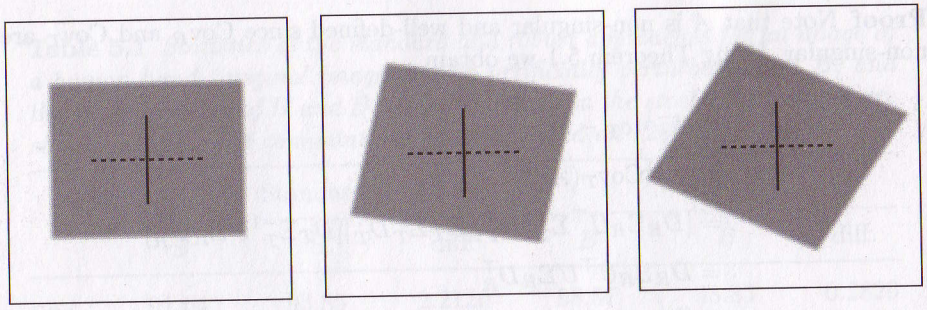

FIG. 5.4 Three different rectangles sharing the same stochastic features.

three different rectangles are depicted. Although these rectangles look quite different, they all share the same stochastic features. As a consequence, the PAT cannot distinguish between these different images.

6

OPTIMAL LINEAR REGISTRATION

In this chapter we investigate the question of how to find an optimal linear registration based on a distance measure $\mathcal{D}$. An analytical solution cannot be expected for the images from our application. Thus, we have to look for a numerical solution. Moreover, since the images under consideration are of high resolution, we focus on fast and efficient schemes, which typically exploit derivatives. Thus, an important property of the distance measure to be discussed is its differentiability.

The choice of an appropriate distance measure is a difficult task. Popular choices to be discussed in the subsequent sections are based on *intensity* (see, e.g., Brown (1992)), *correlation* (see, e.g., Collins & Evans (1997)), or *mutual information* (see, e.g., Viola (1995) or Collignon et al (1995)).

For some particular applications, modifications of these similarity measures have been investigated; see, e.g., Studholme et al (1996) or Roche et al (1999). In addition, the distance measure used in the registration may also be based on particular image features, e.g., edges or surfaces.

We start by introducing a set of feasible transformations, which here are supposed to be affine linear maps, i.e., $\varphi \in \Pi_1^d(\mathbb{R}^d)$; cf., Definition 3.6. A mathematical formulation of the registration problem then reads as follows.

Problem 6.1 *Find* $\varphi \in \Pi_1^d(\mathbb{R}^d)$ *such that* $\mathcal{D}[\varphi] = \min$.

The essential point here is that the set $\in \Pi_1^d(\mathbb{R}^d)$ can be parameterized. For a specific element φ of $\in \Pi_1^d(\mathbb{R}^d)$, we make use of the notation φ_a, where

$$\varphi_{a;\ell}(x) = a_{\ell,0} + \sum_{j=1}^{d} a_{\ell,j} x_j, \quad \ell = 1, \ldots, d.$$

The parameters $a_{\ell,j}$ are gathered together in a vector,

$$a = (a_{1,0}, \ldots, a_{1,d}, \ldots, \quad a_{d,0}, \ldots, a_{d,d})^\top \in \mathbb{R}^n, \quad n = d(d+1).$$

Moreover, we set

$$D(a) := \mathcal{D}[\varphi_a] \quad \text{and} \quad T_a := T \circ \varphi_a. \tag{6.1}$$

Thus, Problem 6.1 may be reformulated in terms of a parameterized finite-dimensional optimization problem.

Problem 6.2 *Find $a \in \mathbb{R}^n$, such that $D(a) = \min$.*

Since differentiability is a major point for the development of fast numerical schemes, we require the template image to be differentiable.

In principle any minimization technique can be used for the minimization of D. However, on the one hand it turns out that *direct methods* or *steepest descent* methods are not fast enough, while on the other hand, second order derivative-based Newton-type methods are not stable for real-life applications. This is because the derivatives of the images have to be approximated from the discrete data. Since these data are typically corrupted by noise, estimating a derivative becomes a delicate matter.

6.1 Intensity-based registration

A straightforward approach is based on the minimization of the so-called *sum of squared differences* (SSD); cf., e.g., Brown (1992) or Čapek (1999).

Definition 6.1 *Let $d \in \mathbb{N}$ and $R, T \in \mathrm{Img}(d)$. The* sum of squared differences (SSD) *distance measure $\mathcal{D}^{\mathrm{SSD}}$ is defined by $\mathcal{D}^{\mathrm{SSD}} : \mathrm{Img}(d)^2 \to \mathbb{R}$,*

$$\mathcal{D}^{\mathrm{SSD}}[R,T] := \frac{1}{2}\|T-R\|_{L_2}^2 = \frac{1}{2}\int_{\mathbb{R}^d} \big(T(x)-R(x)\big)^2 dx.$$

For a transformation $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ we also define

$$\mathcal{D}^{\mathrm{SSD}}[R,T;\varphi] = \mathcal{D}^{\mathrm{SSD}}[R,T\circ\varphi], \tag{6.2}$$

and for a parametric transformation φ_a we set

$$D^{\mathrm{SSD}}(R,T;a) = \mathcal{D}^{\mathrm{SSD}}[R,T\circ\varphi_a]. \tag{6.3}$$

In order to make Newton-type methods applicable, we compute the derivatives of D; cf., eqn (6.3). Elementary computations give

$$\partial_{a_j} D^{\mathrm{SSD}}(R,T;a) = \big\langle (T_a - R), \partial_{a_j} T_a \big\rangle_{L_2},$$
$$\partial_{a_j a_k} D^{\mathrm{SSD}}(R,T;a) = \big\langle \partial_{a_j} T_a, \partial_{a_j} T_a \big\rangle_{L_2} + \big\langle (T_a - R), \partial_{a_j,a_k} T_a \big\rangle_{L_2}.$$

Note that we could also work with subderivatives if the images are non-smooth. However, in all our applications, the images are given in terms of discrete data, and we obtain the continuous images by using an interpolation scheme. It is this interpolated image which has to fulfill the smoothness constraints. Thus, the smoothness restriction is less severe as it might appear at first right.

As already pointed out, a second order derivative-based approach might not be the method of choice. Here, we take advantage of the so-called Gauss–Newton method, where, roughly speaking, the linearization step is performed within the norm. However, Levenberg–Marquardt techniques are also used in the literature; cf., e.g., Thévenaz et al (1998).

