Automated three-dimensional registration of magnetic resonance and positron emission tomography brain images by multiresolution optimization of voxel similarity measures

Colin Studholme,[a] Derek L. G. Hill, and David J. Hawkes
Division of Radiological Sciences, United Medical and Dental Schools of Guy's and St. Thomas' Hospitals, Guy's Hospital, London Bridge, London, SE1 9RT, United Kingdom

(Received 13 February 1996; accepted for publication 4 October 1996)

Approaches using measures of voxel intensity similarity are showing promise in fully automating magnetic resonance (MR) and positron emission tomography (PET) image registration in the head, without requiring extraction and identification of corresponding structures. In this paper a method of multiresolution optimization of these measures is described and five alternative measures are compared: cross correlation, minimization of corresponding PET intensity variation, moments of the distribution of values in the intensity feature space, entropy of the intensity feature space and mutual information. Their ability to recover registration is examined for ten clinically acquired image pairs with respect to the size of initial misregistration, the precision of the final result, and the accuracy assessed by visual inspection. The mutual information measure proved the most robust to initial starting estimate, successfully registering 98.8% of 900 trial misregistrations. Success is defined as providing a visually acceptable solution to a trained observer. A high resolution search ($\frac{1}{16}$ mm step size) of 30 trial misregistrations showed that optimization using the mutual information measure provided solutions with 0.13 mm, 0.11 mm and 0.17 mm standard deviations in the three Cartesian axes of the translation vector and 0.2°, 0.3° and 0.2° standard deviations for rotations about the three axes. The algorithm takes between 4 and 8 minutes to run on a typical workstation, including visual inspection of the result. © *1997 American Association of Physicists in Medicine.* [S0094-2405(97)00601-9]

Key words: automated registration, voxel similarity, multiresolution optimization, magnetic resonance, positron emission tomography

I. INTRODUCTION

Accurate three-dimensional (3D) registration and overlay of magnetic resonance (MR) and positron emission tomography (PET) images of the brain provides important additional information by relating functional information from PET images to the detailed anatomical information available in MR images. While there has been significant progress in recent years in developing semi-automated methods for image registration all currently available methods rely on significant user interaction. More automated techniques to date have lacked sufficient robustness for routine clinical use. Interactive methods based on user guided registration[1] or user identification of point landmarks[2,3,24] are robust but time consuming, require observer skill and are hence prone to observer bias or error. Techniques dependent on the alignment of corresponding surfaces[4] require prior identification and segmentation of those surfaces. Fully automating this process is difficult and some manual editing is usually required. The surfaces visible in each modality may not correspond to the same anatomical surface. With targeted clinical scans only a small area of overlapping surface between the two modalities may be available and only a very small proportion of the total data is used for registration. As a result partial symmetry in the surfaces may lead to incorrect registration, for example with cranio-caudal and lateral rotations. The surface may move between acquisitions or be distorted by MR susceptibility affects, particularly when using the skin surface in the facial region, ears and back of the neck.

Recently there has been renewed interest in methods based on voxel intensity information. Woods *et al.*[5] have proposed a method based on the minimization of the sum of the standard deviation of intensities of PET voxels corresponding to narrow ranges of MR voxel intensities. They have applied this to registration of PET and MR images in which the brain has been segmented. They report that the approach is unreliable when applied to unsegmented MR images.

In this paper we report our work in developing the concept of voxel similarity measures for registration with the aim of providing a robust and accurate fully automated method for the 3D registration of MR and PET images of the head. We use the concept of the feature space, or joint probability distribution, of voxel intensities of the two modalities. We compare the use of cross correlation as used by Apicella[6] for two-dimensional (2D) in-slice alignment of MR and PET images, the variance of corresponding voxel intensities adapted from the Woods measure,[5] the third order moment of the histogram of the joint probability distribution proposed by Hill *et al.*,[7] and measures derived from an information theoretic approach to image registration based on entropy of the joint distribution[8] and mutual information

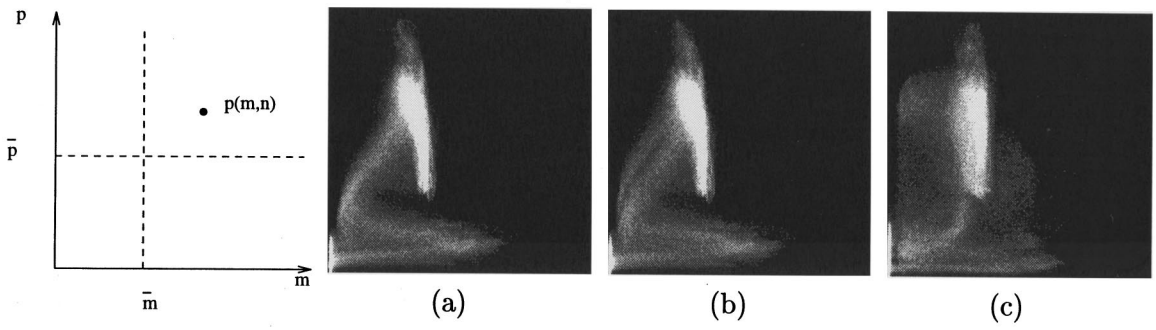

FIG. 1. Joint probability distribution of image intensities created from a MR (horizontal axis) and PET (vertical axis) image pair registered (a) and misregistered by translation along the cranio-caudal axis by 2 voxels (4 mm) (b) and 4 voxels (8 mm) (c).

developed in parallel by Collignon *et al.*[9] and Viola and Wells.[10]

All these measures require iterative search for a global minimum (or maximum) in a six-dimensional parameter space corresponding to the three orthogonal translations and three rotations of rigid body motion. Different authors utilize different optimization schemes in their work. We present a fast, efficient and robust multiresolution optimization approach for searching the six-dimensional parameter space of rigid body registration. Identical optimization is then used for each of the measures to enable direct comparison of their behavior.

Validation of registration algorithms for clinical use is of vital importance yet very difficult. Alternatives reported in the literature include phantom validation,[11] simulation of PET from segmented MR,[12] observer assessment,[13] markers placed on the skin[14] and markers placed within cadaveric heads.[15] Phantom data are never as realistic as patient's scans and simulated data are dependent on MR image interpretation and accurate modelling of the PET imaging process. Markers on the skin may move, stereotactic frames are generally too invasive for these studies and cadaveric head data are unsuitable for PET images. We have therefore undertaken an observer study of minimum detectable misregistration. Observer assessment will therefore provide an upper limit to registration accuracy.[13]

In this paper we report results of registering 10 MR and PET 3D image pairs of the head using 6 similarity measures. We tested and compared the precision and robustness of each measure on 900 misregistrations of the pairs. We assessed the small but significant differences in registration obtained by each measure. Finally we tested robustness to truncation of the axial field of view of the MR data set as might occur in targeted clinical scans.

