2.5 Results of the assessment

This section presents the results of our algorithm under a number of tests. The experiments can be grouped into two major sets, as listed in table 2.3:

Set	Description	Section
1	Visual validation	2.5.1
1	Choice of parameters	2.5.2
1	Comparison with mutual information	2.5.3
2	External validation at Vanderbilt	2.5.4

Table 2.3: List of experiments

The first set, which we published in [42], is restricted to 5 pairs of CT–MR images of different characteristics (table 2.4). The initial experiments were useful to design the initial layout of our algorithm. First, we examined visually the accuracy of the matching, both with the overlapping of crests and with the original images. Then, we tuned several algorithm parameters, and see how it affected the final robustness and accuracy. Finally, we compared our method with another based on mutual information. Images of this section have been kindly given by Dr. P. van den Elsen, from the Utrecht University.

The second set was an external validation of the accuracy, and was done at the Vanderbilt university. We take as a gold reference the transformation which minimises the mean square correspondence error of several stereotactic frame landmarks manually pointed out by experts, as in [102]. We participated in the second phase of the Vanderbilt project [114], to assess the overall accuracy of our method when compared to an other state of the art extrinsic method. Because results had been published, strictly speaking our registration could not be considered blind as the ones in the initial paper. Despite this, since published results could not provide any additional help, in practise our results are as valid as those from the first group. The site at Vanderbilt evaluated our registration with the same criteria as with the previous groups.

The validation procedure for this second set was very demanding: it involved 5 pairs of image modalities for each one of 16 patients. Compared to our previous experiments with one pairs for 5 patients, these images represented a more realistic clinical data, which meant that some parts of the algorithm had to be rebuild in order to gain in generalisation and robustness.

There are two main issues for the validation a registration method:

- **accuracy**, which bounds the error expected when we relate the coordinates of one image into the other.
- **robustness**, which tells the repeatability of the experiment and the reliability of the method under adverse conditions.

The following sections describe the results obtained. **Section 2.5.1** addresses the accuracy issue by the visual validation of the matching between the five pairs of

Dataset number	Modality	Dimensions x, y	Dimensions z	Resolution x, y (mm)	Resolution z (mm)
1	CT	256	100	0.93	1.55
	MR–T1	256	180	0.97	1
2	CT	320	128	0.71	1.5
	MR–T1	256	100	0.9	1.5
3	CT	512	29	0.65	4
	MR–PD	256	26	1.25	4
4	CT	512	29	0.65	4
	MR–T1	256	26	1.25	4
5	CT	512	29	0.65	4
	MR–T2	256	26	1.25	4

Table 2.4: Specifications of the 5 pairs of images used in the first set of experiments.

images. For this purpose we chose the 5 pairs of images from the database in order to represent a variety of conditions. Pairs 1 and 2 represent the best possible conditions: high resolution, good contrast and wide field of view (figs. 2.9 and 2.19). Pairs 3, 4 and 5 (figs. 2.21 and 2.20) are a challenge because they have a low number of thick slices, their contrast is non-homogeneous and their MR acquisition settings do not permit an easy segmentation of the bone.

The first experiment, described in **section 2.5.2**, had the aim of setting up the method basis and tune its quantitative parameters. For this experiment we propose a scheme which consists in measuring how well it is able to recover a trial known transformation. With these values we justify the utility of our hierarchical method.

A second experiment, in **section 2.5.3**, was set to compare the performance of our method to the mutual information's, which is generally considered the most accurate [114] and reliable. The method was the same as the previous section, but this time we used constant values for the parameters, the ones which performed the best, with regard to the performance of the mutual information algorithm.

Finally, **section 2.5.4** reports the accuracy of our method compared to that of the Vanderbilt's golden standard. We give a short explanation about the conditions of the comparison, and we rank it against the results published in the final project paper [114].

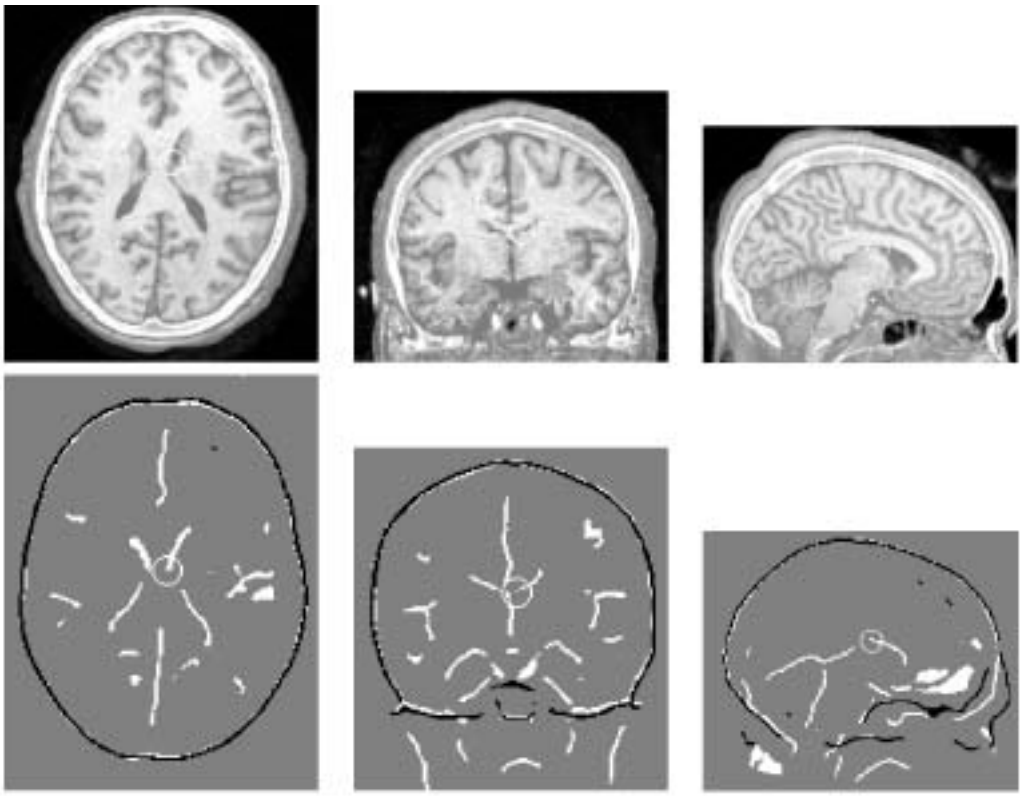


Figure 2.19: Fusion of registered volumes of data set 1. From left to right, first row shows MR with CT bone superimposed. Second row shows CT crest (black) registered with MR valley (white) The white circle marks the intersection of the three orthogonal views. Despite the perfect alignment of the skull, note a small miss-registration at the bottom slices.

2.5.1 Visual validation

The experiment reported in this section consists on the visual inspection of the correctness of the registrations. Figures 2.19 and 2.21 show the mix of the CT and the MR images using a point to point maximum operation. In this visualisation, we have paid special attention to a soft tissue in the inner surface of the bone, the dura, the dura, which appears in black and should be of constant width.

All figures show the MR image with the bone superimposed using the max operator. Although this pattern is visually attractive, sometimes hides possible misalignments. To take them into account, one needs to examine the alignment of crests too. For this reasons, figures include both visualisations.

We have visually inspected the registration for all five data sets and found them to be similarly well aligned both for our method and for mutual information. Differences are bigger for data sets 3, 4 and 5, and are due to two factors. The first is their low

resolution, which magnifies the differences, specially for the z coordinate. The second, also important, is that their settings make the two methods converge to two different solutions which actually are the maxima for their respective alignment function.

Figure 2.20 shows several axial slices of the data set 3. The bone is properly segmented for the medium axial cuts, but in higher planes the bone appears filled by marrow, which is visible in the particular settings of these MR modalities but not in the CT image. Therefore, for these two pairs the segmentation done by the creaseness operator is slightly different for each image.

