Y_y \sin\theta) = W_x \cos\theta + W_y \sin\theta$, etc., the effect of rotating the $2 \times n$ matrix V is to rotate the fitted spline by the same angle.

E. Recomputation of Fig. 2

The reader may find it useful to follow through the algebra of these multiple matrices for the symmetric case set forth geometrically in Fig. 2. We have, for instance, $U(r_{12}) = U(\sqrt{2}) = 2 \log 2 = 1.3863$, etc., resulting in the matrix

$$K = \begin{bmatrix} 0.0 & 1.3863 & 5.5452 & 1.3863 \\ 1.3863 & 0.0 & 1.3863 & 5.5452 \\ 5.5452 & 1.3863 & 0.0 & 1.3863 \\ 1.3863 & 5.5452 & 1.3863 & 0.0 \end{bmatrix}.$$

Note that $5.5452 = 4 \times 1.3863$. The matrix P of 1's and point coordinates is

$$P = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix},$$

combining with K to give a matrix

$$L = \left[\begin{array}{cccc|ccc} 0.0 & 1.3863 & 5.5452 & 1.3863 & 1 & 0 & 1 \\ 1.3863 & 0.0 & 1.3863 & 5.5452 & 1 & -1 & 0 \\ 5.5452 & 1.3863 & 0.0 & 1.3863 & 1 & 0 & -1 \\ 1.3863 & 5.5452 & 1.3863 & 0.0 & 1 & 1 & 0 \\ \hline 1 & 1 & 1 & 1 & & & \\ 0 & -1 & 0 & 1 & & O & \\ 1 & 0 & -1 & 0 & & & \end{array}\right].$$

The matrix V of target point coordinates, augmented, is

$$V = \begin{bmatrix} 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0.75 & 0.25 & -1.25 & 0.25 & 0 & 0 & 0 \end{bmatrix}.$$

The vectors $L^{-1}V^T$ of coefficients W, a_1, a_x, a_y are

$$(0, 0, 0, 0, 0, 1, 0)^T$$

and

$$(-0.0902, 0.0902, -0.0902, 0.0902, 0, 0, 1)^T$$

corresponding to the two rows of V. The first set of coefficients specify the formula for the x-coordinate of the image of (x, y), the second set, those for the y-coordinate.

The meaning of these vectors is as follows. The first corresponds to the function $f_x(x, y) = x$—the identity mapping for the x-coordinate. Indeed, there are no changes of x-coordinate between the left and right configurations of landmarks, and so all of the terms U have coefficients equal to zero. The function $f_y(x, y)$ is a multiple of the expression $\Sigma(-1)^k U(|(x, y) - D_k|)$ with which we are already familiar, together with an affine part equal to the single term y. The terms that are unbounded at infinity, the terms linear in x and y, are the identity mapping. This is because the transformation was generated by the rigid translation of one diagonal with respect to the other, without rotation or change of lengths.

II. Principal Warps as Eigenvectors of $L_n^{-1}KL_n^{-1}$

The matrix $L_n^{-1}KL_n^{-1}$ of the preceding example is very highly patterned. Its numerical value is

$$\begin{bmatrix} 0.09 & -.09 & 0.09 & -.09 \\ -.09 & 0.09 & -.09 & 0.09 \\ 0.09 & -.09 & 0.09 & -.09 \\ -.09 & 0.09 & -.09 & 0.09 \end{bmatrix}.$$

Thus it is of rank one, proportional to the dyadic product $(1, -1, 1, -1)^T(1, -1, 1, -1)$. Patterns of displacement of landmarks of the form $(1, -1, 1, -1)$—equal and opposite translations applied to the two diagonals, or, what amounts to the same thing, translations of one with respect to the other—are the only patterns of displacement *not* annihilated by the matrix $L_n^{-1}KL_n^{-1}$ in this example. All the annihilated patterns of displacement are in fact *affine*, corresponding to adjustments of the constant and the term $a_x x$, $a_y y$ in f which are not bounded at infinity. For instance, the pattern of y-displacements $(1, 0, -1, 0)$ represents an expansion of lengths along the y-axis; the same pattern applied horizontally (i.e., to x-displacements) is a rotation of the y-axis toward the x-axis, transformation of square into rectangle, again representable as a shear.

Thus for the starting configuration of a square (and, in fact, for any starting configuration of four points), *there is only one degree of freedom for nonlinearity of the interpolation*. Any biharmonic interpolation over the corners of a square is the combination of some affine transformation—square to parallelogram—with displacement according to some multiple of Fig. 2 in some direction upon the page. In all nondegenerate cases, this single dimension of terms in the U's can be represented as a displacement of any single landmark holding the others in fixed position. [In the present example, we might imagine an affine transformation that multiplies y-coordinates by 1.5; the transformation we are studying arises by moving the uppermost landmark 1 unit downward after such a shear. Or we might shear the starting square into the parallelogram $(0, -1.5)$, $(1, 0)$, $(0, 0.5)$, $(-1, -1)$, then displace this last corner separately (nonlinearly) to $(-1, 0)$.] In general, *we cannot localize nonlinearity in the four-point interpolation*, any more than we can "localize" the affine part of the transformation that takes parallel lines to parallel lines. It is the same everywhere in the plane: it has no local features.

A. Principal Warps and the Spectrum of the Bending-Energy Matrix

The degeneracy of this nonlinear term vanishes as soon as we add any sort of realistic complexity to the interpolation problem posed. Fig. 4 shows five X's on the left

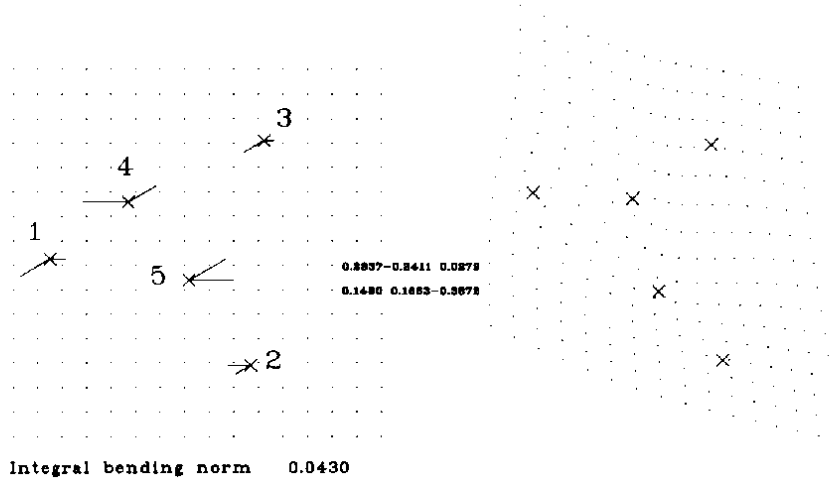


Fig. 4. Thin-plate spline interpolating the correspondence of five pairs of landmarks. The eigenvalues of the bending-energy matrix are listed in the leftmost column of the incorporated table. The coefficients of its eigenvectors are drawn as signed segments out of the landmarks every 30° counterclockwise beginning horizontally. The second and third columns of the tables are the projections of the splines f_x and f_y, the x- and y-components of the computed deformation, upon the two principal warps. The only aspect of this spline map not explicitly shown in the diagram is the affine component (its behavior at infinity), which is discussed in the text.

and five corresponding X's in a somewhat different configuration on the right. The form has narrowed considerably from upper right to lower left, not so much from upper left to lower right; and bends have appeared in both "bars," previously nearly straight, of the original T-shaped configuration. To dissect the features of this and other point-driven deformations efficiently and objectively, we will exploit the eigenstructure of the matrix $L_n^{-1}KL_n^{-1}$. This eigenstructure is coded in Fig. 4 with little lines in the left-hand scene and with certain numbers in the center of the figure.