II. VOXEL SIMILARITY REGISTRATION MEASURES

Given a pair of images to register and a transformation mapping one set of voxels onto the other, we can find, for corresponding voxels in the volume of overlap of the two images, the intensities $m \in M$ in the MR image **m**, and intensities $n \in N$ in the nuclear medicine PET image **n**, where M and N are the sets of all intensity values present in the region of overlap of the MR and nuclear medicine images, respectively. The sets M and N therefore depend on the rigid body transformation between the images. Additionally, we can calculate the probability of the occurrence of individual MR intensities $p\{m\}$, nuclear medicine intensities $p\{n\}$, and corresponding intensity pairs $p\{m,n\}$, that occur within the volume of overlap of the two images for a given transformation. The aim is to provide a similarity measure derived from these occurrences of corresponding intensities to relate the alignment for different transformations and overlaps.

In this section we describe a number of plausible similarity measures within the framework of this 2D distribution of corresponding voxel values (the voxel intensity feature space or joint probability distribution). Figure 1(a) shows the distinctive 2D distribution of intensities for a manually registered MR and PET image pair. Bright areas in the images correspond to large numbers of voxels with those MR and PET intensities. Figures 1(b) and 1(c) show distributions at different translational misregistrations. All the measures evaluated attempt to quantify these observed changes in some way and each makes assumptions about the nature of the distribution. Ideally we would like a measure of voxel similarity to give us a global optimum at registration and be a monotonic function of misregistration.

A. Correlation coefficient

The cross correlation function is commonly used in image matching. Apicella[6] applied the measure to retrieve the orientation of 2D MR and PET slices using Fourier decoupling of rotations and translations. In this work we have evaluated the measure in the spatial domain to solve the full 3D rigid registration problem. In its simplest form the function is dependent both on the volume of overlap and the local intensity of the images. A section of MR image may match two structurally similar regions of PET but the match would be biased toward the PET region with the higher intensity values. A better measure is the correlation coefficient which can be rewritten in terms of the joint probability distribution of voxel values,

$$\gamma(M,N)=\frac{\Sigma_{m\in M}\Sigma_{n\in N}(m-\overline{m}).(n-\overline{n}).p\{m,n\}}{\{(\Sigma_{m\in M}\Sigma_{n\in N}p\{m,n\}(n-\overline{n})^2)(\Sigma_{m\in M}\Sigma_{n\in N}p\{m,n\}(m-\overline{m})^2)\}^{1/2}}, \tag{1}$$

where $\overline{m}$ is the mean MR intensity and $\overline{n}$ is the mean nuclear medicine intensity over the volume of overlap of the two images. At registration $\gamma(M,N)$ will be maximized or minimized if there is a predominantly linear relationship between voxel values in the two modalities at registration. Initial experimentation indicated that when applied to T1 weighted MR images and ^{18}FDG (fluorine-18-fluorodeoxyglucose) PET images, the correlation coefficient provided a maximum at registration. Other MR sequences with different responses to tissue properties may provide a minimum at registration.

B. Minimization of PET intensity variation

The measure proposed by Woods[5] makes the basic assumption that for a given MR value the range of corresponding PET voxel values is a minimum at registration. Here we look at its behavior when applied to matching unsegmented MR brain images with optimization over multiple image resolutions. If $\overline{n(m)}$ is the mean value of corresponding PET voxels for a given MR value m and $\sigma_n(m)$ is the standard deviation of those values, then the normalized standard deviation is defined as

$$\sigma'_n(m)=\sigma_{n\in N}(m)/\overline{n(m)}.$$

In terms of the joint probability density of the voxel values, this is given by

$$\sigma'_n(m)=\frac{1}{\overline{n(m)}}.\sqrt{\sum_{n\in N}p\{m,n\}.(n-\overline{n(m)})^2}.$$

The standard deviation of the distribution of PET (abscissa) intensities for each MR (ordinate) value should be minimized at registration. A weighted sum of this normalized standard deviation for all MR values is a measure of similarity between M and N over their volume of overlap,

$$\sigma''_p(M,N)=\sum_{m\in M}[\sigma'_n(m).p\{m\}]. \tag{2}$$

The weighting ensures that the measure is influenced most strongly by PET intensity variation for the most common MR values.

This measure has been applied by Woods to segmented MR where the measure is calculated at a single scale only for those voxels with MR values corresponding to the brain and PET values above a certain threshold which also should correspond to the brain.

C. Moments of the distribution of values in the intensity feature space

This approach was first proposed by Hill *et al.*[7] from visual examination of the effects of misregistration on the feature space. As the images approach registration, the peaks in the feature space increase in height and the regions of the feature space which contain lower counts decrease in area. In terms of the joint probability distribution there are more larger values in $p\{m,n\}$ and fewer smaller values which can be measured by the higher order moments of this distribution. This results in an increase in skewness in the distribution of values of $p\{m,n\}$. Given $p\{m,n\}$ we calculate the number of occurrences of a particular probability $p\{m,n\}$, $u(p)$. Let w_p be the number of possible discrete probability values p_i, $i=1,\ldots,w_p$. The moment of order k of this distribution can then be evaluated:

$$\mathcal{M}_k(M,N)=\sum_{i=1}^{w_p}u(p_i).p_i^k. \tag{3}$$

This can be normalized by dividing by the zero moment or mass,

$$\mathcal{M}'_k(M,N)=\frac{\Sigma_{i=1}^{w_p}u(p_i).p_i^k}{\Sigma_{i=1}^{w_p}u(p_i)}. \tag{4}$$

Hill proposed the use of the third order moment as a measure of registration, which is the measure we used in these comparisons:

$$\mathcal{M}'_3(M,N)=\frac{\Sigma_{i=1}^{w_p}u(p_i).p_i^3}{\Sigma_{i=1}^{w_p}u(p_i)}. \tag{5}$$

D. Entropy of the intensity feature space

This measure and the following one are derived from communication theory, and describe the dependence of one variable on another. Entropy gives a measure of the average information provided by a set of symbols. In our case the symbols are values occurring in the two images to be registered. We can evaluate the information provided by pairs of values occurring together in the combined image (joint entropy) for a given transformation as used by Collignon[16] and Studholme[8]

$$H(M,N)=-\sum_{m\in M}\sum_{n\in N}p\{m,n\}\log(p\{m,n\}). \tag{6}$$

If we assume there are some shared structures in the two modalities, then when there is misalignment between the images, the combined image will contain two versions of these shared features (e.g., four eyes instead of two). Empirically, bringing the images into alignment will reduce the number of structures in the combined image and reduce the joint entropy. Conversely, by manipulating the transformation in order to reduce the joint entropy we should bring the images into alignment.

