NeuroImage 44 (2009) 839–848

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

A new method for improving functional-to-structural MRI alignment using local Pearson correlation

Ziad S. Saad [a], Daniel R. Glen [a], Gang Chen [a], Michael S. Beauchamp [b], Rutvik Desai [c], Robert W. Cox [a,*]

[a] Scientific and Statistical Computing Core, National Institute of Mental Health, National Institutes of Health, Department of Health and Human Services, 10 Center Dr. Room 1D80 Bethesda, MD 20892-1148 USA
[b] Department of Neurobiology and Anatomy, University of Texas Health Science Center at Houston, Houston TX USA
[c] Department of Neurology, Medical College of Wisconsin, Milwaukee WI USA

ARTICLE INFO

Article history:
Received 2 July 2008
Revised 15 September 2008
Accepted 22 September 2008
Available online 11 October 2008

ABSTRACT

Accurate registration of Functional Magnetic Resonance Imaging (FMRI) T2*-weighted volumes to same-subject high-resolution T1-weighted structural volumes is important for Blood Oxygenation Level Dependent (BOLD) FMRI and crucial for applications such as cortical surface-based analyses and pre-surgical planning. Such registration is generally implemented by minimizing a cost functional, which measures the mismatch between two image volumes over the group of proper affine transformations. Widely used cost functionals, such as mutual information (MI) and correlation ratio (CR), appear to yield decent alignments when visually judged by matching outer brain contours. However, close inspection reveals that internal brain structures are often significantly misaligned. Poor registration is most evident in the ventricles and sulcal folds, where CSF is concentrated. This observation motivated our development of an improved modality-specific cost functional which uses a weighted local Pearson coefficient (LPC) to align T2*- and T1-weighted images. In the absence of an alignment gold standard, we used three human observers blinded to registration method to provide an independent assessment of the quality of the registration for each cost functional. We found that LPC performed significantly better ($p<0.001$) than generic cost functionals including MI and CR. Generic cost functionals were very often not minimal near the best alignment, thereby suggesting that optimization is not the cause of their failure. Lastly, we emphasize the importance of precise visual inspection of alignment quality and present an automated method for generating composite images that help capture errors of misalignment.

Published by Elsevier Inc.

Introduction

In Blood Oxygenation Level Dependent (BOLD) Functional Magnetic Resonance Imaging (FMRI), T2*-weighted volumes (*E*) are acquired to create maps of brain activity. Because T2*-weighted volumes are low-resolution and have poor anatomical contrast, these activation maps are almost always overlaid on a separate, high-resolution T1-weighted structural volume (*S*) collected in the same subject. Inferring any neuroscience conclusion from the activation map therefore depends on a close spatial correspondence between the T2*- and T1-weighted volumes; high precision in this correspondence is particularly critical for pre-surgical mapping (Hirsch et al., 2000; O'Shea et al., 2006; Sunaert, 2006; Yetkin et al., 1996) and cortical surface based analyses (Argall et al., 2006; Dale et al., 1999; Fischl et al., 1999; Van Essen et al., 2000, 2001). In order to ensure the correspondence between anatomy and function, most neuroimaging data analysis packages perform cross-modality registration. Ensuring that *E* is well aligned with *S* is very important.

Because the T2*- and T1-weighted images have very different contrasts, registering them is considered to be "cross-modality" registration (although both volumes are collected using MRI). Automated pair-wise image alignment tools seek improved alignment by minimizing a cost functional that measures the mismatch between the two volumes (or whose negative measures the correspondence between the images). Optimization routines are used to seek the spatial transformation that minimizes the cost functional between the transformed volume and its pair. Achieving alignment is consigned to the successful reduction of the cost functional while avoiding local minima; however, it is known that for many cost functionals, this reduction does not necessarily translate into better alignment. An extreme example is that the Mutual Information (MI) cost of two volumes which are completely out of alignment might be better (lower) than the cost when they were somewhat aligned (Studholme et al., 1999). Such problems, readily detected by a cursory visualization of the results, are addressed in software by using more robust cost functionals and heuristics to restrict the transformation parameter space.

Most current image registration packages use generic cost functionals that do not rely on a specific model of the signal intensity between the volume pair. For example, AFNI's 3dAllineate and FSL's

* Corresponding author. Fax: +1 301 402 1370.
E-mail address: rwcox@nih.gov (R.W. Cox).

1053-8119/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.neuroimage.2008.09.037

FLIRT have variants of MI and Correlation Ratio (CR) cost functionals. As developers of the AFNI neuroimaging data analysis package, our initial impetus for the developments reported herein was a collection of Echo Planar Imaging (EPI) and structural volumes, presented to us by several FMRI researchers who had great difficulty in obtaining decent alignments between their functional time series and their anatomical reference volumes. These datasets had been processed in AFNI (http://afni.nimh.nih.gov) (Cox, 1996; Cox and Jesmanowicz, 1999), FSL (http://www.fmrib.ox.ac.uk) (Jenkinson and Smith, 2001), and SPM (http://www.fil.ion.ucl.ac.uk/spm/) (Collignon et al., 1995), and none of these packages had produced satisfactory results, despite numerous attempts at adjusting the available algorithmic parameters.

Because of the failure of these routines, we developed a specialized cost functional that is optimized for T2*- to T1-weighted image alignment. During this process, we addressed the following questions: do the transformations obtained by minimizing general purpose cross-modality cost functionals reliably result in good alignment of EPI and structural MRI volumes? How can we assess the quality of the alignment over an entire volume, without an objectively computable figure of merit, and considering the minimal anatomical contrast in the EPI data? Answering such questions for real data is difficult. Using simulated data, where the proper alignment is known, does not fully take into account the complex variations in noise, contrast, signal dropouts, and image distortion that occur in practice. In our examination of the registration results, we rely on meticulous visual inspections of the detailed anatomical overlap between the image pairs, facilitated by automatically creating edge-enhanced versions of the volumes and displaying them in a way that allows one to compare alignment results obtained with a variety of cost functionals. While this methodology does not easily allow for precise quantification of misalignment, it does allow for a coarse scale rating of different alignment results.

In the Methods section, we first describe a new cost functional designed for the specific purpose of T2*-to-T1 alignment. We then describe our visual examination methodology for assessing and rating the results of different alignment algorithms, including our statistical approach for interpreting the scores. In the Results section, we illustrate the types of displays the raters used in assessing the registration results, and then present the results of the statistical analysis. We conclude with a discussion of two non-controversial but under-appreciated points: generic image analysis algorithms can often be improved by modality-specific enhancements, and visual examination of transformed images is vital to ensure the integrity of the FMRI data processing stream.

