1.3.1 Image modality: Monomodality versus multimodality and dimensions

Imaging devices are tools designed to present in some graphical form information related to hidden parts of the body. For instance, a radiography plate makes use of the different attenuation index of the tissues when exposed to a source of X–rays. A related modality is the CT or CAT (Computer Axial Tomography), which is able to extract a section perpendicular to the body. Consecutive slices are separated by a distance called inter-slice.

CT is classified as anatomical modality, because it depicts primarily the morphology of the tissues. Other main anatomical modalities are the MRI (Magnetic Resonance Imaging), and ecographies, which are normally seen as video sequences. Functional modalities depict primarily the metabolism of the underlying tissues, e.g. the consume of oxygen or glucose, and permits to distinguish areas morphologically identical. Functional modalities are SPECT (single-photon emission computed tomography) and PET (positron emission tomography), and also some anatomical modalities when an opaque fluid is injected and its absorption studied.

When images belong to the same modality, the registration is called monomodal, otherwise, it is called multimodal. Monomodal registrations appeared before in literature because in general are easier to perform. Images of the same modality will be consecutive in the time, and will show only few differences of interest for the physician, while the rest of the image will usually be similar. Therefore, most equivalent pairs of registered pixels will have similar intensities, perhaps multiplied by a factor to count for global illumination changes.

Time gaps between the images depend on the modality. For normal scanner images, it may be between days to months, because we are assessing the evolution of an illness, or perhaps the result of an intervention.

A different case is the video sequence of up to forty frames per second, which can be used for two different purposes:

assessing changes in the area imaged For instance, a visible fluid has been injected to the patient, and the sequence studies the speed and magnitude of the changes in the vessel tree as it is being propagated. This case is further discussed for SLO, in chapter 3.

volume compounding The device is moving an imaging consecutive slices of the tissue. If the device is calibrated to some external coordinates system, slices can be compounded to form a volume. This is the case of ultrasound image, addressed in chapter 4.

1.3.2 Image contents: imaged area and patient

Head images are more commonly found in literature because the skull is a rigid object, which means that the transformation between the images can be restricted to the rigid type. Another cause is that neuroscience is an upcoming branch, with many new applications and problems. Usual modalities are CT and MR , and not so commonly SPECT and PET.

Of course, the same modalities can be applied to any part of the patient. Maxilar, neck, heart, abdomen and pelvis are commonly screened, and each usually demands an algorithm specially designed.

Ophthalmologic images depict the vessel structure of the fundus of the eyes, and include retinographies (taken with a green filter to discard the red component), angiographies (with a contrast agent opaque to the X-rays) and SLO sequences. Chapter 3 is devoted to ophthalmologic images.

Ecographic or ultrasound images suit specially well for real-time imaging. It is a relatively low cost sensor without over-exposition cautions, and has been miniaturised to access the narrowest regions of the body, such as the vessels. It is also used in surgery for a fast imaging of the underlying structures. Normally images are not taken in single, but rather forming a video sequence interpreted by specialists.

Despite its low signal-to-noise ratio, this modality is gaining popularity in upcoming papers because it is easily employed in the operating theatre, where other more accurate modalities like CT have a prohibitive cost. Chapter 4 is entirely devoted to this aim.