Therefore it is clear that the registration method did not fail, but, rather, the landmark features can not be made to exactly overlap for the two modalities. In section 2.5.4 we will compare our alignment against a golden standard, and will see that the results are accurate for all modalities but one, MR–T1, where the top slices are, in some cases, dissimilar.

However, we think that the registration is still valid because the correlation, which measures the quality of a transformation, takes into account only common structures, which coincide in the rest of the segmentation.

In addition to the images in this section, we have included the appendix B, page 203, with a full report of the visual alignment.

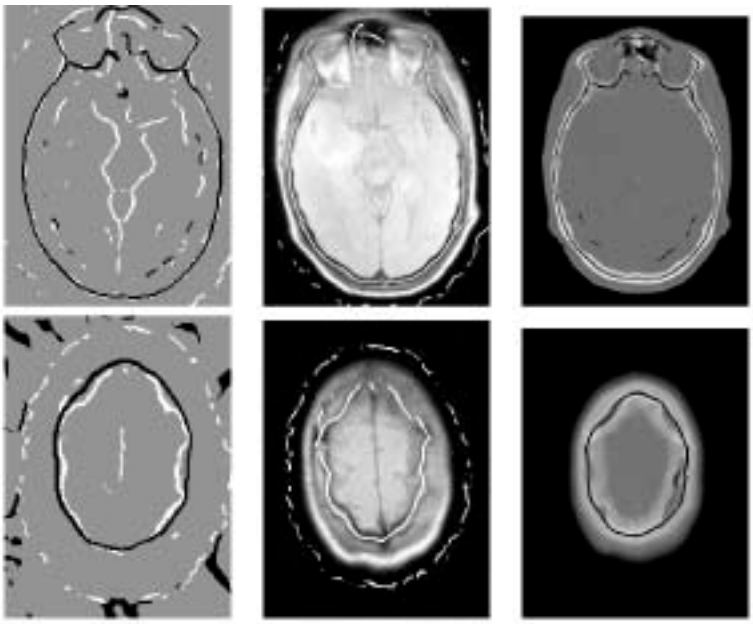


Figure 2.20: Registered creases at two slices of data set 3. The left column shows two axial views of corresponding creases from CT (black) and MR (white). The image below shows mis-registration because, although the crease follows properly the valley, in the MR modality the upper slices do not depict the bone as a valley but as marrow surrounded by a thin layer of bone. This can be clearly seen comparing the segmentation of the bone in the middle column (in white) for the MR image, and the right column (in black) for the CT image.

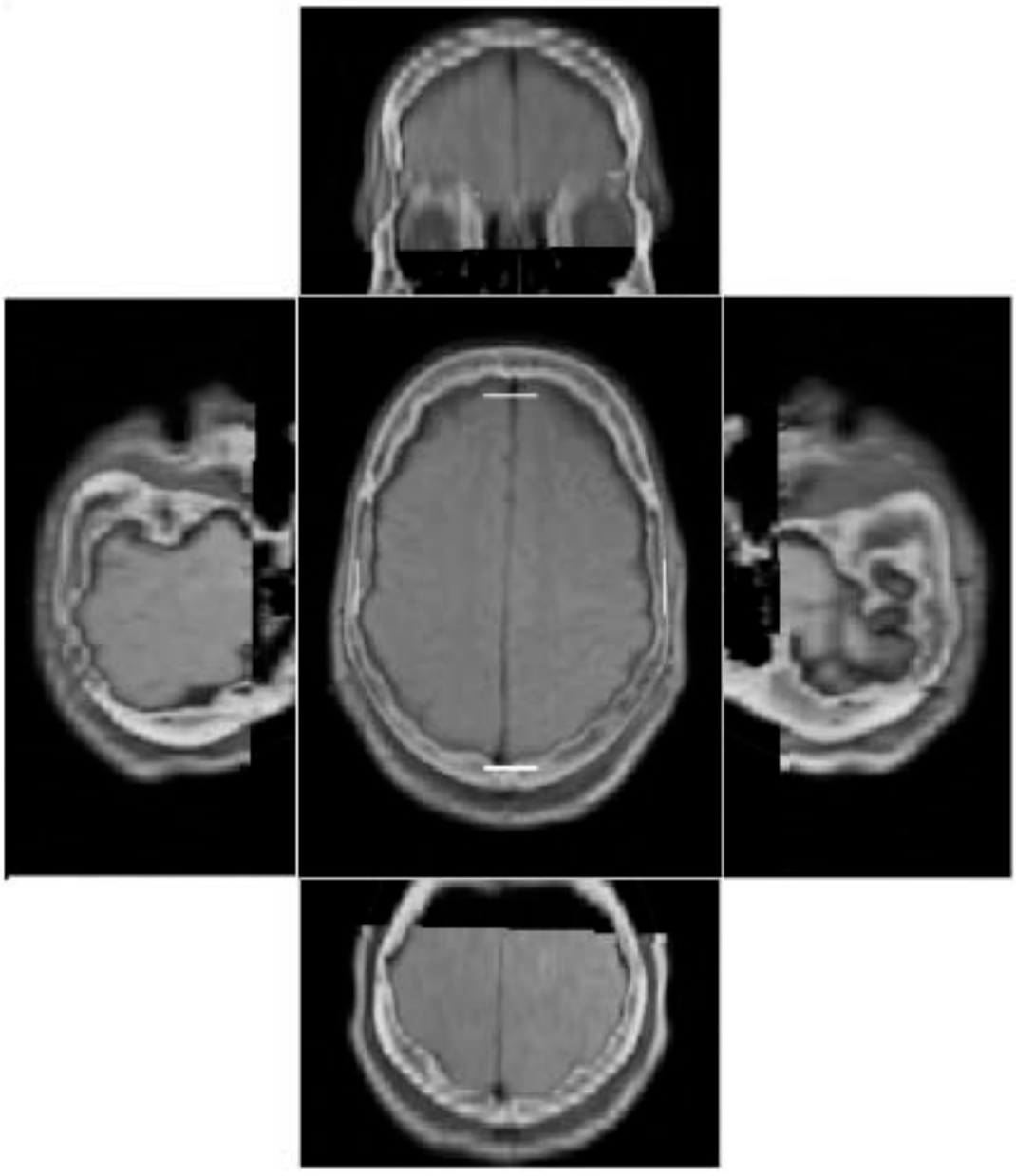


Figure 2.21: Fusion of registered volumes of data set 3 showing MR with CT bone superimposed. Note the thin gap with constant width between the border of the bone in the CT and the brain tissue in the MR. The white lines in the central image shows the placement of the adjacent section.