Methods

Generic cross-modality cost functionals

We denote a single T2*-weighted EPI volume by $E(\mathbf{x})$ and a T1-weighted structural volume by $S(\mathbf{x})$, where $\mathbf{x}$ is the (discretized) spatial coordinate vector. Volume pair registration is generically performed by minimizing some real-valued cost functional $C[E(\mathbf{T}(\mathbf{x};\theta)),S(\mathbf{x})]$ over a subset of proper affine transformations parameterized by a vector θ: $\{\mathbf{T}(\bullet,\theta) \ni \det[\mathbf{T}]>0$ and θ is "reasonable"$\}$. *Ad hoc* constraints are usually placed on the parameter vector θ to limit the search space for the sake of stability and computational efficiency. Joint histogram-based cost functionals, such as Mutual Information (MI) (Wells et al., 1996) and Correlation Ratio (CR) (Roche et al., 1998), are commonly used for cross-modality image registration problems. The underlying idea is that when the images are properly aligned, the voxel values $E(\mathbf{T}(\mathbf{x};\theta))$ and $S(\mathbf{x})$ should be mutually dependent; that is, knowing the value of at a particular location x should maximally reduce the uncertainty about the value of $E(\mathbf{T}(\mathbf{x};\theta))$ (and vice-versa) when $\mathbf{T}(\bullet,\theta)$ correctly aligns the images, since the image intensities would then be derived from the same volume of tissue. The various cost functionals in common use differ primarily in their quantification of "uncertainty reduction". While these methods offer the distinct advantage of being applicable regardless of the types of MRI contrasts in the pair being registered, we found that they often resulted in improper alignment even when volumes were fairly well aligned to begin with. This led us to closely examine their performance and eventually design a T_2*-to-T_1 specific cost functional that we present next.

T_2-to-T_1 specific cost functional: Weighted local Pearson correlation*

For the purpose of functional-structural alignment, we used specific information about the properties of T2*-weighted EPI and T1-weighted Magnetic Resonance (MR) volume images to develop a special purpose cost functional. EPI, with its 2D slice acquisitions and resulting long Repetition Time (TR), has larger image values (is brighter) in CSF than in white matter. Structural volumes are dark in CSF and bright in white matter. Approximately speaking, the contrast between $E(\mathbf{x})$ and $S(\mathbf{x})$ is reversed; however, this negative correlation is most noticeable in the intensity difference between CSF and other tissue types, and is less reliable between white matter and gray matter.

Structural volumes are also bright in fat, whereas EPI volumes are universally acquired with fat suppressing saturation RF pulses. Therefore, all EPI-to-structural registration work must be done with "skull stripped" structural volumes; it is usually necessary to mask off the residual fat signature surrounding the EPI volumes as well.

Our initial efforts at improving EPI-structural alignment used the signed Pearson correlation (PC) between $E(\mathbf{T}(\mathbf{x};\theta))$ and $S(\mathbf{x})$ as the cost functional; minimizing this cost functional tries to find the transformation $\mathbf{T}(\bullet,\theta)$ that yields the most negative correlation between the volume pair. This method worked well in some cases, but not in others. Close examination of the $\{E,S\}$ joint histograms showed that the unreliability of the negative correlation between and in parenchymal tissue was a problem that rendered this method no more reliable (on average) than generic MI- or CR-based registration. In the joint histogram, much of the predictability between E and S was found to originate at the high values of E and low values of S — that is, in CSF overlap. Therefore, we decided to emphasize CSF and to enhance significant negative correlations in our cost functional by making some simple modifications to the PC calculation.

In our new cost functional, we weight bright regions in the EPI volume more heavily (but not exclusively); that is, we compute a spatially weighted correlation coefficient. After some initial experiments, we settled on the weight function $w(\mathbf{x})=\min(1,E(\mathbf{x})/E_{90})$, where E_{90} is the 90% point on the cumulative histogram of the EPI volume, restricted to the EPI brain mask. In addition, to reduce sensitivity to non-uniformity artifacts – particularly important when using multi-coil EPI acquisitions – we first compute this correlation coefficient r only locally, in a neighborhood $N(\mathbf{x})$ about any given point $\mathbf{x}$, then nonlinearly combine a collection of these $r(\mathbf{x})$ values into the cost functional:

$$W(\mathbf{x}) = \sum_{y\in N(\mathbf{x})} w(\mathbf{y}) \qquad \text{[local sum of weights]}$$

$$M(\mathbf{x};F) = \frac{1}{W(\mathbf{x})} \sum_{y\in N(\mathbf{x})} w(\mathbf{y})\cdot F(\mathbf{y}) \qquad \text{[local weighted mean of volume } F\text{]}$$

$$Q(\mathbf{x};F,G) = \sum_{y\in N(\mathbf{x})} w(\mathbf{y})\cdot[F(\mathbf{y})-M(\mathbf{x};F)]\cdot[G(\mathbf{y})-M(\mathbf{x};G)]$$

[local scalar product of F and G]

$$r(\mathbf{x}) = \frac{Q(\mathbf{x};E,S)}{[Q(\mathbf{x};E,E)\cdot Q(\mathbf{x};S,S)]^{1/2}} \qquad \text{[local weighted correlation coefficient]}$$

$$C_{LPC}[E,S] = \frac{\sum_{\mathbf{x}\in P} W(\mathbf{x})\cdot s(r(\mathbf{x}))\cdot|s(r(\mathbf{x}))|}{\sum_{\mathbf{x}\in P} W(\mathbf{x})}$$

[combined correlation coefficients]

where symbols $F()$ and $G()$ denote arbitrary volume images, where the scalar function $s(r)=\mathrm{Tanh}^{-1}(0.9999\cdot r)$ provides a nonlinear stretching

to accentuate larger correlations, and where P is a set of neighborhood centers, chosen so that the union of neighborhoods $\cup_{\mathbf{x}\in P} N(\mathbf{x})$ covers the volume of interest in $S(\mathbf{x})$ (i.e., the brain as output by the skull stripping software, plus a small buffer zone). We call this cost functional the Local Pearson Correlation (LPC). The use of cross-correlation is common as a measure of matching in image processing and in MRI in particular (Collins et al., 1994); however, the LPC approach differs by its use of the weighting function and of localized estimates that are later combined nonlinearly. We emphasize that our algorithm is seeking the transformation $\mathbf{T}(\mathbf{x}, \theta)$ that produces the smallest (most negative) possible $C_{\mathrm{LPC}}[E(\mathbf{T}(\mathbf{x}, \theta)), S(\mathbf{x})]$ (minimizing $C_{\mathrm{LPC}}[E(\mathbf{x}), S(\mathbf{T}^{-1}(\mathbf{x}, \theta))]$ is also an option).

The basic neighborhood $N(\mathbf{x})$ is chosen to be a rhombic dodecahedron (Kepler, 1611) centered at x with size parameter $a = 6.5 \cdot (S \text{ voxel volume})^{1/3}$, with volume $= 2a^3$. The rhombic dodecahedron (*RHDD*) centered at the origin is defined as the set

$$N_{RHDD}(0) = \{(x,y,z) \ni |x| + |y| \le a \text{ and } |x| + |z| \le a \text{ and } |y| + |z| \le a\}$$

The set P, which defines the collection of neighborhoods, is chosen as the set of integer linear combinations of the 3D face centered cubic lattice basis vectors $(a, a, 0)^T$, $(a, 0, a)^T$, and $(0, a, a)^T$, such that each *RHDD* overlaps at least 50% with the volume of interest; this choice provides a non-overlapping set of neighborhoods over which to compute the local correlation coefficients. Rhombic dodecahedra were chosen as the basis neighborhoods because, unlike spheres, they can tile space without overlap or gaps (so that no voxel is uncounted or double counted in the cost functional), because they are "rounder" than cubes, and because they and their lattice are easy to describe and compute.