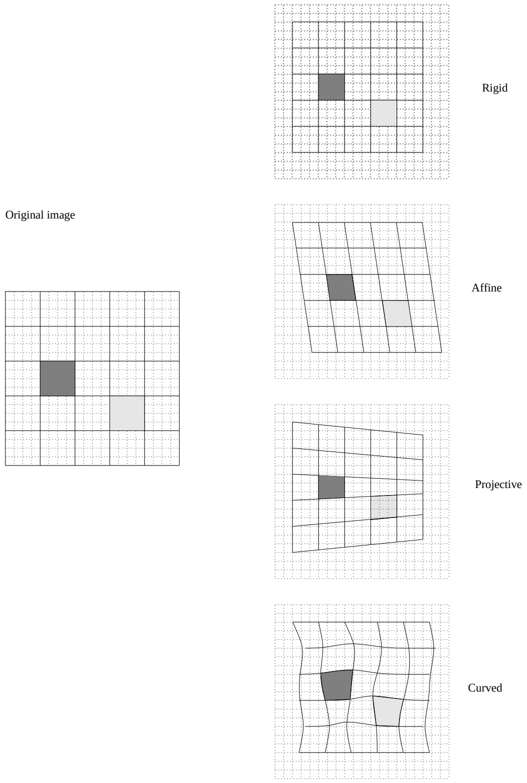

Figure 1.6: Types of global transformations.

1.3.3 Transformation model

In the registration process, one image is called dynamic and it is iteratively transformed until aligned to the other one, called static. A transformation is defined as a mapping of location of points in one image I_1 to new locations in another image I_2.

$$I_2(x,y) = I_1(\mathcal{T}(x,y)) \tag{1.9}$$

This definition extends to any number of dimensions.

There are many types of transformation. The global transformation applies the same equation to all pixels in the image; the equation is usually written in form of matrix multiplication, as shown in equation 1.11.

$$\begin{aligned} C_i &= C_o * M_t && (1.10) \\ \begin{pmatrix} x_i^h & y_i^h & z_i^h & w \end{pmatrix} &= \begin{pmatrix} x_o & y_o & z_o & 1 \end{pmatrix} \left(\begin{array}{ccc|c} & & & 0 \\ & r & & 0 \\ & & & 0 \\ \hline t_x & t_y & t_z & 1 \end{array}\right) && (1.11) \\ C_i &= \begin{pmatrix} x_i & y_i & z_i \end{pmatrix} \equiv \frac{\begin{pmatrix} x_i^h & y_i^h & z_i^h \end{pmatrix}}{w} && (1.12) \end{aligned}$$

The formula reads as following: given an image and a transformation matrix, each coordinate in the resulting image $C_o = (x_o\ y_o\ z_o)$ takes values somewhere at the original image $C_i = (x_i\ y_i\ z_i)$. In general, since C_i will not be an integer values, the value of the image at C_i will have to be interpolated between neighbouring pixels.

Figure 1.6 shows graphically the usual types of transformation. They are defined by a composition of simpler transformation matrices, found in any image analysis book [22].

Rigid map right angles into right angles, and are composition of rotations, translations and scaling.

Affine map parallel lines into parallel lines, and in addition to those of rigid, may have shearing transformations.

Projective map lines into lines.

The curved transformation is a special case, because it can't be represented with a single matrix. It may be represented in a complex form, as a polynomial, and also as a set of *local* transformation matrices. Each apply to a region in the image, arranged in such a way that borders are continuous. We have employed local transformation for correcting distortions in ophthalmologic images. See appendix D for a brief definition, and section 3.5 for a discussion of its benefits and drawbacks.

The transformation should be implemented to be as fast as possible, because it will run once per iteration of the optimisation step. For each pixel transformed from the dynamic image, the cost can be divided in: a) computing the resulting coordinates C_i and b) estimate the value at C_i. The step b) is necessary because in general the resulting coordinates will not have an integer value, where the image is originally defined. The faster interpolation scheme is to take the value of the nearest neighbour; next in speed is to perform bilinear (or trilinear, for 3–D images) interpolation of neighbouring values. More complex forms achieve higher accuracy by means of sync kernels or higher degree polynomials.

Grevera [26] classifies interpolation methods into two groups: scene-based and object-based. Scene-based methods make a straight use of the values of neighbouring pixels, while object-based methods are sensitive to the contents and try to preserve some quality, e.g. shape. In his paper, he compares the estimated data of eight methods to the actual contents, for four volumes depicting different zones.

Although the fastest, the nearest neighbour scheme is usually discarded because it does not warranty the continuity property 1.3 of the comparison function M. According to the definition of nearest neighbour, the resulting image will take the value at the truncated transformed coordinates:

$$D(\mathcal{T}(x,y,z)) = D(\lfloor x' \rfloor, \lfloor y' \rfloor, \lfloor z' \rfloor) \tag{1.13}$$

And therefore M will actually be insensitive to translations up to one pixel.

