Cross Validation of Three Inter-Patients Matching Methods

Jean-Philippe Thirion[1], Gérard Subsol[1] and David Dean[2]

[1]INRIA, Projet Epidaure, 2 004, route des Lucioles, BP 93
06 902 Sophia Antipolis Cedex, France

[2]Department of Anatomy, Case Western Reserve University
10900 Euclid Avenue, Cleveland, Ohio 44106-4930, USA

Abstract. In this paper, we present the cross-validation of three deformable template superimposition techniques a,b and c, used to study 3D CT images of the bony skull. Method (a) relies on the manual identification by anatomists of anthropometric landmarks, method (b) on "crest lines", which have a pure geometric definition, and method (c) is based on 3D non-rigid intensity based matching. We propose to define and compute a distance between methods a, b, c, and also to compute three representations $\bar{I}^a$, $\bar{I}^b$, $\bar{I}^c$ of an "average" skull model based superimposition via these three methods. The overall aim is to determine if the three methods, all developed independently, give mutually coherent image superimposition results.

1 Introduction

The possibility of matching 3D medical images of different patients or of comparing a new patient's image to a diagnostic "normative" image is a challenging issue for the medical community (see [BK89], [CMV94], [CNPE94]). The applications are numerous, ranging from computer-assisted building of 3D electronic atlases, to Computer Aided Surgery (see [CDB+95]) and diagnosis (see [STA95]). More studies such as three dimensional growth of organs (see [Sub95]) or human evolution (see [Dea93]) can be foreseen. But developing useful deformable template-based superimposition techniques is extremely complex, far moreso than for the rigid case: the "ideal" rigid motion can exist, and can be retrieved, up to a residual error transformation (see [PT95]). For inter-patients matching, there can be as many different criteria for image matching as possible applications: the similarities between two patients arise from the sharing of a common phylogeny (i.e., development of the species), and the undergoing of a similar ontogeny (i.e., normal growth and development), both of which are extremely difficult to model. We propose to compare the results of three inter-patients matching techniques, each of them being developed independently and the result of several years of research. First we describe the methods, then we report on cross-validation experiments of both the matching processes and the computation of "average" 3D images.

2 Three superimposition techniques

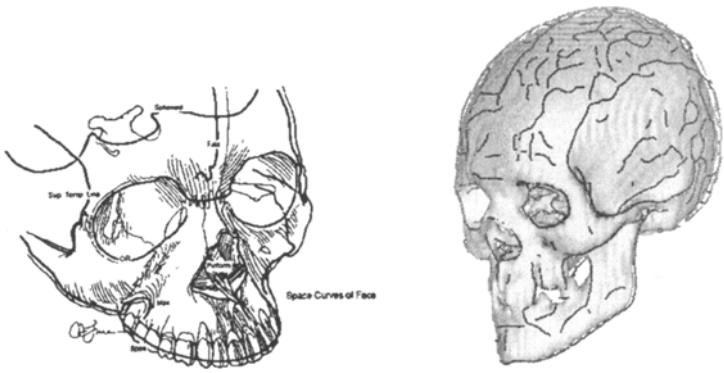

Figure 1. Left, principal anatomical lines (after [BC88]), leading to method a. Right, crest lines extracted by Marching Lines (method b).

Ridge Curve template matching (method a): it has been designed by anatomists, for cranio-facial surgical planning (see [BC88], [Cut91], [Dea93], [CDB+95]). There is a strong need, on one hand, to build an average surface representation of a "normal" patient and on the other hand, to match patients to this average, for diagnosis and simulation. A template of the salient lines of the skull called "ridge curves" (see figure 1, left), has been designed by anatomists. It contains homologous 3D anthropometric landmarks (points), linked by ridge and geodesic curves (see figure 2). The template is matched implicitly by the correspondance of the vertices given by the anatomist. The major advantage is that method a relies on homologous features, that is, points or lines which are anatomically invariant, but a new template has to be manually developed for each organ. The selection of landmarks and the final correction are time consuming processes and there can be inter- or intra-observer variabilities.