For completeness, we describe a few more implementation details. To carry out the minimization of $C[E, S]$ over the parameter vector θ, our software incorporates Powell's derivative-free NEWUOA software (Powell, 2006), modified slightly to allow for range constraints on the solution parameters. The vector θ comprises 3 shifts, 3 rotation angles, 3 scale factors, and 3 shears; the user can fix any of these parameters to zero if desired (e.g., to avoid stretching $E(\mathbf{x})$ in the through-slice direction). Trilinear interpolation in $E(\mathbf{x})$ is used during the alignment procedure; output volumes are produced with cubic interpolation by default. The non-brain tissue voxels are removed from the structural volume prior to registration, using AFNI's 3dSkullStrip program. An intensity- and contiguity-based mask from this masked structural volume is used to define the volume of interest, over which the cost functional is evaluated. The echo-planar image $E(\mathbf{x})$ that is registered to the structural $S(\mathbf{x})$ is a representative volume from the FMRI time series, and is also skull-stripped. In the initial search phase, the images are smoothed and sub-sampled, and several coarse-fit parameter vectors are generated and partially optimized using these smoothed images; in the fine-fit phase, these candidate parameter vectors are further optimized without image smoothing, and the final result is the set of parameters θ that gives the smallest value of $C[E(\mathbf{T}(\mathbf{x};\theta)), S(\mathbf{x})]$ detected. The cost functional is estimated by pairing voxels from both $E(\mathbf{T}(\mathbf{x};\theta))$ and $S(\mathbf{x})$ that fall within the brain masks of each volume. A brain voxel originally at location x in volume S is transformed to location $\mathrm{T}(\mathbf{x};\theta)$ and paired with the voxel at that location in E. However, this location may fall outside the brain mask of E, and the set of such outside-the-mask locations will vary with θ. To diminish the contribution of such voxel pairs, non-brain regions of $E(\mathbf{x})$ are filled with white noise so that they do not contribute appreciably to the cost functional. We chose this approach to "mask" the EPI volume, rather than simply ignore such transformed points, so that each evaluation of any cost functional $C[E(\mathbf{T}(\mathbf{x};\theta)), S(\mathbf{x})]$ will use exactly the same number of voxels. If non-brain voxels in $E(\mathrm{T}(\mathbf{x};\theta))$ were simply excluded when evaluating C, the number of voxels going into the calculation of C would vary as the overlap between $E(\mathbf{T}(\mathbf{x};\theta))$ and $S(\mathbf{x})$ varied with θ, possibly biasing the registration results.

Registration and assessment

We performed alignments using 4 of the cost functionals present in AFNI's 3dAllineate program: MI, CR, LPC, and the Hellinger (HEL) metric (Taneja, 2005) between the joint histogram and the product of the marginal histograms of $E(\mathbf{x})$ and $S(\mathbf{x})$. The same derivative-free optimization routines were used for all 4 cost functionals. The EPI dataset was used as a weighting volume for all of AFNI's registration tests, as described earlier. To verify that our results are not specific to our implementation of histogram-based cost functionals, we also performed the alignments and ratings on the same brain-only volume pairs using SPM's COREG, which uses an unweighted MI-based cost functional (Collignon et al., 1995) and FSL's FLIRT with its default correlation ratio cost functional. COREG uses only 6 parameter rigid-body transformations, while the AFNI and FSL codes use 12 parameter general affine transformations. The quality of these 8 various alignment methods – including the "do nothing" method (ORIG) – was scored as described below. Only the outcomes for AFNI's registration tools are presented herein; however, the results for COREG and FLIRT (with and without a weighting volume) were comparable to the corresponding AFNI MI and CR results, respectively, indicating that the results of our study are not strongly dependent on the software implementations of the cost functional or optimization algorithm.

We did not devise a quantitative metric to compare alignments: our thesis is that no such automated ideal metric yet exists for comparing actual T2*-weighted to T1-weighted images — otherwise, such a metric would be used as the cost functional. Therefore, we used observer judgments of alignment between 27 pairs of EPI and structural images using different registration methods. These datasets are distinct from those on which the LPC cost functional was developed and refined. The whole-brain images in this collection were acquired at 1.5 and 3.0 Tesla from different scanner manufacturers; voxel dimensions averaged about $3\times3\times3$ mm^3 in the EPI volumes and about $1\times1\times1$ mm^3 in the structural volumes. For data collected at the University of Texas, subjects were recruited and informed consent was obtained with the university's Committee for the Protection of Human Subjects. Subjects at the Medical College of Wisconsin gave informed consent in accordance with a protocol sanctioned by the college's Institutional Review Board. Subjects at the National Institute of Mental Health (NIMH) underwent medical screening and provided informed consent in compliance with NIMH-IRP human subjects committee. To reduce bias, the observers (authors ZSS, RWC, and DRG) were blinded to the identity of the registration method used for each dataset. The order of the datasets was randomized between observers to eliminate order effects, and each rater rated brain image pairs independently. Raters evaluated image alignments visually, using a four-point qualitative Likert scale (Likert, 1932) (1 = very poor…4 = good). This rating was based on the match between anatomical features in the T2*- and T1-weighted images, including ventricles, fissures and sulci, and was performed using software tools described in the next section. A score of 1 was used for very poor (often grossly erroneous) alignments. A score of 2 was assigned for errors larger than about 5 mm. A score of 3 was assigned when errors were between 5 and 2 mm. A score of 4 was reserved for pairs with little visible registration error, about 2 mm or less. Sample volume pairs and their corresponding rating can be seen in Figs. 2 and 5.

Out of 648 inter-rater score pairs, there were 13 instances where a particular alignment was given scores differing by 2 points or more. Concerned that these might have reflected errors in entering the scores, these pairs were randomized and re-rated several weeks after the first rating session, with the raters again blind to the registration method and to the previous score. After re-rating, none of the 13 brain pairs had score differences greater than one.

Fig. 1 shows structural (A1) and EPI (A2) images and an overlay of EPI atop the structural (A3). One common mode of assessing the