In our implementations, we have always chosen the linear interpolation because since our algorithms are sensitive to the position rather than the precise value of the pixel, it is not necessary to used more complex methods.

Another interpolation scheme, necessary for voxel-based methods, is explained in section 1.3.5.

1.3.4 Comparison paradigm

We need to measure how similar are two given images or, equivalently, given one image A and two transformations of another image B, we need to decide which transformation is closer to A.

The choice of the similarity metric is the most relevant part of the algorithm. Registration methods can be broadly classified, according to this criteria, as:

- Extrinsic: visible markers are attached to the patient prior to the image acquisition.
- Intrinsic: make solely use of pixels imaging the patient.
  - Segmentation based: only relevant features of the patient are used for comparison.
  - Voxel property based: the full content of the image is used at the comparison step.

Extrinsic methods are often taken for comparison against the others (Golden Standard) because markers provide an easy mean of registration, and the expected accuracy has also been well studied. Read, for instance, the paper [16] for a through study of these methods.

The following sections give more details of the two intrinsic methods, segmentation based and voxel property based.

Segmentation-based methods

Perhaps the most intuitive approach is to select pairs of equivalent features from both images. Identifiable features used for this purpose are called landmarks. Manual extraction of landmarks has been the sole procedure available for some years in medical imaging. In this procedure, a human operator is trained to recognise some sort of landmarks in the images:

- artificial marks attached to the patient. Then the method is extrinsic, because it uses information not belonging to the patient. Extrinsic methods have several disadvantages: first, they are not retrospective and thus they can not be registered to images not containing the same markers. Second, they usually are unfriendly to the patient. An example is the stereotactic frame, a metallic box screwed physically to the skull of the patient used for stereotactic surgery.

  Other markers, non invasive for the patient, can be glued to the skin or attached to head or dental moulds. Once extracted, replacing them in the same position can be difficult because of the elasticity of the body and, again, the high accuracy demanded. For this reason, usually the position given by the user is not taken, and instead the centroid of the neighbourhood is computed, with sub-pixel accuracy.

- Single anatomical landmarks, enhanced in the imaging modality. For instance, for stereotaxis surgery the operator must identify the anterior and posterior commissure, whose position afterwards permits to adjust coordinates within the images to those of a standard atlas. However, this identification can be very difficult, and sometimes not possible, if landmarks have been skipped in between the slices, or largely blurred because the patient moved. Since these methods use only information from the patient, they are called intrinsic.

- Lines or segments enfolding organs. Usually those belonging to the conjunctive tissue, e.g. the dermis of the skin, are easy to segment because their content in fat makes them clearly visible. However, small errors in the segmentation may greatly affect associated measures, like area: a 1-pixel error in a 20×20 rectangle makes a 10% error. Surfaces and volumes are segmented repeating the procedure at consecutive slices.

Equivalent anatomical landmarks appearing in two images can be very difficult to relate with precision. Indeed, images may belong to different modalities, may have different resolution or may not display the feature in the same state, i.e., tumour volumes change between the acquisitions.

Therefore, the graphical interface must be carefully designed to facilitate operations such as zooming, changes in intensity mapping and labelling. Such a repetitive tasks often demand a strong motivation of the operator for results to be of any use, specially in fields demanding high accuracy such as neurosurgery.