Crest lines matching (method b): "Crest lines" are mathematically defined by differential geometry, and coincide with the intuitive definition of "most salient" lines used in method a. The "Marching Lines" algorithm (see [TG95]) automatically extracts crest lines from volumetric images (see figure 1, right). The matching algorithm iteratively computes rigid, affine, and spline transformations, and forms couples of points from the two sets of crest lines, based on their spatial distances after transformation (see [Sub95]). The whole process is fully automatic, leading to a large set of corresponding points within 10 to 20 minutes CPU time (Dec Alpha workstation). The topology of crest lines can vary from patients to patients, the method needs high resolution 3D images and there are no a-priori anatomical information. Their use to build average model representation has been demonstrated in [STA95].

Intensity-based matching (method c): This method has been presented in [Thi95], and [Thi96] and produces a point to point correspondance field between the two 3D images. Its principle is based on an analogy with Maxwell's demons, and relies on polarity (inside/outside) information: the boundaries of the object in the scene image I_s are considered to be hemi-permeable membranes, separating the inside of the object from the outside, and the voxels of the model image I_m are considered to be particles, labeled "inside" or "outside" points. The deformable model I_m is then "diffusing" into the scene image I_s. The approach is entirely automatic and multi-scale. The whole set of voxels is taken into account within 10 to 20 minutes CPU time. No anatomical knowledge is used but the convergence depends on the initial positioning of the two objects, which have to be relatively close.

Each method relies on a privileged set of points. Method a relies on the set A of homologous points (typically several tens of points, or one thousand if the linking curves are taken into account), method b relies on the crest lines (set B, several thousands of points) and method c takes all the voxels with high gradients into account (set C, millions of points). In fact, A,B and C are processes to extract feature points from images and must be distinguished from the matching methods themselves.

3 Cross validation study

Our cross validation relies on two principles: the first is to study the effect of non-rigid deformation on sparse data, and the second is to deny a-priori preference to any of the three methods. The aim is not to determine the "best" of the three methods; that is meaningless. Rather we wish to determine if these three methods produce mutually consistent results. For that purpose, we define a distance between the three methods and compute and compare "average" patients, generated from a reference database of skulls.

3.1 Defining a distance between methods

Let I_1 and I_2 be the images of two specimens. To perform a fair comparison, we compute from the matching result of each technique a B-splines based warp $T^a_{1,2}$, $T^b_{1,2}$, and $T^c_{1,2}$ (see [DSTA95]), with exactly the same parameters (i.e., control points and smoothness constraints). Our distance measurements between methods is close to traditional morphometric studies of skulls (see [ANDB90]) because it is using a sparse set of feature points. It must be contrasted with validation methods based on manually segmented regions (see [GRB93]) or surface distances. We do this to prevent zero distances (between surfaces or regions) while having large local labeling "errors" (i.e. the tip of the nose sliding within the face surface to end up in the middle of the cheek). Our "features of interest" are the points produced by the feature extraction processes A, B or C. We must note that $B_1 = B(I_1)$ is not equivalent to $B_2 = B(I_2)$; in particular,

$T^b_{1,2}(B_1) \neq B_2$. This is true also for c: $(T^c_{1,2}(C_1) \neq C_2)$. However, by definition of a, we have $T^a_{1,2}(A_1) = A_2$ up to the approximation due to warping. A_1 and A_2 are two bijective sets, but correspond to two different inferences of the anatomical features, subject to possible uncertainty and inter-observer variability. In order to compare the different matching methods, one major problem is to reduce the influence of such inferences, for example by considering different types of reference sets A_1, B_1, C_1 in I_1 in our experiments.