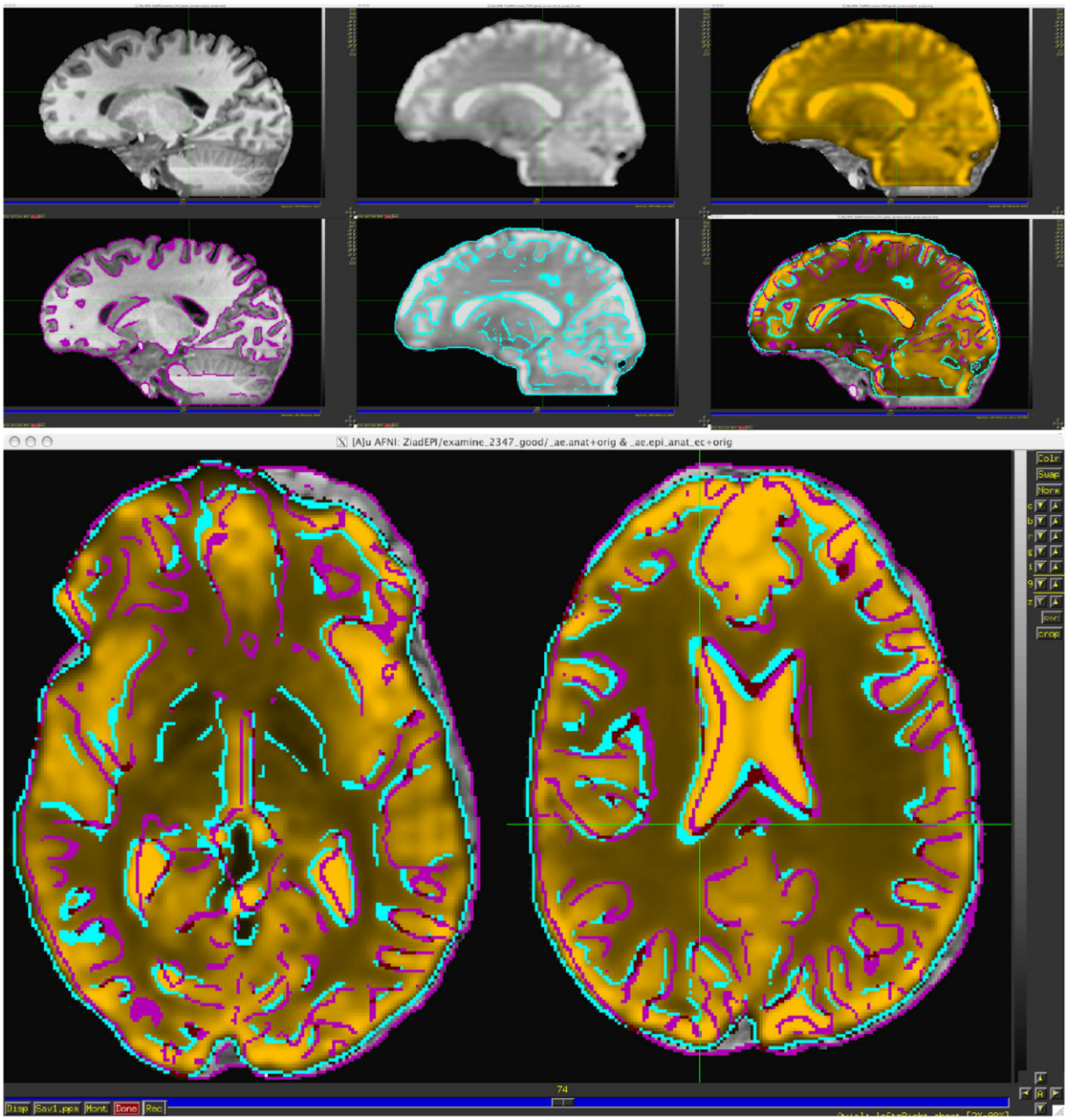

Fig. 1. (A) Structural (A1) and EPI (A2) images, and an overlay of amber-colored opaque EPI atop the structural image (A3). Opacity control is possible in this display mode, but we found that it complicates the composite image and confounds the interpretation. Crosshairs point to the same spatial coordinates in all images. The two horizontal lines show the location of the axial slices shown further below (C). (B) Binary volumes $S^{\#}(\mathbf{x})$ and $E^{\#}(\mathbf{x})$, of detected edges for EPI and structural volumes, respectively. The binary edges are highlighted in purple ($S^{\#}$) and cyan ($E^{\#}$) in B1 and B2, respectively. B3 shows a melded dual-edge-enhanced image created by the addition of $S^{\#}(\mathbf{x})$ and $E^{\#}(\mathbf{x})$ to the color overlay of A3. Dark red is used where structural and EPI edges overlap. (C) Two axial slices corresponding to the horizontal crosshair lines visible in the sagittal views. Alignment ratings for the images shown in this initial alignment state (ORIG) were 3, 3, 3.

alignment between cross-modality data was to place the crosshairs at a particular location in one volume and checking whether the crosshairs in the other volume (at the same (x, y, z) coordinates) correspond to a similar structure. A second mode, which cannot be represented in print, consisted of rapidly alternating the display between one volume and the next. The third common viewing mode was to layer one volume atop another. In Fig. 1(A3), we have the EPI as an opaque amber-toned volume with the structural behind it. Opacity control is possible in this display mode, but we found that it complicates the composite image and confounds the interpretation. In our experience, no single visualization mode is optimal for visually judging the alignment quality. In the first mode, attention is focused to one location in one volume at a time, making it hard to form a judgment for alignment over the whole volume. In the second mode, the difference of contrast and coverage is hard to overcome because of the salience of such features in a dynamic display. The third mode gives an overall sense of overlap between the volumes, but occasionally can be quite misleading, as shown in Fig. 3. Raters were free to use all of these modes in all viewing planes to form their quality evaluations.

Edge enhancement for registration assessment

To aid raters in their visual assessment of alignment, we devised an automated method for generating melded volumes with edge enhancement. Edges are detected in skull stripped EPI and structural

volumes using AFNI's 3dedge3 program, which performs 3D edge detection as implemented in the 3DEdge library by Gregoire Malandain (gregoire.malandain@sophia.inria.fr) (Deriche, 1987; Monga et al., 1991). Detected edges are turned to binary volumes $S^{\#}(\mathbf{x})$ and $E^{\#}(\mathbf{x})$, for EPI and structural volumes, respectively. The binary edges are highlighted in purple ($S^{\#}$) and cyan ($E^{\#}$) in Figs. 1(B1) and (B2), respectively. The melded dual-edge image shown in Fig. 1(B3) is created by the addition of $S^{\#}(\mathbf{x})$ and $E^{\#}(\mathbf{x})$ to the color overlay of Fig. 1 (A3). Dark red is used where structural and EPI edges overlap. Fig. 1(C) shows the two axial slices corresponding to the horizontal crosshair lines visible in the sagittal views. The presence of edges facilitates the comparison of alignments, as the detected edges clearly delineate anatomical features in both image types. Raters looked for EPI and structural edges that overlapped or coursed closely together. Note that not all internal edges are expected to match well; in particular, some edges are present in only one modality. EPIs have lower resolution and contrast, with spatially varying image quality; nevertheless, many sub-cortical and cortical edges are clearly visible in the processed EPI volumes. Based on their knowledge of brain anatomy, raters used the edges that looked plausible as a guide for judging registration. In practice, raters viewed volumes in all 3 cardinal planes (sagittal, coronal, axial), and examined various slices at different locations within each cut-plane direction, before assigning their scores. Raters were able to use any of the other visual checking modes described earlier; however, they usually found the dual-edge-enhanced mode to be the most useful. The process of generating these edge-enhanced images for any pair of volumes, as well as their presentation for viewing is easily automated using the @AddEdge script available in the AFNI distribution. The use of edges in the assessment process does not directly bias the results towards any of the methods used for registration, since none of them explicitly use edge information in their cost functionals.

While dual-edge images are useful for interactively ascertaining alignment quality, they have the drawback of being difficult to render in print where image sizes are significantly smaller than on the screen and where the colors, brightness, and contrast are difficult to control. Consequently, in later Figures we show single-edge-enhanced images where only the EPI edges are overlaid in cyan atop structural images in gray scale. The degree of correspondence between these automatically detected EPI edges and readily visible structural features can also be used to assess image alignment quality; *cf.* Figs. 2 and 3.

Bad convergence or bad functionals?