The points on the left, together with the functions $U(r_{ij})$ describing their spatial relations, are encoded in the partitioned matrix

$$L = \left[\begin{array}{ccccc|ccc} 0.0 & 25.4713 & 31.2510 & 1.2938 & 5.8093 & 1 & 3.6929 & 10.3819 \\ 25.4713 & 0.0 & 24.9811 & 18.8511 & 1.9394 & 1 & 6.5827 & 8.8386 \\ 31.2510 & 24.9811 & 0.0 & 7.0360 & 8.6023 & 1 & 6.7756 & 12.0866 \\ 1.2938 & 18.8511 & 7.0360 & 0.0 & 1.4673 & 1 & 4.8189 & 11.2047 \\ 5.8093 & 1.9394 & 8.6023 & 1.4673 & 0.0 & 1 & 5.6969 & 10.0748 \\ \hline & & & (\text{sym}) & & & O & \end{array}\right]$$

for the ordering of landmarks indicated in Fig. 4. The matrix of landmark coordinates in the right-hand form is

$$V = \begin{bmatrix} 3.9724 & 6.6969 & 6.5394 & 5.4016 & 5.7756 \\ 6.5354 & 4.1181 & 7.2362 & 6.4528 & 5.1142 \end{bmatrix}.$$

The vectors $L^{-1}V^T$ of coefficients W and a are

$$(-0.0380,\ 0.0232,\ -0.0248,\ 0.0798,\ -0.0402;$$
$$1.3552,\ 0.8747,\ -0.0289)^T$$

and

$$(0.0425,\ 0.0159,\ 0.0288,\ -0.0454,\ -0.0418;$$
$$-2.9458,\ -0.2956,\ 0.9216)^T.$$

Let us deal first with the affine part of this map, the function

$$(x',\ y') = (1.3552 + 0.8747x - 0.0289y,$$
$$-2.9458 - 0.2956x + 0.9216y).$$

The constant terms merely refer to a shift between the two images (already corrected in Fig. 4); the linear terms may be collected in a matrix

$$A = \begin{bmatrix} 0.8747 & -.0289 \\ -.2956 & 0.9216 \end{bmatrix}.$$

The singular decomposition of this matrix is

$$A = O_{-53.34^\circ} D_{1.0749,0.7441} O_{44.89^\circ},$$

where the O's are rotation matrices by the angles indicated, and D is a diagonal matrix of singular values. This is to say that A, operating on the left, is extension by a factor 1.0719 in a direction 44.89° clockwise of horizontal, and compression by 0.7441 in the perpendicular direction, followed by a rotation of another 8.45° clockwise. The factor 1.0719 implies some elongation of the form toward the northwest: for instance, the distance from landmark 2 to the midpoint of landmarks 1 and 3 is longer in the right-hand form than in the left. The factor 0.7441, which is almost identical with the ratio of decrease in the distance between landmarks 1 and 3, confirms the compression from southwest to northeast.

Turning now to the remaining terms $\Sigma w_i U(|P_i - (x, y)|)$, we may begin to make sense of these by drawing them out, after the fashion of Fig. 2, as surfaces in their own right. That for the x-coordinate—call it the displacement f_x—is shown in Fig. 5(a); the displacement f_y for the y-coordinate is shown in Fig. 5(b). These pictures should not be considered to represent any particular orthogonal projection; the dropping of the affine part of the mapping is equivalent to not knowing "which way was up" in this style of diagram. The X's, as before, represent the points of the surface which are fixed in position; but now there is no way to draw the armature to which to weld them—no "horizontal." Nevertheless, the nature of the bending of the coordinates separately is immediately clear, as is the contrast between them. The nonlinearity in the x-displacement appears to be concentrated at landmark 4; that for the y-displacement appears to be a larger-scale depression (displacement downward in Fig. 4) of the whole middle of the figure, encompassing both landmarks 4 and 5.

The matrix $L_n^{-1}KL_n^{-1}$ of bending energy as a function of changes in the coordinates of the landmarks on the right is, for this example,

$$\begin{bmatrix} 0.0493 & -.0023 & 0.0329 & -.0744 & -.0055 \\ -.0023 & 0.0389 & -.0004 & 0.0439 & 0.0801 \\ 0.0329 & -.0004 & 0.0219 & -.0485 & -.0059 \\ -.0744 & 0.0439 & -.0485 & 0.1546 & -.0756 \\ -.0055 & -.0801 & -.0059 & -.0756 & 0.1671 \end{bmatrix}.$$

This matrix has three zero eigenvalues, corresponding to patterns of landmark displacement that result in affine transformations, and two nonzero eigenvalues, 0.2837 and 0.1480. The eigenvector corresponding to 0.2837 is

$$(0.2152, -0.3265, 0.1346, -0.6554, 0.6320)^T;$$

that for 0.1480 is

$$(-0.4941, -0.2415, -0.3370, 0.4700, 0.6026)^T.$$

The bending-energy matrix is computed as it applies to patterns of Cartesian displacement of the landmarks in the right-hand frame; but because the components of these

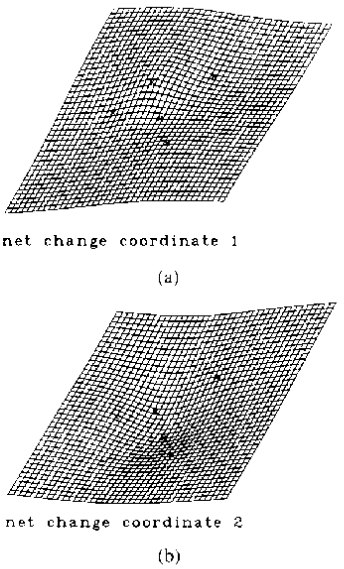
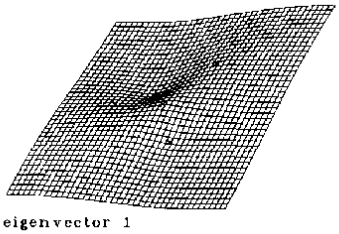


Fig. 5. Display of the affine-free parts of the preceding spline as thin plates. (a) x-coordinate; (b) y-coordinate. The small X's represent the specifications of height of the plates corresponding to displacements from left to right in Fig. 4. As the affine component of this transformation has been deleted, no horizontal "armature" underlying these X's can be drawn.

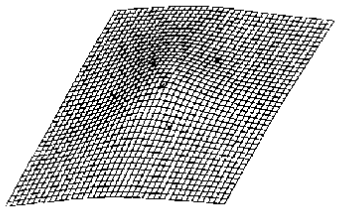
eigenvectors are coefficients for the five functions U based at these five landmarks, we may interpret each as the coefficients of a thin-plate spline of its own, attached to a base plane at the left-hand landmarks and flat at infinity. These functions $f_{0.2837}$, $f_{0.1480}$ are displayed in Figs. 6(a) and (b) as lofted into the third dimension—true thin-plate splines—after the fashion of Fig. 2.

The surface in Fig. 6(a) appears to be more bent than that in Fig. 6(b). To be precise, it requires more bending energy, 0.2837 versus 0.1480, for the same net amplitude of vertical displacements (measured as the sum of squares of coordinate changes at the landmarks). As is plain in the figure, the landmarks whose contrasting changes drive the spline in Fig. 6(a) lie closer together, and so the splined surface must change its slopes at higher rates, thus increasing the quadratic variation which, integrated out to infinity, corresponds to the eigenvalue of 0.2837 reported. That these f's are eigenvectors implies that the coefficients are identical with the displacements they afford landmark by landmark. Then we may draw them as well as displacements superimposed over the landmarks in their own plane. In Fig. 4, these loadings are shown as little segments attached to the X's themselves. The loadings of the first eigenvector run left or right, those of the second, at 30° counterclockwise of these. The segments are signed, so that positive and negative coefficients run in opposite directions out of the landmark X. For instance, the pattern of loadings of the eigenvector $f_{0.2837}$—absolutely largest, but with opposed signs, at landmarks 4 and 5—is visualized in Fig. 4 by the opposed horizontal segments attached to those central points.

Thus the eigenvectors of the bending energy matrix $L_n^{-1}KL_n^{-1}$, interpreted as deformations, are a canonical de-



(a)



(b)

Fig. 6. Display of the principal warps of the configuration at the left in Fig. 4 as thin plates of their own. (a) The principal warp of eigenvalue 0.2837; compare Fig. 5(a). (b) The warp of eigenvalue 0.1480: compare Fig. 5(b).

scription of the modes according to which points are displaced irrespective of global affine transformations. I call them the *principal warps* of the configuration of landmarks on the left-hand side. They are computed as features of bending at *successively higher levels of bending energy*; by the identification of bending energy with second derivatives of in-plane displacement, they correspond to features of deformation at *successively smaller physical scales*. This is plain in Fig. 6: the first eigenvalue corresponds to a relatively *small* feature—differences in the displacements of the two nearest points of the form, landmarks 4 and 5; the last nonzero eigenvalue corresponds to a relatively *large*, but still not global (affine) feature—the deviation of landmarks 4 and 5, together, from the average displacement of the three landmarks at the outside "corners." As an affine transformation would move the center of this triangle in accordance with the displacement of those corners, this pair of principal warps may be considered a simple sum-and-difference transformation of the original basis for displacement space (at landmarks 4 and 5 separately). But the *difference* between 4 and 5 has a higher eigenvalue (greater stiffness) than the *sum* of 4 and 5, as it represents a feature at smaller geometric scale. By its dependence on the matrix K of quantities $U(r)$, functions of adjacency, the principal warps of the spline are inextricable from the geometry of the landmark configuration itself.