A first order Taylor expansion gives

$$\begin{aligned} D^{\mathrm{SSD}}(R,T;a+b) &= \tfrac{1}{2}\|T_{a+b} - R\|_{L_2}^2 \\ &\approx \tfrac{1}{2}\|T_a - R + \nabla_a T_a^\top b\|_{L_2}^2. \end{aligned} \tag{6.4}$$

Minimizing the term on the right hand side with respect to b, we discover a linear least squares problem. Thus, for a fixed a, the optimal solution is characterized by the *normal equations*

$$M(a)b = f(a),$$

where $M(a) := (m_{j,k}(a)) \in \mathbb{R}^{n\times n}$, $f(a) = (f_j(a)) \in \mathbb{R}^n$, with

$$f_j(a) = \big\langle T_a - R, \partial_{a_j} T_a \big\rangle_{L_2}, \tag{6.5}$$

$$m_{j,k}(a) = \big\langle \partial_{a_j} T_a, \partial_{a_k} T_a \big\rangle_{L_2}. \tag{6.6}$$

The overall algorithm for computing an optimal parameter a^* is summarized in Algorithm 6.1. The computationally expensive parts are the $d(d+3)/2$ inner products for the computation of $f(a)$ and $M(a)$ and the computation of $T\circ\varphi_a$. Here, an interpolation scheme has to be exploited. From a theoretical point of view, this approach requires at least a quadratic interpolation scheme. However,

Algorithm 6.1 *Gauss–Newton method for the minimization of D^{SSD}; cf., eqn (6.3).*

Set $k = 0$, choose initial guess $a^{(k)}$.
While not STOP,
  compute $f(a^{(k)})$ and $M(a^{(k)})$,
  cf., eqns (6.5) and (6.6), respectively;
  solve the system of linear equations
  $M(a^{(k)})b = f(a^{(k)})$;
  update $a^{(k+1)} = a^{(k)} + b$, $k \mapsto k+1$;
end.

The iteration is stopped once the norm of the update b is brought below a certain tolerance (here, $\mathrm{tol}_1 = 10^{-5}$) or the relative decrease in the objective function is too slow, $f(a^{(k)}) - f(a^{(k+1)}) \le \mathrm{tol}_2 f(a^{(k+1)})$. In our implementation we used $\mathrm{tol}_2 = 10^{-6}$.

numerical experience provides evidence that d-linear interpolation schemes as introduced in Section 3.1.3 can be used with success, too.

In the case of discrete images, it is more stable to replace the normal equations by the solution of the finite-dimensional least squares problem using a QR-decomposition; cf., e.g., Golub & van Loan (1989, §5.3.4).

From a mathematical point of view, the type of parameterization of the transformation φ is of no particular importance. Thus, any of the restricted models introduced in Section 3.3.1 can be treated in a similar fashion.

6.2 An example of intensity-based affine linear registration

The results of an optimal intensity-based affine linear registration are shown in Fig. 6.1. Here, the intensity values of the template image have been modified and the minimization is performed with respect to both the geometrical parameters

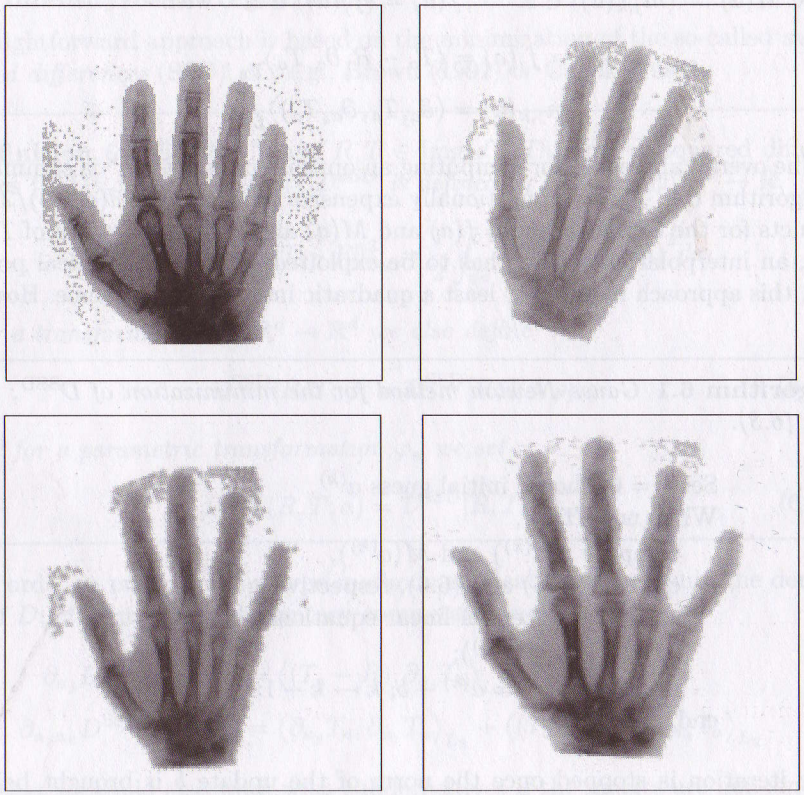

FIG. 6.1 Optimal intensity-based affine linear registration; reference (TOP LEFT), template (TOP RIGHT), template after rigid registration (BOTTOM LEFT), template after affine linear registration (BOTTOM RIGHT).

and the parameters of an affine linear model for the gray values. Two registration results are shown, one for a rigid registration and one for a general affine linear map.

6.3 Correlation-based registration

Registrations based on modifications of the so-called *correlation* have been studied by various authors; see, e.g., Collins & Evans (1997). Here we use the definition introduced by Gonzales & Woods (1993, p. 583).

Definition 6.2 *Let $d \in \mathbb{N}$ and $R, T \in \mathrm{Img}(d)$. The* correlation *between R and T is given by*

$$\mathrm{Corr} : \mathrm{Img}(d)^2 \times \mathbb{R}^d \to \mathbb{R}, \quad \mathrm{Corr}_{R,T}(y) := \int_{\mathbb{R}^d} R(x)T(x-y)dx.$$

The correlation may also be viewed as the L_2-inner product between R and $T(\cdot - y)$. If, in particular, R and T are normalized, such that they are of unit length, the correlation is the cosine of the angle between the two images. Maximization of the correlation with respect to y gives an image $T(\cdot - y)$ which is close to R in the sense that R and $T(\cdot - y)$ are maximally linearly dependent.

The usual normalization (see, e.g., Gonzales & Woods (1993, p. 583) is by statistics of the first kind. To this end, we assume that the support of all images under consideration is contained in a region $\Omega \subset \mathbb{R}^d$. For simplicity and without loss of generality we assume $\Omega =]0,1[^d$ and thus $|\Omega| := \int_\Omega dx = 1$.

Definition 6.3 *Let $d \in \mathbb{N}$ and $B \in \mathrm{Img}(d)$ be an image. The* expectation value *μ and the* standard deviation *σ of B are defined by*

$$\mu(B) := |\Omega|^{-1} \int_\Omega B(x)dx \quad \text{and} \quad \sigma(B) := \mu\Big((B - \mu(B))^2\Big).$$

Definition 6.4 *Let $d \in \mathbb{N}$. The* correlation coefficient *is defined by $\gamma : \mathrm{Img}(d)^2 \times \mathbb{R}^d \to \mathbb{R}$,*

$$\gamma(R,T;y) := \left\langle \frac{R - \mu(R)}{\sigma(R)}, \frac{T_y - \mu(T_y)}{\sigma(T_y)} \right\rangle_{L_2},$$

where $T_y(x) = T(x-y)$ and $\mu(B)$ and $\sigma(B)$ are defined in Definition 6.3.

Using this normalization the correlation coefficient is just the cosine of the angle between $R - \mu(R)$ and $T_y - \mu(T_y)$.