An alternative is to make the segmentation an automatic part of the registration algorithm. Artificial marks are easy to detect, specially if they follow some geometric pattern. But this is not the case of anatomical landmarks. If single target points are

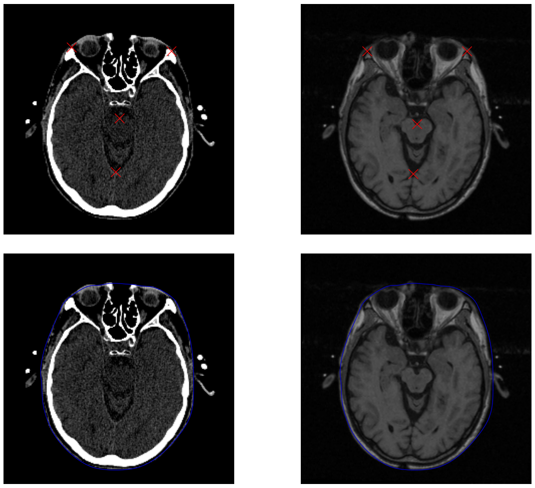

Figure 1.7: Segmentation of individual landmarks (top) or surfaces (bottom) can be a difficult and tedious task even for trained users.

to be found, the algorithm must deal with variations both in patient anatomy and in image conditions. For this reason, anatomical landmarks are usually pointed out interactively.

Once the image has been segmented, the problem is to find the transformation α minimising its mean distance. When the segmentation is composed by two lists of k corresponding landmarks pairs, $P_A = \{A_i\}, P_B = \{B_i\}$, this is a classic mathematical problem known as the Orthogonal Procrustes problem and resolved e.g. [2] by means of singular values decomposition.

$$Dist(P_A, P_B) = \sum_k |(A_i - T_\alpha(B_i))|^2 \tag{1.14}$$

for each point,

$$Res_i = |(A_i - T_\alpha(B_i))| \tag{1.15}$$

gives the error associated to each point. In general, the distance will not be zero and it is interesting for the user to see the contribution of each point to the general error.

A different case is that of surface segmentation. As seen in figure 1.7, the comparison is not made between individual points but between the global shape of the figures. Automatic segmentation algorithms must deal with at least the same difficulties as the user does:

- Sensitivity to intensity changes, both global and local. For instance, simple threshold will often fail because some areas in the body may not give proper signal in the image, even for monomodal series.

- Different appearance for different modalities: some areas appearing in one modality may not appear at all in the other.
- Specificity to the imaged area: the segmentation procedure will not be portable to other areas.
- Changes in the content: images taken after an operation show very different shapes, due to the craniotomy.

One criticism often made to segmentation-based algorithms is that results will be at most as good as the segmentation errors permit, i.e., errors in the segmentation determine the final accuracy. While this statement might be true for individual landmarks, we think it doesn't hold for surface matching as long as there is enough consistent information.

That means that the optimisation step must be designed to deal with occasional failures, originated by the effects in the previous list: a number of segments will match to void areas, while others wrongly match non-corresponding segments.

A landmark of choice for head image registration is the skull: since it is a rigid structure, its transformation fulfils the rigid assumption. This idea was first proposed by van den Elsen in [102]. However, the detection of the skull is not a trivial task, much less with the demanded sub-pixel accuracy. For this purpose she made use of the geometric properties of the bone in the images: in CT , it produces a strong signal, while in MR its lack of mobile protons produces weak signal. Therefore, to segment valid landmarks one must detect ridges in the CT and valleys in the MR .

There are many definitions of creases, some based on differential geometry. Since the registration is based on the accuracy of the segmentation step, creaseness detectors should give an accurate, continuous and robust response. The paper [56] studies the performance of several definitions for conditions similar to those of medical imaging and, despite their failure for a number of cases, it concludes they are equally useful for registration purposes.

Our registration algorithm was initially based on the scheme proposed by van den Elsen, and after some modification we used it for CT to MR image registration. The algorithm is fully described in chapter 2, chapter 3 for the 2–D case and chapter 4 for the 2D–3D case.

Once the surfaces have been extracted, the next step is to search for the transformation which best aligns hem. In this case we do not have pairs of corresponding points, but rather we are interested in minimising the mean distance between the surfaces. A popular algorithm is the Iterative Closest Point (ICP), described in section 1.3.6.