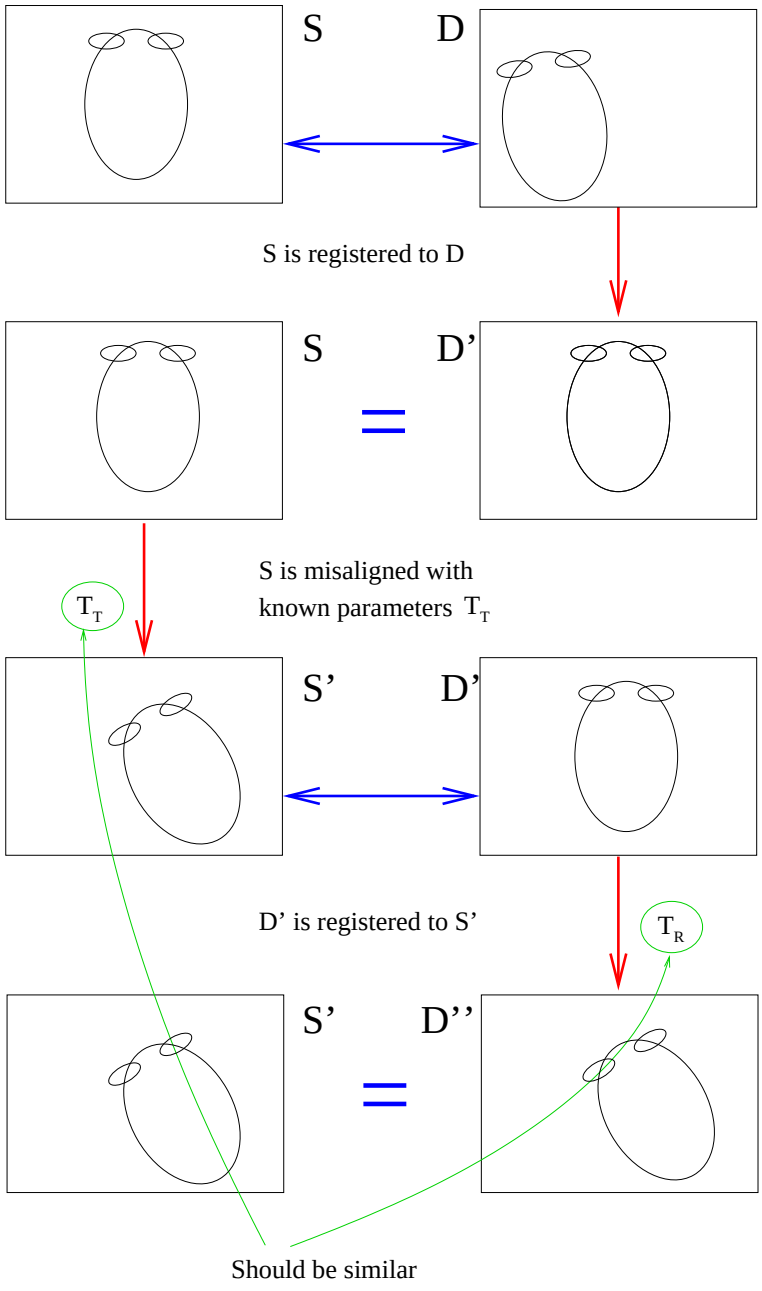


Figure 2.22: Experiment to check for the robustness of the registration method: after the images have been registered, one is miss-aligned by a known random transformation. The algorithm is run again, and the result transformation must be similar to the trial.

2.5.2 Experimental choice of algorithm parameters

We have carried out a set of experiments to assess the properties of the method. The aim is too see whether the algorithm:

- depends on some parameter in some ill-posed fashion.
- is able to recover large misregistrations, or perhaps it is sensitive to a particular range of transformation values. (robustness)
- the final alignment is invariant to the relative initial position of the images. (repeatability)

An ideal experiment would consist on comparing the transformation achieved by our method to some other registration values with higher accuracy. But this information was difficult for us to obtain because we did not have contact with a hospital with such facilities. Therefore, we had to make up an experiment to simulate a database with dozens of registered images.

The structure of the experiment, shown in figure 2.22, is as follows:

run the described method to register the source images S (CT) and D (MR). Call the registered image S and D_R, and T_0 the transformation.

the static image (S) image is transformed with random known parameters T_T; let S_T be the transformed image.

the registration algorithm is applied to S_T and D_R, giving the transformation parameters T_R.

measure the similarity of the given and recovered transformation (mean distance, MD):

$MD = 0$
for each pixel S_i in S:
  if S_i is not void
    $i_t = i * T_T$
    $i_r = i * T_R$
    $MD = MD + |i_t - i_r|$

Figure 2.23: Experiment to assess the robustness

An error-free method would return T_R equal to T_T, while a large difference would correspond to mis-registration. An important issue is the method to assess the quality of a recovered transformation; a simple subtraction of the six parameters is difficult to quantify, because of their different unit measures.

We have summarised all the information into a single number, in distance units, which is meant to be the expected distance between corresponding features. We measure the distance between coordinates of each pixel transformed by T_R and T_T, and take their mean. To discard pixels belonging to uninteresting regions, we select only those pixels belonging to the skull by means of a value threshold.

This is reasonable because we do not have a medical partner to select specific regions of interest, and if we had selected the full contents of the image the empty pixels at the corners would have distorted the measure. We have called this measure MD.

We have build a set of 50 trial transformations with random parameters distributed uniformly with increasing magnitude. For rotations, the range is $\pm[4-25]$ deg, and $\pm[4-25]$ mm for translations, which is about 10% of the total length. See table 2.7, left columns, for a sample of the list of transformations tried. The experiment has been applied to pairs 1 and 3 from table 2.4, belonging to different modalities, to ensure the generality.

The first goal was to optimise the parameters in our registration method. There are several main issues to be investigated experimentally:

- the need of a hierarchical approach.
- the number of seeds needed at each level.
- whether more iterations in the convergence step influence the final accuracy.

An additional run was made changing the alignment function. Instead of correlation the creaseness image, we measure the mean distance between the extracted surfaces. As explained in page 46, this can be made very efficiently by means of chamfer matching, which actually leads to the same optimisation scheme applied to transformed images.

We have tested 4 different combinations of parameters, and checked for their statistics. We did not test all the combination of parameters because the results achieved were already significant. Table 2.5 shows the statistics of the experiments. Labels in the table have the following meaning:

Ftol the tolerance threshold of the optimisation method at the highest resolution level. It determines the number of iterations to align the images. A low value would force the algorithm to give up before the proper alignment has been achieved, but on the other hand a too high value would refine the alignment with no visible improvement.

Seeds per level how many results are kept from one level to the next. Number are listed from highest to lowest resolution.

Median the median of the mean distances MD for those transformations which have successfully converged.

time to compute the results, not including the creaseness extraction.

Success number of registration successfully recovered, i.e., with a MD lower than 10mm, out of 50. The threshold of 10mm has been chosen arbitrarily.

And the four tested configurations are:

Method A passes only one seed through the hierarchy. That means that a false maximum may misguide the search in the following levels, but has the advantage to be very fast.

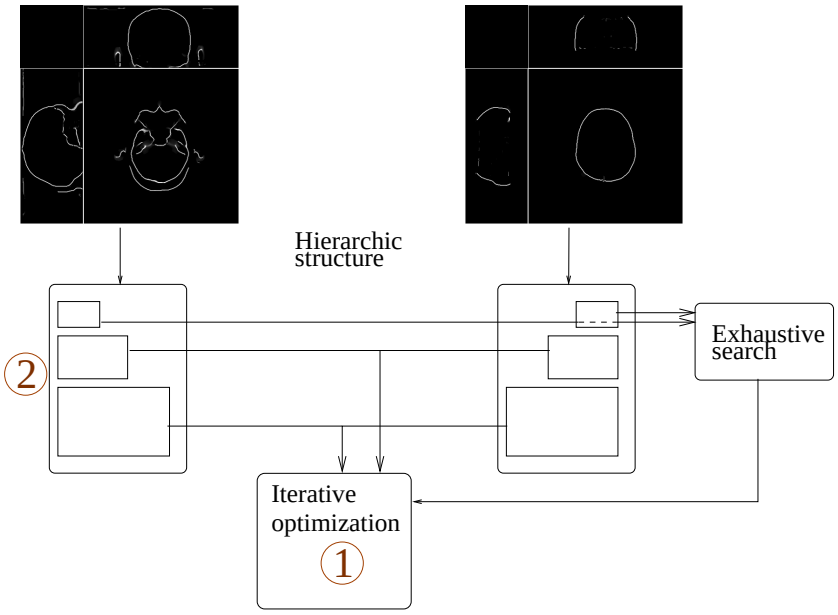


Figure 2.24: Parameters optimised in the experiment: ① $ftol$ and ② seeds per level.

Method B should provide the best results for both robustness (multiple seeds are passed) and accuracy ($ftol$ higher than for method A.

Method C starts the optimisation without any initial search. Thus, convergence for images not initially close is not bound to succeed.

Method D tries the chamfer function to measure the alignment, to see if provides any improvement with respect to the simple cross-correlation.

Table 2.5 shows the results for 4 choices of parameters. Methods A and B show very similar results. Both achieve to recover all the trial transformations, the only difference being the time and the median. The better median of the second has been obtained at the price of more steps (determined by $ftol$), and therefore the time needed is about four times higher. Part of this time has been spent also in the optimisation of the full set of seeds, in contrast with the single seed in method A.