E. Mutual information

The measure of mutual information is used in communication theory to describe the information carried by a communications channel, by relating the information content of the transmitted and received symbols.[17] It has been proposed independently as a measure of image registration in various medical applications by Collignon *et al.*[9] and Viola and Wells.[10,18] As with the joint entropy, the information content is derived from the occurrence of values in the two images. The difference is that mutual information relates the joint entropy to the entropies of the modalities separately,

$$I(M;N)=H(M)+H(N)-H(M,N), \tag{7}$$

where $H(M)$ and $H(N)$ are the marginal entropies derived from the probabilities of occurrence of intensities in the overlapping portions of the images given by

$$H(M)=-\sum_{m\in M} p\{m\}\log p\{m\} \tag{8}$$

and

$$H(N)=-\sum_{n\in N} p\{n\}\log p\{n\}, \tag{9}$$

respectively. Using these $I(M;N)$ can be rewritten giving,

$$I(M;N)=\sum_{m\in M}\sum_{n\in N} p\{m,n\}\log\frac{p\{m,n\}}{p\{m\}p\{n\}}. \tag{10}$$

In terms of registration, by maximizing mutual information we minimize the information in the combined image (joint entropy) with respect to that provided by the two images separately (marginal entropies). To align the images we must evaluate and compare the measure derived from different orientations, and therefore overlaps, of the two image volumes. Because the regions of the two images being compared change with overlap, the information provided by the two images also changes. The joint entropy $H(M,N)$ is then not only a function of how well the images match in the overlap, but also by how much information is provided by the two modalities in the overlap. This means that $H(M,N)$ can be minimized simply because $H(M)$ or $H(N)$ is small in the region of overlap. $I(M;N)$ should provide a better measure of alignment than $H(M,N)$ alone because it simply represents the information shared between the two modalities for a given overlap.

III. REGISTRATION PROCEDURE

A. Multiresolution sampling

Medical image data are commonly sampled at different intervals within plane and between plane. In all the tests each pair of images was first resampled to cubic voxels. Where we needed to increase sampling to achieve this we used tri-linear interpolation and where we needed to decrease sampling we used voxel averaging by the nearest integer factor followed by tri-linear interpolation. This base sampling rate in effect determines the precision of the final estimate and the total processing time. An *octree* of lower resolution versions were then created from this using the same approach, first creating lower resolution images differing by a factor of 2 in sampling, and then forming an intermediate resolution by tri-linear interpolation from the next highest resolution. Although not as accurate as a Gaussian resampling scheme, we have found that this approach had a negligible influence on the optimization results when compared with Gaussian resampling. Its major practical advantage is that it requires minimal processor and memory resources, which is an important factor in the use of the software at many clinical installations.

B. Evaluation of $p\{m,n\}$

A given rigid body transformation defines a mapping of one discretely sampled set of voxels onto the other. We use the MR as the reference image and for each MR voxel find the intensity at the corresponding location in the PET volume. Each of these voxel value pairs are used to form a discrete representation of the joint probability distribution, $p\{m,n\}$. The binning of intensity values is achieved as follows. At the highest sampling resolution, the full intensity range of the image of MR intensities R_m is found and this range is repeatedly divided by 2 until it is less than 128,

$$bm=\frac{R_m}{2^i} \quad \text{for } i \text{ such that } 64\leqslant bm<128.$$

This number is then taken as the number of intensity bins for MR to form $M=\{m_1,m_2,..,m_{bm}\}$. The process is then repeated for the PET intensities forming $N=\{n_1,n_2,..,n_{bn}\}$, thus defining the discrete binning of the joint probability distribution. At lower data resolutions fewer voxels are available to form the estimate of $p\{m,n\}$ and so the number of bins in the histogram was reduced proportionately. At a 4 mm data resolution, with $\frac{1}{8}$ the number of voxels of the 2 mm base image, $\frac{1}{8}$ of the number of bins was therefore used. Experiments on a small number of data sets indicated that starting with a limit of 256 or 64 bins at the highest resolution made no significant difference to the performance of any of the measures. Large numbers of bins can lead to significant computation times for evaluation of the measures from $p\{m,n\}$.

Since the volumes are discretely sampled, some form of interpolation is required when evaluating PET intensity corresponding to a particular voxel in the reference MR image. In our implementation we used nearest neighbor interpolation for evaluation at each resolution down to and including the base sampling rate. Where increased precision was being assessed we then continued optimization by evaluating measures using tri-linear interpolation of the lowest level in the octree.

C. Optimization of measures

The *capture range* is an important feature of this type of registration scheme. The similarity measures give an indication of how well matched the data are in the volume of overlap. If the proportion of overlapping data is small, then the measures can give a misleading indication of registration.

An extreme example of this arises if the images are so misaligned that the only overlap is a small uniform region from one image, such as air surrounding the patient, overlaying a completely unrelated but also uniform region in the second image. The consequence of this is that at some large misregistration, all the algorithms can give an incorrect global optimum. There will always be a limited range of transformations for which a similarity measure is a monotonic function of misregistration (unlike registration schemes based on the alignment of equivalent corresponding features). This capture range is a function of the similarity measure, the image content, and the field of view, and cannot be determined *a priori*. Optimization algorithms employing multiple starting estimates would require assumptions to be made about the shape and scale of the capture range to ensure all starting estimates fall within this region. As a result we have employed a simpler registration scheme which takes a starting transformation estimate $T_0=\{t_x,t_y,t_z,\theta_x,\theta_y,\theta_z\}$, assumes that it is within the capture range of the registration measure and simply improves the registration in steps.

We evaluate the chosen similarity measure for a set of 13 transformations $\mathcal{T}(T_0)$. These are the current starting estimate and the starting estimate with increments and decrements of each of the 3 translations ($\pm\delta t$) and 3 rotations ($\pm\delta\theta$). We can look for a better estimate of the registration transformation T_1 for a given measure S such that for $S \in \{\sigma_p''(M,N),H(M,N)\}$,

$$T_1 = \min_{T\in\mathcal{T}(T_0)} \{S(\mathbf{m}(V_m \cap TV_n),\mathbf{n}(V_m \cap TV_n))\} \tag{11}$$

and for $S \in \{\gamma(M,N),\mathcal{M}_3'(M,N),I(M;N)\}$,

$$T_1 = \max_{T\in\mathcal{T}(T_0)} \{S(\mathbf{m}(V_m \cap TV_n),\mathbf{n}(V_m \cap TV_n))\}. \tag{12}$$

If $T_{n+1}{}^{a} \neq T_n$ then we can repeat the search with $\mathcal{T}(T_{n+1})$ until $T_{n+1}=T_n$. The step sizes $\{\pm\delta t,\pm\delta\theta\}$ can then be reduced and the search continued, the minimum values of $\{\pm\delta t,\pm\delta\theta\}$ determine how close we get to the optimum transformation. We extend this simple optimization approach by applying it to multiple resolution versions of the images and linking the step size to the resolution of the data. At level l in the octree, the isotropic voxel dimension, or resolution, is r_l. At a resolution r_l we set δt to r_l and the rotational step size (in degrees) $\delta\theta=k\times r_l$. Experimentation indicated that changing the value of k between 0.5 and 2.0 had no significant effect on registration results, and so we arbitrarily set $k=1.0$ for the work presented in this paper.