Voxel-based methods

Methods belonging to this family rely on the following idea: a modality can be modelled as function with domain at a set of tissues and image scalar values. The object imaged I can be seen as a non-overlapping set of points, each set belonging to a different tissue.

$$I = \cup\{t_1, t_2, \cdots, t_n\} \tag{1.16}$$

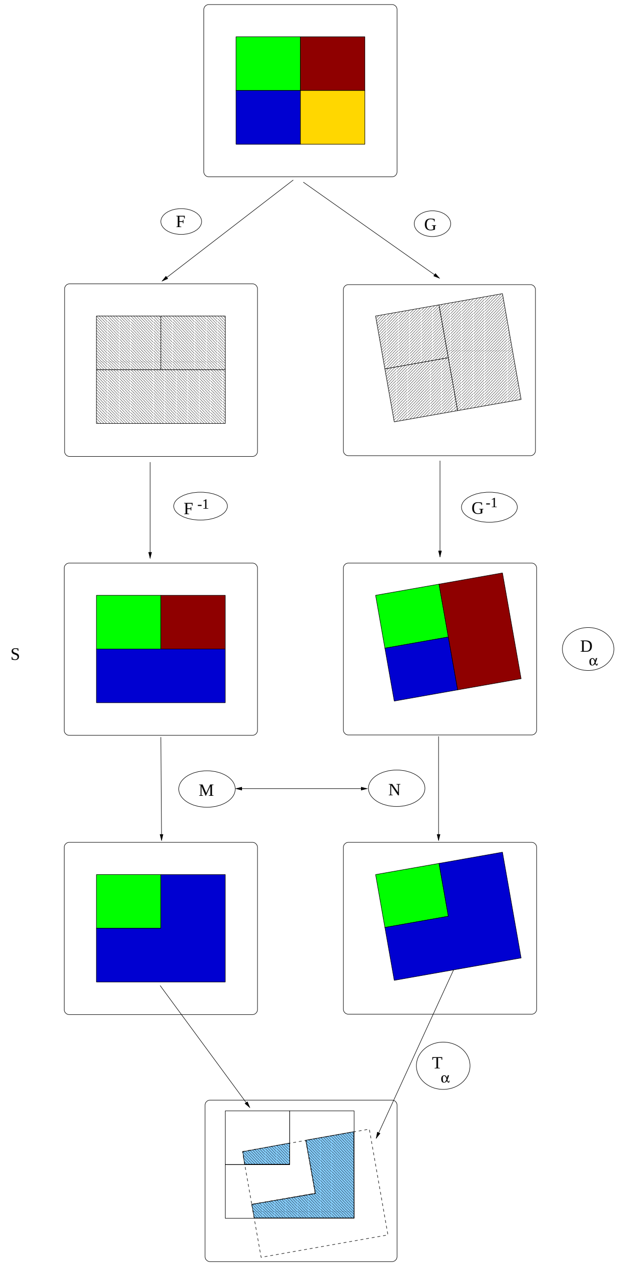

Figure 1.8: Voxel-based registration scheme: for each modality, a voxel is assigned to a region according to its grey level. Aligning then consists in maximising the intersection of the two sets of regions.

And the two modalities take values on I:

$$\begin{aligned} F :& I \to \{F(t_1), F(t_2), \cdots, F(t_n)\} \equiv I_F \\ G :& I \to \{G(t_1), G(t_2), \cdots, G(t_m)\} \equiv I_G \end{aligned} \tag{1.17}$$

The set of pixels $F(t_i)$ will have grey levels within a certain interval. In general, n and m will be different numbers and F and G will not be injective ($F(t_i) = F(t_j), i \neq j$). Should we be able to find the inverse functions F^{-1} and G^{-1}, we could build for each image S and D a set of regions:

$$\begin{aligned} F_{-1} &= \cup\{F^{-1}(I_F)\} \\ G_{-1} &= \cup\{G^{-1}(I_G)\} \end{aligned} \tag{1.18}$$

Then after mapping each set into a common reference:

$$\begin{aligned} M :& \{F_{-1}\} \to \{F'_{-1}\} \\ N :& \{G_{-1}\} \to \{G'_{-1}\} \end{aligned} \tag{1.19}$$

Registration occurs at

$$\max_{\alpha} \ \{M(F_{-1})\} \cap \{N(G^{\alpha}_{-1})\} \tag{1.20}$$

where G^{α}_{-1} is the mapping of dynamic image transformed by the parameters α

If F and G are the same function, thus the two images belong to the same modality, we may skip the rest of the formulae and simple compare the grey level of each individual pixel. Note, however, than even images of the same object taken with the same device may differ when taken along the time, because the distortions in the image brightness, caused by a low-frequency inhomogenity field, vary.

