November 1988

Preliminary Clinical Evaluation of Elastic Matching System

Robert Dann
University of Pennsylvania

John Hoford
University of Pennsylvania

Stane Kovacic
University of Pennsylvania

Martin Reivich
University of Pennsylvania

Ruzena Bajcsy
University of Pennsylvania

University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-88-35.

Preliminary Clinical Evaluation of Elastic Matching System

Abstract

In order to evaluate the performance of our elastic matching system, we have created a digitized atlas from a young normal male brain, using 135 myelin-stained sections at 700 *micron* spacing. Software was written to enter and edit regional anatomic contours, which were stacked and aligned to create a 3D atlas. We then evaluated the matching system by comparing computer generated contours with expert-defined contours for several subcortical structures, based on CT scans from six neurologically normal patients. The error in positioning, as defined by the distance between the centers of gravity, averaged 4.2 *mm* for the computer and 1.7 *mm* for the worst expert's reading, with the computer-drawn region frequently inscribed within that of the expert. Comparison was also made for each structure by determining the volume of overlap and the volumes not overlapping. On average, the computer's agreement with the experts was about 20% less than the agreement among the experts. This was a preliminary test of the system using only subcortical structures. The results are promising, and techniques are being implemented to overcome the current deficiencies.

Comments

University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-88-35.

PRELIMINARY CLINICAL EVALUATION OF ELASTIC MATCHING SYSTEM

Robert Dann, John Hoford, Stane Kovacic, Martin Reivich and Ruzena Bajcsy

MS-CIS-88-35
GRASP LAB 142

Department of Computer and Information Science
School of Engineering and Applied Science
University of Pennsylvania
Philadelphia, PA 19104

November 1988

Acknowledgements: This work was in part supported by: NSF grant DCR-8410771, Airforce grant AFOSR F49620-85-K-0018, Army/DAAG-29-84-K-0061, DAA29-84-9-0027 NSF-CER/DCR82-19196 Ao2, DARPA/ONR NIH grant NS-10939 -11 as part of Cerebro Vascular Research Center, NIH 1-RO1-NS-23636-01,NSF INT85-14199, NSF DMC85-17315, ARPA N0014-85-K-0807, NATO grant No. 0224/85 by DEC Corp., IBM Corp. and LORD Corp.

Preliminary Clinical Evaluation of Elastic Matching System

Robert Dann, John Hoford, Stane Kovacic

Martin Reivich, Ruzena Bajcsy

General Robotics and Active Sensory Perception Laboratory

UNIVERSITY of PENNSYLVANIA

Department of Computer and Information Sciences

Philadelphia, Pennsylvania 19104-6389

Abstract

In order to evaluate the performance of our elastic matching system, we have created a digitized atlas from a young normal male brain, using 135 myelin-stained sections at 700 *micron* spacing. Software was written to enter and edit regional anatomic contours, which were stacked and aligned to create a 3D atlas. We then evaluated the matching system by comparing computer generated contours with expert-defined contours for several subcortical structures, based on CT scans from six neurologically normal patients. The error in positioning, as defined by the distance between the centers of gravity, averaged 4.2 *mm* for the computer and 1.7 *mm* for the worst expert's reading, with the computer-drawn region frequently inscribed within that of the expert. Comparison was also made for each structure by determining the volume of overlap and the volumes not overlapping. On average, the computer's agreement with the experts was about 20 % less than the agreement among the experts. This was a preliminary test of the system using only subcortical structures. The results are promising, and techniques are being implemented to overcome the current deficiencies.

Acknowledgement: This work was in part supported by: NSF grant DCR-8410771, Airforce grant AFOSR F49620-85-K-0018, Army/DAAG-29-84-K-0061, NSF-CER/DCR82-19196 Ao2, DARPA/ONR NIH grant NS-10939 -11 as part of Cerebro Vascular Research Center, NIH 1-RO1-NS-23636-01,NSF INT85-14199,NSF DMC85-17315, ARPA N0014-85-K-0807,NATO grant No.0224/85. by DEC Corp., IBM Corp. and LORD Corp.

1 Introduction and motivation

While human observers are able to successfully interpret qualitative information in radiographic data, objective measurements of regional size, density or volume pose particular problems. For certain types of images, like positron emission tomography (PET) scans, which represent function, e.g., the distribution of flow or metabolism, it is useful to have independent anatomic information. To interpret such scans, many groups have chosen to collect anatomic images (computer tomography - CT, or magnetic resonance scans - MRI) which are in some way aligned with the PET scans. Regions of interest (ROI's) are first defined on the anatomic images and then superimposed on the PET scans to extract functional data within those anatomic regions [1,2,3]. Other groups define ROI's directly on the PET images, using a reference atlas or template of ROI's based on average anatomy [4,5]. The goal of our work is to assist in both approaches, that is to develop improved methods to perform quantitative analysis of images that are considered to reflect anatomy (CT and MRI), and to define standardized anatomic ROI's that may be used to extract data from images that represent brain function, like PET or single photon emission computed tomography (SPECT).

In order to evaluate the performance of our matching system [6,7], we have created a digitized atlas from a young normal male brain. This 3D anatomy atlas can be matched to the anatomic scans, and then superimposed on an aligned set of PET scans. This paper reports our progress towards the matching of the atlas with CT scans, while the accuracy of matching the atlas with PET scans remains as a future project.

The rest of this paper is organized as follows: section 2 provides a brief overview of the matching process; section 3 describes the construction of the anatomy brain atlas; section 4 details the materials used, the experiments involved, reports the computer performance illustrated with the inter- and intra-observer variability, and shows the limitation of the current system; the final section consists of conclusions.