The method C has failed for 8 cases out of 50, thus showing the need of the initial search to ensure convergence. For those transformations which actually converged, the resulting accuracy is, as expected, as good as with method B, because the parameter $ftol$ is the same.

Matching with chamfer distances (D) has a success rate comparable to methods A and B. Although it has the same settings, it converges faster than B, thus indicating that the function is smoother. But this property also seems to affect the sharpness of the maximum: the median error is broader, and it is less repeatable than A or

	A Single seed		B Multiple seed		C No hierar.		D Chamfer	
Tolerance ($ftol$)	10^{-3}		10^{-5}		10^{-5}		10^{-3}	
Seeds per level	{1,1,1,1,1}		{1,3,4,4,8}		–		{1,3,4,4,8}	
Dataset	1	3	1	3	1	3	1	3
Median (mm)	0.59	1.4	0.54	0.95	0.55	0.99	1.38	1.21
Time (min)	2.61	3.12	8.65	13.15	5.43	7.11	4.31	8.12
# Success (50)	50	50	50	50	42	42	49	50

Table 2.5: Performance of our method for 50 trial misregistrations and two datasets.

B. Therefore, the chamfer distance is not of any use for this modalities. A possible application would be for modalities not showing corresponding features, e.g. MR and PET. The bone extracted in MR ought to be matched against the single feature visible in PET, the boundaries of the head. Since both do not match exactly, for this case the chamfer distance would be a good measure of the closer possible distance.

We conclude from these statistics than the creaseness segmentation is accurate enough to provide a good alignment, and that the hierarchical scheme is necessary to ensure convergence. Also, a number of parameters have some influence on the final accuracy and robustness, and thus they need to be tuned. But the final numbers are fairly good in all cases, which means that its dependency to the parameters to optimise is well defined.

2.5.3 Comparison with a mutual information algorithm

The goal now is to validate the performance of our method against another, which we can take as reference. We have chosen the *normalising mutual information* (MI) because it is representative of the voxel-based methods and its results are the best according to today literature [114]. For this purpose, we were fortunate to have access to software implementing this measure, written by Dr. C. Studholme under the direction of Dr. D. Hawkes, from the Computational Imaging Science Group in Radiological Sciences at UMDS, Guy's & St Thomas' Hospital, London.

The experiment to compare both methods is the same as that in the previous section, i.e., we assess the ability to recover a known random misregistration. Since MI employs the full contents of the image, we had to make sure that the transformations applied did not discard out of limits the contents of the image.

For this purpose, we did not actually compute any transformed image, but instead we applied the misregistration matrix within the algorithm, thus leaving available all the initial information. Also, the initial transformation to align the images before the experiment starts (T_0 in algorithm 2.23) was computed using the same method that had to be evaluated afterwards, this is, MI for the MI experiments.

We have set our method to the parameters labelled as 'single seed' in table 2.5, in order to achieve times comparable to those of MI . The experiment included 50 random trial miss-registrations applied to 5 pairs of images, as specified in table 2.4. The transformations were generated randomly with increasing range of values, as

Dataset number	Method	Error Mean (mm)	Error Max (mm)	$N < 10$ (50)	Mean Time (min)
1	C	0.59	1.47	50	2.61
CT MR–T1	MI	2.25	4.72	50	6.78
2	C	0.35	0.83	49	3.48
CT MR–T1	MI	1.73	3.61	50	5.97
3	C	1.4	3.8	50	3.12
CT MR–PD	MI	4.12	9.69	50	3.13
4	C	1.92	4.49	40	3.86
CT MR–T1	MI	5.34	10.62	50	3.49
5	C	1.53	3.86	41	3.46
CT MR–T2	MI	6.42	14.11	50	2.97

Table 2.6: For each data set we compare the global results of the creaseness method (C, first row) and mutual information (MI, second row). The mean and max columns refer to the mean and maximum errors of the 50 trial transformations done for each pair. The next column gives the number of the transformations recovered within a distance of 10 mm from the trial.

specified in the previous section.

Our algorithm takes 3–5 *min* to converge on a PC *Pentium* at 350 MHz with 256 Mb of memory running under Linux, plus approximately 5 $min. \times 2$ images to extract the creases. Results for the MI algorithm were obtained on the same computer and took from 3 to 6 *min* to complete.

We have distributed the results of this experiment into two tables. The first gives statistical figures, while the second lists some of the tried transformations for illustration purposes. Table 2.6, right hand-side, provides statistical results for all 5 pairs of images. Tables 2.7 and 2.8 give a sample with the narrower and wider tried mis-registrations and the resulting error for both methods, for data set 1, 3 and 5.

Our method gives best results for data sets 1 and 2 because the segmentation process is more accurate due to the quality and number of slices of the images. In these sets, it converges more confidently and with a higher repeatability than the mutual information method. Results from data sets 3, 4 and 5 are less clear: failures are equally high for both methods, and the lower mean of the creaseness method is less significant.

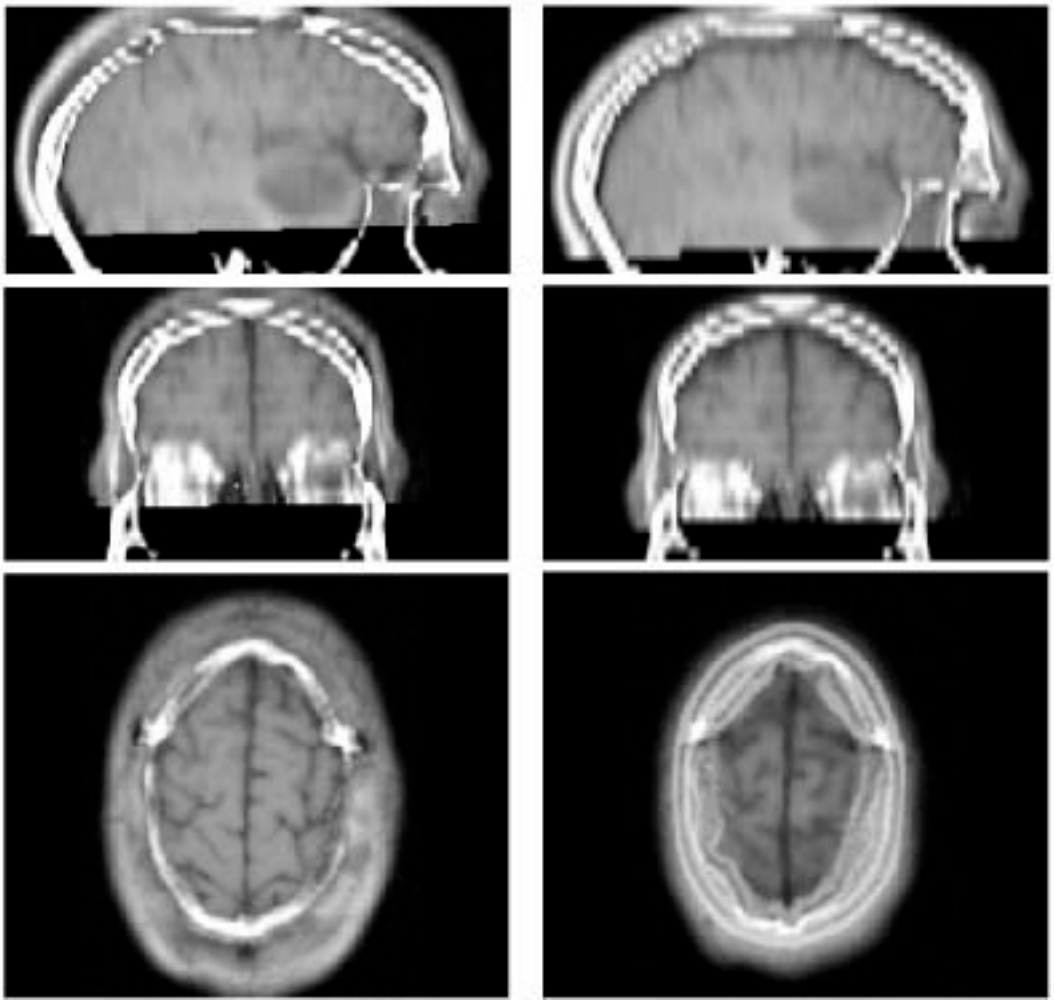
The explanation is that the overlapping area of the two original images is narrow, and once the trial transformation T_T has been applied, the resulting images may be too distant to converge. Since the creaseness method has only the extracted crests to compare, when the exhaustive search at the top of the pyramid (coarser resolution) is computed, the overlapping section may not be enough to permit a successful recover. These numbers could be improved if the initial search would be done at lower level, with higher resolution, but then the price would be an increase of the computing time. Such increase is not necessary for registration of normal clinical images, which have a transformation range much more restricted.