Given two images S and D (for static and dynamic), taking at each pixel $c = \{i, j, k\}$ the values S_c and D_c, we define the following alignment measures:

Correlation

$$\underset{\alpha}{\mathrm{COR}} \ (S, D) = \sum_c S_c \ D_c \tag{1.21}$$

Normalised correlation

$$\underset{\alpha}{\mathrm{NCOR}} \ (S, D) = \frac{\sum\limits_c (S_c - \overline{S}) \ (D_c - \overline{D})}{\sqrt{\sum\limits_c (S_c - \overline{S})^2} \ \sqrt{\sum\limits_c (D_c - \overline{D})^2}} \tag{1.22}$$

where $\overline{S}$, $\overline{D}$ are the mean values of the images at the intersection. And S_c stands for the value of S in the pixel coordinates c.

Sum of absolute differences

$$\underset{\alpha}{\mathrm{SAD}}\ (S,D)=\sum_{c}|S_c-D_c| \tag{1.23}$$

When the two images belong to the same modality and they contain Gaussian noise, the correlation values for a transformation will be proportional to the alignment. Many papers exploit additional properties of the correlation at the Fourier domain: [1, 41, 4, 91] expand the Fourier-shift theorem to achieve invariance properties of the cross-correlation for translation, rotation and scaling. Unfortunately, these properties are not easily extended to 3–D volumes because they work under polar coordinates.

Cross-correlation and related measures do not work well for multimodality images, because there is no linear correlation between the grey levels of corresponding voxels. First attempts to overcome this problem consisted of simulating a modality with data from the other. Then the two images were similar, and the measures listed above could be computed. In [103], van den Elsen proposes a mapping from the CT image to the MR image: background voxels are mapped to zero intensity, soft tissues have a linear ascending scale, soft bone has a linear descending scale and hard bone is mapped to zero intensity. Results for two pairs of images seem comparable or even better than those of skin markers. See figure 1.9 for a sample image of this idea.

The papers from Wells et al. [112] and Maes et al [54] extended this idea in a much powerful way. They included the estimation of the correspondence between the two image as parameters of the search algorithm. Images are considered as channels of information, and then registration becomes maximisation of the mutual information. A major advantage of this approach is its generality: in theory, it can be applied without modification to any combination of modalities and to any parts of the body. They have become seminal papers for many others published afterwards; they are further explained in the next section.

1.3.5 Short description of the mutual information method

Mutual information methods present the generic scheme in figure 1.8. The key idea is to measure the dispersion of the mutual 2–D histogram, which counts the occurrences of pairs of values (s, d), one for each image. Hill and colleagues' first idea [92] was successful but required the manual specification of relevant histograms regions. A related paper was presented by Woods and colleagues in [117]. In this section, we will follow the scheme presented by Maes in [54], which is mathematically equivalent to that of Wells [112], but permits a much simpler implementation.