Sampling resolution was reduced by a factor of $\sqrt{2}$ from 2 mm down to $8\sqrt{2}$ mm. This rate of sampling reduction was found to give improved performance on some of the data sets compared to a simple quad-tree reduction. Experimentation with starting estimates provided by manual estimates of registration indicated that using starting resolutions coarser than $8\sqrt{2}$ mm led to an increase in failures with some of the measures, presumably due to stepping out of the capture range.

One measure of optimization efficiency is the number of evaluations of the similarity measure required, which can be expressed in terms of the equivalent number of evaluations at the highest resolution (i.e. evaluations of a resolution containing $\frac{1}{8}$ the number of voxels to transform would count as $\frac{1}{8}$ of an evaluation). Over the many tests carried out on the data using the standard algorithm a mean of around 100 evaluations was recorded ranging from 50 to around 150 (for those optimizations which were successful). Typical processing times were between 3 and 7 minutes on a Sun SPARC station 5/70 (Sun Microsystems, Mountain View, CA), dominated by evaluations at the finest resolution (2 mm).

TABLE I. Transformation parameters estimated by manual point landmark identification.

	Manual MR-PET (^{18}FDG) Registration								
	Translation (mm)			Rotation (deg.)			RMS Num.		MR voxel
Patient	t_x	t_y	t_z	θ_x	θ_y	θ_z	error	points	size (mm)
A	2.9	23.3	8.4	11.7	−1.9	−0.5	3.9	12	0.86×0.86×2.5
B	−2.6	1.8	7.6	24.2	1.1	8.7	3.6	12	0.90×0.90×1.5
C	−9.6	−9.8	22.8	21.6	−0.8	6.1	3.0	14	0.94×0.94×1.5
D	−3.5	7.0	15.4	16.1	−2.7	3.8	2.9	14	0.94×0.94×1.5
E	−1.8	18.3	0.6	11.3	0.2	0.6	2.7	12	0.94×0.94×1.5
F	−3.4	−11.0	0.0	18.4	−0.8	5.9	3.5	12	0.94×0.94×1.5
G	1.0	7.0	−8.7	18.6	4.3	−5.7	3.1	11	0.90×0.90×1.2
H	2.1	11.8	0.4	14.1	−0.9	−0.4	4.3	14	0.86×0.86×1.2
I	3.5	12.8	−7.3	18.3	0.5	3.7	2.9	10	0.86×0.86×1.2
J	0.7	22.6	−4.9	12.7	−3.0	2.3	2.5	11	0.86×0.86×1.2

IV. METHOD

A. The data

The basis for our registration tests were a set of 10 routinely acquired image pairs of the brain, chosen from 3 clinical protocols to represent a typical range of image data acquired at our site. All data consisted of nominally transaxial slices but with a range of orientations typical in routine clinical scanning. All the PET images were acquired on a Siemens/CTI scanner (Knoxville, TN). The PET images were all static, summed from 6 dynamic ^{18} FDG acquisitions reconstructed to 31 slices of 128 by 128 voxels. The voxel size was 2.0×2.0×3.375 mm and the point spread function has full width at half maximum of approximately 8.0 mm. The MR acquisitions came from 3 different scanners, 1 2D spin echo image (patient A) from a 1.5T Philips Gyroscan S15/HP, 5 3D gradient echo images (patients B to F) from a 1.5T GE Signa, and 4 3D gradient echo images (patients G to J) from a 1.5T Philips Gyroscan ACS II. All were T1 weighted and intended to show good grey/white matter delineation. Large MR volumes were used to give a good range of tissue types (white matter, grey matter, skull, scalp etc.) for registration.

All the image pairs were first manually registered by interactive location of between 10 and 14 corresponding point landmarks. These estimates are shown in Table I in the form of 3 translations and 3 rotations to map each PET image to MR coordinates (where the x-axis is the patient from right–left, the y-axis from front to back and the z-axis from feet to head). Alignment was achieved by minimization of the sum

of squared distances between corresponding points. The root mean square (rms) error between the points for this estimate is included in the table.

To check registration accuracy these manual point based registration estimates were inspected as described in Sec. IV B. All registrations were confirmed visually to be within about 3 mm over the brain volume. These manual results provide a valuable indication of the scale of typical misregistration encountered between brain images in clinical practice. The in-plane translational misalignment of the original images were within the range ± 25 mm (roughly $\frac{1}{5}$ of the field of view).

B. Assessment of registration

Numerically the 6 parameter solutions can be assessed by comparison with those provided by manual registration. Inspection of the manual point based estimates still indicated small but visually detectable misregistrations. As a result we chose to use expert visual inspection of the automated results to select a good registration for comparison, rather than directly comparing automated estimates with the manual point based estimates. We employ a number of visualization tools to assess accuracy based around an interactive display of orthogonal slices. A color overlay of PET intensity onto grey level MR combined with the depiction of interactively selected iso-intensity PET boundaries onto MR grey levels can give a sensitive indication of misregistration. In our protocol, to assess registration quality the observer uses orthogonal slices of the combined data intersecting in the following regions:

(1) The interthalamic area.
(2) The center of each orbit.
(3) The posterior part of temporal lobes.
(4) A transaxial slice above the lateral ventricles.

The observer is also encouraged to view other areas if thought necessary. Studies such as those by Pietrzyk[13] have shown ''misalignment of 4 mm was detected unambiguously.'' We have performed similar studies using our software and higher resolution image data to assess five experienced observers' ability to detect misregistration of MR and PET images of the head.[19] This has shown that all observers can detect an X or Y translational misregistration of 2 mm or more, a Z axis translational misregistration of 3 mm or more, rotations about the Z axis of 2° or more and rotations about the X axis or Y axis of 4° or more (4° corresponds to a displacement of 7 mm at 100 mm from the axis of rotation).

C. Experiments

1. Direct image registration

The first test was to register each of the image pairs as delivered by the scanners. Initially the center of the two volumes were aligned and the slice orientation was that of the scan acquisition. From this starting estimate, for each of the 10 image pairs, each of the measures was optimized with respect to the transformation. To ensure precision of the result, tri-linear interpolation was then used to further improve the estimate, reducing the step sizes down to $\frac{1}{16}$ mm and $\frac{1}{16}°$ by factors of $\sqrt{2}$. Visual inspection was used to distinguish good results from obvious failures.

2. Assessment of robustness to starting estimate

We defined a series of known misregistrations from the manual point based estimates shown in Table I. In total, 90 misregistrations were defined. Thirty of these corresponded to a misregistration by translation of 10 mm and a rotation of 10°, 30 of 20 mm and 20° and 30 of 30 mm and 30°. Each set of 30 was randomly distributed over the surface of spheres in translational and rotational parameter space so that a set distributed over a wide range of possible combinations of rotations and translations was tested. From Table I it can been seen that misregistrations of 30 mm and 30° are at the extremes of likely initial misregistrations. The same set of 90 misregistrations was then used as starting estimates for each measure. For these estimates optimization continued down to a step size of $\sqrt{2}$ mm with sampling of 2 mm.