Definition 6.5 *Let $d \in \mathbb{N}$ and $R, T \in \mathrm{Img}(d)$. The* correlation-*based distance measure $\mathcal{D}^{\mathrm{corr}}$ is defined by $\mathcal{D}^{\mathrm{corr}} : \mathrm{Img}(d)^2 \to \mathbb{R}$,*

$$\mathcal{D}^{\mathrm{corr}}[R,T] := \left\langle \frac{R-\mu(R)}{\sigma(R)}, \frac{T-\mu(T)}{\sigma(T)} \right\rangle_{L_2},$$

where μ and σ are defined by Definition 6.3. For a transformation $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ we also define

$$\mathcal{D}^{\mathrm{corr}}[R,T;\varphi] = \mathcal{D}^{\mathrm{corr}}[R, T \circ \varphi], \tag{6.7}$$

and for a parametric transformation φ_a we set

$$D^{\mathrm{corr}}(R,T;a) = \mathcal{D}^{\mathrm{corr}}[R, T \circ \varphi_a]. \tag{6.8}$$

Since

$$2\mathcal{D}^{\mathrm{SSD}}[R,T] = \|R\|_{L_2}^2 + \|T\|_{L_2}^2 - 2\langle R, T\rangle_{L_2}$$

and

$$\begin{aligned}\sigma(R)\sigma(T) \cdot \mathcal{D}^{\mathrm{corr}}[R,T] &= \langle R - \mu(R), T - \mu(T)\rangle_{L_2} \\ &= \langle R, T\rangle_{L_2} - \mu(R)\mu(T)\end{aligned}$$

we see that there is a strong connection between the minimization of $\mathcal{D}^{\mathrm{SSD}}$ and the maximization of $\mathcal{D}^{\mathrm{corr}}$. If in particular the transformation is restricted to pure translation, we have $\det(\nabla\varphi) = 1$ and the approaches coincide. This follows from $\|R\|_{L_2}, \sigma(R)$, and $\mu(R)$ and $\|T\|_{L_2}, \sigma(T)$, and $\mu(T)$ being constant, respectively.

6.4 Mutual information-based registration

Since 1995, *mutual information* has been used in image registration. This approach was proposed independently by Viola (1995) and Collignon et al (1995) and has been used since then by many authors, e.g., Kim et al (1997), Maes et al (1997), Gaens et al (1998), Meihe et al (1999), or Abram (2000).

6.4.1 *Mutual information*

The basic idea is the maximization of the so-called *mutual information* of the images with respect to the transformation. Mutual information is an entropy-based measure with a widespread use in information theory. The precise definitions of the distance measure $\mathcal{D}$, the mutual information MI, and the entropy H are summarized as follows.

Definition 6.6 *Let $q \in \mathbb{N}$ and ρ be a density on $\mathbb{R}^q$, i.e., $\rho : \mathbb{R}^q \to \mathbb{R}$, $\rho(x) \geq 0$, and $\int_{\mathbb{R}^q} \rho(x)\,dx = 1$. The (differential)* entropy *of the density is defined by*

$$H(\rho) := -\mathbb{E}_\rho[\log\rho] = -\int_{\mathbb{R}^q} \rho \log\rho \; dg.$$

Definition 6.7 *Let $d \in \mathbb{N}, R, T \in \mathrm{Img}(d)$. The* mutual information (MI) *distance measure $\mathcal{D}^{\mathrm{MI}}$ is defined by $\mathcal{D}^{\mathrm{MI}} : \mathrm{Img}(d)^2 \to \mathbb{R}$,*

$$\mathcal{D}^{\mathrm{MI}}[R,T] := H(\rho_R) + H(\rho_T) - H(\rho_{R,T}),$$

where ρ_R, ρ_T, and $\rho_{R,T}$ denote the gray-value densities of R, T, and the joint gray-value distribution, respectively.

For a transformation $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ we also define

$$\mathcal{D}^{\mathrm{MI}}[R,T;\varphi] = \mathcal{D}^{\mathrm{MI}}[R, T \circ \varphi], \tag{6.9}$$

and for a parametric transformation φ_a we set

$$D^{\mathrm{MI}}(R,T;a) = \mathcal{D}^{\mathrm{MI}}[R, T \circ \varphi_a]. \tag{6.10}$$

The basic idea of mutual information is illustrated by Fig. 6.2. Here, the transformed templates $T \circ \varphi$, where φ is essentially a rotation of degree α, and the joint gray value density $\rho_{T,T\circ\varphi}$ are displayed. This figure shows that the density is very "sharp", when $T_\varphi = T$ and becomes "smeared out" when α increases. Since the mutual information essentially measures the entropy of the joint density, it is maximal if the images are maximally related.

The entropy is the expectation of the negative logarithm of the density. Thus we may also write

$$\mathcal{D}^{\mathrm{MI}}[R,T] = -\mathbb{E}_{\rho_{R,T}}\left[\log\frac{\rho_{R,T}}{\rho_R\rho_T}\right],$$