Distance definitions For a couple of images (I_1, I_2), we can define a relative distance $d_Z(x, y)$ between two matching methods x and y, relative to set Z of n_Z reference points in image I_1. We choose to compute the average distance between the transformed points $T^x_{1,2}(Z)$ and $T^y_{1,2}(Z)$:

$$d_Z(x, y) = \frac{1}{n_{Z_1}} \sum_{\mathbf{z} \in Z_1} ||T^x_{1,2}(\mathbf{z}) - T^y_{1,2}(\mathbf{z})|| \tag{1}$$

where $T^x_{1,2}$ (resp. $T^y_{1,2}$) is the transformation between I_1 and I_2 obtained with the method x (resp. y), $||\cdot||$ is the Euclidean norm, and $Z_1 = Z(I_1)$. It is easy to verify that d_Z corresponds to the mathematical definition of a distance. We can also compute the median distance, that is, the distance $m_Z(x, y)$ for which 50% of the matched points in Z have a distance $||T^x(\mathbf{z}) - T^y(\mathbf{z})||$ less than m_Z, or the maximal distance: $M_Z(x, y) = max\{||T^x(\mathbf{z}) - T^y(\mathbf{z})||, \mathbf{z} \in Z\}$.

To show visually how close the three methods are, we propose to represent them as the vertices of a triangle, the lengths of the edges representing the relative distances between methods (see figure 4). This can be extended to an arbitrary number N of superimposition methods, and leads to a N-simplex representation.

3.2 Computation using a 3D CT skull image database

As we have access to a large number of specimens, we can compute a more robust estimate of the distance $\bar{d}_Z$ between methods a, b, c, by averaging the results over the specimens. We select a particular image I_i, which is considered to be the reference: all the images $I_{j \neq i}$ are matched with I_i, and the average distance $\bar{d}_{Z,i}$ is simply the average of the $n - 1$ distances d_Z obtained by comparing the $n - 1$ skull images $I_{j \neq i}$ with a reference skull I_i:

$$\bar{d}_{Z,i}(x, y) = \frac{1}{n \times n_Z} \sum_{j \neq i}^{n} \sum_{\mathbf{z} \in Z_i} ||T^x_{ij}(\mathbf{z}) - T^y_{ij}(\mathbf{z})|| \tag{2}$$

where, for example, $T^x_{ij}(\mathbf{z})$ is the transformation between the reference image I_i and image I_j, using method x and applied to a 3D point $\mathbf{z}$ of $Z_i = Z(I_i)$.

4 Validation of average patients computation

Using methods a and b, we have previously computed "average" images from the same 3D CT skull database (see [CDB+95] for method a, [STA95] for method b). The resulting average images $\bar{I}^a$ and $\bar{I}^b$, appeared quite similar on visual inspection. We are now giving quantitative evaluation by measuring the average distances between $\bar{I}_a$, $\bar{I}_b$, $\bar{I}_c$.

4.1 Computing average patients

To measure the average model of n skull's images, using a method x we:

- 1. compute the $n-1$ deformations T_{ij}^x from a reference image I_i toward the other $n-1$ skull images $I_{j\neq i}$.
- 2. subtract the similarity transforms S_{ij}^x (that is, rotation, translation and scaling) from each of these deformations, to get only anatomically meaningful deformations $\tilde{T}_{ij}^x$.
- 3. compute the average transformation: $\bar{T}_i^x = \frac{1}{n}\sum_j \tilde{T}_{ij}^x$ (including $\tilde{T}_{ii}^x = Identity$).
- 4. apply the average transformation $\bar{T}_i^x$ to I_i (or to $Z_i \subset I_i$) to get the average image $\bar{I}_i^x = \bar{T}_i^x(I_i)$ (or the set of average features $\bar{Z}_i = \bar{T}_i^x(Z(I_i))$).