Visual inspection of the direct image registration results and the point based estimates provided an indication of the quality of results and a good registration was selected, giving a visually defined ''gold standard'' for each of the image pairs. We then count the number of optimization results for each of the 90 misregistrations falling within 3 mm and 4° of the visually selected gold standard. These ranges were selected from a study of the skilled observers' ability to detect misregistration.[19] Although this does not give a direct measure of accuracy it gives a good indication of robustness to different magnitudes of initial misregistration.

3. Assessment of limits on precision

To test the limits to precision, optimization of the 30 misregistrations of 10 mm and 10° were continued for patient I reducing the step size to $\frac{1}{16}$ mm by factors of $\sqrt{2}$. The mean registration parameters and their distribution about the mean were computed and compared. Using data from patient I was an arbitrary choice from the 4 data sets which gave a reasonable initial registration for each measure.

4. Assessment of the effects of axial truncation

To test robustness to truncation, three equal axial segments of 72 mm (from patients B to J) and 70 mm (for patient A) axial extent were extracted from the top, middle and base of the MR data sets of patients. The registration experiments were repeated for the 10 mm and 10° random misregistrations for each of the measures.

V. RESULTS

A. Direct image registration

Initial experimentation on the image data indicated that all the measures, other than mutual information, $I(M;N)$, and correlation coefficient $\gamma(M;N)$, exhibited much improved behavior if low intensities, corresponding to air in the two modalities, were excluded from the evaluation. A similar approach has been used by Woods[5] to select intensities from

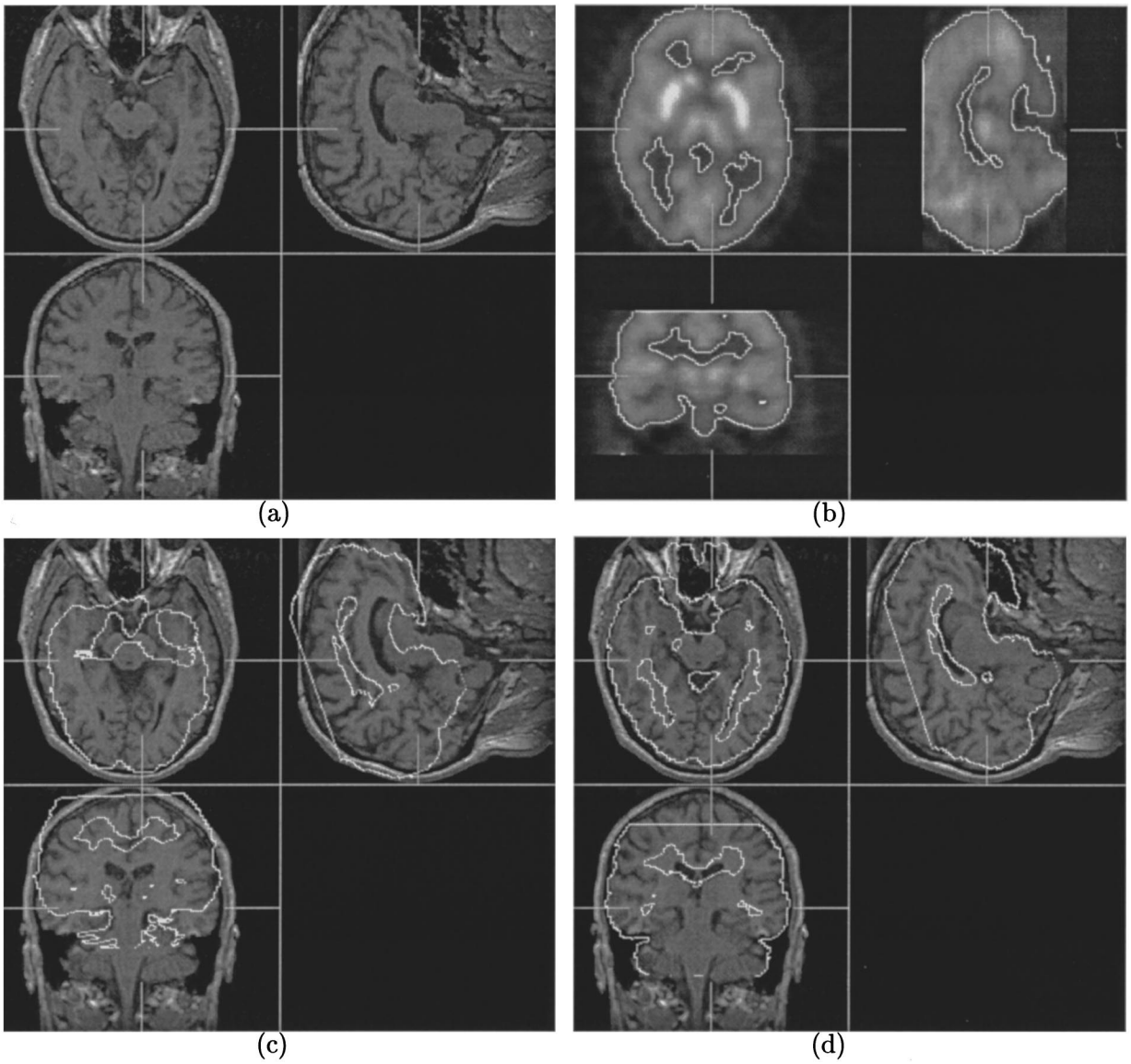

FIG. 2. Orthogonal slices through (a) MR data for patient D, (b) PET data (with a PET iso-intensity boundary overlaid), MR data with overlay of PET iso-intensity boundary for (c) poor registration estimate provided by optimization of correlation coefficient, (d) visually acceptable registration estimate provided by mutual information.

PET to match with segmented MR. As a result, for all the results presented here we have not used MR intensities below 10% of the maximum value and PET tracer values below 20% of the maximum value when evaluating $\sigma''_p(M,N)$, $H(M;N)$ and $\mathscr{M}'_3(M,N)$. Using this approach, alignment starting from the orientation delivered by the scanners showed that all measures achieved visually acceptable registrations except for the correlation coefficient measure, which failed for 5 of the patients, and the moment of the histogram of the feature space measure, which failed for 1 patient. Examples of two of these results for patient D are shown in Figure 2. Failure of the correlation coefficient measure may have been due in this case to increased PET signal in the sinuses, caused either by unusually high ^{18}FDG uptake in this region, or an error in the attenuation assumption in the PET reconstruction, resulting in significant differences in brain boundary delineation in MR and PET. This is illustrated in the region between the eyes in Figures 2(c) and 2(d), where the threshold chosen matches the MR brain boundary well (note that the threshold chosen here does not delineate corresponding internal structures, such as ventricles). It is clear that on its own this result is not a good comparison of the robustness of the different measures.

B. Robustness to starting estimate

In this test we were interested in the large scale behavior of the optimization for each of the registration measures. The solutions found can be divided into those which were close to those found in the direct image registration results, and

TABLE II. The percentage of estimates falling within 3 mm and 4° of the fine resolution optimization estimate, for 30 initial misregistrations of 10 mm and 10°, 30 of 20 mm and 20° and 30 of 30 mm and 30°.