Creaseness measure seems to have a maximum in the optimising function narrower than MI, as can be seen in table 2.7: for no transformation at all ($N = 0$), MI has an error only slightly better than the mean for all the mis-registrations, while the creaseness method is almost error-free.

We tried to relate the error for each transformation to the magnitude of the misregistration, i.e., more severe misregistrations would be expected to have higher error than those smaller, but no relationship was found. This can be double-checked looking at the numbers in tables 2.7 and 2.8: the MD is not higher for bottom, more severe, misregistrations.

It is interesting to note that the proper convergence of the algorithm is independent of the specific tried transformation, but it depends a lot on the quality of the original images. Some transformations were successful for datasets $1 - -3$, but failed for the other pairs.

Another remark is that the recovered parameter $\Delta\theta_x$ has an error statistic generally two or three times worse than the others. This fact, which we have not investigated in deep, may be attributed to the particular shape of the head or to the different sampling. The translation in z has similar problems, but these are caused by the different sampling of the z axis with respect to the x and y. See that this problem does not apply to dataset 1, which has a more similar sampling.



Creaseness Mutual information

Figure 2.25: Registration of data set 4 exhibits the clearest differences between the two methods. These three views (sagittal, coronal and axial) show that it is not easy to decide visually which is better. Our registration is clearly too low along the z axis, because of the same problems as in data set 3. In contrast, the mutual information transformation seems to position the bone too high, even overlapping the skin on the top of the head.

		Original transformation						Error in recovered transformation												
								Creaseness							Mutual Information						
	N	θ_x	θ_y	θ_z	t_x	t_y	t_z	$\Delta\theta_x$	$\Delta\theta_y$	$\Delta\theta_z$	Δt_x	Δt_y	Δt_z	MD	$\Delta\theta_x$	$\Delta\theta_y$	$\Delta\theta_z$	Δt_x	Δt_y	Δt_z	MD
Dataset 1	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0.72	0.12	0.89	0.12	0.25	0	**1.82**
	1	-1.21	-3.01	0.33	3.82	-0.34	0.52	0.35	0.22	0	0.09	0.09	0.01	**0.6**	0.13	0.26	1.16	0.2	0.03	0.11	**1.98**
	2	3.54	-1.5	0.66	-1.7	-3.83	-2.2	0.02	0.36	0.08	0.09	0.05	0.03	**0.55**	1.21	0.42	1.2	0.18	0.17	0.08	**2.72**
	3	-1.54	-2.03	-2.42	0.04	-1.58	-0.86	0.11	0.44	0.11	0.08	0	0.06	**0.68**	0.17	0.24	0.2	0.16	0.11	0.11	**0.58**
	4	-0.69	0.13	0.77	-0.7	0.35	-0.42	0.1	0.1	0.11	0.05	0.02	0.04	**0.27**	0.15	0.11	1.13	0.18	0.1	0.48	**1.97**
	45	13.86	7.42	-3.78	15.71	7.04	-5.73	0.1	0.4	0.19	0.06	0.05	0.03	**0.65**	1.21	0.03	0.34	0.09	0.29	0.29	**1.8**
	46	-15.81	-13.44	11.78	6.66	-11.71	9.41	0.12	0.24	0.03	0.05	0.01	0.02	**0.38**	0.18	0.2	1.4	0.09	0.16	0.22	**2.25**
	47	14.08	12.99	-1.15	-13.41	-12.16	4.03	0.18	0.32	0.1	0.05	0.03	0.02	**0.53**	1.19	0.01	0.52	0.15	0.22	0.09	**1.77**
	48	-8.2	-12.44	0.2	-11.2	5.8	12.92	0.01	0.05	0.1	0.06	0.04	0.04	**0.19**	0.15	0.15	0.14	0.05	0.24	0.02	**0.42**
	49	11.96	-13.56	10.96	12.36	-7.44	3.4	0.11	0.43	0.19	0.07	0.06	0.08	**0.71**	12.32	0.55	1.18	0.17	0.18	0.1	**2.24**
	Mean values							0.11	0.26	0.09	0.06	0.04	0.03	**0.59**	1.74	0.21	0.82	0.14	0.18	0.15	**2.25**
Dataset 3	0	0	0	0	0	0	0	0	0.01	0	0	0	0	**0.01**	1.97	0.42	0.6	0.19	0.06	0	**3.49**
	1	-1.21	-3.01	0.33	3.82	-0.34	0.52	0.23	0.19	0.07	0.02	0.02	0	**0.5**	2.76	0.14	1.04	0.24	0.09	2.29	**5.27**
	2	3.54	-1.5	0.66	-1.7	-3.83	-2.2	1.68	1.42	0.66	0.28	0.1	1.25	**3.93**	1.63	0.54	0.78	0.45	0.54	2.26	**3.99**
	3	-1.54	-2.03	-2.42	0.04	-1.58	-0.86	0.08	0.29	0.06	0.04	0.1	0.34	**0.59**	2.98	0.24	0.32	0.21	0.17	2.68	**5.24**
	4	-0.69	0.13	0.77	-0.7	0.35	-0.42	0.37	0.07	0.05	0.01	0	0.67	**0.81**	1.64	0.13	0.95	0.07	0.02	0.42	**3.42**
	45	13.86	7.42	-3.78	15.71	7.04	-5.73	1.47	0.07	0.07	0.15	0.11	1.07	**2.5**	1.03	0.27	0.46	0.04	0.21	2.6	**3.01**
	46	-15.81	-13.44	11.78	6.66	-11.71	9.41	0.13	0.31	0.03	0	0.04	0.16	**0.55**	3.52	0.88	1.93	1.22	0.48	3.59	**7.4**
	47	14.08	12.99	-1.15	-13.41	-12.16	4.03	0.59	0.45	0.07	0.05	0.39	0.57	**1.29**	1.25	0.23	0.4	0.34	0.16	2.22	**2.8**
	48	-8.2	-12.44	0.2	-11.2	5.8	12.92	0.17	0.35	0.05	0.1	0.1	0.43	**0.71**	2.95	0.21	0.8	0.7	0.05	2.58	**5.25**
	49	11.96	-13.56 1	0.96	12.36	-7.44	3.4	1.11	0.57	0.15	0.18	0.13	0.36	**2.11**	1.16	0.38	0.7	0.64	0.19	2.47	**3.75**
	Mean values							0.58	0.37	0.12	0.08	0.1	0.49	**1.4**	2.09	0.34	0.8	0.41	0.2	2.11	**4.12**

Table 2.7: First and last 5 results (for light and heavy misregistration) of the robustness experiment. Left column shows trial parameters in *mm* and *deg*, following columns give the absolute difference to the recovered parameters for the two methods. The last row is the mean of the errors for each parameter and for all transformations applied.