4.2 Comparing the averaging techniques

We can then measure the distance between methods, $\hat{d}_{Z,i}(x,y)$, using the averaged deformations $\bar{T}_i^x$ and $\bar{T}_i^y$. The distance $\hat{d}_{Z,i}(x,y)$ is equivalent to the average of the distances between the averaged features of the different patients:

$$\hat{d}_{Z,i}(x,y) = \frac{1}{n_{Z_i}} \sum_{\mathbf{z}\in Z_i} ||\bar{T}_i^x(\mathbf{z}) - \bar{T}_i^y(\mathbf{z})|| \tag{3}$$

The average deformation $\bar{T}_i^x$ can be computed from the set of deformations T_{ij}^x in several ways. If it is done from point to point correspondences, that is, $\bar{T}_i^x(\mathbf{z}) = \frac{1}{n}\sum_j \hat{T}_{ij}^x(\mathbf{z}), \mathbf{z} \in Z_i$, then it is easy to demonstrate with the triangular inequality that $\hat{d}_{Z,i}(x,y) \leq \bar{d}_{Z,i}(x,y)$, where $\bar{d}$ is the average distance between methods already described. Although each method has a dedicated way to compute average patients, we have verified this inequality experimentally: the distances $\hat{d}(a,b)$, $\hat{d}(b,c)$, $\hat{d}(a,c)$ between the average patients $\bar{I}^a$, $\bar{I}^b$, $\bar{I}^c$ are lesser than the averaged distances $\bar{d}(a,b)$, $\bar{d}(b,c)$, $\bar{d}(a,c)$ between methods. As we can see measures obtained with average patients are more robust, which justifies, at least computationally, the production of average skull images (Many researchers have produced atlases of ideal patients, see [TT88], [ANDB90], [CDB+95])

4.3 Influence of the reference patient

Our averaging methods arbitrarily adopts the frame of a reference image I_i. We can eliminate this partially from the computation of $\bar{d}_Z$, for example by considering the deformations in a circular permutation of the skulls ($I_0 \rightarrow I_1, I_1 \rightarrow I_2 \ldots I_n \rightarrow I_0$). However, the result must still be expressed with the same given set of "features of interest", for example $A_0 \subset I_0$. Unfortunately, in the case of "average" patients, we cannot perform this circular permutation. We are obliged to compute, for a method x, the average image $\bar{I}_i^x$ (or the average model $\bar{Z}_i \subset \bar{I}_i^x$). A matching T_{ij}^x is still performed with respect to the same reference specimen I_i. Ideally, we would compute the average of $\bar{I}_i^x$ for each images being the reference $I_i, i \in [1,n]$, but this is too expensive computationally (n^2). Our experiments verify that in method c, the choice of the reference specimen I_i has a negligible influence on the computation of $\bar{I}^c$.

5 Experiments

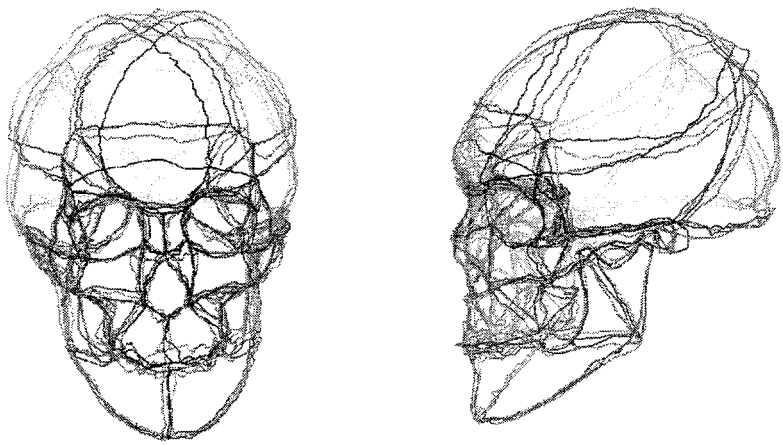

Figure 2. The templates A_2, A_3, A_4 projected on the template A_1 of the reference specimen, using method c (demons). It also illustrates visually $\bar{d}(a,c)$

We have also checked visualy the coherence of methods a, b and c, by projecting in the space of a reference image I_1 all the templates obtained with A: $T_{ji}^c(A_j), j \in [1,4]$ (see figure 2). The four templates are visually very close one to another. They are for the salient lines of the skull, but can differ significantly in smooth areas, ignored by the ridge curve template.