2 Matching system

The input to our matching system are CT scans and PET scans of the same subject. The matching is based on the anatomy brain atlas (or a set of atlases), which represents a three-dimensional anatomic model of a healthy human brain. The overall matching process consists of:

- preprocessing of CT and PET images,
- localization of brain in CT and in PET images,
- global alignment or registration of CT brain, PET brain and anatomy brain atlas images,

- individualization of anatomy brain atlas by elastically mapping it onto CT brain images,
- analysis of anatomy in CT images and analysis of metabolism in PET images using individualized brain atlas.

All code is written in the C programming language, running under UNIX operating system on a Hewlett-Packard workstation. The developed matching process is three-dimensional without giving preference to the slicing plane, since plane of section can vary from patient to patient. First, preprocessing and brain localization take place. This gives us 3D images of the patient's CT and PET brain with cubically shaped voxels of size ($1.0\,mm \times 1.0\,mm \times 1.0\,mm$), which are used in global alignment. Next, the 3D PET brain image is aligned with the 3D CT brain image by translation and rotation, and the 3D atlas image is aligned with the CT brain image by translation, rotation and scaling. To derive transformation parameters, we approximate each brain by an ellipsoid-like scatter of particles uniformly distributed in space and find the centroid and the covariance matrix for each of the brain images. The translational difference between two brains is eliminated by aligning the centers of mass. For rotation and scaling correction we use the method of principal axes. Although this method is not as accurate as the method reported by Pelizzari [8], we accepted its result as the initial approximation of the elastic matching. After global alignment, the idealized atlas is transformed into an individualized atlas by elastically deforming it to match CT brain. Deformation proceeds step-by-step using coarse-to-fine strategy. The outer edge of the atlas brain is matched to the outer edge of the CT brain and the atlas ventricles are matched to the CT brain ventricles [7]. Other anatomic structures are deformed as a side-effect of ventricles and outer edge matching. Finally, the deformed atlas is used to facilitate the analysis of the patient's CT images and/or PET images. The atlas can be visualized as two-dimensional sections through the tissue. Each structure may be either displayed as a solid or as a contour, optionally superimposed on the corresponding CT and PET sections.

3 Brain atlas

The process of constructing an idealized brain atlas requires, among other things, images of a normal brain at sufficiently fine spacing to eliminate, or at least minimize, artifacts introduced by interpolation. It is also desirable that the brain sections be obtained at a standard plane of section, to facilitate the definition of structures by reference to existing atlases. The danger of using a single brain, even if it is normal, is that it might, by chance, represent an extreme of the normal distribution. The atlas that we have created from this brain must, therefore, be considered as only the initial step towards a final atlas (or atlases) that will evolve with time. It is commonly felt [9] that the best atlas is one based on average normal MRI scans. Our software has the potential to provide such an atlas by elastically matching the initial atlas to a large number of MRI scans and then creating an atlas based on the average sizes and positions of anatomic structures.

There were no existing atlases that met all of our criteria, so we chose to create one from the brain of a normal 31-year old male in the Yakovlev Collection of the Armed Forces Institute of Pathology. This brain was embedded in celloidin and cut parallel to the plane bisecting the anterior and posterior commissures. Sections were 35 μ thick, and every 20*th* section was stained for myelin. (Contiguous sections were stained with cresyl violet, so the potential exists to go back to them in order to define cytoarchitectural regions). This resulted in tissue sections that were 700 μ apart, which corresponded to 0.89 *mm* in vivo after correcting for the shrinkage due to the embedding and staining processes. Each section was photographed from a fixed location, with a ruler included for later scaling. The negatives were enlarged onto dimensionally stable acetate sheets, and these were then digitized using a video camera connected to a Vax computer. Figure 1 shows one of 135 digitized images.