An important question is whether the under-performance of a method is one of optimization (Jenkinson et al., 2002) (Jenkinson and Smith, 2001), rather than an inappropriate cost functional. In the former case, the registration may be stuck in a local minimum and a superior algorithm might achieve better results. In the latter case, better optimization will not solve the problem. We compared a particular cost functional's value when it was used to align the volume to its value computed at the alignment obtained when LPC was used to produce a better alignment. We calculated the following as the percent change in a cost functional between the two alignment results

$$\Delta C_X = 200 \cdot \left(C_{X|LPC} - C_{X|X}\right) / \left|C_{X|LPC} + C_{X|X}\right|$$

where $C_{A|B}$ is the cost functional A estimated between volumes that were aligned by minimizing cost functional B; ΔC_{MI}, ΔC_{HEL}, and ΔC_{CR} were estimated for each of the 27 cases studied. To further understand the differences between cost functionals we mapped the cost functionals space along the two dimensions that typically experience the most movement: Translation along the inferior to superior direction (Tz), and Rotation along the left to right axis (Rx). Starting with the alignment parameters obtained with LPC, we varied Tz by up to ±10 mm and Rx by up to ±10° in increments of 0.5. The anatomical volume was repositioned per the new transform and the various cost functionals recomputed. This resulted in a 40×40 map of each cost functional about the location of optimal alignment as determined by LPC and by visual inspection. Such a map reveals the shape and smoothness of a cost functional's space in the vicinity of the optimal alignment, albeit for only two of the 12 dimensions possible.

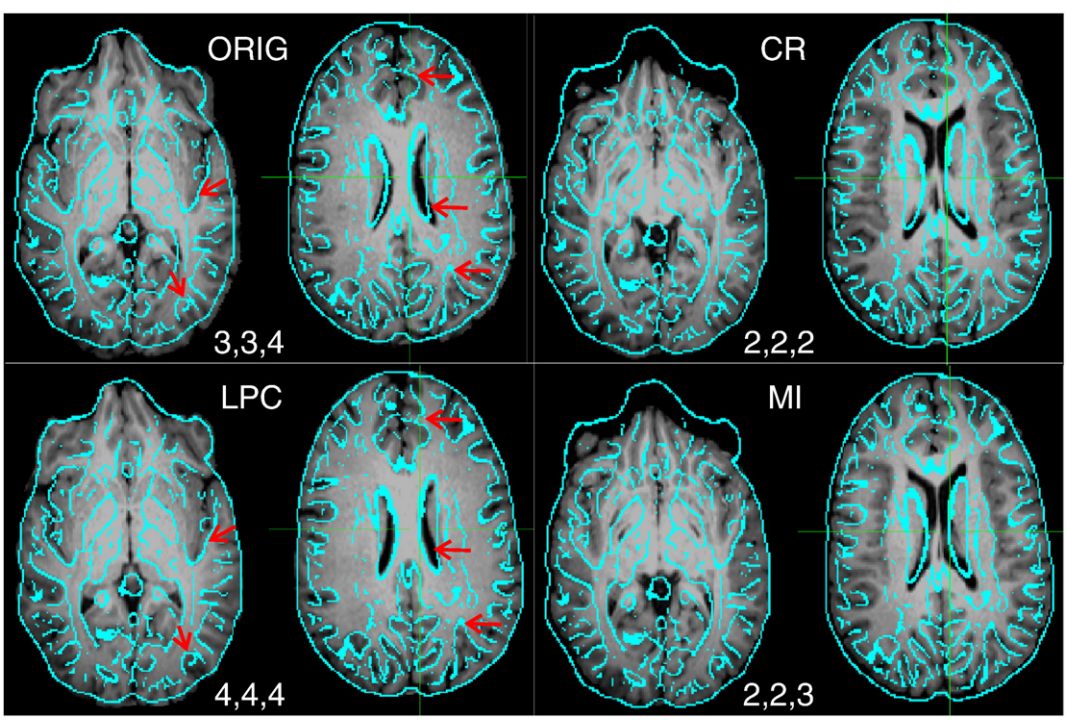

Fig. 2. Melded axial slices showing single-edge enhanced images extracted from the same *S* and *E* volume pair under 4 alignment outcomes: ORIG (as acquired), CR, MI, and LPC. Ratings from each of the three raters are shown below the alignment method's label. Note the better correspondence with LPC of EPI edges, with the anatomical features of the anatomical image in grayscale. Red arrows point to corresponding central and peripheral zones where improvements in the alignment between ORIG and LPC are evident.