5.1 Cross-validation

The voxel sizes of our skull images are $1 \times 1 \times 1.5$ mm ($175 \times 200 \times 140$ voxels). To compare the influence of the features of interest (see figure 4, we have measured

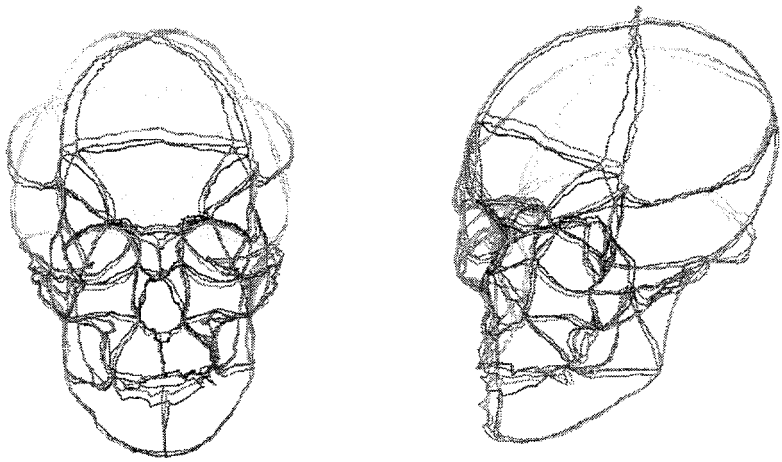

Figure 3. Illustrates visually distances $\hat{d}$: superimposed average templates $\bar{A}^a$, $\bar{A}^b$, $\bar{A}^c$, obtained independently by methods a,b,c.

the average distances between methods using: $B_1 = B(I_1)$, the set of crest lines in image I_1; Bf_1, the set B_1 where only the crest lines with high curvatures are kept; A_1 the template in image I_1 defined by an anatomist. The template or crest lines of the occipital and parietal parts of the skull are much less "well" defined than the face (i.e. no salient lines). This is illustrated by the good performances obtained using the filtered crest lines (Bf), which range from $3mm$ to $6mm$. Figure 4 is a pictorial representation of the average distances, which emphasize this. We can also see that no method is inconsistent with respect to the two other ones.

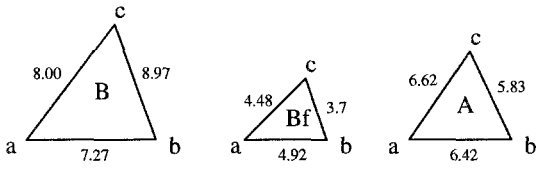

Figure 4. Simplex representation of the average distances between methods a (template), b (crests) and c (demons), in millimeters, for three types of features: B (crest lines), Bf (filtered crest lines), and A (anatomical template).

5.2 Average patients

We have computed the average patients with the three methods (see figure 3), and also the distances between average patients, using different features (see figure.5). The three methods give very close results, and the templates are much smoother than those of each individual skulls. Again, results are more stable around anatomically salient lines. The distances $\hat{d}$ between average patients

are much smaller than the average distances $\bar{d}$ between methods. This result support the production of average patients from a computational viewpoint. It should also be noted that the features of interest have only small influence on the results, which suggests that the average is stable in all portions of the skull.

methods	$\hat{d}_B$	$\hat{d}_{Bf}$	$\hat{d}_A$	m_B	m_{Bf}	m_A	M_B	M_{Bf}	M_A
$\hat{d}(a,b)$	2.92	2.59	3.37	2.58	2.45	2.98	11.3	8.58	11.5
$\hat{d}(a,c)$	3.96	3.78	3.61	3.65	3.77	3.12	11.3	8.86	12.7
$\hat{d}(b,c)$	3.33	3.34	3.09	3.18	3.19	2.83	8.87	8.87	8.31

Figure 5. Distances between average features, obtained with method a (templates), b (crests) and c (demons) in millimeters, with three types of features of interest B, Bf and A, and three distances: $\hat{d}$: average distance, m median, M maximal distance.