		Original transformation						Error in recovered transformation												
								Creaseness							Mutual Information						
	N	θ_x	θ_y	θ_z	t_x	t_y	t_z	$\Delta\theta_x$	$\Delta\theta_y$	$\Delta\theta_z$	Δt_x	Δt_y	Δt_z	MD	$\Delta\theta_x$	$\Delta\theta_y$	$\Delta\theta_z$	Δt_x	Δt_y	Δt_z	MD
Dataset 5	0	0	0	0	0	0	0	0	0	0	0	0.01	0	**0.01**	3.34	0	0.84	0	0.5	0	**5.72**
	1	-1.21	-3.01	0.33	3.82	-0.34	0.52	0.35	0.02	0.05	0.01	0.03	0.31	**0.62**	3.71	0.21	1.4	0.18	0.47	0.98	**6.71**
	2	3.54	-1.5	0.66	-1.7	-3.83	-2.2	0.82	0.35	0.17	0.02	0.11	0.49	**1.52**	0.14	0.54	1.32	0.2	0.08	0.7	**3.21**
	3	-1.54	-2.03	-2.42	0.04	-1.58	-0.86	0.78	0.24	0.07	0.04	0.18	0.16	**1.33**	4.41	0.18	0.44	0.04	0.33	1.11	**7.19**
	4	-0.69	0.13	0.77	-0.7	0.35	-0.42	0.89	0.16	0.13	0	0.18	0.47	**1.53**	3.31	0.13	0.95	0.18	0.34	0.58	**5.77**
	45	13.86	7.42	-3.78	15.71	7.04	-5.73	0.66	0.4	0.07	0.07	0.16	0.26	**1.23**	0.94	0.15	0.22	0.15	0.84	0.85	**2.03**
	46	-15.81	-13.44	11.78	6.66	-11.71	9.41	1.79	0.12	0.15	0.29	0.56	0.62	**3.03**	4.35	1.12	2.35	0.72	0.21	2.09	**8.47**
	47	14.08	12.99	-1.15	-13.41	-12.16	4.03	**	**	**	**	**	**	****	0.24	0.23	0.46	0.03	0.16	0.22	**1.17**
	48	-8.2	-12.44	0.2	-11.2	5.8	12.92	1.57	0.22	0.06	0.17	0.45	0.62	**2.64**	4.56	0.03	1.21	0.26	0.51	1.64	**7.52**
	49	11.96	-13.56	10.96	12.36	-7.44	3.4	0.85	0.37	0	0.13	0.21	0.47	**1.54**	0.03	0.16	1.18	0.27	0.32	0.72	**2.7**
	Mean values							0.86	0.21	0.08	0.08	0.21	0.38	**1.53**	2.5	0.28	1.04	0.2	0.38	0.89	**6.42**

Table 2.8: Same figures than table 2.7. ** indicates unsuccesfull convergence.

2.5.4 The Vanderbilt database

It has been proved [15] that the accuracy of retrospective registration can be consistently assessed visually by a human observer. But since our site is outside a medical centre, we were unable to measure the rate of mis-registration by our own means. Even if we had access to original images from a medical site, developing a full assessment would require a long and intensive collaboration with a medical team (large number of images, validated several times) which we could not afford. Dr.Fitzpatrick's project kindly provided us with this chance. We acknowledge Dr. J.M. Fitzpatrick, head of the project "Evaluation of retrospective image registration" (project number NIH R01 NS33926-01) from Vanderbilt University for providing us with their image database.

The objective of the study undergone at Vanderbilt's University [114, 113] was to perform a comparison of the accuracy of some automatic retrospective registration methods. The University provided the image database to the sites where the actual registration had to be taken, and results were given back using an objective protocol.

The comparison was made against a golden standard, a prospective marker-based method, and markers were removed from the images to ensure the blindness of the experiment. Also, the communication protocol was based exclusively on the Internet and designed to detect any inconsistency in the results transformations.

The database consisted in images from seven patients. For each, the CT image was to be matched against three MR modalities: MR–T1, MR–T2 and MR–PD, and also against a PET image. At a second phase, not included in the papers published, nine extra patients were added (numbered 1–9 B), with a extra MR-RAGE modality. Characteristics are summarised in table C.1, page 209. In total, 70 image pairs.

The accuracy of the methods is evaluated at multiple volumes of interest (VOI), chosen between areas in the brain of neurological interest as follows: the coordinates of centroid of the VOI in the MR image (c) are transformed to the CT using a known (see below) Golden Standard transformation (c'), and then back to the MR using the evaluated method (c''). The difference between c and c'', called the target registration error *TRE*, is evaluated at 6 zones for each pair of images to be registered.

The Golden Standard transform against whom the other are compared employed four fiducial markers attached to the bone. These markers, filled with a liquid visible in the imaged modality, are rigidly attached to the bone and thus are considered to provide a very good accuracy. For each image, the markers were segmented and their coordinates recorded with sub-pixel accuracy by means of a location technique, published in [107]. Then, the two lists of corresponding coordinates were aligned as in [2], which gave the Golden Standard transformation parameters between the two source images. To ensure the blindness of the study, the markers were removed from the image by substituting their intensity values by neighbouring empty voxels.

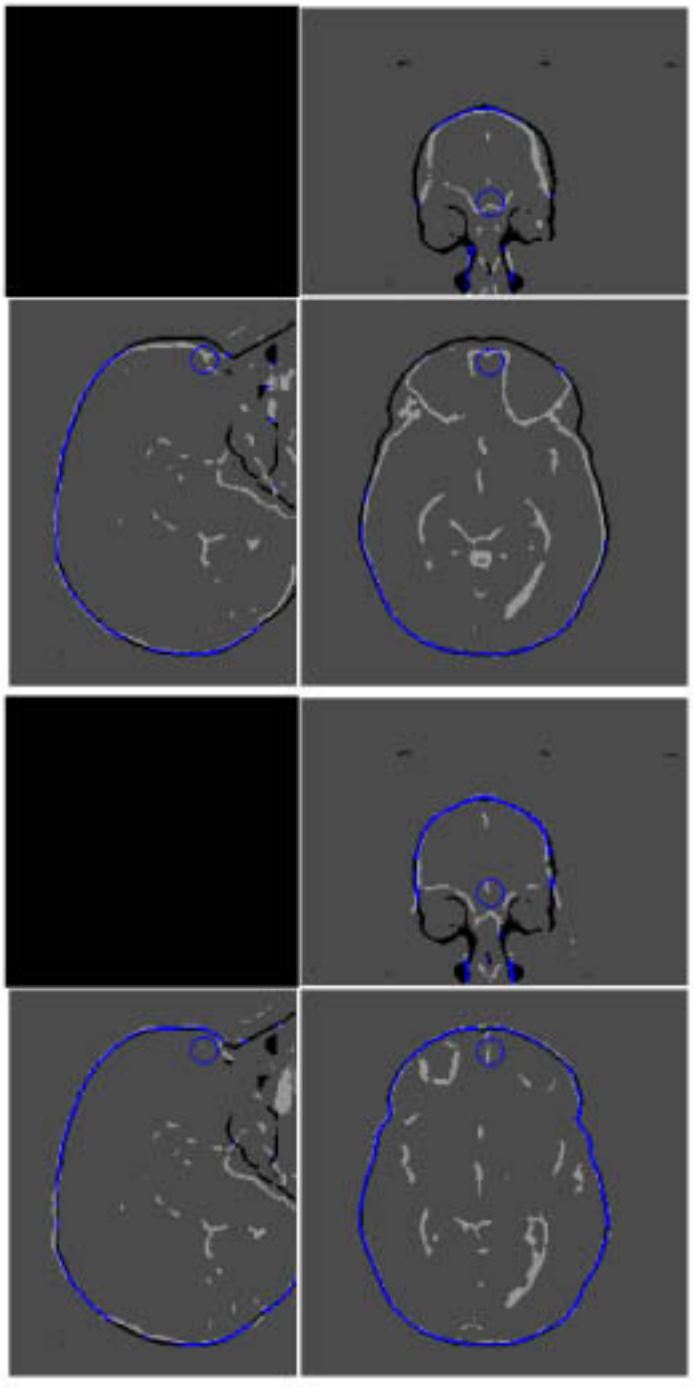


Figure 2.26: For some pairs of images, further optimisation of alignment including scaling parameters improves results spectacularly with respect to the rigid solution. Top: the rigid registration misses most of the frontal bone, which can be seen (bottom) to match perfectly on the scaled solution.