To this point the analysis involves only the starting configuration of landmarks; it would be the same whatever the positions of their homologs in the right-hand frame in Fig. 4. Those positions, of course, affect the coefficients of f_x and f_y. We have already analyzed the affine part of those functions. The rest of the information about this deformation is encoded in the forms of Figs. 5(a) and (b), and its features are expressed in the relationships between Fig. 5 and Fig. 6. Plainly Fig. 5(a), the surface f_x of affine-free x-displacement, resembles an inverted version of Fig. 6(a), the first principal warp, and Fig. 5(b), the surface f_y of affine-free y-displacement, resembles an inverted version of Fig. 6(b), the second principal warp. This suggests that we expand the functions f_x, f_y of the actual thin-plate deformation (Fig. 4) in terms of the principal warps: we have, in fact,

$$f_x = -0.2411 f_{0.2837} + 0.1663 f_{0.1480},$$

and

$$f_y = 0.0279 f_{0.2837} - 0.3872 f_{0.1480}.$$

These coefficients combine with the eigenvalues to make up the integral bending:

$$0.2837(.2411^2 + .0279^2) + 0.1480(.1663^2 + .3872^2) = 0.0430.$$

Note that the x-displacements are spatially more concentrated, emphasizing the discrepancy between the displacements of landmarks 4 and 5. (The upper one has moved considerably to the right between the frames, the lower one not so much.) The y-displacement emphasizes instead the displacement of both central points downward relative to the remaining landmarks—the bending of the previously straight bar (landmarks 1-4-3) of the T.

These coefficients, which are printed adjacent to the corresponding eigenvalues in Fig. 4, represent the decomposition of deformations I am recommending. Each principal warp is a geometrically independent mode of affine-free deformation at its own geometrical scale (which may be taken as the inverse of its eigenvalue, its "bending energy"). For configurations of three landmarks, all transformations can be modeled as affine, and there are no principal warps. For four landmarks, there is only one warp, the single eigenvector shown for the case of a starting form which is square in Figs. 2 and 3. For more than four landmarks, the bending-energy matrix has a nontrivial spectrum which is of great practical interest.

For describing deformations, this spectrum serves a role analogous to that of the more familiar orthogonal decompositions of single pictures or outline forms. In the single picture, higher terms of an orthogonal decomposition represent features of progressively smaller scale. Likewise, the higher terms of the bending-energy spectrum represent aspects of deformation of progressively smaller scale: specifications of warping more and more local. Whenever homologous landmarks or other data permit the use of the deformation model, the general shape comparison can be described and measured more efficiently by these features of deformation than by comparing features measured upon the forms separately in accordance with *a priori* intuitions. Information about deformation of an image is separable in principle from information about content of the image as it is deformed. The ordinary orthogonal decompositions of separate images, such as the Karhunen–Loéve

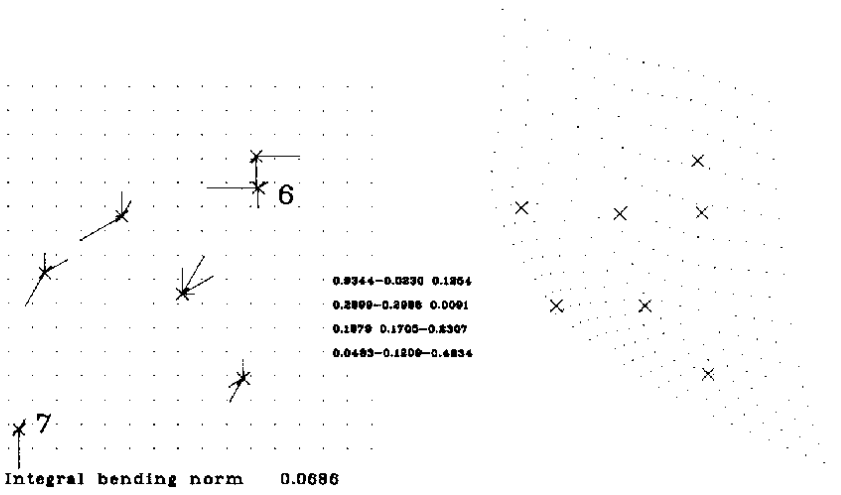


Fig. 7. Augmentation of Fig. 4 by two additional landmarks, one near an existing landmark and the other far away. Note the wide range of the spectrum.

or the Fourier, do not respect this separation, even for images as abstracted as the labeled points or continuous curves that underlie these splines. The principal warps of this presentation are orthogonal components of the *deformation*, not of the image. The landmarks, in their pairing, sample the deformation; they do not necessarily represent features of either image separately. The principal warps, then, are a basis for the representation of shape *change*; they become useful for the representation of shapes that are all deformations of one another or of a suitable primitive capable of this sort of paired pointwise labeling. Their efficiency for this purpose is due to their finite spectrum of spatially separated features. We will return to this matter in Sections IV and V.

B. Effect of Variations in Landmark Spacing

We can learn more about the behavior of these spline interpolants by extending the previous example with an additional pair of landmarks chosen to demonstrate two extremes of landmark spacing. In Fig. 7, we have added a sixth landmark rather near landmark 3 at upper right, and a seventh landmark quite distant from all the others, at lower left.

As is indicated in the table within the figure, the spectrum of the bending energy matrix now ranges twenty-fold, from 0.0483 to 0.9344. The eigenvector of highest bending energy, sketched at the left with horizontal segments, is mainly a contrast between displacements of landmarks 3 and 6, those closest together (at upper right), together with a small weight for the previously central landmark 4, closest of the others to this pair. Inspection of this eigenvector as a thin plate, Fig. 8(a), indicates that its principal feature is a slope at upper right, limited to the region of landmarks 3 and 6. This principal warp specifies mainly the *vertical directional derivative* of the interpolating spline f in that vicinity. As tabulated in Fig. 7, the large loading of the y-displacement upon this eigenvector corresponds to the considerable discord between the separation of these two landmarks on the *right* in the figure and the separation implied by the spline based on using one of these points, together with the other five, but not the other. That is, the map with the vertical directional derivative constrained at landmark 3 (or 6), as shown, is highly bent in that vicinity. The pair of points at upper right is highly informative about bending energy—it is a small-scale feature of the deformation—and, in the instance, there is considerable information at this scale in the y-component of the deformation observed.

The second eigenvector of this landmark configuration, Fig. 8(b), is quite similar to the first eigenvector of the five-point analysis, Fig. 5(a). It is mainly a contrast of displacements between landmark 4 and landmark 5. Similarly, the third eigenvector of the seven-point configuration, Fig. 8(c), is quite comparable to the second eigenvector of the five-point configuration, representing a central "peak" of its thin plate somewhat broader than the crimp shown in Fig. 8(b) for the second eigenvalue—joint displacement of landmarks 4 and 5 with respect to those surrounding.

Of larger geometrical scale (lower bending energy) than any other deformation is the last principal warp, shown in Fig. 8(d). It is, in fact, the same transformation as the simple nonlinear warp of a square: compare Fig. 8(d) to Fig. 2. The form of this surface is clear as well in Fig. 7, where it is coded in the signed lengths of the vertical segments out of the landmarks. In this approximately square configuration, the extremes of one "diagonal" are displaced downward, while all other points, lying upon the opposite diagonal, are displaced upward. This is the only eigenvector to which landmark 7 contributes to any extent. The large projection of f_y on this gentlest principal warp owes to the massive upward translation of this landmark between left and right configurations.

The displays of affine-free net displacement (Fig. 9(a),

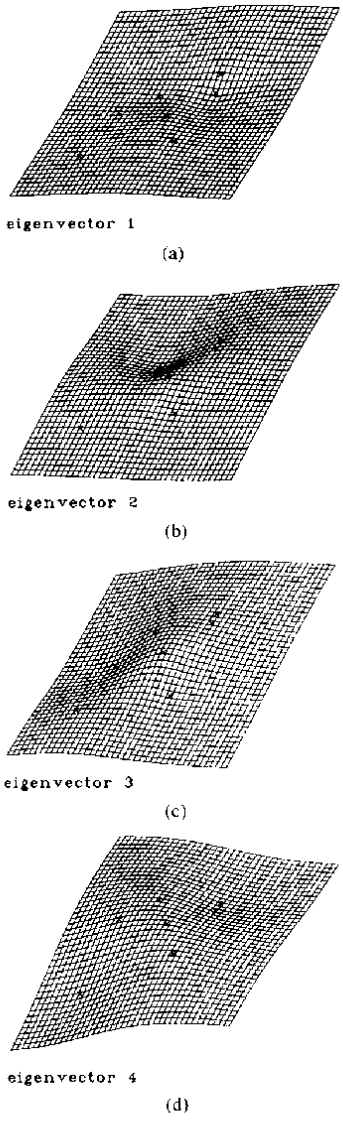


Fig. 8. Thin plates [(a)–(d)] representing the four principal warps of the deformation in the previous figure. The principal warp of largest eigenvalue, panel (a), denotes the clamping of a derivative; the thin plate for the last principal warp, panel (d), resembles the simple four-constraint spline in Fig. 2.

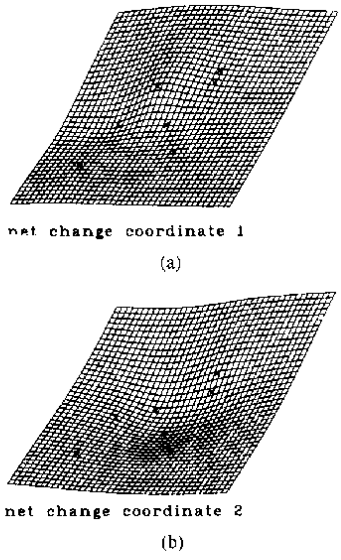


Fig. 9. Affine-free parts of the transformation in Fig. 7. (a) x-coordinate; (b) y-coordinate.