	Robustness to starting estimate Initial misregistration		
Measure	10 mm and 10°	20 mm and 20°	30 mm and 30°
$\gamma(M,N)$	87.3%	58.3%	57.7%
$\sigma''_p(M,N)$	92.0%	91.0%	88.7%
$\mathscr{M}'_3(M,N)$	99.7%	73.3%	31.3%
$H(M,N)$	87.3%	28.3%	6.7%
$I(M;N)$	100.0%	99.7%	97.0%

those which have escaped from the capture range of the measure and resulted in significant misregistration. Table II summarizes the number of the solutions lying within 3 mm and 4° of the visually defined ''gold standard'' for each of measures from the 900 random starting estimates.

These results show that mutual information performed extremely well with only a very small number of failures for all of the experiments (11 out of 900). The PET intensity variance measure performed reasonably well with only 85 failures out of 900 trials. The third order moment of the intensity feature space performed well for small initial misregistrations but showed poor results for larger initial misregistrations. The other two measures performed significantly worse, particularly for starting estimates further away from registration.

C. Precision of registration

The results of the experiment to test the limits of precision for the different measures for patient I are shown in Table III. There is a small but significant difference in the mean registration transformation for each measure even when they do apparently produce good registrations as was the case for patient I.

We examined the clustering of the solutions as follows. As it is difficult to visualize the distributions of the registration solutions in a six-dimensional parameter space we projected the solutions into two dimensions. Residual translation and rotation vectors were computed for each solution by taking the root sum of squares of the rotational and translational displacements from the point based solution. These scatter plots were generated for each of the 6 measures and are presented in Figure 3, which clearly shows the consistent bias provided by the different measures.

Optimization using the mutual information measure was continued to a finer step size and did not produce any significant difference in the means and standard deviations. These figures therefore represent fine structure of the measure in parameter space and not the limited resolution of the search.

D. Effects of axial truncation

The results for all 10 image pairs using axially truncated MR are summarized in Table IV. The correlation coefficient measure failed completely to produce any acceptable registrations on the truncated images. One possible explanation of this difference in behavior is that the measure is dominated by the air-patient boundary. Removal of this boundary (in this case in the Z-axis) results in a failure. The remaining measures behaved significantly better on the central and upper volumes of the images. Mutual information again performed best but about 20% of the registrations still failed for the middle and upper slices and the measure performed poorly for the lower slices. The poor results for the lower slices may be explained by the effect of a larger range of tissue types and MR intensities in the neck with no corresponding PET intensity.

For both the top and base of the data sets there was evidence of bias in the axial direction, with brain activity from PET registering with scalp from MR at the top and with muscles of the eye, face and back of the neck at the base of the truncated data sets. The bias in these solutions is illustrated for patient I in the scatter plots of the rotational and translational vectors in Figure 4. There is a consistent misregistration upwards of the PET data for the upper segment of 2 mm for the mutual information measure. Conversely, there is consistent misregistration downwards of the lower segment of 4 mm.

VI. CONCLUSION

Using a simple and computationally efficient optimization approach over multiple resolutions, we have compared the robustness and precision of a number of voxel similarity measures for MR and PET brain image registration when presented with a typical range of initial misregistrations. The

TABLE III. Mean and standard deviations of 30 high precision registration trials for patient I from 10° and 10 mm random misregistrations (see text).

	Precision from random starts Translations (sd)			Rotations (sd)		
Measure	t_x mm	t_y mm	t_z mm	θ_x°	θ_y°	θ_z°
$\gamma(M,N)$	2.83 (0.23)	13.44 (0.17)	−8.85 (0.23)	19.00 (0.37)	0.27 (0.36)	1.67 (0.40)
$\sigma''_p(M,N)$	3.25 (0.10)	12.92 (0.15)	−8.55 (0.22)	20.10 (0.31)	−0.53 (0.21)	3.34 (0.15)
$\mathscr{M}'_3(M,N)$	2.95 (0.19)	14.30 (0.25)	−7.98 (0.36)	18.88 (0.28)	−0.89 (0.42)	3.15 (0.22)
$H(M,N)$	3.04 (0.18)	15.60 (0.32)	−8.23 (0.35)	20.30 (0.47)	−0.95 (0.40)	3.35 (0.34)
$I(M,N)$	3.39 (0.13)	13.97 (0.11)	−8.76 (0.17)	19.80 (0.21)	−0.84 (0.30)	3.58 (0.22)