Patient	CT-PD	CT-PDr	CT-T1	CT-T1r	CT-T2	CT-T2r	CT-RA	RA-T2
1–A	3.5	1.5	3.6	3.2	1.7	3.3		
2–A	2.4	1.3	3.6	3.9	1.8	2.8		
3–A	1.7	2.4	4.5	5.6	1.6	1.4		
4–A	3.0	2.1	9.3	6.3	1.2	1.8		
5–A	2.0	1.8	1.3	1.7	3.0	1.6		
6–A	2.8	2.6	4.5		1.2	0.5		
7–A	2.2	1.5	2.1	2.6	1.0	1.4		
1–B	2.7		3.0		4.1		5.8	7.9
2–B	2.5		4.4		2.1		1.1	4.8
3–B	3.0		3.9				10.2	
4–B	2.4		2.8		1.7		7.9	7.2
5–B			4.5		2.0		1.6	3.6
6–B			3.1		3.1		1.7	4.3
7–B			1.8		2.6		2.4	5.1
8–B			8.0		1.4		2.5	2.9
9–B			2.4		1.9		3.7	3.5
Median	2.51	1.73	3.09	3.41	1.89	1.62	2.7	4.54

Table 2.9: Numerical results of the Vanderbilt tests: for each patient and pair of images, we give the median of the error once run our method reported at the VOIs. For each pair we have framed the highest value. Void cells denote non-available modality pairs.

A secondary goal of the project was to evaluate the importance of correcting geometrical distortions in the original images. Such distortions, which often occur , are caused by defects in the calibration of the machine and have the bothersome effect to widen the registration error. To account for this distortions, a COMPASS stereotactic frame was attached also to the patient, and its coordinates in the image were then employed to correct the voxel spacing provided by the scanner. The resulting image is called with the prefix *rect*.

The original study comprised 14 different registration techniques from 11 researches in several countries. It includes manual, voxel-based and segmentation-based techniques. For some of them, the mean error was consistently around 1 mm, and therefore the authors concluded that automatic registration had the potential to produce satisfactory results.

After the papers had been published, the Vanderbilt University offered to validate the additional submissions in a similar procedure. For our team it was very important to test the algorithm under the variety of settings provided because, as a matter of fact, leaded to several improvements, as reported in previous sections.

For instance, we discovered that for a particular MR modality, MR_T1, the assumption that the skull appears as a valleys does not hold for the upper slices. To solve this, we selected a lower scale for the computation of derivatives, to segment the two layers of bone, and then the method converged to one of them. See figure 2.27 for the visual explanation.

Modality	Surface group		Creaseness registration		Volume group		N
Pair	Mean (std.)	% > 10	Mean (std.)	% > 10	Mean (std.)	% > 10	
CT–T1	5.7 (7.8)	11.7	3.8 (3.8)	9.2	2.9 (2.4)	1.2	7
CT–PD	5.8 (8.0)	11.3	2.4 (0.8)	0	2.9 (2.5)	1.6	7
CT–T2	6.3 (7.9)	12.3	2 (1.0)	0	2.4 (1.4)	0	7
CT–T1 rect.	6.1 (8.3)	13.1	4 (2.2)	0	2.0 (2.5)	1.9	6
CT–PD rect.	5.7 (7.8)	12.0	1.8 (0.7)	0	1.8 (2.0)	0.0	7
CT–T2 rect.	6.1 (7.6)	12.1	1.9 (1.2)	0	2.1 (1.6)	0.0	7

Table 2.10: Global results compared in groups, as in [113]. The error threshold of 10 mm is set in order to have some estimation of misregistered results.

Another item we explored was to include the scaling factor along each axis as a further parameter to optimise. To avoid too large degrees of freedom, as a first step we registered the two images without them, and then we ran again the algorithm with the scaling using the previous solution as a single seed. Results were better for all cases; for some of them, the correlation value raised up to 20% higher of the value obtained without scaling.

Table 2.26 gives a visual example of the improvement obtained with the scaling optimisation. However, we could not validate these results because the Vanderbilt protocol to transmit results to between our and their site considers exclusively rigid transformations.

Also, we run the whole set of experiments using the chamfer distance as the measure of matching between the two crease surfaces. Although it worked well for most cases, in others the algorithm failed absolutely to converge to a proper solution. Therefore, we did not further investigate the accuracy.

The global results of our algorithm are presented in table 2.9; they have been published at Vanderbilt site at [14]. Appendix C lists full transformations and statistics. In addition to the median for each modality, it is important to take into account the maximum value as well, because it gives and indication of the worst mis-registration to be expected. Modalities CT–PD and CT–T2 present best median values, and also lowest maximum values. The T1 modality suffers from the segmentation problem, as reported previously and, although it never fails, for some patients the alignment is not precise.

The results of pairs CT to MR–Rage are good for most patients, although for others, such as 1–B, 3–B and 4–B, the median is so high it probably would not be suitable for clinical routine. The same should hold for most registration of Rage to T2.

We have ranked our results against those of other research groups participating in the Vanderbilt project. The first comparison is made in table 2.10: we have grouped the methods which participated in the initial evaluation phase according to their paradigm: voxel-based and segmentation-based, same as in [113].

In order to make comparisons fair, our results have been restricted to the same modalities and group of patients as the two groups. This grouping permits to make general evaluations: for instance, segmentation-based have a worst profile than voxel-

based, in mean error as well as in maximum value. The accuracy of the results of our method is similar to that of voxel-based.

The last comparison is made individually for all the methods included in the original publication. Each method has been abbreviated as in the papers: Barillot et al (BA, [40]), Collignon et al (CO, [5]), van den Elsen et al (EL, [103]), Harkness (HA, [71]), Hemler et al (HE,[29]) Hill et al (HI, [31, 30, 93]), Maintz et al (MAI [57, 58], [56]), Malandalain et al (MAL, [60, 61, 62]), Noz et al (NO, [55]), Pelizzari (PE, [71]) and Robb et all (RO, four methods [27]).

Because of the low number of patients, a ranking of the methods is not statistically significant but still it is interesting to make some remarks. The method most similar to ours is that of MAI: it is segmentation-based, but uses the boundaries as surfaces to match. Their results consistently have a mean error of around 4 to 5 mm, and are independent of the modality. Conversely, our results are not, but the error is lower for all but MR–T1. Of course, differences may be caused by multiple factors, like the optimisation method and the accuracy of edginess operators, but it seems that for registration purposes the centre of skull may be more suitable than the edges of the head.

An open issue is the reason why segmentation-based methods do not improve accuracy for rectified modalities, something that voxel-based methods do. West, in [113], left the question unanswered but in further personal discussions by e-mail he suggested the improvement may be hidden by the high error of segmentation-based methods, while it would show for the other, more accurate, methods. However, this explanation does not seem to apply to ours: indeed, it ranks as one of the most accurate for MR–T2, but the error is not lower for the corresponding rectified modality.

In our opinion, the reason is that the optimisation space has broad maximum for the z translation, caused by the particular shape of the head. One pixel translation in the x or y axis misaligns the whole structure, while the same amount of translation in the z axis remains hidden because the shape changes relatively less along the axial direction. In addition, the MR rectified modality changes about 0.01 mm in the x and y voxel scaling, but 0.09 mm in the z axis, precisely where our method is less sensitive. That would explain why voxel-based methods, which do not depend on the shape of segmented features, improve.