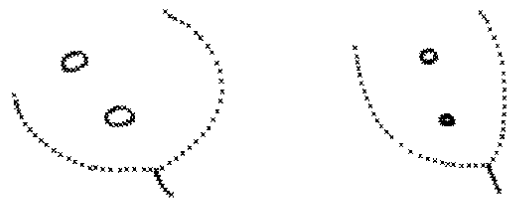


Fig. 10. The pair of "flowers" from which the landmarks of Fig. 4 were taken. The landmarks are the trifurcation of the "stem," the ends of the "petals," and the tops of the "seeds." The size of the X's is intended to suggest the precision with which these edges were located. Figs. 10–13 and 16–21 are after [7] and are used by permission.

x-coordinate; Fig. 9(b), y-coordinate) confirm that, as tabulated in Fig. 7, the y-transformation is a superposition of principal warps 1, 3, and 4 at roughly equal energies ($0.1254^2 \times 0.9344 = 0.0147$, $0.2307^2 \times 0.1879 = 0.0100$, $0.4834^2 \times 0.0483 = 0.0113$). The x-transformation is mainly a multiple of principal warp 2 (old principal warp 1), just as it was before augmentation of the scene by landmarks 6 and 7.

III. Principal Warps and the Interactive Refinement of Warping Functions

A. Edge Information

The landmarks of the scene analyzed in Figs. 4–6 were extracted from the pair of simulated edge-images (actually, doodles on a desk pad) shown in Fig. 10. (The width of the X's at each digitized point is presumed to encode the uncertainty of edge location.) If these were botanical data, the landmarks would be the "stem" of the flowers, the ends of the two "petals," and the uppermost points of the pair of "seeds." When the warping function of Fig. 4 is applied to every point in the left-hand image, there results the warped image shown in Fig. 11. Here, the large X's locate the five landmarks used to drive the thin-plate spline. The edge points of Fig. 10 are copied into this diagram by middle-sized X's. Finally, the small X's in the right-hand image represent the edge-points of the left-hand flower after transformation by the map $f = (f_x, f_y)$ based on these five landmarks.

It is apparent that this map fails to do justice to the form of the flowers: the petals on the right do not overlie the images of the petals on the left. We may begin to remedy this failure by somewhat arbitrarily selecting points along the petals near the middle of their arcs, left and right, and using them to drive the seven-landmark spline mapping shown in Fig. 12. This augmentation of the data greatly improves the apparent goodness-of-fit of the warped left image to the actual right image, at little cost in bending energy (net deformation). The units of this quantity, as

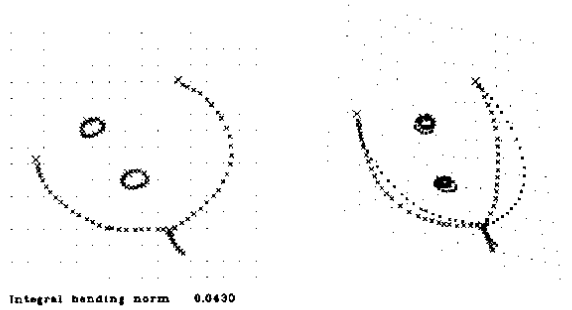


Fig. 11. Result of applying the warp of Fig. 4 to the left-hand form in Fig. 10. Large X's: landmarks driving the spline map. Middle-sized X's: observed edge data. Small X's: image of the left-hand flower after warping in accordance with the correspondence of five landmarks. There is considerable lack of fit: the five-landmark map does not fairly represent the relation between the forms as a deformation.

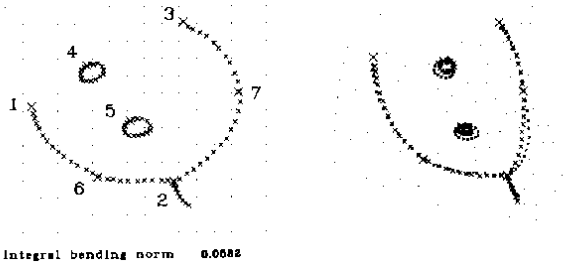


Fig. 12. Refinement of the map by matching midpoints of petals. Notice the great increase in visual quality of fit with only modest increase in net bending.

printed under the grid, are arbitrary, a function of the scale of my digitizing tablet.

The eigenanalysis of the seven-landmark map, Fig. 13, indicates one principal warp rather stiffer than the others (eigenvalue 0.3732, versus 0.1982 for the next stiffest). This most bent eigenvector mainly taps landmarks 1, 4, 5, and 6—the smallest quadrilateral of landmarks in this scheme, at lower left. From the display of these warps as physical thin plates, Fig. 14(a), we recognize the discordant-diagonal construction of Fig. 2. Principal warp 2 [Fig. 14(b)] is a slightly less bent surface, while warps 3 and 4 [Figs. 14(c)–(d)], both quite gentle, appear to be permuted versions of one another. From the pattern of segments out of landmark 7 in Fig. 13, we see that this landmark, so considerably displaced between Figs. 11 and 13, contributes mainly to this last principal warp, which accounts for half of the actual bending observed: 0.0275 ($0.1035 \times [0.3325^2 + 0.3948^2]$) out of 0.0582. This dominant feature of the deformation is the shift of landmarks 4, 5, and 2 downward-rightward with respect to the others: a relative translation of two diagonals of a square, as in Fig. 3. Because the function f_y loads mainly on principal warp 4, the thin plate for f_y in Fig. 15(b) strongly resembles that of the principal warp, Fig. 14(d).

B. *Deficient Landmarks*

In passing from Fig. 11 to Fig. 13, we have augmented our store of information about the correspondence of points between the two forms, but we have inserted a bit of misinformation as well. We do not know precisely which point of the curve between landmarks 1 and 2 on the right should be considered to correspond to the landmark 6 we selected on the left, and likewise which point of the arc from 2 to 3 corresponds to landmark 7. I will refer to landmark 7 in explaining the procedure by which such ambiguities are resolved.

Landmark 7 on the right could have been chosen to be any point of the appropriate arc. As the "best" point (in a sense to be explained in the next paragraph) is likely not too far from the point actually chosen, we can model this "freedom of choice" as the freedom of point (x_7', y_7') to vary along the line tangent to the right-hand curve near the starting guess. For ease of exposition, take this tangent line to be vertical; then the "data" are limited to the x-coordinate of landmark 7, while the y-coordinate y_7' may be set subject to any reasonable criterion. Because the digitized location of landmark 7 on the right has only one valid Cartesian coordinate, not two, I refer to its as a *deficient landmark*.

In the context of these splines, a criterion which immediately suggests itself is to place the point (x_7', y_7') so that *the net bending energy of the resulting spline is least.* In effect, we are using the energy of the spline as a measure of information content (measured by squared second derivatives) of the deformation as it deviates from the affine condition, the map with all second derivatives zero. We seek the representation of the map which has the least information consistent with what we actually know about the data (in this case, the coordinate x_7', but not the coordinate y_7').

The computation to be performed may be intuited graphically. Fig. 15 presents the thin plates corresponding to the complete nonlinear (affine-free) part of the transformation in Fig. 13. We see in Fig. 15(b) that the y-deformation is bent somewhat upward at landmark 7; therefore, its relaxation toward a state of lowered bending energy will push it downward. Landmark 7 loads most heavily on principal warp 4 [Fig. 14(d)], and secondarily on principal warp 2 [Fig. 14(b)]. Its movement downward rapidly decreases the bending energy associated with principal warp 4, but also increases, albeit more slowly, that associated with eigenvector 2 [since it is already tacked down by that landmark: Fig. 14(b)]. At the computed optimum of this shift, the amount of downward displacement of the coordinate y_7' will just balance the decrease in bending energy of the fourth principal warp against the increase in bending energy of the second (each squared, then weighted by its eigenvalue).