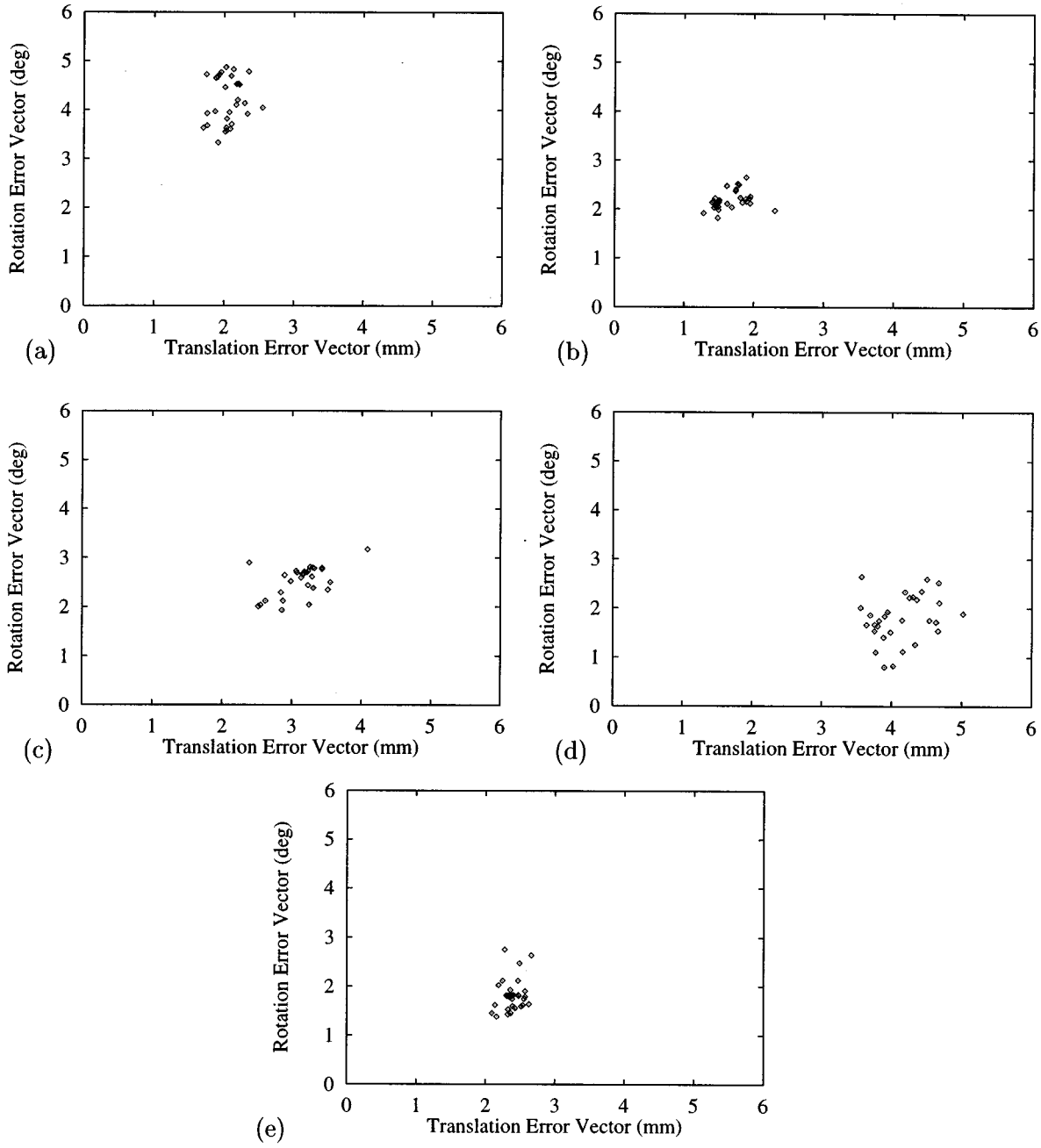

FIG. 3. Translation (mm) and rotation (°) error vectors between the manual estimate and automated estimate from random starts for patient I following optimization of correlation (a), PET variance (b), third order moment (c), entropy (d) and mutual information (e).

range of misregistrations was guided by that obtained on 10 clinically acquired MR and PET image pairs and included displacements and rotations of up to nearly 30 mm and 30°. We found that the mutual information measure was robust and precise. Of 900 initial misregistrations of 10 MR and PET image pairs, only 11 (1.2%) failed to produce a visually acceptable result. The measure of PET intensity variation performed reasonably well with 85 (9%) failures of 900 trials. The correlation coefficient measure proved less reliable, particularly for certain image pairs, with more failures for starting estimates further from registration. The third-order moment of the intensity feature space and the jointen-

TABLE IV. Number of successful registrations from 300 trials, judged as within 3 mm and 4° of the direct high resolution result obtained with the mutual information measure for half the axial MR volume located at the bottom, middle and top of the original MR volume.

	Effects of axial truncation		
Measure	Lower slices	Middle slices	Upper slices
$\gamma(M,N)$	0.0%	0.0%	0.0%
$\sigma''_p(M,N)$	0.0%	37.3%	33.3%
$\mathscr{M}'_3(M,N)$	0.0%	47.3%	42.0%
$H(M,N)$	0.0%	24.3%	46.7%
$I(M,N)$	36.7%	82.0%	82.7%

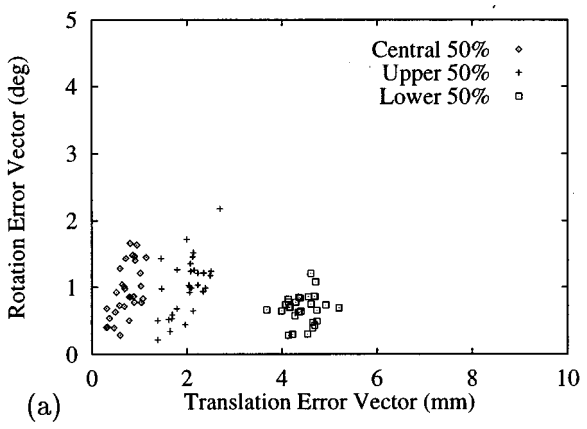

FIG. 4. Translation (mm) and Rotation (°) error vectors between the full volume estimate and estimates from random starts using 50% of MR slices for patient I following optimization of mutual information (a).

tropy showed an even more significant fall in performance as the starting estimate was degraded. This contrasts with the results of direct registration of the images as delivered by the scanners. Direct registration failed to distinguish between the performance of any of the measures, except the cross correlation which failed for half of the image pairs. This underlines the need to assess registration performance with a range of starting estimates as well as a number of different image pairs. The difference in behavior between the mutual information and joint entropy measures reflects the normalization effect of the marginal entropies of the modalities in the region of overlap.

We have shown that the different measures, when they do produce an acceptable registration, converge to a solution which is precise. For the mutual information measure when the search is continued down to a step size of $\frac{1}{16}$ mm (0.0625 mm) and $\frac{1}{16}°$ (0.0625°), the standard deviation of thirty evaluations of the translational estimates are 0.13 mm, 0.11 mm and 0.17 mm in the X, Y and Z axes respectively and 0.21°, 0.30° and 0.22° for rotations about the X, Y and Z axes. As expected there are small but significant differences between each measure. These differences, for the one data set studied, are of the order of 1 mm and 1°.

We have no means of knowing whether these estimates are distributed about the ''true'' registration but they are visually acceptable. A previous study has confirmed that visual assessment will detect misregistrations greater than 2 mm of transaxial translation, 3 mm of axial translation, 2° of axial rotation, or 4° of cranio-caudal or lateral rotation. These figures determine the upper limit of accuracy in this present study.

The concept of capture range is important. When a local optimum exists which produces a visually acceptable registration there may well be a global optimum or other significant local optima corresponding to significant misregistrations. Ensuring that the optimization process does not erroneously find these false solutions requires care when defining the maximum allowable initial misregistration of the starting estimate and the coarsest step size of the multiresolution search. These in turn will depend on the corresponding information content of each image and empirically we have found that for MR and PET whole brain images the mutual information measure performs well with a maximum step size of $8\sqrt{2}$ mm (11.3 mm) and initial misregistrations within 30 mm translation and 30° rotation.

The performance of the mutual information measure depends on there being sufficient corresponding intensity information to achieve an optimum at registration. Clearly there could be situations when this does not occur. The experiments on truncated data illustrate these limitations. In this case mutual information is maximized when the upper and lower truncated data sets are erroneously axially translated by a few millimeters. Interestingly, registration of these truncated data sets is also very difficult for all but the most highly skilled observers, implying that high level anatomical information may well be required in this particular task. The algorithm does appear to follow reasonably well the performance of the human visual system when the observer does not have additional anatomical information.

The presence of large space occupying lesions or significant intensity artifacts (streak artifacts or intensity shading) may also affect the robustness and accuracy of the measure. All our image data were collected to study neuro functional parameters and significant space occupying lesions were not present. The MR images were processed as they were delivered by the MR scanners. Residual image distortion was not corrected. Except for the possible presence of susceptibility effects in and around the air sinuses phantom experiments have indicated that MR distortion should be less than about 1.5 mm throughout the volume of interest. Image scaling was not a parameter in our registration algorithm and we relied on effective quality control of the scanners used. We did not detect significant scaling errors in the data acquired for this study.