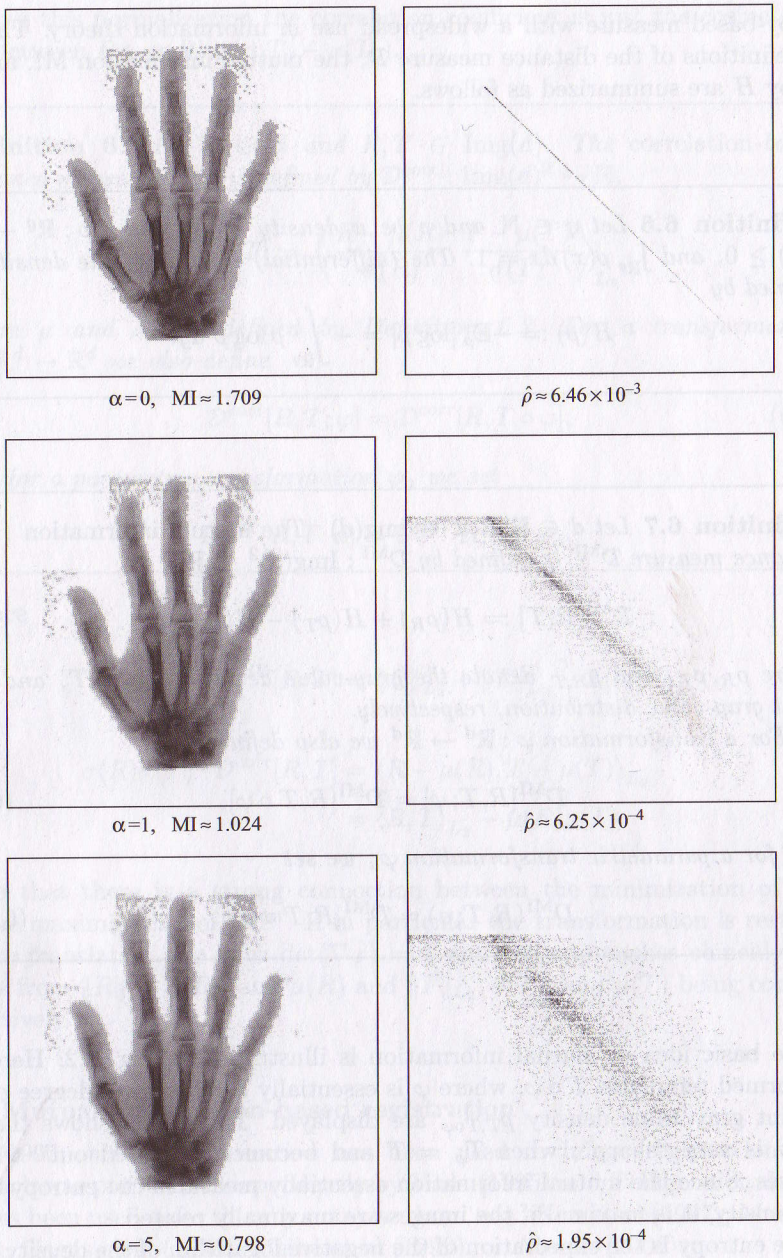

FIG. 6.2 LEFT: deformed template, RIGHT: log-plot of the joint density ρ_{T,T_φ}, where φ is a rotation of $\alpha, \alpha = 0, 1$, and 5 degrees, $\mathrm{MI} := \mathcal{D}^{\mathrm{MI}}[T, T \circ \varphi], \hat{\rho} := \|\log(1 + \hat{\rho}_{T,T_\alpha})\|_\infty$. For illustration purposes, the density has been re-scaled and background values have been neglected.

since

$$\begin{aligned}
&\mathbb{E}_{\rho_{R,T}}\left[\log\frac{\rho_{R,T}}{\rho_R\rho_T}\right] \\
&\quad= \int_{\mathbb{R}^2}(\log\rho_{R,T}(g_1,g_2) - \log\rho_R(g_1) - \log\rho_T(g_2))\rho_{R,T}(g_1,g_2)\; d(g_1,g_2) \\
&\quad= -H(\rho_{R,T}) - \int_{\mathbb{R}}\int_{\mathbb{R}}\rho_{R,T}(g_1,g_2)\; dg_2\; \log\rho_R(g_1)\; dg_1 \\
&\qquad - \int_{\mathbb{R}}\int_{\mathbb{R}}\rho_{R,T}(g_1,g_2)\; dg_1\; \log\rho_T(g_2)\; dg_2 \\
&\quad= -H(\rho_{R,T}) - \int_{\mathbb{R}}\rho_R(g_1)\log\rho_R(g_1)\; dg_1 - \int_{\mathbb{R}}\rho_T(g_2)\log\rho_T(g_2)\; dg_2,
\end{aligned}$$

where Fubini's theorem has been used.

However, we are interested in a differentiable distance measure, and for non-smooth gray value densities, the mutual information is not differentiable. In order to circumvent this disadvantage, we approximate the joint density by a smooth, i.e., differentiable, approximation using Parzen window techniques.

Here we summarize the approach presented in Wells et al (1996). The basic idea is to estimate the density ρ_S using a given sample $X = (X_1, \ldots, X_n)$ and the entropy H_S given a sample $Y = (Y_1, \ldots, Y_m)$, where $X_j \in \Omega$ and $Y_k \in \Omega$ are independently and identically distributed (i.i.d.) on Ω. The mutual information, which is not directly accessible, is computed in terms of these estimates.

In the first step, the densities ρ_R, ρ_T, and $\rho_{R,T}$ are estimated using the so-called *Parzen window method* (see, e.g., Duda & Hart (1973)),

$$\rho_S(z) \approx \rho_S(z, X) := \frac{1}{n}\sum_{j=1}^{n} g_q\big(\Sigma_S, (z - S(X_j))\big), \tag{6.11}$$

where $g_q : \mathbb{R}^{q\times q} \times \mathbb{R}^q \to \mathbb{R}$ is chosen to be a q-variate Gaussian density

$$g_q(\Sigma, z) := (2\pi)^{-q/2}(\det\Sigma)^{-1/2}\exp(-z^\top\Sigma^{-1}z/2), \tag{6.12}$$

i.e., $\Sigma \in \mathbb{R}^{d\times d}$ and $z \in \mathbb{R}^d$. For $S = R, T$, we have $q = 1$, and for $S = (R,T)^\top$, we have $q = 2$. The question of how to find the optimal values for Σ is discussed later. Note that Σ_R, Σ_T, and $\Sigma_{(R,T)^\top}$ are the covariance of the Parzen window and not of the density to be estimated.

In the second step, the entropies are estimated using a Monte Carlo method, i.e.,

$$H(\rho_S) = \mathbb{E}_{\rho_S}[-\log \rho_S] \approx -\frac{1}{m}\sum_{k=1}^{m} \log \rho_S(S(Y_k)).$$

Combining these two approximation steps, one obtains the following estimates $\tilde{H}(\rho_S; X, Y)$ for the entropies $H(\rho_S)$,

$$\begin{aligned} H(\rho_S) &\approx -\frac{1}{m}\sum_{k=1}^{m} \log \rho_S(S(Y_k)) \\ &\approx -\frac{1}{m}\sum_{k=1}^{m} \log \rho_S(S(Y_k), X) \\ &= -\frac{1}{m}\sum_{k=1}^{m} \log\left(\frac{1}{n}\textstyle\sum_{j=1}^{n} g_q(\Sigma_S, (S(Y_k) - S(X_j)))\right) \\ &=: \tilde{H}(S; X, Y). \end{aligned}$$

In Viola (1995, p. 49f.) it is argued that a measure for the quality of the Parzen estimate is to evaluate the standard deviation normalized by the mean. Viola gives the estimation

$$\frac{\sigma(P^*(x, X))}{\mathbb{E}[P^*(x, X)]} \approx \sqrt{\frac{k}{n}}\sqrt{\frac{k - P^*(x, X)}{P^*(x, X)}},$$

where P^* is the Parzen window approximation based on the sample X, n is the length of the sample, and k is a normalization constant, chosen such that P^* integrates to one; cf., Viola (1995, p. 50). Note that this estimate neglects the dependence of the density on the bandwidth σ. Considering also this dependency, the rate of convergence is only $n^{-2/5}$, following standard arguments; cf., Dümbgen (2001).

Estimating a univariate density ρ from the i.i.d. sample $X_j, j = 1, \ldots, n$, using the kernel estimator

$$\hat{\rho}_{n,\sigma}(y) := \frac{1}{n}\sum_{j=1}^{n} g(\sigma, y - X_j),$$

where $g = g_1$, we have

$$\begin{aligned} \mathbb{E}_\rho[\hat{\rho}_{n,\sigma}(x)] &= \frac{1}{n}\mathbb{E}_\rho\left[\sum_{j=1}^{n} g(\sigma, x - X_j)\right] = \mathbb{E}_\rho[g(\sigma, x - X_1)] \\ &= \int_{\mathbb{R}} g(\sigma, x - y)\rho(y)dy \\ &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-y)^2/(2\sigma)}\rho(y)dy \\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-z^2/2}\rho(x - \sqrt{\sigma}z)dz \\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-z^2/2}\big(\rho(x) - \sqrt{\sigma}\rho'(x)z + \sigma\rho''(x - \theta z)z^2\big)dz \\ &= \rho(x) + \mathcal{O}(\sigma), \\ \mathrm{Var}_\rho[\hat{\rho}_{n,\sigma}(x)] &= \frac{1}{n}\mathrm{Var}_\rho[g(\sigma, x - X_1)] \\ &= \frac{1}{n}\int_{\mathbb{R}} g(\sigma, x - y)^2\rho(y)dy \\ &= \frac{1}{n}\int_{\mathbb{R}} \frac{1}{2\pi\sigma} e^{-(x-y)^2/\sigma}\rho(y)dy \\ &= \frac{1}{n\sqrt{\sigma}}\int_{\mathbb{R}} \frac{1}{2\pi} e^{-z^2}\rho(x - \sqrt{\sigma}z)dz \\ &= \mathcal{O}((n\sqrt{\sigma})^{-1}). \end{aligned}$$