Similarly, we may inquire as to the possible effect on the bending of the spline of allowing landmark 6 on the right to slide along the tangent line there. A glance at Fig. 15 indicates that we are not likely to more it too far. This landmark tacks both coordinate sheets *down*, and by

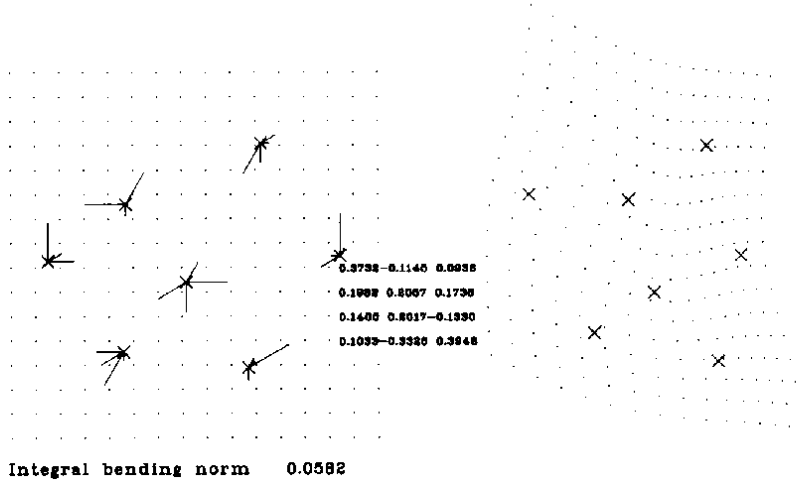


Fig. 13. Spectral analysis of the seven-landmark map. One principal warp is much stiffer than the others, but nearly half of all the bending is associated with the least stiff eigenvector, the relative translation of the two "diagonals" of the form.

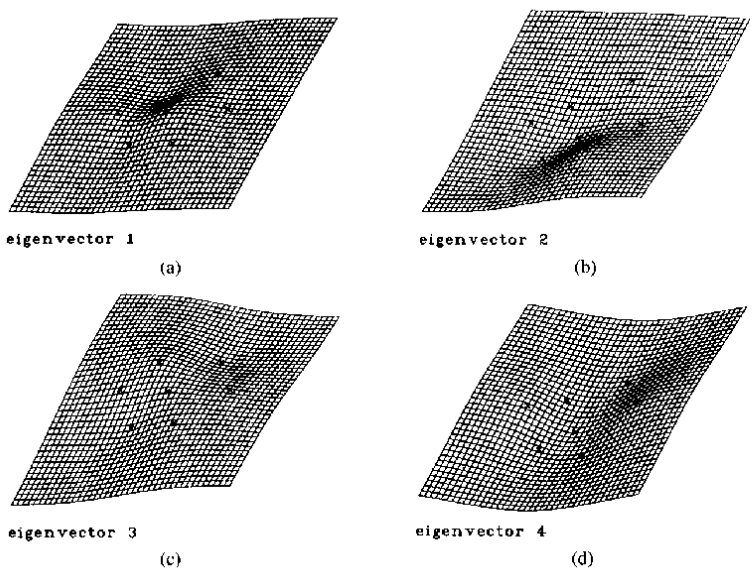


Fig. 14. Thin plates corresponding to the four principal warps of Fig. 13.

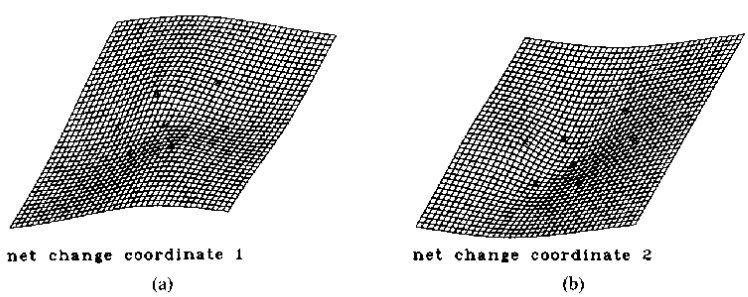


Fig. 15. Affine-free parts for the coordinates of Fig. 13, drawn as thin-plate splines.

roughly the same amount. The direction of the tangent line, along which we must move this landmark, is $(-1, 1)$; then any improvement in flatness of the x-sheet induced by a shift will be obviated by increased bending of the y-sheet, and vice versa. Because these contributions go as the squares of the projections on the eigenvectors, the effect of freeing landmark 6 to slide will very likely be null.

To accommodate this relaxation procedure in algebra, we must allow the homolog (x_i', y_i') of (x_i, y_i) to be any point of the form

$$(x_i', y_i') = ([x_i']_0 + t_i r_i, [y_i']_0 + t_i s_i),$$

where $([x_i']_0, [y_i']_0)$ is the point actually digitized, now merely representative of its tangent line; r_i and s_i are direction cosines of the line along which (x_i', y_i') is varying; and t_i is the amount of shift along the tangent line, determined so as to minimize the net bending energy. If k homologs are freed to slide along lines in this way, the matrix V actually covers an affine k-flat (k-dimensional vector subspace shifted away from the origin):

$$V = V(t_{j_1}, \cdots, t_{j_k})$$

$$= \begin{bmatrix} x_1' \cdots [x_{j_1}']_0 + t_{j_1} r_{j_1} \cdots [x_{j_k}']_0 + t_{j_k} r_{j_k} \cdots x_n' \\ \\ y_1' \cdots [y_{j_1}']_0 + t_{j_1} s_{j_1} \cdots [y_{j_k}']_0 + t_{j_k} s_{j_k} \cdots y_n' \end{bmatrix}.$$

As the t_{j_i} vary, the integral $I_f = I_f(t_{j_1}, \cdots, t_{j_k}) = V(L_n^{-1} K L_n^{-1}) V^T$ varies about a nonnegative minimum as a positive-semidefinite quadratic form in $t_{j_1} \cdots t_{j_k}$. The minimizing of $I_f(t_{j_1}, \cdots, t_{j_k})$ is numerically very tractable for $k \leq n - 3$.

The result of this relaxation in the present example is shown graphically in Fig. 16. The previous locations of landmarks 6 and 7 at the right are shown with large + signs. When each is freed to slide along the tangent line indicated, the computed positions of least bending energy are shown by the new large X's on the right. As expected, landmark 7 has moved considerably downward, whereas landmark 6 has hardly moved at all. The spline map that results is shown in Fig. 17. It is no less consistent than Fig. 12 with the information we actually have about biological homology. But its bending energy is only 0.0574. We have "saved" unnecessary bending induced by the passage from five landmarks to seven.

Fig. 16. Relaxation of landmarks 6 and 7. Their original positions on the right (Fig. 13) are indicated by large + symbols; the directions along which they are free to slide were entered by hand. The minimization of bending energy places these landmarks at the points along their tangent lines marked by X's.

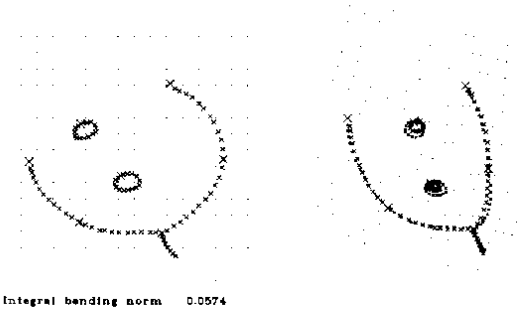


Fig. 17. Deformation using the new positions of landmarks 6 and 7 on the right. Note, by comparison with Fig. 12, the absence of the impression of "bending" of the grid lines about landmark 7.

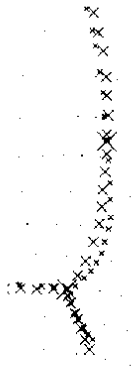


Fig. 18. Enlargement of Fig. 17, showing an apparent deviation of tangent lines between landmark 7 and the stem.

C. *Iterative Refinement of Deformations*

In Fig. 18, an enlargement of the region near landmark 7, there is apparent a systematic deviation of the tangent to the right-hand form below landmark 7 from the image of the tangent in the left-hand form. We may attempt to further refine the mapping in this region by choosing yet another intermediate landmark. In Fig. 19 I have added, arbitrarily, a landmark 8 slightly below 7 on the left and a bad guess at a homolog for it on the right. The addition of the eighth landmark pair adds 0.0035 units of bending energy to the computed spline—not an inconsiderable

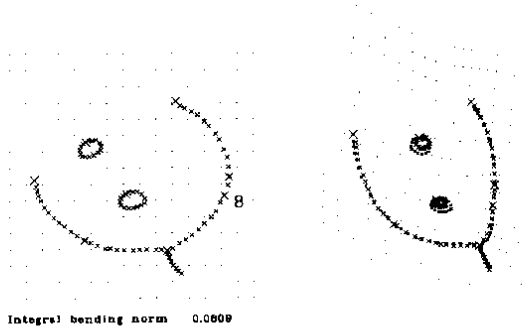


Fig. 19. Another refinement: an eight-point spline incorporating a two-point specification (both position and directional derivative) in the vicinity of landmark 7.