Processing times, excluding data transfer from scanners to workstation but including visual inspection of the results, were between 4 and 8 minutes per study. This compares favorably with typical times in our laboratory for fully interactive registration, interactive point based registration, surface matching and minimization of corresponding PET intensity variation (the latter two include steps for user guided MR brain segmentation). The time is compatible with routine clinical use in centers where efficient data transfer procedures are implemented.

This registration method is also being applied successfully to MR and CT images of the head[23] and work is in progress in MR, CT and PET images of the neck, thorax and pelvis. Further work is required to examine absolute accuracy using an appropriate gold standard such as that proposed in the Vanderbilt study,[20] to quantify in more detail the effects of image artifact, and to compute local rigid body registration in the presence of patient movement, severe geometric distortions and intensity inhomogeneity. Results of the algorithms performance on many more data sets at multiple clinical sites will be reported in due course. Work is also underway to improve robustness and application to other image types, for example, by incorporating label-

ling of connected regions[21] and encoding spatial location of intensity.[22]

ACKNOWLEDGMENTS

This work was funded by UK EPSRC. We are grateful for the support and encouragement of our clinical colleagues in this work, in particular Dr. Wai-Lup Wong, Dr. Joseph Wong, Dr. Iain Cranston and for the technical assistance of the Radiographic staff of Guy's and St. Thomas' Hospitals in London. We acknowledge useful discussions with Andre Collignon and Dirk Vandermeulen of KUL, Leuven, Belgium and Sandy Wells of MIT and Brigham and Women's Hospital Boston, MA, in particular on the subject of mutual information. We also thank Dr. John Little of UMDS for assistance in the Mathematical formulations.

[a]Electronic mail: C. Studholme@umds.ac.uk

[1]U. Pietrzyk *et al.*, ''An interactive technique for 3-dimensional image registration: Validation for PET, SPECT, MRI and CT brain studies,'' J. Nucl. Med. **35**, 2011–2018 (1993).

[2]A.C. Evans, S. Marrett, J. Torrescorzo, S. Ku, and L. Collins, ''MRI-PET correlation in three dimensions using a volume-of-interest (VOI) atlas,'' J. Cereb. Blood Flow Metab. **11**, A69–A78 (1991).

[3]D.L.G. Hill *et al.*, ''Registration of MR and CT images for skull base surgery using point-like anatomical features,'' Br. J. Radiol. **64**, 1030–1035 (1991).

[4]D.L.G. Hill *et al.*, ''Accurate frameless registration of MR and CT images of the head: applications in planning surgery and radiation therapy,'' Radiology **191**, 447–454 (1994).

[5]H. Jiang, R.A. Robb, and K.S. Holton, ''New approach to 3-D registration of multimodality medical images by surface matching,'' Proc. SPIE **1808**, 196–213 (1992).

[6]R.P. Woods, J.C. Mazziotta, and S.R. Cherry, ''MRI PET registration with automated algorithm,'' J. Comput. Assist. Tomogr. **17**, 536–546 (1993).

[7]A. Apicella, J.S. Kippenham, and J.H. Nagel, ''Fast multi-modality image matching,'' Proc. SPIE **1092**, 252–263 (1989).

[8]D.L.G. Hill, C. Studholme, and D.J. Hawkes. ''Voxel similarity measures for automated image registration,'' Proc. SPIE **2359**, 205–216 (1994).

[9]C. Studholme, D.L.G. Hill, and D.J. Hawkes, ''Multiresolution voxel similarity measures for MR-PET registration,'' Proceedings of Information Processing in Medical Imaging, Ile de Berder, 1995, edited by Y. Bizais, C. Barillot, R. Di Paola (Kluwer Academic, Dordrecht, 1995), pp. 287–298.

[10]A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, and G. Marchal, ''Automated multimodality image registration using information theory,'' in Ref. 8, pp. 263–274.

[11]P.A. Viola and W.M. Wells, ''Alignment by maximisation of mutual information,'' Proceedings of the 5th International Conference on Computer Vision, 1995 (IEEE, New York, 1995).

[12]T.G. Turkington *et al.*, ''Accuracy of registration of PET, SPECT and MR images of a brain phantom,'' J. Nucl. Med. **34**, 1587–1594 (1993).

[13]P. Neelin, J.E. Crossman, D.J. Hawkes, Y. Ma, and A.C. Evans, ''Validation of an MRI/PET landmark registration method using 3-D simulated PET images and point simulations,'' Comput. Med. Imaging **17**, 351–356 (1993).

[14]U. Pietrzyk *et al.*, ''Three-dimensional alignment of functional and morphological tomograms,'' J. Comput. Assisted Tomogr. **14**, 50–59 (1990).

[15]T.G. Turkington *et al.*, ''Accuracy of surface fit registration for PET and MR brain images using full and incomplete brain surfaces.'' J. Comput. Assist. Tomogr. **19**, 117–124 (1995).

[16]P.F. Hemler, P.A. Van den Elsen, T. Sumanaweera, S. Napel, J. Drace, and J.R. Adler, ''A quantitative comparison of residual errors for three different multimodality registration techniques,'' in Ref. 8, pp. 251–262.

[17]A. Collignon, D. Vandermeulen, P. Suetens, and G. Marchal, ''3D multimodality medical image registration using feature space clustering,'' Proceedings of CVR Med '95, Nice, FR, in Lecture Notes in Computer Science (Springer-Verlag, Berlin, 1995), Vol. 905, pp. 195–204.

[18]F.M. Reza, *An Introduction to Information Theory* (Dover, New York, 1994), pp. 104–106.

[19]W.M. Wells, P. Viola, and R. Kikinis, ''Multimodal volume registration by maximization of mutual information,'' Proceedings of the 2nd annual international symposium on Medical Robotics and Computer Assisted Surgery, Baltimore, 1995 (Wiley, New York, 1995), pp. 55–62.

[20]J. Wong, C. Studholme, D.J. Hawkes, and M.N. Maisey, ''Evaluation of the limits of visual detection of image registration in a brain F-18 FDG PET-MRI study,'' J. Nucl. Med. **37**, 208 (1996).

[21]J. West *et al.*, ''Comparison and evaluation of retrospective intermodality registration techniques,'' J. Comput. Assisted Tomogr. (in press).

[22]C. Studholme, D.L.G. Hill, and D.J. Hawkes, ''Incorporating connected region labelling into automated image registration using mutual information,'' Proceedings of Mathematical Methods in Biomedical Image Analysis, San Francisco, 1996 (IEEE Computer Society Press), pp. 23–31.

[23]C. Studholme, D.L.G. Hill, J. Wong, M.N. Maisey, and D.J. Hawkes, ''Registration measures for automated 3D alignment of PET and intensity distorted MR images,'' Proceedings in Image Fusion and Shape Variability Techniques, Leeds, 1996 (Leeds University Press), pp. 186–193.

[24]C. Studholme, D.L.G. Hill, and D.J. Hawkes, ''Automated 3D registration of truncated MR and CT images of the head,'' Proceedings of the British Machine Vision Conference, Birmingham, British Machine Vision Association, 1995, edited by D. Pycock (BMVA Press), pp. 27–36.