Hence,

$$\begin{aligned} &\mathbb{E}_\rho\left[(\hat{\rho}_{n,\sigma}(x) - \rho(x))^2\right]^{-1/2} \\ &= \left((\mathbb{E}_\rho[\hat{\rho}_{n,\sigma}(x)] - \rho(x))^2 + \mathrm{Var}_\rho[\hat{\rho}_{n,\sigma}(x)]\right)^{-1/2} \\ &= \left(\mathcal{O}(\sigma^2) + \mathcal{O}((n\sqrt{\sigma})^{-1})\right)^{-1/2} \\ &= \mathcal{O}(n^{-2/5}), \quad \text{for optimal } \sigma = C \cdot n^{-2/5}. \end{aligned}$$

In Viola (1995), no analysis with respect to the approximation order of the Monte Carlo approach for the entropy estimation is given.

For the choices of the variances in the Parzen window functions, Viola suggests a cross-validation approach based on the sample X. However, concrete formulas and the approximation order of these estimations are missing. Moreover, the assumption that the covariance matrix Σ_{R,T_a} of the Parzen window for the joint density $S = (R, T_a)^\top$ is diagonal is a severe restriction on the images under consideration.

It is worthwhile noticing that modifications of the mutual information may also be used. Studholme et al (1996) for example use

$$\mathcal{D}^{\mathrm{MIN}}[R,T] := \frac{H(\rho_{R,T})}{H(\rho_T)+H(\rho_R)}.$$

6.4.2 *Gradient of mutual information*

For the computation of the Gâteaux derivative of the Parzen window estimation-based approximation of the mutual information $\mathcal{D}^{\mathrm{MI}}[R,T;\varphi]$ with respect to a perturbation ψ, we introduce the following abbreviations:

$$\begin{aligned}
T[\varphi] &:= T\circ\varphi, \quad \sigma := \Sigma_{T[\varphi]} \in \mathbb{R}, \quad H_T[\varphi] := H(T[\varphi]),\\
g_{j,k}[\varphi] &:= g_1\big(\sigma, (T[\varphi](Y_k) - T[\varphi](X_j))\big),\\
\Sigma &:= \Sigma_{(R,T[\varphi])^\top} \in \mathbb{R}^{2\times 2}, \quad H_{R,T}[\varphi] := H((R,T[\varphi])^\top),\\
G_{j,k}[\varphi] &:= g_2\big(\Sigma, (R(Y_k)-R(X_j), T[\varphi](Y_k)-T[\varphi](X_j))^\top\big),
\end{aligned}$$

where σ and Σ are assumed to be independent on φ. Hence

$$\begin{aligned}
dT[\varphi;\psi] &= \left\langle \nabla T\big|_\varphi, \psi\right\rangle_{\mathbb{R}^d},\\
g_{j,k}[\varphi] &= \frac{1}{\sqrt{2\pi\sigma}}\cdot\exp\big(-(T[\varphi](Y_k)-T[\varphi](X_j))^2/(2\sigma)\big),\\
dg_{j,k}[\varphi;\psi] &= -g_{j,k}[\varphi]\big(T[\varphi](Y_k)-T[\varphi](X_j)\big)\cdot\sigma^{-1}\big(dT[\varphi;\psi](Y_k)-dT[\varphi;\psi](X_j)\big),\\
H_T[\varphi] &= \frac{1}{m}\sum_{k=1}^m \log\left(\frac{1}{n}\sum_{j=1}^n g_{j,k}[\varphi]\right),\\
dH_T[\varphi;\psi] &= \frac{1}{m}\sum_{k=1}^m\sum_{j=1}^n \frac{dg_{j,k}[\varphi;\psi]}{\sum_{\ell=1}^n g_{\ell,k}[\varphi]},\\
G_{j,k}[\varphi] &= \frac{1}{2\pi\sqrt{\det(\Sigma)}}\\
&\quad\cdot\exp\left(-\frac{1}{2}\begin{pmatrix} R(Y_k)-R(X_j)\\ T[\varphi](Y_k)-T[\varphi](X_j)\end{pmatrix}^\top\cdot\Sigma^{-1}\begin{pmatrix} R(Y_k)-R(X_j)\\ T[\varphi](Y_k)-T[\varphi](X_j)\end{pmatrix}\right),
\end{aligned}$$

$$\begin{aligned}
dG_{j,k}[\varphi;\psi] &= -G_{j,k}[\varphi]\begin{pmatrix} R(Y_k)-R(X_j)\\ T[\varphi](Y_k)-T[\varphi](X_j)\end{pmatrix}^\top\\
&\quad\cdot\Sigma^{-1}\begin{pmatrix} 0\\ dT[\varphi;\psi](Y_k)-dT[\varphi;\psi](X_j)\end{pmatrix},\\
H_{R,T}[\varphi;\psi] &= \frac{1}{m}\sum_{k=1}^m \log\left(\frac{1}{n}\sum_{j=1}^n G_{j,k}[\varphi]\right),\\
dH_{R,T}[\varphi;\psi] &= \frac{1}{m}\sum_{k=1}^m\sum_{j=1}^n \frac{dG_{j,k}[\varphi;\psi]}{\sum_{\ell=1}^n G_{\ell,k}[\varphi]}.
\end{aligned}$$

Assuming that $\Sigma = \Sigma_{(R,T[\varphi])^\top}$ is also diagonal, $\Sigma = \mathrm{diag}(\sigma_1,\sigma_2)$, we finally obtain

$$\begin{aligned}
d\mathcal{D}^{\mathrm{MI}}[R,T;\varphi;\psi] &= dH_T[\varphi;\psi] - dH_{R,T}[\varphi;\psi]\\
&= \frac{1}{m}\sum_{k=1}^m\sum_{j=1}^n\left(\frac{dg_{j,k}[\varphi;\psi]}{\sum_{\ell=1}^n g_{\ell,k}[\varphi]} - \frac{dG_{j,k}[\varphi;\psi]}{\sum_{\ell=1}^n G_{\ell,k}[\varphi]}\right)\\
&= \frac{1}{m}\sum_{k=1}^m\sum_{j=1}^n\left(\frac{g_{j,k}[\varphi]}{\sigma\sum_{\ell=1}^n g_{\ell,k}[\varphi]} - \frac{G_{j,k}[\varphi]}{\sigma_2\sum_{\ell=1}^n G_{\ell,k}[\varphi]}\right)\\
&\quad\cdot\big(T[\varphi](Y_k)-T[\varphi](X_j)\big)\big(dT[\varphi;\psi](Y_k)-dT[\varphi;\psi](X_j)\big).
\end{aligned}$$

If in particular the transformation φ is parametric, i.e., $\varphi=\varphi_a$, the derivatives can be computed explicitly.