5.3 Sensitivity to the choice of the reference specimen

We also tested the sensitivity of the choice of the reference specimen with method c (see [Sub95] for similar results with method b). We must be very careful on the choice of the set of features of interest Z. We compute:

- $\bar{A}_1$, using image I_1 as the reference, and A_1 as features, where A_1 is the template, semi-automatically defined by the anatomist in image I_1.
- $\bar{A}_2$, using image I_2 as the reference, and A_2 as features. By definition for method a, $A_2 = T^a_{12}(A_1)$ (this is not true for methods b and c). Hence we can compare directly $\bar{A}_1$ and $\bar{A}_2$
- $\overline{T^c_{12}(A_1)}_2$, using image I_2 as reference, and where $T^c_{12}(A_1)$ is the template A_1, projected onto image I_2 by method c.

We compute the distance $d = \bar{d}_{\mathbf{z}\in A_1}(\bar{A}_1, \bar{A}_2) = \bar{d}_{\mathbf{z}\in A_1}(\bar{A}_1, \overline{T^a_{12}(A_1)}_2)$, which is the average distance between two estimates of average features, using two different reference specimens I_1 and I_2. We also compute $d' = \bar{d}_{\mathbf{z}\in A_1}(\bar{A}_1, \overline{T^c_{12}(A_1)}_2)$ and the results are presented in Figure 6.

	d_A	m_A	M_A
d	4.83	4.09	22.9
d'	1.25	1.13	4.47

Figure 6. Distances between average skulls obtained with two different reference images. d: with two manual template selections A and one application of method c. d': with a single template selection A and two times the application of c.

It should be noticed that d represents the differences in using two distinct reference specimens I_1 and I_2, but also two independent inferences A_1 and A_2

of the segmented anatomical features. This is equivalent to the concatenation of the "errors" of two manual selections, A_1 and A_2, and of one non-rigid matching with c. On the other hand, for d', only method c is used (twice). A_1 is used solely as a list of "features of interest" to produce an average distance. Hence d' is more representative of the perturbation introduced by the choice of the reference specimen alone than distance d. We note that this error ($1.3mm$) is of the same order as the image resolution, which shows that the choice of the reference specimen has very little influence on the result of averaging.

Our personal interpretation of the results are the following: the three methods give mutually coherent results, with an average difference for feature location of $3mm$ to $4mm$ where the skull is highly curved. In smoother places, this average precision is reduced to $6mm$ to $9mm$, but for specific points (outliers) the distance grows to $3cm$ or $4cm$. The computation of "average patients" gives more robust and coherent results, everywhere on the skull surface, with an "average" difference of $3mm$ to $4mm$ and up to $1cm$ for outliers. The computation of the average patient does not depend too much on the choice of the reference specimen.

6 Conclusion

The comparison of three image superimposition methods on a set of four 3D CT-scans of different dry skulls was performed, in order to verify that three methods, all developed independently, give mutually coherent results. To our opinion, the answer is yes. We have established quantitative measurements, such as distances between the individual features produced by these three methods and average features computed by superimposition, which could be used to compare other non-rigid techniques. We intend to use these methods, independently or combined, to classify groups or sub-groups of patients, and possibly to automatically diagnose pathologies.

Acknowledgement: Many thanks to Dr. Bruce Latimer, Curator of Physical Anthropology, Cleveland Museum of Natural History, for the access to the Hamann-Todd collection and also to Dr. Jon Haaga, Chair and Professor, and Elizabeth Russel, Technician, Department of Radiology, Case Western Reserve University, for providing CT-scans of these specimens.

References

ANDB90. A. H. Abbot, D. J. Netherway, D. J. David, and T. Brown. Application and Comparison of Techniques for Three-Dimensional Analysis of Craniofacial Anomalies. *Journal of Craniofacial Surgery*, 1(3):119–134, July 1990.