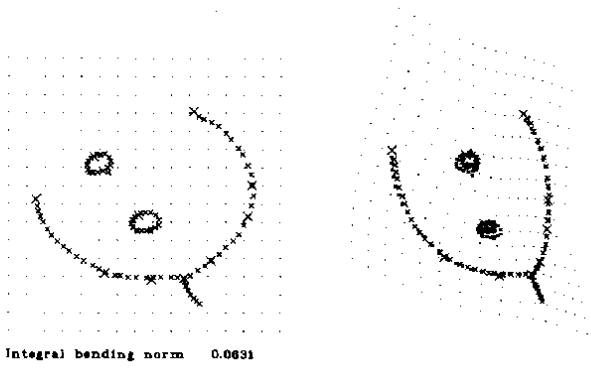


Fig. 20. Relaxation of Fig. 19 so that only the direction of the tangent at landmark 7 is specified, not the magnitude of the directional derivative along that tangent.

amount. While the deviation of tangent direction below landmark 7 has been corrected, this warping function is obviously inappropriate in the vertical direction. The grid lines of the right-hand form are bent apart in-between landmarks 7 and 8: the directional derivative has been specified inappropriately. Again, we have allowed ourselves to be misled by the pairing of Cartesian coordinates. What we know, in fact, takes the form of a constraint on the directional derivative of f at landmark 7. The direction into which the map takes the tangent to the curve on the left at landmark 7 is known—the ratio $df_y/ds : df_x/ds$ along the homologous arc of the left-hand flower—but not the magnitude of the derivative in this direction.

Landmark 8 must be freed to relax so that the projection on the eigenvector of highest bending energy [the constraint on the directional derivative: compare Fig. 8(a)] is close to zero, where its square will be balanced by increasing contributions to all the other, less stiff eigenvectors. The result of freeing both this landmark and (again) landmark 7 is shown in Fig. 20: the additional bending energy required is reduced from 0.0035 units to 0.0022.

The right-hand petals still fail to line up just above the "stem," because the angles at the stem have changed considerably beyond what is consistent with the curving of petals farther away. We can fix this by incorporation of two landmarks near the stem on opposite sides, both freed to slide. (The effect is to specify the effect of the affine derivative on angles, but not to constrain its effects upon lengths.) The addition of this local feature "costs" 0.0035 more bending (Fig. 21). The net effect on the outline of relaxing all five intercalated landmarks along the petals is to somewhat sensitively estimate the homology map *along* those arcs that is consistent with the pairing of petal ends and stems and that otherwise yields the minimum of bending over the whole picture given the positions of the seeds.

This solution for relaxation of an entire curve (up to a certain informal tolerance) may be compared to the method of "snakes" of [14]. Kass *et al.* likewise deform a smooth curve to fit image contours subject to generalized point constraints; they proceed by minimizing an energy functional which incorporates bending energy (in this case, one-dimensional) as one of its terms. Our landmarks are a special case of their point constraints, and the stem of our flower is a special case of their pointwise balance between "plate" and "membrane" energetics. Nevertheless there are striking formal differences between these approaches. Kass *et al.*'s energy term is a *line* integral limited to the curve itself, but allows "forces" to be applied at all points of the curve. The bending-energy of this paper is instead a double integral over the entire picture. The procedure here, unlike Kass', culminates in a finite-dimensional feature space (so that the "higher-level" processing may proceed directly). Kass *et al.* have no term for long-distance forces such as span the interior of the biological images here. The method of [21] lifts this limitation, and allows "forces" at all points, at the cost of being iterative and local. Like that method, the present method incorporates knowledge of interiors and handles multiple disconnected curves without further modification: their relative positions are already explicit in the global bending energy. One can imagine many useful hybrids between these two methods that are suited to intermediate applications.

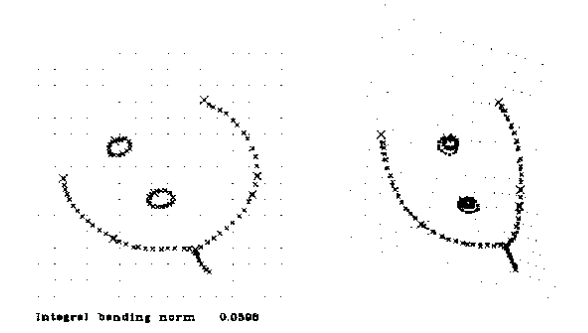
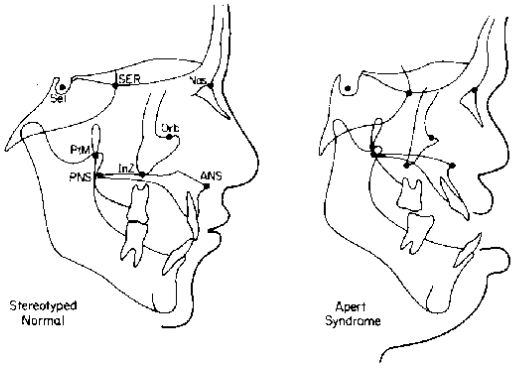
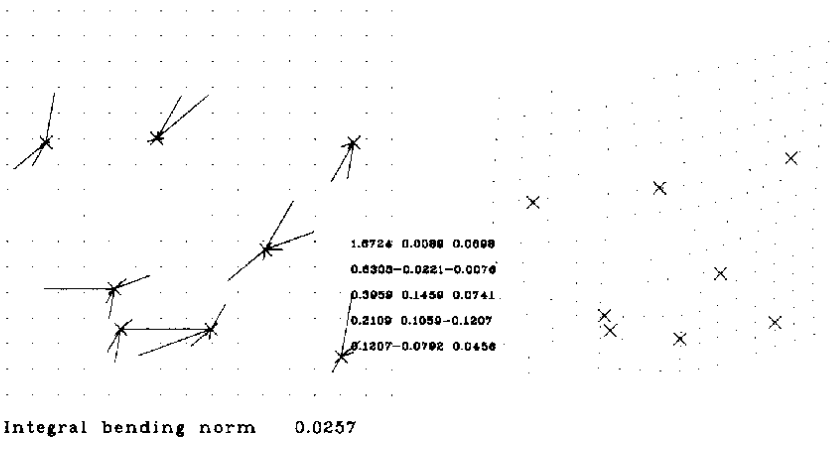


Fig. 21. Two more deficient landmarks specifying the angles of the affine derivative at the stem. The result of this editing has been the construction of a homology function along the arc of the petals such that the net bending energy of the spline it induces, given the locations of the seeds, is minimized.

Further adjustment of the warping function to match the scales of the "seeds" to each other costs another 0.0637



(a)



(b)

Fig. 22. Apert Syndrome as a deformation of normal. (a) Stereotyped tracings of X-rays, with landmarks, for normal children (left) and 14 cases of Apert Syndrome (right). Landmarks: ANS, Anterior Nasal Spine; Sel, Sella; SER, Sphenoethmoid registration point; Nas, Nasion; Orb, Orbitale; InZ, Inferior Zygoma; PtM, Pterygomaxillary Fissure; PNS, Posterior Nasal Spine. (b) Thin-plate spline using the eight landmarks shown. The principal warps are coded by segments counterclockwise from horizontal at 20° increments.

units of bending energy [7, Fig. 22] more than all the bending in Fig. 21. The shrinkage of the scale of the seeds is massively inconsistent with the behavior of the map all around the periphery of the flower, and it is not helpful to model it as a deformation of the area in-between the seeds and the petals, where there are, after all, no data.

IV. Example: Apert Syndrome

Apert Syndrome is a congenital craniofacial anomaly characterized by underdevelopment of the maxilla (upper jaw) apparently consequent upon abnormalities of the sutures joining the bones at the base of the brain. This example depicts abnormality in a mean configuration of eight landmarks from this part of the skull. The data come from *lateral cephalograms* (X-rays of the head from the side) traced and digitized by hand. The analysis (Fig. 22) is of averaged forms from 14 Apert patients treated at the Institute of Reconstructive Plastic Surgery, New York University, and the corresponding averaged form for normal Ann Arbor youth. The landmarks named in the figure caption have operational definitions as in [18].

The affine part of the mapping has principal strains of 0.882 and 0.690 in directions respectively along and perpendicular to (0.937, 0.350) on the left, (0.902, 0.431) on the right. This represents a change of proportion by some 21 percent involving compression aligned along a direction through ANS passing somewhat anterior to SER. The general appearance of this affine transformation is hinted at toward the upper right corner of the right-hand grid of Fig. 22.

The deviations of the observed data from this uniform change are shown for each Cartesian coordinate in Fig. 23. As viewed by eye, the dominant feature of x-nonlinearity is the "upward" deviation at landmark SER (i.e., its relative displacement *forward*). The y-nonlinearity appears to have a crimp in the vicinity of PtM-PNS and a dip (relative motion of landmarks *downward*) posterior and superior to ANS.

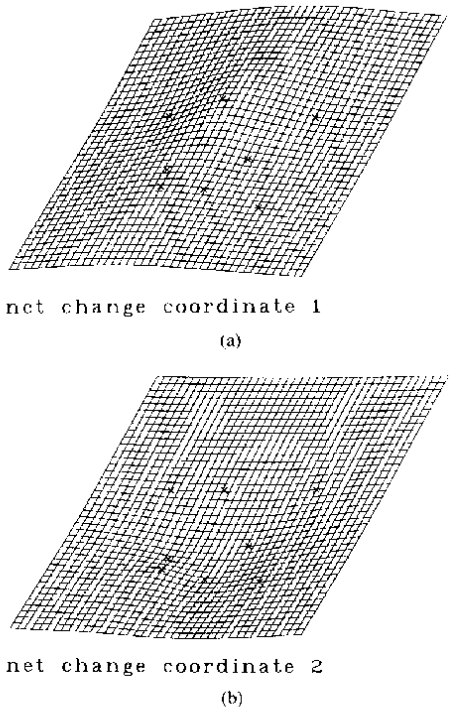


Fig. 23. Net x- and y-deviations from the uniform shear for the Apert deformation.