Example 6.1 For an affine linear registration based on mutual information we set

$$\varphi_a(x) = \begin{pmatrix} a_1 & a_2\\ a_3 & a_4\end{pmatrix}\begin{pmatrix} x_1\\ x_2\end{pmatrix} + \begin{pmatrix} a_5\\ a_6\end{pmatrix}$$

and with

$$\begin{aligned}
&\psi_1(x) = (b_1x_1,0)^\top, \quad \psi_2(x) = (b_2x_2,0)^\top, \quad \psi_3(x) = (0,b_3x_1)^\top,\\
&\psi_4(x) = (0,b_4x_2)^\top, \quad \psi_5(x) = (b_5,0)^\top, \quad \psi_6(x) = (0,b_6)^\top,
\end{aligned}$$

$$\begin{aligned}
\Delta w_{j,k}(a) &:= \frac{g_{j,k}[\varphi_a]}{\sigma\sum_{\ell=1}^n g_{\ell,k}[\varphi_a]} - \frac{G_{j,k}[\varphi_a]}{\sigma_2\sum_{\ell=1}^n G_{\ell,k}[\varphi_a]},\\
\Delta T_{j,k}(a) &:= T(\varphi_a(Y_k)) - T(\varphi_a(X_j)),
\end{aligned}$$

we thus have

$$\partial_{a_1} D^{\mathrm{MI}}(R,T;a) = d\mathcal{D}^{\mathrm{MI}}[R,T;\varphi_a,\psi_1]$$
$$= \frac{1}{m}\sum_{k=1}^{m}\sum_{j=1}^{n}\Delta w_{j,k}(a)\Delta T_{j,k}(a)\big(\partial_{x_1}T(\varphi_a(Y_k))Y_{k,1} - \partial_{x_1}T(\varphi(X_j))X_{j,1}\big),$$

$$\partial_{a_2} D^{\mathrm{MI}}(R,T;a) = d\mathcal{D}^{\mathrm{MI}}[R,T;\varphi_a,\psi_2]$$
$$= \frac{1}{m}\sum_{k=1}^{m}\sum_{j=1}^{n}\Delta w_{j,k}(a)\Delta T_{j,k}(a)\big(\partial_{x_1}T(\varphi_a(Y_k))Y_{k,2} - \partial_{x_1}T(\varphi(X_j))X_{j,2}\big),$$

$$\partial_{a_3} D^{\mathrm{MI}}(R,T;a) = d\mathcal{D}^{\mathrm{MI}}[R,T;\varphi_a,\psi_3]$$
$$= \frac{1}{m}\sum_{k=1}^{m}\sum_{j=1}^{n}\Delta w_{j,k}(a)\Delta T_{j,k}(a)\big(\partial_{x_2}T(\varphi_a(Y_k))Y_{k,1} - \partial_{x_2}T(\varphi(X_j))X_{j,1}\big),$$

$$\partial_{a_4} D^{\mathrm{MI}}(R,T;a) = d\mathcal{D}^{\mathrm{MI}}[R,T;\varphi_a,\psi_4]$$
$$= \frac{1}{m}\sum_{k=1}^{m}\sum_{j=1}^{n}\Delta w_{j,k}(a)\Delta T_{j,k}(a)\big(\partial_{x_2}T(\varphi_a(Y_k))Y_{k,2} - \partial_{x_2}T(\varphi(X_j))X_{j,2}\big),$$

$$\partial_{a_5} D^{\mathrm{MI}}(R,T;a) = d\mathcal{D}^{\mathrm{MI}}[R,T;\varphi_a,\psi_5]$$
$$= \frac{1}{m}\sum_{k=1}^{m}\sum_{j=1}^{n}\Delta w_{j,k}(a)\Delta T_{j,k}(a)\big(\partial_{x_1}T(\varphi_a(Y_k)) - \partial_{x_1}T(\varphi(X_j))\big),$$

$$\partial_{a_6} D^{\mathrm{MI}}(R,T;a) = d\mathcal{D}^{\mathrm{MI}}[R,T;\varphi_a,\psi_6]$$
$$= \frac{1}{m}\sum_{k=1}^{m}\sum_{j=1}^{n}\Delta w_{j,k}(a)\Delta T_{j,k}(a)\big(\partial_{x_2}T(\varphi_a(Y_k)) - \partial_{x_2}T(\varphi(X_j))\big),$$

where $X_j = (X_{j,1}, X_{j,2})^\top$ and $Y_k = (Y_{k,1}, Y_{k,2})^\top$, respectively.

Wells et al (1996) proposed a gradient-based steepest descent method with fixed step size α for the minimization of D; cf., Algorithm 6.8.

Algorithm 6.8 *Stochastic maximization algorithm of Wells et al (1996).*
Repeat:
Collect sample $X = (X_1, \ldots, X_{n_x})$ from Ω.
Collect sample $Y = (Y_1, \ldots, Y_{n_y})$ from Ω.
Update $a \mapsto a + \lambda \nabla_a D(a)$.

The steepest descent method in Algorithm 6.8 is based on different samples in different iteration steps. This implies that the objective function changes from step to step. A convergence proof for the modified algorithm is missing. Finally, as is well-known, the convergence rate of the steepest descent method might be arbitrarily slow, a fact that has already been observed for quadratic optimization problems; cf., e.g., Fletcher (1987).

Our implementation is based on a Levenberg–Marquardt-type technique; cf., Algorithm 6.9. To this end, let $f : \mathbb{R}^n \to \mathbb{R}$, $f(a) := -D^{\mathrm{MI}}(R,T;\varphi_a)$. For a given a and λ, we replace f by the quadratic model

$$q_a(\Delta a) := f(a) + \nabla f(a)\Delta a + \tfrac{1}{2}\Delta a[\nabla f(a)\nabla f(a)^\top]\Delta a,$$

where $[\nabla f(a)\nabla f(a)^\top]$ gives an approximation to the Hessian matrix $\nabla^2 f(a)$.

Remark 6.1 Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a function and $h : \mathbb{R}^n \to \mathbb{R}$ be defined by $h(a) = \frac{1}{2}\|f(a)\|^2_{\mathbb{R}^m}$. The $(p,q)^{th}$ entry of the Hessian matrix of h is given by

$$\frac{\partial^2 h}{\partial a_p \partial a_q}(a) = \sum_{j=1}^{m}\left(\frac{\partial f_j}{\partial a_p}(a)\frac{\partial f_j}{\partial a_q}(a) + f_j(a)\frac{\partial^2 f_j}{\partial a_p \partial a_q}(a)\right).$$

If $\|f(a)\|_{\mathbb{R}^m}$ is small enough, we have $\nabla^2 h(a) \approx \nabla f(a)\nabla f(a)^\top$; see also Nocedal & Wright (1999, §10.2).