BC88. F. L. Bookstein and C. B. Cutting. A proposal for the apprehension of curving cranofacial form in three dimensions. In K. Vig and A. Burdi, editors, *Cranofacial Morphogenesis and Dysmorphogenesis*, pages 127–140. 1988.

BK89. R. Bajcsy and S. Kovačič. Multiresolution Elastic Matching. *Computer Vision, Graphics and Image Processing*, (46):1–21, 1989.

CDB+95. C. Cutting, D. Dean, F .L. Bookstein, B. Haddad, D. Khorramabadi, F. Z. Zonneveld, and J.G. Mc Carthy. A Three-dimensional Smooth Surface Analysis of Untreated Crouzon's Disease in the Adult. *Journal of Craniofacial Surgery*, 6:1–10, 1995.

CMV94. G. E. Christensen, M. I. Miller, and M. Vannier. A 3D Deformable Magnetic Resonance Textbook Based on Elasticity. In *Applications of Computer Vision in Medical Image Processing*, pages 153–156, Stanford University (USA), March 1994.

CNPE94. D.L. Collins, P. Neelin, T.M. Peters, and A.C. Evans. Automatic 3d intersubject registration of mr volumetric data in standarized talairach space. *J. of Computer Assisted Tomography*, 18(2):192–205, March 1994.

Cut91. C. B. Cutting. Applications of computer graphics to the evaluation and treatment of major craniofacial malformations. In J. K. Udupa and Herman G. T., editors, *3D Imaging in Medicine*, chapter 6, pages 163–189. CRC Press, 1991.

Dea93. D. Dean. *The Middle Pleistocene Homo erectus/Homo sapiens Transition: New Evidence from Space Curve Statistics.* PhD thesis, The City University of New York, 1993.

DSTA95. J. Declerck, G. Subsol, J.Ph. Thirion, and N. Ayache. Automatic retrieval of anatomical structures in 3D medical images. In N. Ayache, editor, *CVRMed'95*, volume 905 of *Lecture Notes in Computer Science*, pages 153–162, Nice (France), April 1995. Springer Verlag.

GRB93. J. C. Gee, M. Reivich, and R. Bajcsy. Elastically Deforming 3D Atlas to Match Anatomical Brain Images. *Journal of Computer Assisted Tomography*, 17(2):225–236, March 1993.

PT95. X. Pennec and J.Ph. Thirion. Validation of 3-D Registration Methods based on Points and Frames. In *Proceedings of the 5th Int. Conf on Comp. Vision (ICCV95)*, pages 557–562, Cambridge, (USA), June 1995.

STA95. G. Subsol, J.Ph. Thirion, and N. Ayache. A General Scheme for Automatically Building 3D Morphometric Anatomical Atlases: application to a Skull Atlas. In *Medical Robotics and Computer Assisted Surgery*, pages 226–233, Baltimore, Maryland (USA), November 1995.

Sub95. G. Subsol. *Construction automatique d'atlas anatomiques morphométriques à partir d'images médicales tridimensionnelles.* PhD thesis, Ecole Centrale Paris, December 1995.

TG95. J-P Thirion and A Gourdon. Computing the differential characteristics of isointensity surfaces. *Computer Vision and Image Understanding*, 61(2):190–202, March 1995.

Thi95. J-P. Thirion. Fast Non-Rigid Matching of 3D Medical Images. In *edical Robotics and Computer Aided Surgery (MRCAS'95)*, pages 47–54, Baltimore (USA), November 1995.

Thi96. J-P. Thirion. Non-Rigid Matching Using Demons. In *Computer Vision and Pattern Recognition (CVPR'96)*, San Francisco (USA), June 1996. (to appear).

TT88. J. Talairach and P. Tournoux. *Co-Planar Stereotaxic Atlas of the Human Brain.* Georg Thieme Verlag, 1988.