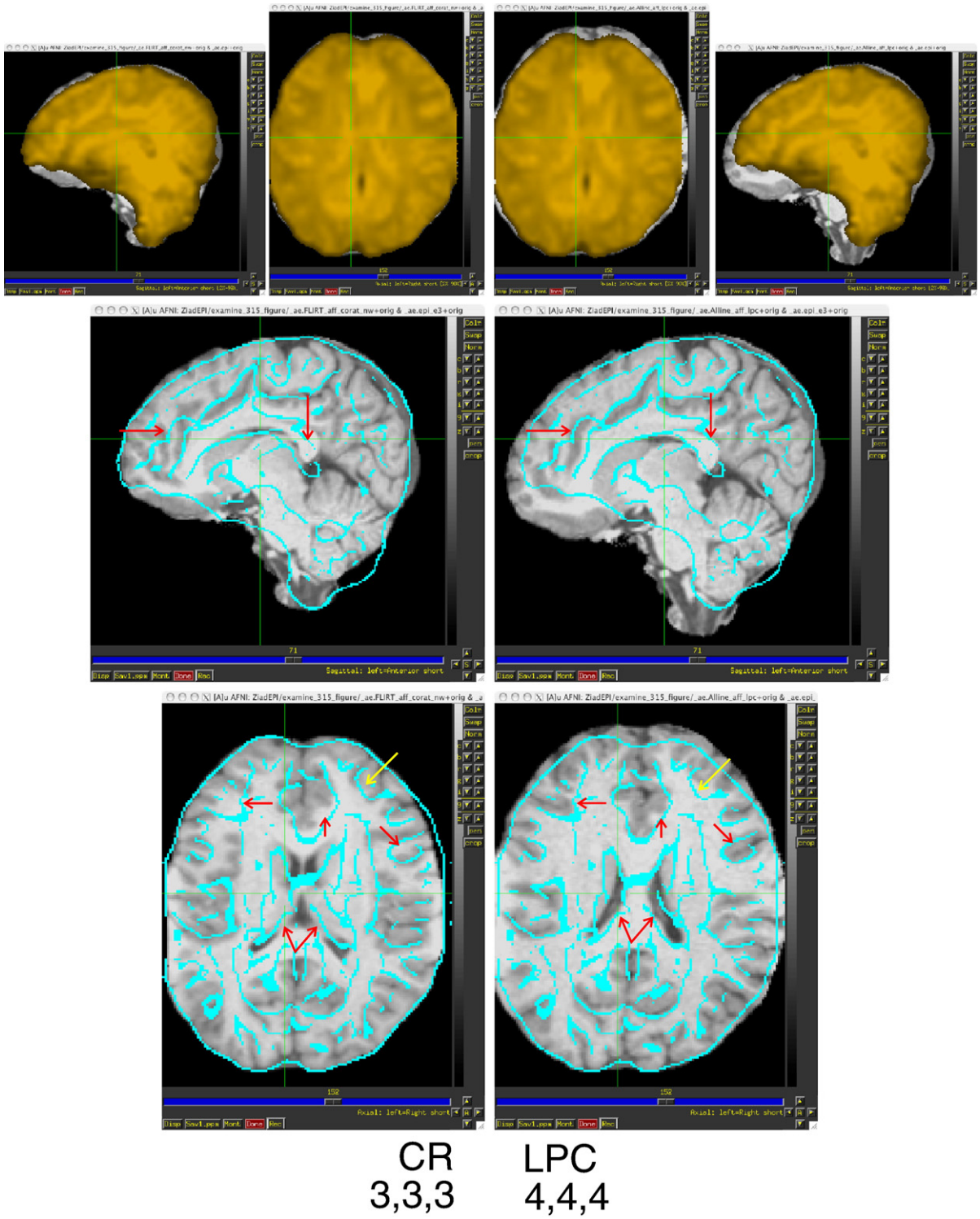

Fig. 3. Top Row: A standard overlay of EPI in amber atop a structural image in gray scale after alignment using the CR cost functional (left) and LPC (right). Middle and Bottom Rows: EPI edges only, overlaid in cyan atop T1-weighted images. Red arrows mark corresponding zones in the two aligned cases that can be used for comparing the quality of registration. The yellow arrow points to a location where the EPI edge alignment may have worsened. Note that although the CR result appears better in the display mode of the first row, it was scored worse than the LPC result after scrutiny of the overlap between internal features. Matching brain outlines is not always a good strategy.

Statistical score analysis

The collective set of ratings (27 image pairs × 3 raters × 5 alignment methods = 405 values) was statistically analyzed using contingency table methods (Agresti, 2002).

We first employed the following log-linear (or Poisson regression) model to account for the variability of the frequencies of each scoring case by a three-way table with the registration method (X) and rater (Z) as nominal variables, and the rank (Y) of a method as an ordinal variable:

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \mu_i v_j + \varphi_k v_j + \lambda_{ik}^{XZ} + \eta_{ik} v_j \quad (1)$$

where m_{ijk} is the expected number among the total number, N, of registered volumes in the j-th rank ($j = 1, 2, 3, 4$), using the i-th method

(i=1, 2, …, 5), rated by the k-th rater (k=1, 2, 3); λ is the overall effect (constant), λ_i^X, λ_j^Y, and λ_k^Z are the main effects corresponding to the three variables; interactions between method and rating are modeled as the product of a method-dependent factor μ_i and quality rank v_j (v_j=1 for "very poor", 2 for "poor", 3 for "moderate", and 4 for "good"), between method and rater by the product of a rater-dependent factor φ_k and quality rank v_j, between method and rater by λ_{ik}^{XZ}, and between the three variables by the product of a factor η_{ik} and quality rank v_j. The observed number of registered volumes in the j-th rank using the i-th method rated by the k-th rater is assumed to follow a Poisson distribution with expected value m_{ijk}. Since each rater scored all the combinations (135 volume pairs, 27 volumes×5 methods), the rater main effect λ_k^Z is set to 0. Also, for the same reason, the interaction between rater and method, λ_{ik}^{XZ}, was redundant. We therefore dropped those two components from Eq. (1) to obtain Eq. (2):

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \mu_i v_j + \varphi_k v_j + \eta_{ik} v_i \tag{2}$$

Model (2) was fitted and analyzed using the generalized linear model function in R (R Development Core Team, 2008) through the analysis of deviance on the rating data. The result indicates a non-significant rater effect φ_k on rating (deviance=0.044, df=2, p=0.978, based on an approximate χ^2 distribution), consistent with Fig. 4, which graphically shows little difference between inter-rater means and standard errors of the mean (SEMs). Also, no significant interaction was found between rater and method η_{ik} on rating (deviance=0.145, df=8, $p\approx$1.0, based on an approximate χ^2 distribution). Thus rater (Z,k) was dropped entirely to form a simplified model for the expected distribution of counts in each entry in a 2-way contingency table:

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \mu_i v_j \tag{3}$$

Model comparison through the analysis of deviance shows that the simplification from (2) to (3) is well justified (likelihood ratio statistic=0.189, df=10, $p\approx$1.0, based on an approximate χ^2 distribution). A further model reduction, where the interaction term between alignment method and rating $\mu_i v_j$ is assumed to be zero in model (3), results in a significantly worse fit, which suggests that there is a strong association between method and rating (likelihood ratio statistic=80.34, df=4, $p\approx$0.0, based on an approximate χ^2 distribution).

One interest of the analysis lies in quantitative estimates of the registration method's differential effect on rating, parameterized by μ_i in model (3). If a contrast between two method effects, μ_a–μ_b, turns out positive and statistically significant, we can reasonably infer that the ath method provides a better registration than the bth method; moreover, the relative success of a method can be quantified by its odds of being identified with one unit of higher score (e.g., good instead of moderate), estimated to be exp(μ_a–μ_b) times higher for the ath method than for the bth method.

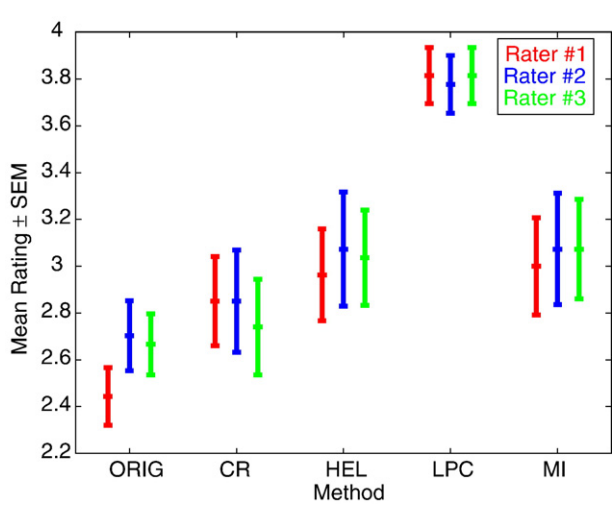

Fig. 4. Rating statistics for the various methods used. Bars represent the mean±SEM rating from each of the raters for each of the cost functionals. Also see Table 1 for results of a more detailed statistical analysis of the scores.

Results and discussion

Comparisons of generic functionals to LPC

Fig. 2 shows melded images for the data as acquired (ORIG), and after alignment with CR, LPC, and MI cost functionals, respectively. The scores given by the three raters are shown in white under the alignment cost functional label. Examination of the edge-highlighted views clearly shows that LPC resulted in an improved alignment from the ORIG case; improvement is evident at internal structures such as the ventricles, and at sulcal edges. Red arrows point to corresponding central and peripheral zones where improvements in the alignment between ORIG and LPC are evident. Alignments with CR and MI methods were rated as poor, as indicated by the lack of correspondence between EPI internal edges and the anatomy. In this example, one can observe that CR and MI alignments are poor without the use of edge enhancement, since there is an obvious volume contour mismatch in the frontal areas between EPI and anatomy. However, judging alignments primarily on brain outlines can be very misleading, as shown in Fig. 3, which shows two alignments of another {E, S} volume pair. In the top row, the CR alignment on the left appears better based on the matching of brain outlines, but it was the LPC alignment on the right that was deemed superior by all raters after examining the edge-enhanced images such as those shown in the lower part of Fig. 3. Red arrows point to image zones that show an improved alignment with LPC. A yellow arrow is used to point to a location that appears to have gotten worse in the LPC case. This mismatch between the EPI edge and the nearby structural edge could be due to nonlinear image distortion in the EPI volume, or to the fact that not all edges in the EPI data correspond to obvious structural features. The EPI edges a few cm posterior to the yellow arrow match the sulci visible in the structural image quite well, so if the problem is due to nonlinear image warping, the distortion must be quite local. The mismatch in the exterior brain contours in inferior and frontal areas is likely due to susceptibility induced signal dropout in the EPI volume.