The next step is the minimization of the quadratic model, subject to $\|\Delta a\|_{\mathbb{R}^n} \le h$. The restriction results in

$$[\nabla f(a)\nabla f(a)^\top + \lambda I_n]\Delta a = -\nabla f(a),$$

where h and λ are related, and for a meaningful choice of h we have that $[\nabla f(a)\nabla f(a)^\top + \lambda I_n]$ is positive definite; see, e.g., Fletcher (1987, §5.2). The next iterate is given by $a' := a + \Delta a$. We denote the iterates in the k^{th} iteration by $a^{(k)}$ and $\lambda^{(k)}$, respectively. If $f(a^{(k+1)} \ge f(a^{(k)})$, the k^{th} step is not successful. In this situation, we increase the value of $\lambda^{(k)}$, i.e., shrink the size h of the trust region in our model; see Algorithm 6.9 for details.

Note that the objective function f also depends on the samples X and Y which change from step to step. Thus, we measure convergence numerically by the variance of the last m parameters $a^{(k-m+1)}, \ldots, a^{(k)}$, where k denotes the iteration and typically $m = 10$.

6.5 An example of mutual information-based affine linear registration

The result of a mutual information-based optimal affine linear registration is shown in Fig. 6.3. Here, the intensity values of the template image have been modified. The gray values of the initial image have been re-scaled and inverted.

Algorithm 6.9 *Minimization of Parzen window approximation of mutual information* $f(a) := -D^{\mathrm{MI}}(R, T; \varphi_a)$ *with respect to the parameter* a *using a Levenberg–Marquardt-type technique.*

Set $\mu = 1$, set $\beta > 1$, e.g., $\beta = 5$. Choose initial $a \in \mathbb{R}^n$, e.g., $a = (1, 0, 0, 1, 0, 0)^\top$ for affine linear registration.
Repeat
  Compute $f(a)$ and $\nabla f(a)$, set $\lambda = \mu$.
  Compute Δ_a from $[\nabla f(a) \nabla f(a)^\top + \lambda I_n] \Delta_a = -\nabla f(a)$.
  While $f(a + \Delta a) \geq f(a)$,
    $\lambda = \beta\lambda$.
    Compute Δ_a from $[\nabla f(a) \nabla f(a)^\top + \lambda I_n] \Delta_a = -\nabla f(a)$,
  end.
  Set $\mu = \lambda/\beta$, $a = a + \Delta_a$.
Until convergence.

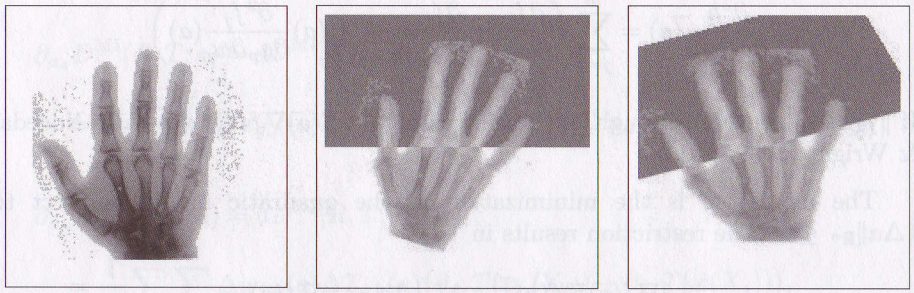

FIG. 6.3 Optimal mutual information-based affine linear registration; reference (LEFT), template (MIDDLE), template after registration (RIGHT).

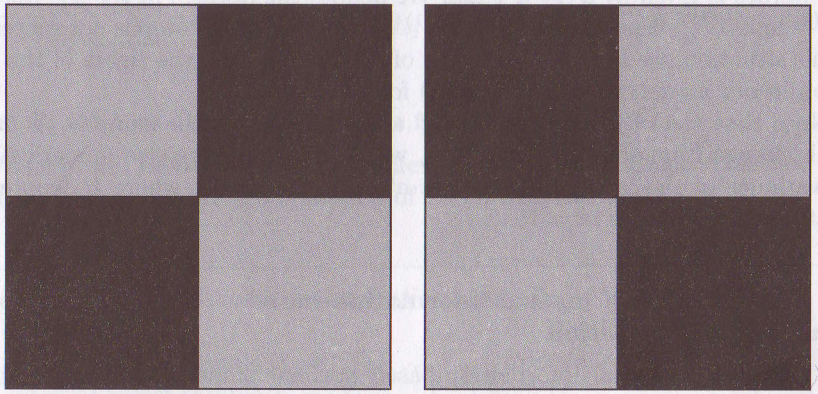

FIG. 6.4 Intensity and mutual information-based registration.

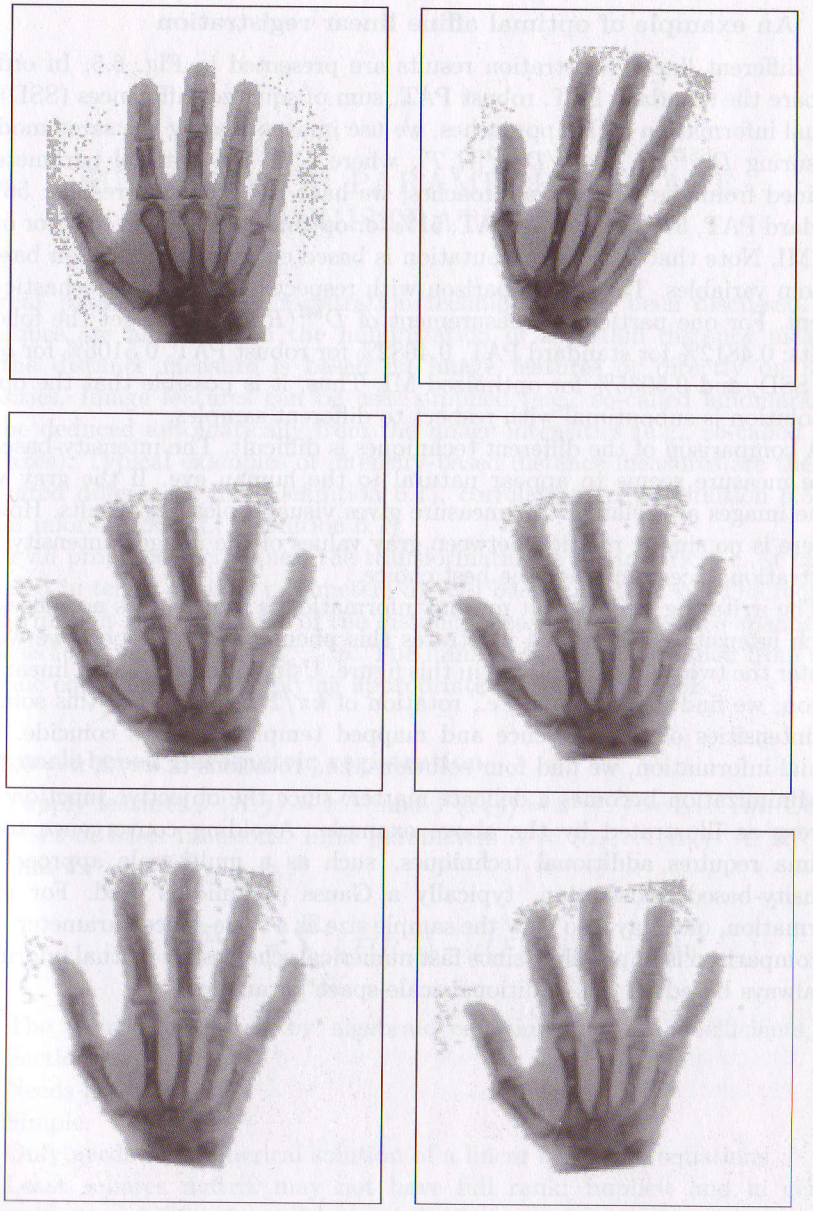

FIG. 6.5 Four different linear registrations: reference (TOP LEFT), template (TOP RIGHT), standard PAT registration, (MIDDLE LEFT), robust PAT registration (MIDDLE RIGHT), affine linear SSD registration (BOTTOM LEFT), and affine linear MI registration (BOTTOM RIGHT).