The decomposition of this deformation into principal warps is the roster of vector multiples listed on Fig. 22(b) for the five principal warps shown in Fig. 24 as splined surfaces. The bending energies associated with these principal warps, products of the eigenvalues by the summed squared loadings, are 0.0083, 0.0003, 0.0106, 0.0054, 0.0010, totalling 0.0257 as noted in Fig. 22(b). The fifth principal warp, Fig. 24(e), although familiar from Fig. 2, explains little of the net bending energy: the Apert deformity is not usefully considered to incorporate any aspect of the rigid shift of one diagonal or edge of the structure with respect to all the other landmarks. The point ANS, which contributes substantially to bending energy only through principal warp 5, is located approximately where it is left by the global affine term. The meaning of the most energetic principal warp [Fig. 24(a)] is clear by comparison to Fig. 8(a). The most local feature of the Apert deformation is the relative displacement of the pair at closest separation: PtM and PNS. In this example, that principal warp incorporates a signal primarily in the vertical direction.

There remain for consideration the two principal warps of moderate geometric scale, Figs. 24(c) and (d). The sharper of this pair, explaining the most observed bending, represents our familiar "pure inhomogeneity" applied to the anterior part of the scene: landmarks SER and ANS (and a bit of PNS-PtM as well) translated with respect to the line Sel-Orb-InZ in between them. Graphically, the general topography of x-nonlinearity, Fig. 23(a), may be imagined a weighted sum of the sharper ridge of Fig. 24(c) with the gentle dome of Fig. 24(d). The general facies of the y-nonlinearity, Fig. 23(b), is, apart from the clamped derivative at PtM-PNS, a weighted *difference* of these surfaces.

In respect of these eight projected landmark locations, we have expressed the deformity which is the sample mean Apert Syndrome as the combination of four features of deformation: a general affine term representing compression along an axis at about 60° clockwise of the cranial base; a highly localized change in the (projected) relation between PtM and PNS; and two localized relative displacements, SER to the right with respect to its neighbors, and SER and Orbitale to the right and downward with respect to their neighbors.

V. Discussion

The methods proposed here for decomposition of deformations relate to many classic problems in computer vision and image analysis. Throughout I have emphasized the importance of a feature space *for* deformation. Solutions of the stereopsis problem or the matching problem which are not tractable analytic functions of the data, which do not reside in a finite-dimensional function space, do not supply the necessary features. The principal warps instead embody the multivariate distribution of configurations of landmarks in a very helpful form. In this context the problem of landmark number is the biologist's, not the computer scientist's: our human ability to name homologous points is limited. The representation by pure thin-plate splines is very useful for some purposes (such as biometric analysis) and less useful for others (such as computer animation). Of course, some of the extended splining techniques have biometric applications as well: for instance, Terzopoulos' [20] "controlled-continuity spline" may model the sort of discontinuity generated when a biological structure is torn in the course of serial sectioning.

In the context of biological and medical measurement, the principal warps drive a conventional multivariate statistical analysis of their variances and covariances (cf. [8]). Choice of the bending energy rather than any generalization incorporating first-order terms is related to some aspects of this subject-matter, notably, the pervasive biological fact of *allometry* [9], which is the tendency of changes in proportions of form to be spatially graded. Biological forms tend to vary in proportions not only globally but also regionally. The pure thin-plate spline incurs little cost for deformations gently graded from region to region: it relaxes local first derivatives to averages at larger scale [cf. Figs. 7(c), (d), 24(d), (e)]. Methods of matching which instead incorporate first derivatives in the cost function, such as Broit's (see [10] or [1]), are substantively misleading. After "correction" of a uniform shear at the outset, these produce warping functions which relax back to similarities inside regions which are clearly related by affine changes instead. The thin-plate spline induces the correct relaxation toward the local, rather than the global, mean.

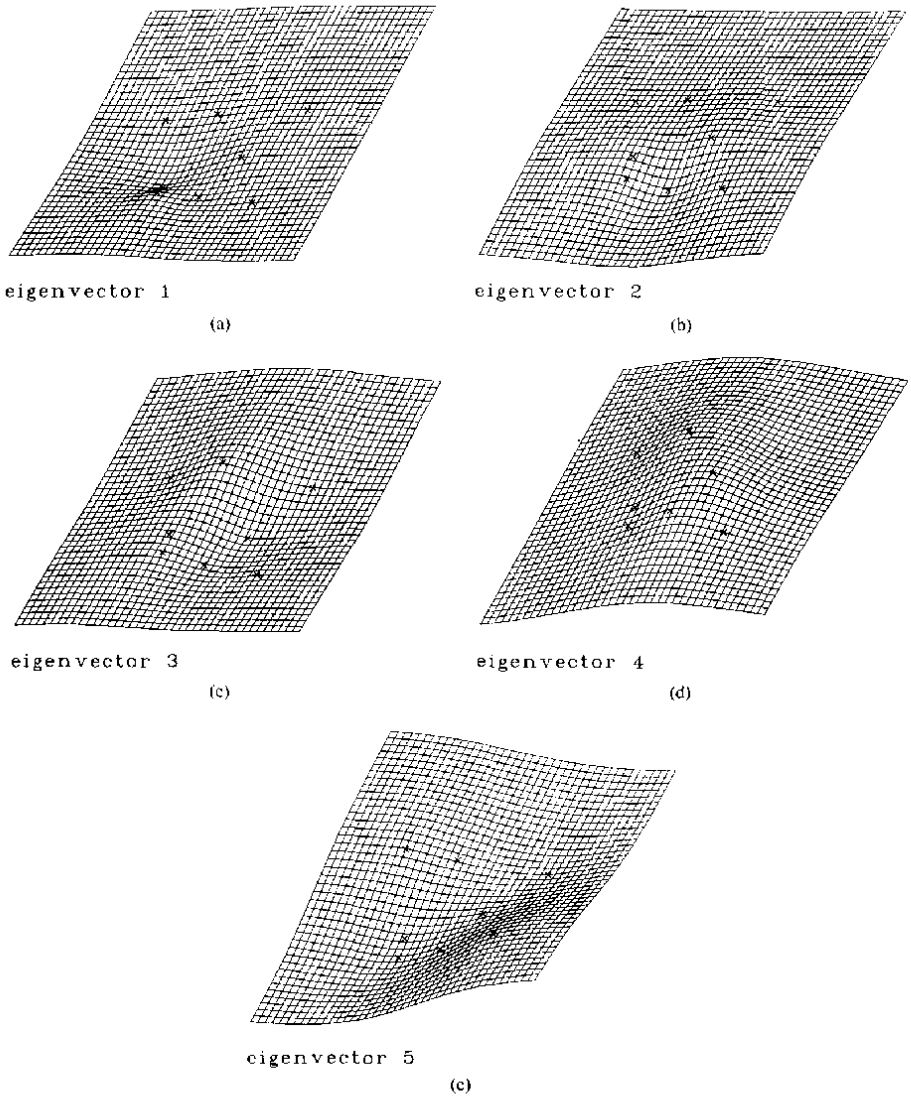


Fig. 24. The five principal warps for the configuration of eight landmarks of Fig. 22.

Considerations of energetics aside, the measurement of changes in landmark configurations has at least four points of contact with traditional concerns of computer vision: image discrimination, landmark identification, description of actual deformations, and instantiation of primitives.

A. Image Discrimination

Statistical methods exist for the multivariate analysis of arbitrary changes in landmark configurations in two and three dimensions [5]. The techniques apply to all the conventional biometrical designs: describing the difference between two samples of forms, the mean change in a single sample observed over time (e.g., biological evolution, or human growth), the difference in changes observed in two samples (such as patient groups subjected to different treatments), or the relation of a single form to a normative mean (its "diagnosis"). If landmarks can be chosen consistently on the left-hand form, the space of the decompositions explored here is a natural context for interpretation of all these multivariate findings, because features of the three-dimensional plates are much clearer than the same features displayed in the equivalent two-dimensional grid diagrams. The analysis of Apert form in the previous section exemplifies just such a thrust.

One major application of image analysis is to the problem of medical diagnosis. Often the nature of the scene is known (that is, we know we are looking at a human head) up to particulars of the deformation which are the precise subject of medical concern. In the presence of homologous landmarks, the joint distribution of the principal warps, together with the affine component, provides an

efficient coding of features for diagnostic decision-making. For instance, the example of Fig. 22 could easily lead to a linear discriminator for the Apert-Normal distinction; to a scalar quantity for the severity of any individual case of Apert; or to an index of relative efficacy for surgical corrections.