Images S and D are modelled as random variables with probability distributions $p_S(s)$ and $p_D(d)$ and joint probability distributions $p_{SD}(s, d)$. The mutual information measures the distance between the later and the join distribution associated to their complete independence, $p_S(s) \cdot p_D(d)$.

$$I(S,D)=\sum_{s,d}p_{SD}(s,d)\log\frac{p_{SD}(s,d)}{p_S(s)\,p_D(d)} \tag{1.24}$$

In the image processing field, $p_{SD}(s, d)$ is defined as the normalised number of

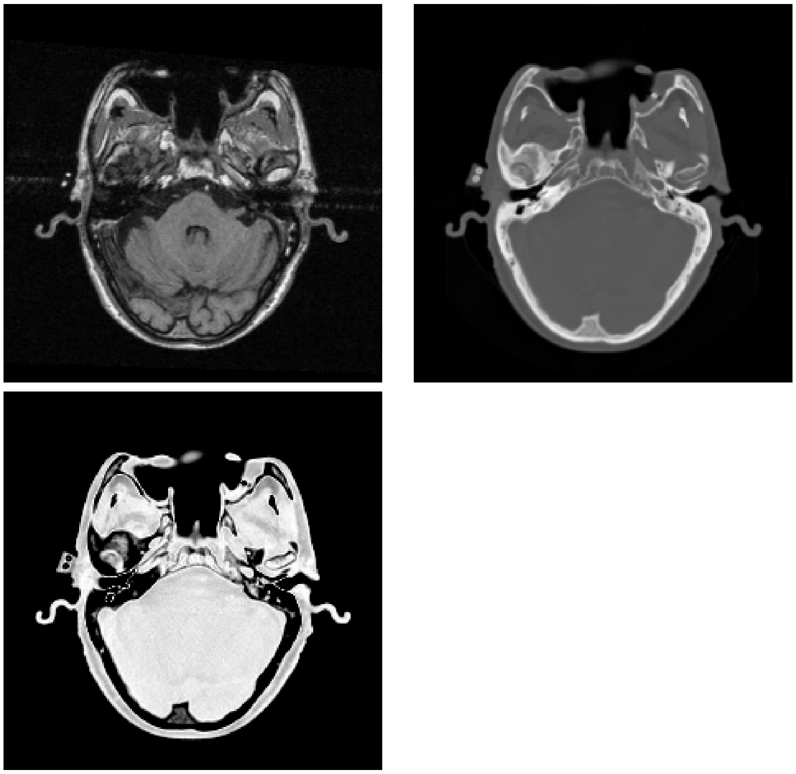

Figure 1.9: CT image (top right) with grey values remapped (bottom) to resemble MR (top left) as proposed in [103].

occurrences of the value s of a pixel $\mathbf{x}$ in S overlapping its corresponding pixel in D with value d, i.e., the normalised histogram of the overlapping part of S and D.

The meaning of I can be explained in terms of entropy. The entropy of variable, $H(S)$, measures the amount of uncertainty about the variable, while $H(S|D)$ is the uncertainty left when D is known. Entropy is related to M.I. by the following equations:

$$\begin{aligned} I(S,D) &= H(S) + H(D) - H(S,D) && (1.25) \\ &= H(S) - H(S|D) && (1.26) \end{aligned}$$

This can be though as following. For the CT and MR case, we know that high intensity values in CT, mostly due to the bone tissue, are likely to be mapped to low intensity values in MR. Therefore, if we know that a pixel has low intensity in MR , the uncertainty of its values in the CT image is greatly reduced. The MI measure models this statement, but since it does not make limiting assumptions, it can potentially use all the information available in both images.

Another related measure $Y(S,D)$ is called *normalising mutual information*, and it is the one we have used in this paper for comparison purposes because it is less sensitive to the size of the area overlapped by the two images. For a complete review of all these measures, see for instance [94].

$$Y(S,D) = \frac{H(S) + H(D)}{H(S,D)} \tag{1.27}$$

It is interesting to note that the measure is highly sensitive to the method used to interpolate the transformed image. The reason is, the actual value of the voxel is not meaningful, but its relationship to the tissue it represents. A linear interpolation method will create continuous but not real values when trying to estimate non existing voxels. For instance, zooming out a zone in a CT image containing air (0 grey level) and bone (1200) will create 500-intensity tissues which did not exist in the original image.

The problem was solved by rearranging the transformation algorithm and the computing of the histogram. Instead of explicitly computing the transformation of the dynamic image, the histogram is filled with the coefficients of the trilinear interpolation. That makes appear only the occurrence of the true values, but weighted to their distance to the interpolated coordinates.