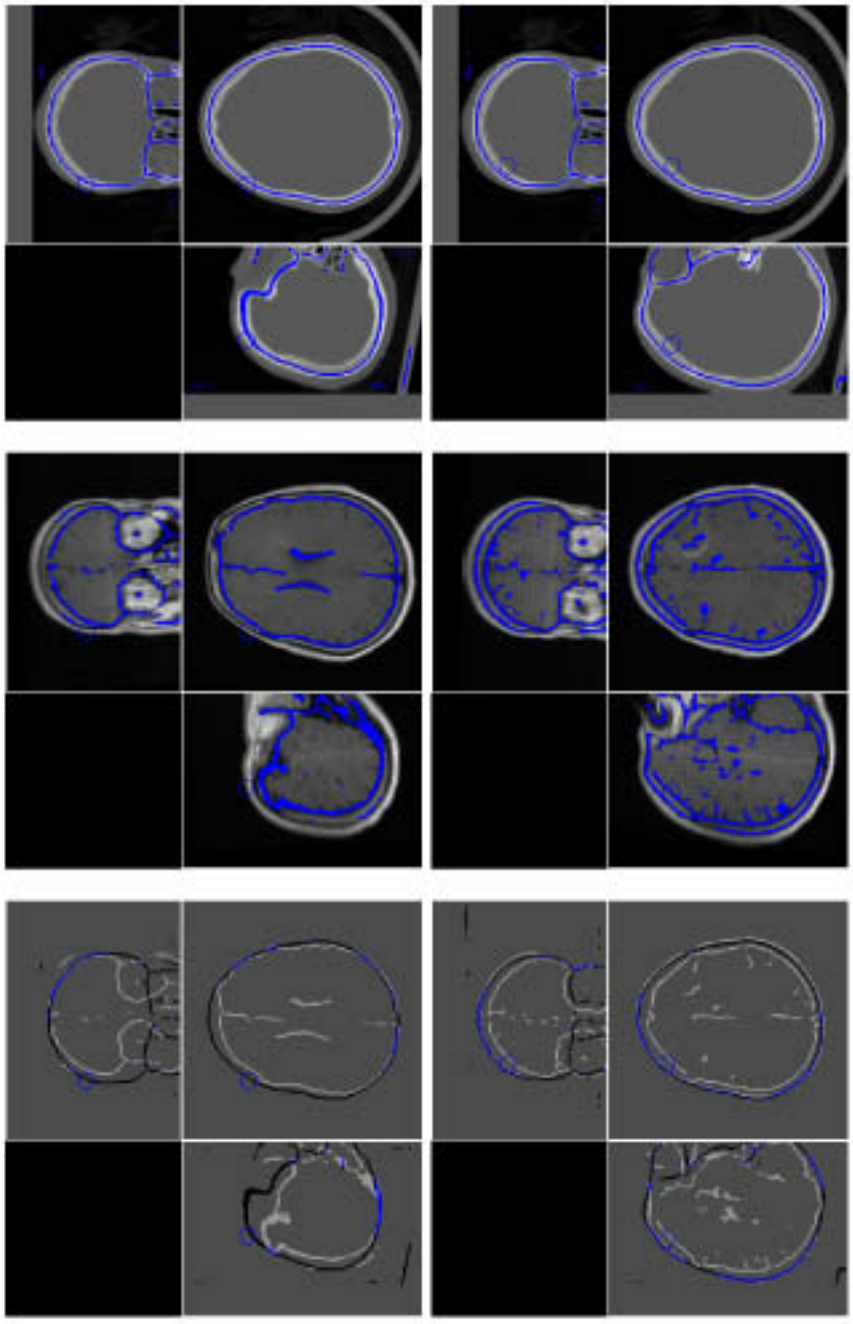


Figure 2.27: Top: Patient 6–B for the CT–MR_T1 registration gave the worst results due to a poor segmentation. After adjusting the scale (bottom), the bone in the higher axial slices was segmented properly, and images could be aligned.

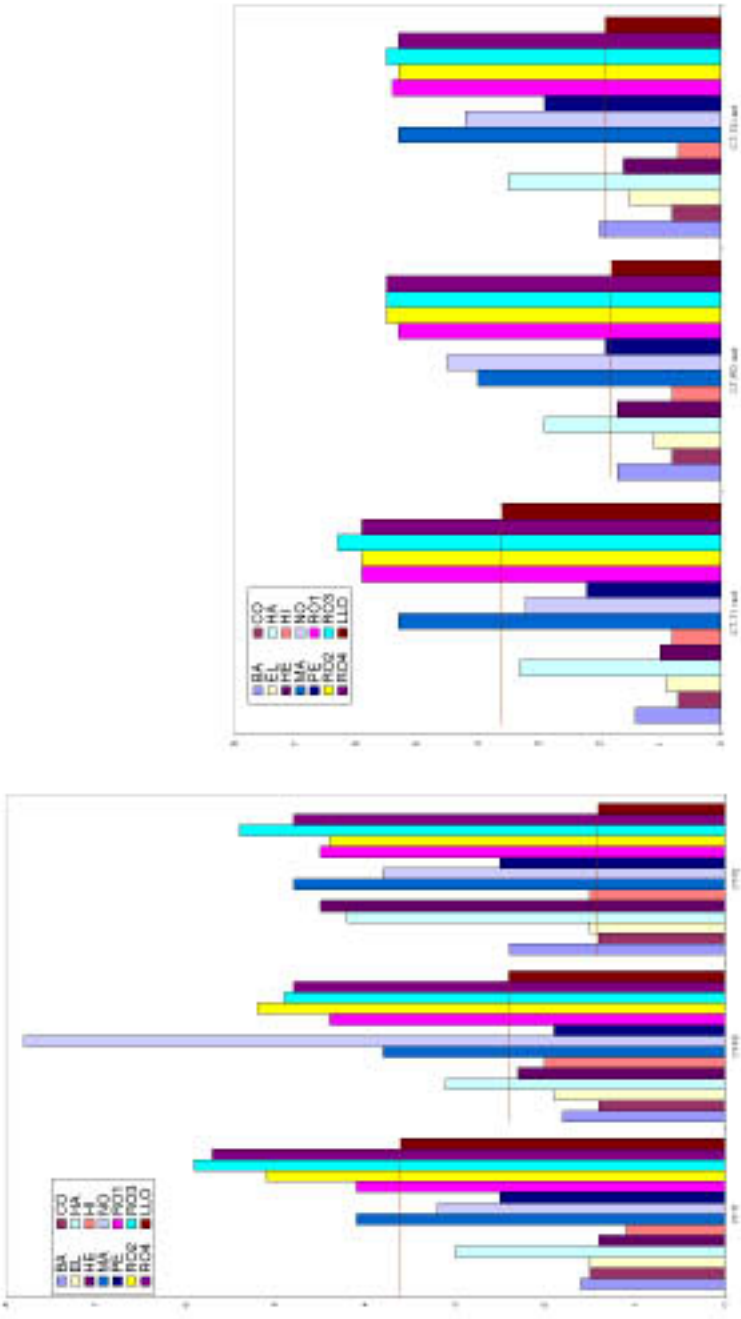


Figure 2.28: Comparison of all methods for each pair of modalities. Abbreviations are the same as in [113], reproduced in page 71. Our method is the last bar in each row, named as *LLO*.

2.6 Conclusions

In this chapter we have presented in detail an algorithm for brain image registration. We have evaluated its performance under a variety of conditions, and validated it with a large image database. For one modality, CT to MR–T2, our method achieves the best results compared to the state of the art (even mutual information), which indicates that the alignment optimisation step of the algorithm is properly designed for this problem. Also, it compares favourably to another segmentation-based method employing the boundaries as anatomical landmarks, which indicates that the creaseness operator to extract the skull is precise and robust.

A major problem occurs with some images in MR–T1 modality. For these, the assumption that the skull depicts as a valley is no longer valid, specially at distal axial slices, because the marrow shows bright between two layers of bone. For these cases, the algorithm aligns the segmented surface in the CT image to one of these layers while keeping the proximal slices registered. Therefore, the final accuracy is not as good as for the other modalities.

Since our method takes into account only the rigid transformations of the skull and ignores the soft tissues, we think it can very well suit the applications which need to compute only the registration of the skull. For instance, in studies which need to compare images taken over a long time, our method can be useful because the skull is the only undeformable structure of the head, while others are often changed by chronic diseases. A further step could be to apply intensity-based registration algorithms to specific parts of the resulting registered images. This second registration would give a measure of their relative movement and of their volume changes.

We have not explored yet a further improvement consisting on extracting the creaseness features at different scales, and combine the resulting creaseness images simply by computing the correlating simultaneously at all the scales. Then, the alignment measure would take into account common features other than the skull. However, this scheme would possibly have an alignment function not so well posed, and therefore an initial alignment by means of the original algorithm would be necessary before further refinement.