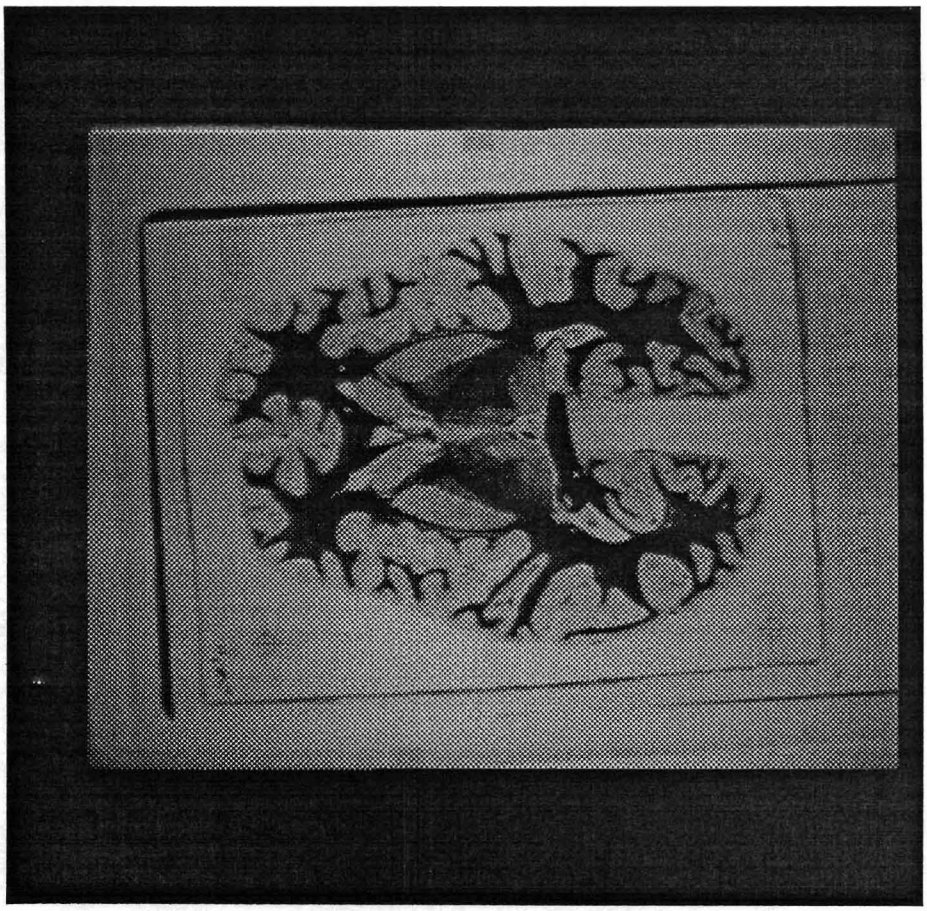

Figure 1. Digitized picture of one section of the anatomy brain atlas.

The images of the rulers were carefully measured, and any slight differences in any of the photographic steps were corrected by appropriate software scaling. The digitized images, which contained fiducial markers (burr holes drilled through the brain while in situ) were aligned with interactive software and stacked to create a cube of data. This cube was then resliced using

1.0 *mm* voxels, which allows reformatting in other image planes without loss of resolution.

The atlas was created by manually tracing various structures on the original digitized images, using a mouse to draw the contours. One experienced neuroanatomist drew outlines of several structures and his outlines were checked independently by another, also experienced neoroanatomist. Software to do this was written in the C programming language on a Hewlett-Packard workstation, running under UNIX and X-Windows. Because of the myelin staining, the boundaries of subcortical structures and gray/white interfaces are clearly defined. Cortical regions may be assigned using standard reference sources [10, 11], and the outer edge of the brain and the ventricles are readily apparent.

Two distinct atlases are created by coloring the structures differently. In the atlas that is used for elastic matching, each structure should be assigned a color (grey scale values) that corresponds to the color that structure would be on a CT scan. While there is noise in the CT scans, the atlas brain ventricles and the rest of the atlas brain are colored uniformly using an average grey value of the corresponding regions in the CT brain (figure 2-left). In the other atlas (Figure 2-right), which is used for display and analysis, the structures can be assigned arbitrary colors to best display the results.

Figure 2. The brain atlas. Only one horizontal section is shown. The left picture shows one section of the atlas which is used for matching, the right picture shows the corresponding section of the atlas which is used for display and analysis.

4 Evaluation of the matching

4.1 Materials used

We used CT scans obtained from GE9800 scanner, located at the Hospital of the University of Pennsylvania. Each patient was represented by 26 to 28 slices. Slices were 5 *mm* apart with pixel size of (0.41 *mm* $\times$ 0.41 *mm*). To provide a preliminary test of how well the matching system performs, we planned to analyze the CT scans of six young normal males corresponding closely to the subject whose brain was used to create the atlas. Unfortunately it was impossible to obtain finely spaced CT scans on young controls, so the subjects that are included in this report include three male and three female patients, whose neurological status was either normal or whose pathology did not involve major distortions of normal anatomy. We evaluated the matching system by comparing computer-generated contours with expert-defined contours. Since subcortical structures are relatively well-defined, we felt that there would be more variability in the way experts would draw cortical ROI's. If the matching software did not work well with subcortical regions, it would almost certainly do even worse on cortical regions, so we chose to perform this initial test using only subcortical structures. The computer and several experts defined the contours of a variety of these subcortical structures.

4.2 Region of interest analysis

One neuroradiologist, who was particularly interested in and knowledgeable about anatomy, drew contours on all structures at all levels for all subjects. Another neuroradiologist also outlined structures for two of the patients, while a neurologist experienced in interpreting CT and PET scans outlined all structures in all subjects. Finally, a neuroanatomist, also experienced in interpreting tomographic images, did the complete analysis three times, separated by at least two weeks, in order to establish a measure of intra-observer variability. All readings were done independently and without discussion, using the same image contrast and brightness settings. There was no way to know objectively where the true edges of the structures were located and we assumed, that the experts' readings were scattered around the truth. The readers agreed in advance upon which structures were present on each CT slice. The analyses were based only on these sections, although in some cases the computer produced regional contours for these structures on other section images as well. The computer matching was performed on the same six subjects and the regional contours defined in this process were compared with the experts' readings.

4.3 Elastic matching results

We included in this preliminary test the following subcortical regions: body of the caudate nucleus, head of the caudate nucleus, lenticular nucleus (putamen and globus pallidus), thalamus, brainstem (not divided) and cerebellum. In the subsequent descriptions the experts' readings are denoted by WA, JC, DG, and RD. C is the computer-defined reading and patients are denoted by P1, P2, P3, P4, P5 and P6. The regional indices are: 111-left body of caudate nucleus, 112-right body of caudate nucleus, 131-left head of caudate nucleus, 132-right head of caudate nucleus, 135-left lenticular nucleus, 136-right lenticular nucleus, 117-left thalamus, 118-right thalamus, 201-brainstem, 301-cerebellum.