6.6 An example of optimal affine linear registration

Four different linear registration results are presented in Fig. 6.5. In order to compare the standard PAT, robust PAT, sum of squared differences (SSD), and mutual information (MI) approaches, we use images sharing the same modality. Measuring $D^{\mathrm{SSD}}(R,T;a^*)/\mathcal{D}^{\mathrm{SSD}}[R,T]$, where a^* is the optimal parameter set obtained from the different approaches, we have the following results: 56% for standard PAT, 57% for robust PAT, 51% for optimized SSD, and 62% for optimized MI. Note that our MI computation is based on an approximation based on random variables. Thus a comparison with respect to MI has a stochastic component. For one particular measurement of $D^{\mathrm{MI}}(R,T;a^*)$ we get the following results: 0.4812% for standard PAT, 0.4682% for robust PAT, 0.5106% for optimized SSD, and 0.5035% for optimized MI. Thus, it is possible that the optimal MI solution is suboptimal with respect to different samples.

A comparison of the different techniques is difficult. The intensity-based distance measure seems to appear natural to the human eye. If the gray values of the images are related, this measure gives visually pleasing results. However, if there is no simple relation between gray values of the images, intensity-based registration is certainly not the best choice.

The irritating point about mutual information is that it does not necessarily match intensities. Figure 6.4 illustrates this phenomenon. Suppose we want to register the two images displayed in this figure. Using intensity-based linear registration, we find two minima, i.e., rotation of $k\pi/2, k = 1, 3$. For this solutions, the intensities of the reference and mapped template images coincide. Using mutual information, we find four solutions, i.e., rotations of $k\pi/2$, $k = 0, 1, 2, 3$.

Minimization becomes a delicate matter, since the objective function is not convex, as illustrated by the above example. Avoiding convergence to local minima requires additional techniques, such as a multi-scale approach. For intensity-based registration, typically a Gauss pyramid is used. For mutual information, one may also view the sample size as a scale-space parameter. A direct comparison is impossible, since fast numerical schemes for mutual information are always based on this additional scale-space parameter.

7

SUMMARY OF PARAMETRIC IMAGE REGISTRATION

Different parametric image registration techniques have been discussed. The techniques are all based on the minimization of a certain distance measure, and the distance measure is based on image features or directly on image intensities. Image features can be user supplied (e.g., so-called landmarks) or may be deduced automatically from the image intensities (e.g., so-called principal axes). Typical examples of intensity-based distance measures are the sum of squared differences (cf., Definition 6.1), correlation (cf., Definition 6.5), or mutual information (cf., Definition 6.7).

For all proposed techniques, the transformation is parametric, i.e., it can be expanded in terms of some parameters α_j and basis functions ψ_j. The required transformation is a minimizer of the distance measure in the space spanned by the basis functions ψ_j, $j = 1, \ldots, n$. The minimizer can be obtained from some algebraic equations or by applying appropriate optimization tools.

Landmark-based parametric registration

- Supply features $\mathcal{F}(R,j) = x^{R,j}$ and $\mathcal{F}(T,j) = x^{T,j}$, $j = 1, \ldots, m$. Choose a set of basis functions. Find parameters $\alpha = (\alpha_1, \ldots, \alpha_n)^\top \in \mathbb{R}^n$ such that for $\varphi = \sum_{j=1}^n \alpha_j \psi_j$,

$$\mathcal{D}^{\mathrm{LM}}[\varphi] = \sum_{j=1}^{m} \|\mathcal{F}(R,j) - \varphi(\mathcal{F}(T,j))\|_f = \min.$$

  The solution is given by algebraic equations for the coefficients; cf., Section 4.2.
- Needs landmarks.
- Simple.
- Only needs the numerical solution of a linear system of equations.
- Least squares matrix may not have full rank; implicit and in general unknown additional conditions on the features.
- Results might be arbitrarily awful.

Landmark-based smooth registration

- Supply features $\mathcal{F}(R,j) = x^{R,j}$ and $\mathcal{F}(T,j) = x^{T,j}$, $j = 1, \ldots, m$. Choose a regularizer $\mathcal{S}^{\mathrm{TPS}}$ (cf., eqn (4.22)) and a regularizing parameter $\alpha \geq 0$.

For $\alpha = 0$, find φ such that

$$\mathcal{S}^{\mathrm{TPS}}[\varphi] = \min \text{ subject to } \varphi(\mathcal{F}(T, j)) = \mathcal{F}(R, j), \quad j = 1, \ldots, m.$$

Alternatively, for $\alpha > 0$, find φ such that

$$\alpha \mathcal{S}^{\mathrm{TPS}}[\varphi] + \mathcal{D}^{\mathrm{LM}}[\varphi] = \min .$$

The solutions are given by algebraic equations for the coefficients in a radial basis expansion; cf., Section 4.3.4.

- Needs landmarks.
- Only needs the numerical solution of a linear system of equations, essentially m unknowns and m equations; system is always non-singular.
- Physically meaningful transformation, minimizes curvature.
- Results may be bad.

Principal axes-based registration

- Compute $\varphi \in \Pi_1^d(\mathbb{R}^d)$, such that $\mathcal{F}(T \circ \varphi) = \mathcal{F}_R$, where $\mathcal{F}_B$ is a feature vector containing the center of gravity, the standard deviations, and the principal axis based on an appropriate density class, e.g. Gauss or Cauchy density; cf. Chapter 5. A solution can be deduced from an eigenvalue decomposition of the moment matrices of R and T, respectively; cf., Theorem 5.2.
- Simple, fast, easy to understand and to interpret.
- Needs moment matrix and eigenvalue decompositions of two d-by-d matrices.
- Not suitable for multimodal densities/images.
- Ambiguous results.
- Very few registration parameters.

Optimal parametric registration

- Choose an appropriate distance measure $\mathcal{D}$. Choose a set of basis functions. Find parameters $\alpha = (\alpha_1, \ldots, \alpha_n)^\top \in \mathbb{R}^n$ such that for $\varphi = \sum_{j=1}^n \alpha_j \psi_j$,

$$\mathcal{D}[\varphi] = \min .$$

A numerical solution can be obtained by using optimization methods, e.g., a Gauss–Newton or Levenberg–Marquardt method.

- General, flexible.
- No physical, meaningful transformation.
- Optimization can be very slow, in particular for high-dimensional spline spaces.

Part II

Non-parametric image registration