Fig. 4 shows the rating statistics for the various methods used. Bars represent the mean±SEM rating from each of the raters for each of the cost functionals. The LPC method had the highest average score and the smallest SEM. Fig. 5 shows the frequency with which a score was assigned to a particular method, and also shows sample images from each score. The LPC method was assigned the highest score 88% of the time, with MI and HEL distant seconds at 50% and 47%, respectively. Implementations of histogram-based methods in FSL (CR) and SPM (MI) performed no better than their AFNI counterparts. The four image pairs represent sample alignment cases where the scoring was unanimous. The 'very poor' case is clearly far out of alignment. In the 'poor' and 'moderate' cases, red ellipses mark zones where edge mismatch led raters to assign the score. The small red mark provides a 1 cm scale for these images.

Fig. 6 shows the extent of inter-rater agreements and disagreements. Pairs of raters gave the same score 66% of the time (sum of diagonal entries) and scores differed by 1 point (off diagonal) 34% of the time.

Table 1 summarizes the statistical analysis of the model Eq. (3): (a) LPC is superior to all other methods (*cf.* the LPC column/row). For example, the likelihood that LPC gives a 'good' registration instead of a 'moderate' one is estimated to be exp(1.370)≈3.94 times higher than

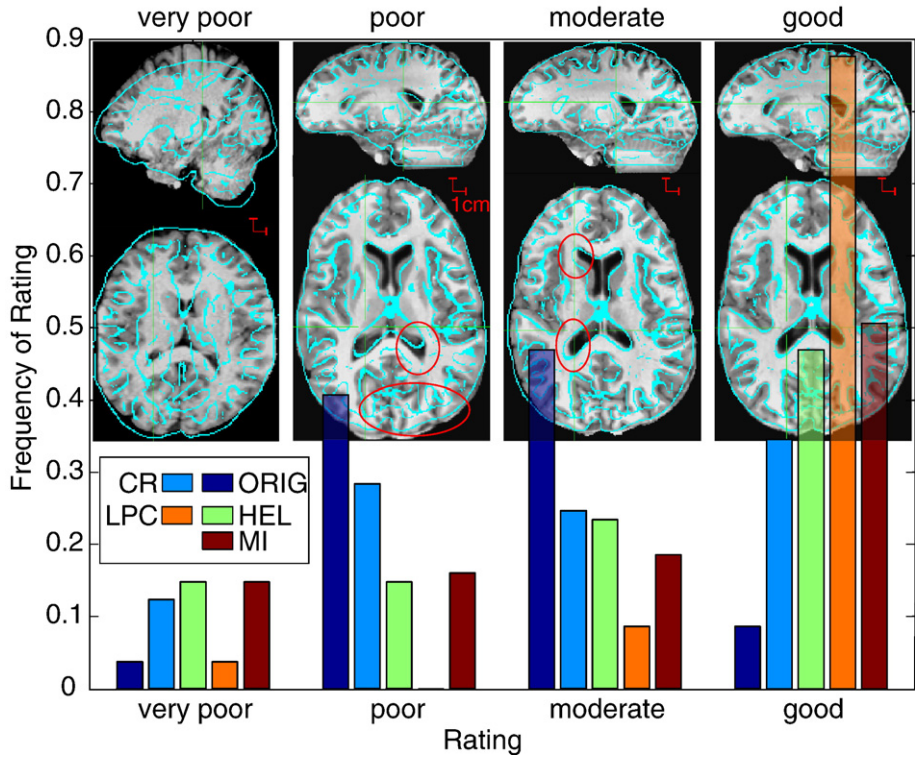

Fig. 5. Frequency with which a score was assigned to a particular method. The raw counts from this figure form the contingency table analyzed using Eq. (3). The four image pairs represent sample alignment cases where the scoring was unanimous. The 'very poor' case is clearly out of alignment. In the 'poor' and 'moderate' cases, red ellipses mark some of the zones where edge mismatches led raters to assign the score. The small red right-angle provides a 1 cm scale.

MI. (b) HEL and MI are significantly better than ORIG (*cf.* the ORIG column/row). For example, MI is exp(0.226)≈1.25 times as likely as ORIG to register an image with one unit of higher score. (c) CR, HEL, and MI are not significantly different from each other (*cf.* the CR and HEL columns/rows).

In addition to the blinded ratings and statistical tests obtained on the 27 sample datasets, we tested LPC alignment on 22 additional dataset pairs judged to be of good quality. We found that LPC alignment (scored by ZSS only and with no randomization) improved from an average of 2.9 (pre-registration) to 4.0. For those high quality datasets, the other cost functionals performed equally well. Furthermore, various beta testers have successfully applied LPC to more than 100 new cases. While no systematic ratings have been done on those additional data pairs, we conclude that the algorithm and its implementation are reasonably robust. (Readers that have AFNI installed on their computers can easily test the LPC registration method using the script *align_epi_anat.py*, which is a part of AFNI's distribution.)

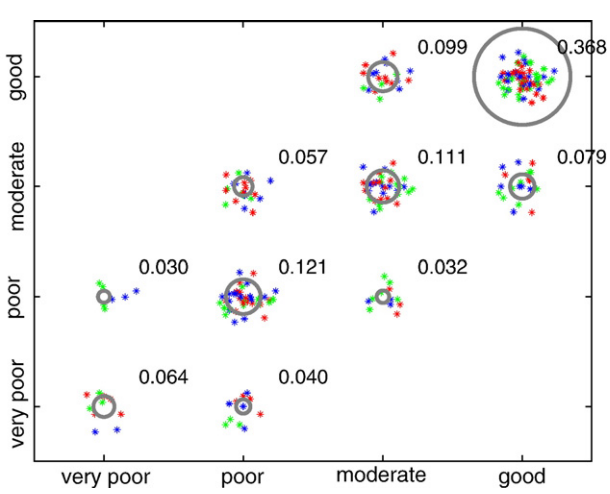

Fig. 6. Scatter plot of the ratings by each pair of raters. A small random value was added to each score to make the individual points visible; otherwise, all points at a particular rating would overlap because of the discrete-valued scoring. The circles are sized to reflect the fraction of data points in each score combination. The Spearman correlation coefficients for the 3 pairs of raters are: $\rho_{1,2}=0.84$, $\rho_{1,3}=0.78$, $\rho_{2,3}=0.84$.

Bad convergence or bad cost functionals?