For each individual ROI, the volume and the center of gravity were calculated for the experts and for the computer. Tables 1 and 2 demonstrate the computer performance by comparing the regional volume. The values in Table 1 show that the volume defined by the computer tends to be smaller than the volume defined by the experts. In table 2 we see that of 60 structures analyzed, there were 28 structures with the computer defined volume in the range of experts' variability.

Table 1. Comparison by volume for one patient.

Structure	Patient P6				
	Low	Mean	High	C	IN / OUT
111	409	1071	1634	537	IN
112	516	1012	1323	675	IN
131	2418	2928	3486	1878	OUT
132	2334	2920	3737	2301	OUT
135	4581	4906	5315	5305	IN
136	4177	4760	5777	5480	IN
117	5567	6416	7056	6358	IN
118	6101	6490	7377	6133	IN
201	14680	16308	18162	23873	OUT
301	104841	107151	108592	127647	OUT

Comparing the computer-generated volume with the experts-defined volume for one of the patients. "IN" indicates that the volume generated by the computer "C" was within the range of variability among the experts. "Low" is the smallest and "High" is the largest volume among the six experts' readings with the mean volume "Mean". All values are in mm^3.

Table 2. Comparison by volume for all patients.

Structure	P1	P2	P3	P4	P5	P6
111	OUT	OUT	OUT	OUT	IN	IN
112	OUT	IN	OUT	IN	IN	IN
131	IN	IN	IN	OUT	IN	OUT
132	IN	IN	OUT	OUT	IN	OUT
135	OUT	OUT	OUT	IN	IN	IN
136	IN	IN	OUT	IN	OUT	IN
117	OUT	IN	IN	OUT	IN	IN
118	IN	IN	OUT	OUT	OUT	IN
201	OUT	OUT	OUT	OUT	OUT	OUT
301	IN	OUT	OUT	OUT	OUT	OUT

Comparing the computer-generated volume with the experts-defined volume for all six patients. "IN" indicates that the volume generated by the computer was within the range of the experts.

Tables 3 and 4 show how the coordinates of computer-defined center of gravity for each individual ROI compare with those defined by the experts. Table 3 shows the positioning errors for one patient, where $d_x(i)$, $d_y(i)$, and $d_z(i)$, are the positioning errors of reading i with respect to the n other readings, in direction x, y and z, respectively:

$$d_\alpha(i) = \frac{1}{n}\sum_{j=1}^{n} |\alpha_i - \alpha_j|, \quad (\alpha = x, y, z). \tag{1}$$

Table 4 summarizes the positioning errors for all six patients, where D(i) is the positioning error of reading i with respect to the n other readings,

$$D(i) = \frac{1}{n}\sum_{j=1}^{n} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}\,. \tag{2}$$

The mean error in positioning by the computer for all structures, as defined by the average distance in *mm* between the computer-placed center of gravity and the experts' defined centers of gravity was 4.2 *mm* and for the worst expert's reading 1.7 *mm* .

Table 3. Positioning errors for one patient.

Patient P1						
Structure	$d_x(E)$	$d_y(E)$	$d_z(E)$	$d_x(C)$	$d_y(C)$	$d_z(C)$
111	0.7	1.7	0.0	1.3	3.0	0.0
112	0.9	0.5	0.0	2.6	5.7	0.0
131	0.2	0.3	0.5	0.9	1.5	2.2
132	0.0	0.5	0.5	0.9	2.8	1.9
135	0.4	1.0	0.2	0.9	0.6	2.3
136	2.3	2.6	1.1	0.7	3.0	1.1
117	0.9	1.2	0.4	1.5	4.1	1.4
118	0.7	0.5	0.6	0.5	2.9	0.3
201	0.8	0.4	0.3	4.7	2.5	0.2
301	0.1	1.3	1.4	4.6	1.8	1.1

Computer-generated positions were compared with the experts-defined positions and the differences $d_x(C)$, $d_y(C)$, $d_z(C)$, are shown above in columns 5-7. In addition, each expert was compared with the other experts and the differences of the worst expert $d_x(E)$, $d_y(E)$, $d_z(E)$, are shown in columns 2-4. All numbers are in units of mm.

Table 4. Positioning errors for all six patients.

Structure	P1 D(E), D(C)	P2 D(E), D(C)	P3 D(E), D(C)	P4 D(E), D(C)	P5 D(E), D(C)	P6 D(E), D(C)
111	1.9, 3.3	1.8, 3.0	2.6, 10.3	3.0, 16.6	2.8, 6.8	6.7, 7.0
112	1.1, 6.3	1.8, 2.6	1.3, 11.6	3.8, 13.0	2.4, 6.3	5.1, 7.3
131	0.6, 2.8	1.2, 2.0	0.8, 2.2	1.1, 2.8	0.7, 1.7	1.5 2.7
132	0.7, 3.5	1.3, 2.0	1.2, 3.3	1.0, 2.7	1.0, 3.0	1.3, 1.9
135	1.1, 2.6	1.4, 2.7	1.5, 2.4	2.3, 3.1	1.3, 1.8	2.0, 3.7
136	3.6, 3.3	2.3, 2.7	1.6, 5.5	2.8, 4.2	1.4, 3.9	1.7, 2.7
117	1.6, 1.6	1.9, 4.5	1.5, 1.3	1.9, 1.9	2.6, 3.0	0.8 3.5
118	1.1, 3.0	3.0, 4.5	1.1, 2.7	1.3, 2.1	1.3, 5.5	0.8, 3.5
201	0.9, 5.4	1.0, 1.7	1.4, 4.4	1.6, 4.4	1.0, 2.4	1.1, 2.4
301	2.0, 5.1	0.6, 2.7	2.0, 7.9	0.5, 6.7	1.1, 5.6	0.5, 5.2

Comparing the computer-defined positions with the experts for all six patients. For each individual ROI, the center of gravity was calculated for the experts and for the computer. The computer generated positions are compared with the experts' positions, and the distances D(C) are tabulated above, together with the worst expert distances D(E) to the remaining experts. The numbers reported are in units of mm.