1.3.6 Optimisation method

The comparison function together with the transformation parameters form a search space which we must iteratively scan for the optimum value.Optimisation methods must deal in general with the following items:

- Each iteration consists of a transformation followed by a comparison. The computational cost can be very high, both in time and in memory. For instance, two float images of $256 \times 256 \times 180$ need at least $90Mb$ of memory, which four years ago was an expensive good. Any operation in the Fourier domain, as images are implemented with double precision, increases four times this requirements.
- the function is not monotonic: it local maxima occur and the search algorithm is bound to get trapped.
- the search space has 3 dimensions for 2-D images, and 6 for 3–D images (3 rotations plus 3 translations).
- the accuracy of the transformation affects the final solution. Fast transformation schemes would not give enough accuracy, and then the comparison function simply is not sensitive to fine adjustments of the transformations parameters.
- for most functions the derivatives will not be available.
- the maximum value may be at a sharp peak, difficult to detect since nearby parameters take much lower values.

This set of properties make exhaustive search unfeasible and therefore some heuristic to optimise the alignment values is needed. To our opinion, the most demanding of problems listed is that of local maxima. Many optimisation schemes exist, but no matter how sophisticated they are there is no warranty not to be missing a better solution.

Many papers, including ours, use the well-documented optimisation algorithms from Press et al, [80]. A short list of them is: Powell's, Downhill Simplex, Brent's, Levenberg-Marquardt optimisation and gradient descend methods. All these methods are highly deterministic, this is, for an initial seed, they are constrained to follow a fully determined path. To mitigate the problem of local maxima, a usual solution is to build a hierarchical scheme, which, in addition to smooth the function profile, may be computationally attractive.

Non deterministic search methods, contrary to the previous, introduce a weighted random element to permit apparently bad choices at some iteration.Examples include genetic programming and simulated annealing. Although the random element accomplishes the desired effect to avoid local convergence, this element must be tuned to permit that the deterministic part finally leads to the solution.

1.4 Objectives of the thesis

Creases are recurrent features amongst medical and non-medical images. Examples amongst the first class are skull and bones in CT and MR images, vessels in most modalities (angiographies, retinographies), but in general any line or segment brighter or darker than the background. We have investigated a registration algorithm based on the creaseness image of the original images. We present the following novelties:

- We employ a new operator to extract creases from the images, which gives better performance than others.
- The registration algorithm is based on a hierarchical structure, which greatly enhances the final time.
- For CT to MR volume registration, we run two different alignment schemes: a broad fast approximation to compare principal axis of extracted features, and an iterative search on a correlation function to refine the results.
- The accuracy of our algorithm for CT to MR registration has been validated at an external independent university.
- The algorithm also suits to register 2–D slices of ophthalmologic images. We have studied its performance for series of video images.
- We have explored an upcoming research field: composition and registration of ultrasound images. After a lengthy work of design and calibration of our own system to acquire the images, we experimented with two registration problems: 3D ultrasound to 3D MR , and 2D ultrasound frames to 3D MR volumes.

1.5 Organisation of the thesis

This thesis presents a method for medical image registration demonstrated on three different types of images. Chapter 2 describes in detail the algorithm as it was originally designed: creaseness extraction, optimisation, validation of results and comparison to another method.

Chapter 3 addresses the second medical application of the registration algorithm: the alignment of long video sequences of retinographic images. We study the robustness and accuracy of the algorithm for a number of cases, with the goal to achieve an implementation suitable for a real application.

The third type of images we study are the ecographies taken during interventions in neurosurgery. In chapter 4 we fully describe a system to process sequences of these images to compound a volume image. Details of mathematical issues necessary for the calibration of the transducer are given, but the novelty is 2D–3D registration of the ultrasound images to an MR volume, which may be of interest for a future application to measure the brain shifting.

Finally, chapter 5 addresses the conclusions extracted from the experiments, and proposes future continuation lines.