B. Landmark Identification

Some points biologically homologous between images are clearly identifiable by local image processes; others are not [cf. Fig. 22(a)]. When landmarks can be localized to edges, the method of Section III applies to choose a least bent candidate; when the choice is between two or more possibilities at finite separation, the features of the principal warp analysis can be used to compute a likelihood-ratio criterion, to be combined with other evidence (e.g., of local pattern strength), driving a rational higher-level choice between them.

C. Description of Actual Deformation

The approach to two- and higher-dimensional signal-matching in [21] does not provide a framework for description of the matching operation per se. In many applications the primary concern is this description, rather than explicit unwarping. For instance, the calibration of medical imaging devices requires analysis of regional aspects of the warping actually observed of a "phantom" (e.g., a rectilinear array of beams). In other cases, it is the deformation itself which must be measured and diagnosed, as for the heartbeat or the respiratory cycle. The algebra of discriminating deformations from diverse starting forms is essentially the same as that already introduced here.

D. Instantiation of Primitives

It is not always feasible to observe all the details by which instances differ; at least, not all at once. Fortunately, in biomedical imaging, details are ordinarily strongly intercorrelated. In that case, via regression analysis upon principal warp scores, one may pass from observations of a few of these landmark locations to shrewd computation of the expected positions of the rest, and thus know where to look for them. This technique has immediate applications to the problem of *stereotaxy* [6], the prediction of the anatomy of a patient's brain from a few external measurements together with a standard Atlas. The analysis of principal warps permits the design of cost-efficient schemes for the order of collecting landmark data as an alternative to brute-force least-squares fits (see [2]).

E. Three Dimensions

The formalism of the thin-plate spline copies over to three-dimensional data without any major changes. The fundamental solution of the biharmonic equation, which was $|r|^3$ in one dimension and $r^2 \log r^2$ in two dimensions, is now $U(r) = |r|$. Even though this is not differentiable at the landmarks, the superposition of several of these together with affine terms continues to minimize the integral over all space of the sum of all squared second derivatives. There are three areas of the extension of this technology from two dimensions to three that nevertheless will require considerable imagination. It is not clear how to draw a "thin-hyperplate spline," as the equivalent of Fig. 2 is a slightly bent hyperplane in Euclidean four-space. For two-dimensional data, points freed to roam the plane would (in the absence of other energy terms) merely adopt a position corresponding to the spline through the others, as the plate is the minimum of bending energy for any sample of however many points exactly upon it; but for three-dimensional data, the category of deficient landmarks splits into two varieties. Some have two meaningful coordinates (points restricted to curves, such as nerve tracts or blood vessels) while others have only one (points free to wander over whole surface neighborhoods, such as the skull or the skin). Especially for surface data in three dimensions, combinations of the global thin-plate algebra with the functionals underlying Kass *et al.*'s "active surface fitting" would seem to be very promising.

As these refinements are pursued, the decomposition of deformations may be expected to supplement the more usual procedure of extracting features from images separately for many conventional applications. Both thrusts are concerned with the sorting of information by geometric scale and with the efficient extraction of quantitative summaries for particular problems of discrimination and seriation. The explicit quantitative analysis of deformations has made considerable inroads in mathematical biology since D'Arcy Thompson's original suggestion of the method in 1917; it is time it was exploited in image analysis as well. The algebraic technology for describing features of deformations proposed here is a first step in that direction.

Acknowledgment

D. Ragozin of the University of Washington at Seattle called my attention to the elegant algebraic properties of thin-plate splines when I was Visiting Professor of Statistics there in 1985. The Tenth International Conference on Information Processing in Medical Imaging (Zeist, The Netherlands, June 1987) listened to an earlier version of this argument [7] and made many helpful suggestions regarding the exposition. This manuscript benefited from conversations with P. Sampson (Seattle) and with S. Pizer and W. Oliver of the Department of Computer Science at the University of North Carolina, Chapel Hill, and from a most conscientious anonymous reviewer supplying an embarrassing number of citations to related literature.

References

[1] R. Bajcsy, R. Lieberson, and M. Reivich, "A computerized system for the elastic matching of deformed radiographic images to idealized atlas images," *J. Comput. Assisted Tomography*, vol. 7, pp. 618-625, 1983.

[2] R. Bajcsy and F. Solina, "Three-dimensional object representation revisited," in *Proc. First Int. Conf. Computer Vision*, IEEE Comput. Soc., Catalog No. 87CH2465-3, 1987, pp. 231-240.

[3] F. L. Bookstein, *The Measurement of Biological Shape and Shape Change* (Lecture Notes in Biomathematics, vol. 24). New York: Springer-Verlag, 1978.

[4] ——, "Transformations of quadrilaterals, tensor fields, and morphogenesis," in *Mathematical Essays on Growth and the Emergence of Form*, P. L. Antonelli, Ed. Edmonton, Alta.: University of Alberta Press, 1985.
[5] ——, "Size and shape spaces for landmark data in two dimensions," *Stat. Sci.*, vol. 1, pp. 181–242, 1986.
[6] ——, "Morphometrics for functional imaging studies," *J. Cerebral Blood Flow and Metabolism (Assessment of Goals and Obstacles in Data Acquisition and Analysis from Emission Tomography)*, J. C. Mazziotta and S. H. Koslow, Eds., vol. 7, pp. S23–S27, 1987.
[7] ——, "Toward a notion of feature extraction for plane mappings," in *Proc. Tenth Int. Conf. Information Processing in Medical Imaging*, C. de Graaf and M. Viergever, Eds. New York: Plenum, 1988, pp. 23–44.
[8] ——, "Comment on D. G. Kendall, 'A survey of the statistical theory of shape,'" *Stat. Sci.*, May 1989.
[9] F. L. Bookstein, B. Chernoff, R. Elder, J. Humphries, G. Smith, and R. Strauss, *Morphometrics in Evolutionary Biology: The Geometry of Size and Shape Change, with Examples from Fishes*. Philadelphia, PA: Acad. Natural Sci., 1985.
[10] C. Broit, "Optimal registration of deformed images," unpublished doctoral dissertation, Univ. Pennsylvania, 1981.
[11] J. Duchon, "Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces," *RAIRO Analyse Numérique*, vol. 10, pp. 5–12, 1976.
[12] J. M. Fitzpatrick and M. R. Leuze, "A class of one-to-one two-dimensional transformations," *Comput. Vision, Graphics, Image Processing*, vol. 39, pp. 369–382, 1987.
[13] R. Franke, "Scattered data interpolation: Tests of some methods," *Math. Computat.*, vol. 38, pp. 181–200, 1982.
[14] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vision*, vol. 1, pp. 321–331, 1987.
[15] J. Meinguet, "Multivariate interpolation at arbitrary points made simple," *Zeitschrift für Angewandte Mathematik und Physik (ZAMP)*, vol. 30, pp. 292–304, 1979.
[16] ——, "An intrinsic approach to multivariate spline interpolation at arbitrary points," in *Polynomial and Spline Approximation*, B. Sahney, Ed. Dordrecht, The Netherlands: Reidel, 1979, pp. 163–190.
[17] ——, "Surface spline interpolation: Basic theory and computational aspects," in *Approximation Theory and Spline Functions*, S. P. Singh *et al.*, Eds. Dordrecht, The Netherlands: Reidel, 1984, pp. 127–142.
[18] M. L. Riolo, R. E. Moyers, J. S. McNamara, and W. S. Hunter, "An atlas of craniofacial growth," Center for Human Growth and Development, Univ. Michigan, 1974.
[19] D. Terzopoulos, "Multilevel computational processes for visual surface reconstruction," *Comput. Vision, Graphics, Image Processing*, vol. 24, pp. 52–96, 1983.
[20] ——, "Regularization of inverse visual problems involving discontinuities," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-8, pp. 413–424, 1986.
[21] A. Witkin, D. Terzopoulos, and M. Kass, "Signal matching through scale space," *Int. J. Comput. Vision*, vol. 1, pp. 133–144, 1987.

Fred L. Bookstein received the B.S. degree in mathematics and physics from the University of Michigan, Ann Arbor, in 1966, the M.A. degree in sociology from Harvard University, Cambridge, MA, in 1971, and the Ph.D. degree in statistics and zoology from the University of Michigan in 1977.

At present he is a Research Scientist with the Center for Human Growth and Development at that University. His research concentrates on the invention of methods for measuring complex systems with hidden parameters. Much of this research deals with morphometrics, the statistical study of biological form and its changes, but the formal strategies required to analyze biological images efficiently have counterparts throughout the natural and social sciences, whenever crucial determinants of processes must be reconstructed from indirect indicators of outcomes. His current projects include the anatomy of gross craniofacial birth defects and the optimizing of their surgical repair; the psychological sequelae of prenatal exposure to alcohol; description of the anatomical variability of normal human brains, and the interplay between their anatomical and metabolic images; and the import of random walks as a model for evolutionary history. The topic of the present paper, feature extraction for changes in biologically labeled Cartesian point data, is also the subject of a forthcoming book, his third.