Comparison was also made for each structure by determining the volume of overlap and the volumes not overlapping. We defined the relative overlap $R\ (i, j)$ of two readings i and j as the volume of the intersection divided by the total volume defined by either of the two readings (their union):

$$R\ (i\ , j) = \frac{volume\ of\ intersection}{volume\ of\ intersection + volume\ out\ of\ intersection}. \tag{3}$$

For example, if a region defined by reader 1 contained 900 voxels, and the region defined by reader 2 contained 1100 voxels, 800 of which overlapped with reader 1, then the denominator that we used was 1200 voxels (800 intersection + 100 reader 1 only + 300 reader 2 only). The relative overlap in this example was, therefore, $R(1, 2) = 800 / 1200 = 0.667$. This is very conservative measure and gives lower values than if we had chosen to define overlap relative to only the first reader, which in this case would have been 800 / 900, or 0.889 . There was no a priori reason to assume that reader 1 was necessarily more correct than reader 2, so the 300 voxels defined by reader 2 but not by reader 1 should also be accounted for, which is the approach we have taken. Intuitively, one might expect that the two volumes of equal size of 1000 voxels with one half of overlapping voxels (500) should overlap one half. Yet, by our definition we get only 0.33.

Using this measure, the computer was pairwise compared with all of the experts. Moreover, each of the experts was pairwise compared with all of the other experts. Tables 5 and 6 show these comparisons for one patient. The last column contains the overall relative overlaps, using the definition of relative overlap of reading i with respect to the n readings j:

$$R\ (i\ ) = \frac{1}{n}\sum_{j=1}^{n} R(i,j) \tag{4}$$

Tables 7 and 10 are the composite of the overall relative overlaps, ones averaged over the patients and ones averaged over the readers. Table 7 shows that the average relative overlaps varied from 0.29 to 0.70. We can also see that, on average, the computer's agreement with the experts was about 20 % less than the agreement among the experts, with the worst computer results for the body of caudate.

Table 5. The computer agreement with the experts.

Patient P6						
Structure	R(C,WA)	R(C,JC)	R(C,RD1)	R(C,RD2)	R(C,RD3)	R(C)
111	0.47	0.47	0.36	0.23	0.22	0.35
112	0.37	0.55	0.33	0.31	0.33	0.38
131	0.43	0.49	0.40	0.38	0.41	0.42
132	0.52	0.62	0.57	0.53	0.53	0.55
135	0.62	0.69	0.56	0.56	0.71	0.63
136	0.59	0.59	0.61	0.67	0.75	0.64
117	0.72	0.69	0.68	0.73	0.72	0.71
118	0.66	0.66	0.65	0.65	0.68	0.66
201	0.63	0.65	0.63	0.56	0.57	0.61
301	0.68	0.68	0.68	0.68	0.68	0.68

For each individual structure and for each expert, the value reported is the relative overlap between the computer and one of the experts. R(C) in the last column is the overall relative overlap of the computer with the experts.

Table 6. The agreement between the expert WA and the other experts.

Patient P6					
Structure	R(WA,JC)	R(WA,RD1)	R(WA,RD2)	R(WA,RD3)	R(WA)
111	0.41	0.66	0.46	0.43	0.49
112	0.39	0.53	0.66	0.67	0.56
131	0.67	0.67	0.77	0.73	0.71
132	0.64	0.62	0.79	0.76	0.71
135	0.72	0.66	0.73	0.77	0.72
136	0.65	0.67	0.66	0.74	0.68
117	0.83	0.74	0.83	0.81	0.80
118	0.76	0.77	0.77	0.85	0.79
201	0.79	0.80	0.75	0.77	0.78
301	0.89	0.91	0.91	0.92	0.91

For each individual structure, the number reported is the relative overlap between the expert WA and each of the other experts. R(WA) in the last column is the overall relative overlap of the expert WA with the other experts.

Table 7. Composite by reader. For each individual structure, for each reader and for the computer, the value reported is the overall relative overlap, averaged over the six patients.

Structure	Readers					C
	WA	JC	RD1	RD2	RD3	
111	.61	.58	.62	.65	.61	.30
112	.68	.57	.63	.69	.68	.35
131	.74	.71	.74	.76	.76	.52
132	.73	.71	.74	.74	.76	.50
135	.75	.72	.78	.78	.78	.64
136	.72	.68	.75	.76	.75	.64
117	.75	.75	.77	.79	.75	.67
118	.74	.74	.78	.78	.78	.62
201	.80	.80	.83	.83	.83	.59
301	.90	.88	.90	.91	.91	.70

Figure 3 demonstrates how the experts and the computer typically drew contours on the same CT slice, and Figure 4 shows the computer-generated outlines in comparison with the unions defined by the experts.

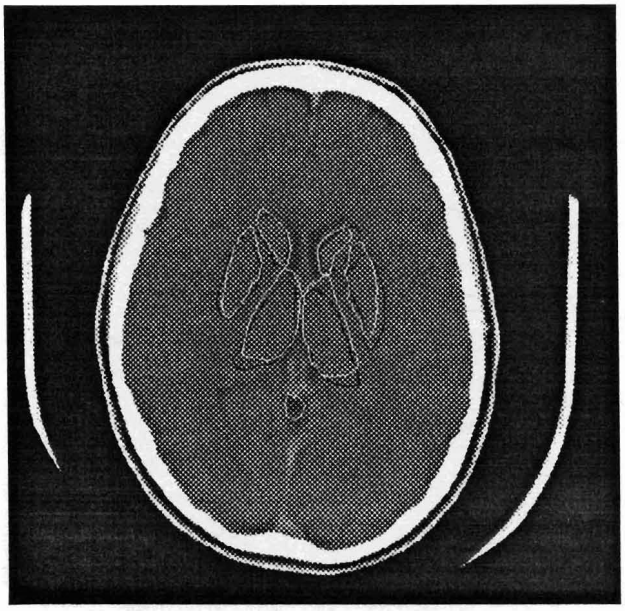

Figure 3. Comparing typical expert and typical computer. One of the experts generated outlines (shown in darker) and the computer generated outlines superimposed on the corresponding CT slice.

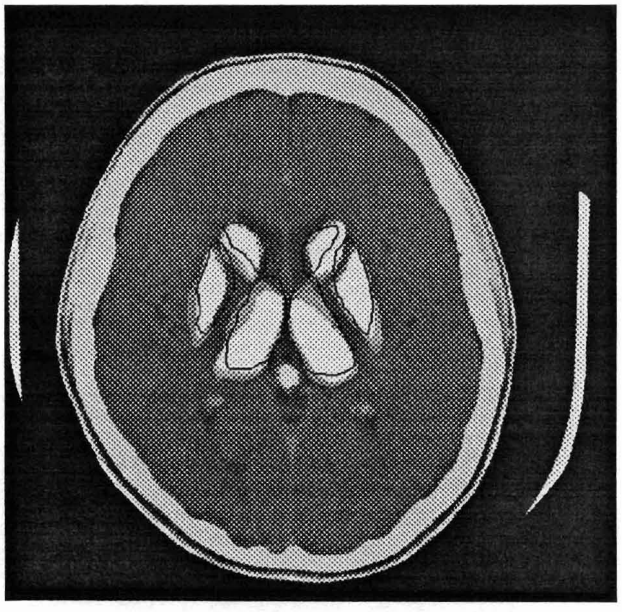

Figure 4. Comparing typical computer with the experts. The computer-generated outlines are superimposed on the corresponding CT slice, along with the union of ROI's defined by the experts.

4.4 Intra-observer reproducibility

As mentioned, one reader drew contours three separate times. Figure 5 shows just an example of the three readings superimposed on the corresponding CT slice. Table 8 shows the pairwise comparison of the three readings by RD for one patient. The final results are summarized in Table 9. It is of interest that the best results for any region averaged less than 93 % total agreement, even for a structure as large and well-defined as the cerebellum, and only about 82 % for the thalamus. There was considerable variability in the placement of ROI's over the body of the caudate nucleus, with some cases where there was less than 50 % agreement between individual readings. This is a somewhat surprising finding, and reflects primarily the difficulty in seeing an edge on the CT where there is both considerable noise and similar gray values both inside and outside the structure. Another factor in the intra-observer variability is the ability to maintain a constant set of mental criteria, since the image parameters and the knowledge of anatomy remain constant. The greatest agreement apparently occurred with the largest, best-defined structures, and the worst agreement with the smallest, ill-defined structures. Since edge definition appears to be the major problem, it is not surprising that the smallest regions produce the largest errors, since misplacement of the edge by even one pixel would have a greater relative effect on small regions. It is impossible from our data to determine how much of the error is due to size, how much to poor constrast, and how much to inconsistent criteria. The data in Table 8 from the three readings of RD suggest that no learning took place from the first to last reading, which validates the implicit assumption that they were independent measurements.

Table 8. Intra-observer reproducibility for one patient.

Patient P1				
Structure	R(RD1,RD2)	R(RD1,RD3)	R(RD2,RD3)	R(RD)
111	.61	.50	.78	.63
112	.74	.67	.76	.72
131	.76	.77	.84	.78
132	.66	.65	.79	.70
135	.82	.78	.79	.80
136	.70	.73	.72	.72
117	.70	.76	.84	.77
118	.73	.79	.82	.78
201	.85	.88	.85	.86
301	.91	.92	.93	.92

Pairwise comparison of the three readings by the reader RD for the patient P1. R(RD) in the last column is the average relative overlap of the three readings.

Table 9. Intra-observer reproducibility for all six patients.

Structure	Patients						MEAN
	P1	P2	P3	P4	P5	P6	
111	.63	.69	.68	.75	.73	.59	.68
112	.72	.70	.69	.79	.71	.63	.71
131	.78	.80	.74	.80	.82	.76	.78
132	.70	.82	.76	.84	.81	.78	.79
135	.80	.83	.81	.84	.86	.74	.81
136	.72	.81	.80	.86	.85	.77	.80
117	.77	.81	.86	.86	.78	.84	.82
118	.78	.73	.86	.87	.79	.84	.81
201	.86	.87	.83	.89	.87	.84	.86
301	.92	.93	.90	.93	.94	.93	.93

Summary for all six patients. Individual values are the averages of the three readings for each patient. The last column shows the average intra-observer reproducibility for each particular structure, averaged over the six patients.

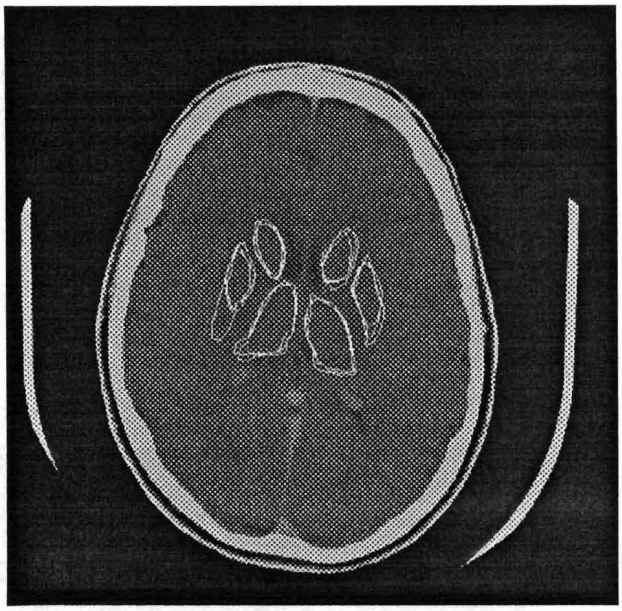

Figure 5. Typical example showing intra-observer variability. The three readings of the same expert were superimposed on the corresponding CT slice.

4.5 Experts' comparisons

Table 10 shows the relative overlap for each structure, using the means of all expert comparisons. It also shows that there is no difference in the experts' performance among the six patients analyzed, which supports the assumption we made that each patient was sufficiently normal not to affect the way the experts evaluated the scans. There is less difference among the experts for the head of the caudate, lenticular nucleus, and thalamus, while there is more disagreement among the experts over the placement of the body of the caudate. This finding is due, no doubt, to the same factors discussed in the previous section. A variable not present in the intra-observer evaluation, but present here, is the different anatomic knowledge possessed by each reader. From the anatomy, one would expect more variability in the way different people would perceive the thalamus, yet this was not the case according to the data in Table 7. It would seem, therefore, that each reader had essentially the same knowledge base, or at least that any differences had little effect on the analysis we did.

The task of drawing ROI's with the mouse combined both perceptual and motor skills with visual feedback. Our expert-expert results should not be surprising, since it has been shown in similar tasks that adults make average errors of 10–20 % in copying clearly defined objects like squares and diamonds [12]. When the uncertainty of edge detection is added, as we did, even less agreement is to be expected. Figure 6 demonstrates the variability among the experts.

Table 10. Composite by patient. Individual values in columns 2-7 are the means of all readers. The last column is the average over the patients.

Structure	Patients						MEAN
	P1	P2	P3	P4	P5	P6	
111	.65	.66	.61	.63	.66	.47	.61
112	.73	.65	.67	.68	.65	.54	.65
131	.74	.73	.71	.77	.76	.72	.74
132	.68	.73	.73	.80	.74	.74	.74
135	.79	.76	.77	.72	.77	.72	.76
136	.69	.72	.74	.73	.78	.72	.73
117	.75	.75	.81	.77	.71	.81	.77
118	.75	.69	.80	.80	.73	.81	.76
201	.83	.81	.80	.84	.82	.80	.82
301	.87	.91	.90	.90	.90	.92	.90

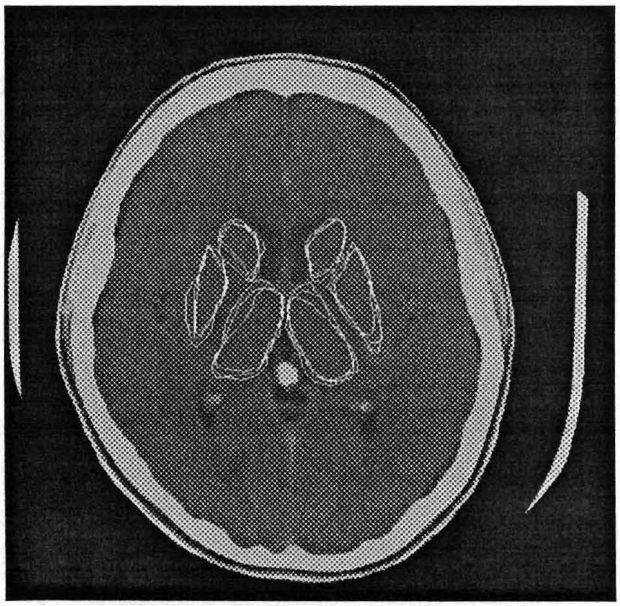

Figure 6. Typical example showing inter-observer variability. The three different experts' readings superimposed on the corresponding CT slice.

4.6 Limitations of current system

The major limitations to the system as it is currently implemented come from two major sources - bone and ventricles. The atlas, in its present form, was derived from brain sections that did not contain all of the structures that would be present in vivo, like the falx and bony structures. This means that the atlas is not colored appropriately and the two brain representations do not correspond. The best solution to this may be to use MRI scans to refine the current atlas.

Another problem with the atlas coloring is that it does not include choroid plexus. Even normal CT scans contain choroid within the ventricles, and although we currently do some preprocessing to remove calcifications from the CT images, their variable size and position make it difficult to create an atlas to correspond to their CT appearance. This can also be resolved by interactively removing the calcified, or even normal, choroid from the CT's prior to performing the matching. The worst errors in the current study occurred in the patient with both the largest ventricles and the greatest extent of calcified choroid plexus.

In some ways we biased this study against the computer. A more realistic evaluation of a finished system will have to include cortical ROI's, which will not be agreed upon in advance, but will be selected by each observer independently. We expect that this will result in considerably more variability among the experts, and our atlas-based software may prove to be both more objective and more reliable, although this remains to be seen.

Since MRI scans provide better definition of the gray/white interface than does CT, they should lead to more correct anatomical images. This should reduce the inter- and possibly intra-observer variability in defining ROI's, while the effect on the computer matching is unpredictable.

It may also be possible to improve the computer's performance by applying 2-D elastic matching to each slice after the overall 3D elastic matching of the brain as a whole. Other interactive pre-processing and post-processing may also be appropriate.

5 Conclusions

The current findings are promising. For most regions the errors in positioning by the computer were on the order of 2-4 mm. For many structures the computer's ROI was placed within the experts', so when applied to the analysis of functional images like PET or SPECT the computer might actually be more likely to extract correct data.

The finely spaced brain atlas that we have created can be elastically matched to multiple CT or MRI scans. We know that there is variability from person to person in both gross and microscopic anatomy [13]. In addition, there are well-documented interhemispheric differences, although the magnitude of these asymmetries has not been clearly established [14, 15]. This

software gives us the ability to create an anatomic stereotactic atlas like the one called for by Mazziotta and Koslow [9], since we can take a large number of MRI scans and map them all onto a single standard representation, eliminating the size and shape variations from person to person. A wide range of morphometric analyses may then be possible after this anatomic normalization.

Acknowledgements

We gratefully acknowledge the assistance of Mr. Mohamad Haleem, of the Armed Forces Institute of Pathology, Washington, D.C., without whose help we could not have made the brain atlas. We would like to thank Dr. John Chawluk, Dr. Bill Armington, and Dr. Debbie Gusnard, all from the Hospital of the University of Pennsylvania, for their help in outlining regions of interest on the CT scans. We appreciate the assistance of Ms. Laura Thompson and Dr. Robert Zimmerman, also of the Hospital of the University of Pennsylvania, in providing the CT scans to us. We thank to Howard Choset, Craig Meyer, Steven Sherin for their active participation in the project. We are thankful to Dr.Alan Rosenquist for his interset in our work and generous help. We thank to Dr. Max Mintz, Raynold McCendall, and Ales Leonardis for helpful discussions. We also thank Dr. Robert Livingston, of the University of California, San Diego, whose digitized brain atlas we used for testing and comparison. The research work leading to this paper was supported by NIH grants 5-27941 and 5-23913.

References

1. Mountz JM, Stafford-Schuck K, and Koeppe R, ''Comparison of MRI and the stereotactic method for localization of brain structures on O-15-H2O PET scans,'' *J Nucl Med*, (28) p. 702 , 1987.
2. Pelizzari CA, Chen GTY, Halpern H, Chen CT, and Cooper MD, ''Three dimensional correlation of PET, CT and MRI images,'' *J Nucl Med*, (28) p. 682 , 1987.
3. Martin WRW,, Grochowski E, Palmer M, and Pate BD, ''Correlation of structural and functional images in the same patient,'' *J Nucl Med* , (28) p. 634 , 1987.
4. Dann R, Muehllehner G, and Rosenquist A, ''Computer-aided data analysis of ECT data,'' *J Nucl Med*, (24) p. 82 , 1983.
5. Jones S, Greenberg J, Dann R, Robinson G, Kushner M, Alavi A, and Reivich M, ''Cerebral blood flow with the continuous infusion of oxygen-15-labelled water,'' *J Cereb Blood Flow Metab*, (5) pp. 566-575 , 1985.
6. Broit C, ''Optimal Registration of Deformed Images,'' Doctoral dissertation, University of Pennsylvania, Philadelphia, August 1981.
7. Bajcsy R and Kovacic S, ''Multiresolution Elastic Matching,'' Technical Report MS-CIS-87-94 GRASP LAB 123, Dept. of Computer and Information Science, University of Pennsylvania, Philadelphia, 1987.
8. Pelizzari CA, Chen GTY, Spelbring DR, Weichselbaum RR, and Chen CT, ''Accurate Three-Dimensional Correlation of CT, PET and MR Images of the Brain,'' *Submitted to J. Comp. Assist. Tomogr.*, , 1988.
9. Mazziotta JC and Koslow SH, ''Assessment of goals and obstacles in data acquisition and analysis from emission tomography: report of a series of international workshops,'' *J Cereb Blood Flow Metab*, (7) pp. 1-31. , 1987.
10. Matsui T and Hirano A, *An atlas of the human brain for computerized tomography,* Igaku-Shoin, Tokyo, 1978.
11. DeArmond SJ, Fusco MM, and Dewey MM, *Structure of the brain: a photographic atlas,* Oxford University Press, New York, 1976.
12. Laszlo JI and Broderick PA, ''The perceptual-motor skill of drawing,'' pp. 356-373 in *Visual Order (N.H. Freeman and M.V. Cox, Eds.)*, Cambridge University Press, Cambridge, 1985.
13. Geschwind N and Levitsky W, ''Human brain: left-right asymmetries in temporal speech region,'' *Science*, (161) pp. 187-187 , 1968.
14. LeMay M and Culebras A, ''Human brain: morphologic differences in the hemispheres demonstrable by carotid arteriography,'' *N Engl J Med*, (287) pp. 168-170 , 1972.
15. Galaburda AM, LeMay M, and Kemper TL et al, ''Right-left asymmetries in the brain,'' *Science* , (199) pp. 852-856 , 1978.