Changes of cost functionals evaluated at the LPC solutions (ΔC_{MI}, ΔC_{HEL}, and ΔC_{CR}) were positive in 80 out of the 81 estimates, and ranged between −1% and +150%. In all but one case, MI, HEL, and CR cost functionals increased from their minimal values (became worse) at the alignment parameter vector computed with LPC, regardless of whether the alignment with LPC was judged better by the raters. (The single tiny observed decrease was observed with the CR cost functional.) These findings indicate that the problem of obtaining better alignments with current joint-histogram based functionals is not one of optimization. The problem lies with the cost functionals not being globally minimal at the better alignment. This also helps explain why, in some cases, alignments got worse with non-LPC methods, even when *S* and *E* were originally in good alignment. Despite this observation, it is still possible that non-LPC methods may exhibit better minima that are just not being reliably found by the optimization. This could be the case if these cost functionals were highly oscillatory in the vicinity of the optimal alignment. However, Fig. 7 indicates otherwise. The figure is a juxtaposition of 16 cost functional maps created using four functionals on four pairs of datasets. Columns show maps for the same cost functional and rows show maps from different pairs. Each of the sixteen maps show a cost functional, normalized to [0,1] as a function of ΔRx and ΔTz. The lowest value of the map (dark red) coincides with the best (lowest) cost functional in the range of parameters tested. The white dot in the

Table 1
Statistical comparison of the ratings of 5 registration methods: μ_i is the effect of registration method X_i on the ratings; the two-sided p-values are computed from the z-scores for the given μ_i contrast (*e.g.*, the p-value for MI to be superior to CR is only 0.15, even though $\mu_{MI}>\mu_{CR}$)

Method X_i	ORIG		CR		HEL		MI		LPC	
	$\mu_i-\mu_1$	p	$\mu_i-\mu_2$	p	$\mu_i-\mu_3$	p	$\mu_i-\mu_4$	p	$\mu_i-\mu_5$	p
i=1: ORIG	–	–	−0.19	0.21	−0.39	0.01*	−0.41	8E-3**	− 1.78	3E-11***
i=2: CR	0.19	0.21	–	–	−0.20	0.19	−0.23	0.15	−1.60	2E-9***
i=3: HEL	0.39	0.01*	0.20	0.19	–	–	−0.03	0.87	−1.40	2E-7***
i=4: MI	0.41	8E-3**	0.22	0.15	0.03	0.87	–	–	−1.37	3E-7***
i=5: LPC	1.78	3E-11***	1.59	2E-9***	1.39	2E-7***	1.37	3E-7***	–	–

Significance codes.
* $p<0.05$.
** $p<0.01$.
*** $p<0.001$.

center of each map indicates the parameters for which alignment was optimal per the LPC solution, which in each case resulted in a very good alignment. Iso-contours are added at levels 0.7, 0.4 and 0.1 to help visualize the terrain. The first row shows results from a simulated pair, where the "EPI" volume was a grayscale-inverted version of the anatomy with additional noise and smoothing by a 4 mm FWHM Gaussian kernel. This simulated case is intended to show the map for a case where near perfect alignment is possible. All four maps reach a minimum at the expected location, with the CR method showing the widest contours.

The second row was generated using one of the high-quality dataset pairs. In this case, all cost functional maps reached their minima very close to the optimal alignment position. This was expected given that the volume pair aligned very well with all methods, but it is instructive to note that the cost functional terrains appear equally smooth. The third row was generated using the same image pair illustrated in Fig. 3, which was in very good alignment from the start (score of 4,4,4), although there was significant EPI signal drop out in anterior and inferior regions. Therefore the unknown ideal alignment should be in the vicinity of the white dot. The maps show that all non-LPC maps do not reach a minimum in the vicinity of optimal alignment nor do they exhibit the high-frequency structure of many local minima that could easily derail an optimization routine.

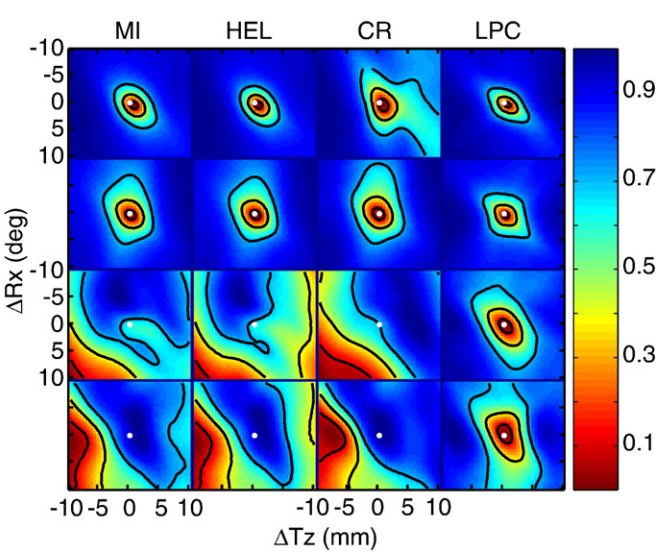

Fig. 7. 4×4 matrix of maps of cost functionals, individually normalized to the range [0,1], versus the ΔRx rotation and ΔTz translation parameters. The optimal alignment, found with the LPC method, is at ΔRx=ΔTz=0 and is marked by a white dot in each map. Iso-contours are drawn at values 0.7, 0.4, and 0.1, with the latter hidden at times under the white dot. Rows contain maps from different image volume pairs. The top two rows were from pairs where all methods succeeded in registering the volumes. The bottom two rows were from pairs where only LPC resulted in a good registration.

While there is a local minimum (light blue) for cost function MI in the vicinity of the optimal alignment, a well-implemented optimization would likely not get stuck there. Maps from the last row were assembled from a pair with an EPI dataset of poor contrast that aligned well only with LPC. In this case we can see a minimum in the CR map; however, it occurs at parameters that would result in a poor alignment between $E(\mathbf{x})$ and $S(\mathbf{x})$.

As expected from the visual inspection and optimization studies, non-LPC cost functionals do not always reach a minimum at the location of alignment. Such behavior has been reported for MI-based methods, but for different dataset pair types (Penney et al., 1998; Pluim et al., 2000). However, the maps for all cost functionals are smooth, revealing one clear minimum in the range of parameters examined. This reinforces the argument that better optimization would not yield better alignment. It is difficult to pinpoint exactly why the non-LPC cost functionals failed with some datasets. One possible explanation is that the presence of certain artifacts in the image, such as significant non-uniformity or signal dropouts, could make two images appear more predictive of each other, per the cost functional, when they are out of alignment. Recent improvements (Gan et al., 2008), which add gradient information in addition to image intensity to MI cost functionals are promising, and may prove to work as well as LPC on the difficult cases.

The performance of the LPC functional was improved with the incorporation of the weighting scheme. That same weighting scheme was also used to produce the results of the other AFNI cost functionals, although it may not be optimal for them. However, in our attempts to improve the alignment results, we began by trying various weighting schemes with joint histogram-based methods, but we found no single scheme that resulted in improved alignment in the majority of cases studied. The lack of robust improvement with simple modifications of joint histogram-based methods led us to the development of LPC.

The LPC cost functional has been tested and optimized for within-subject registration. We do not yet know how well this method would perform when S and E are from different subjects, or if S were a T1-weighted template image averaged from many subjects, or how well it would work with high-order non-affine transformations. Although the LPC method has some built-in allowance for image non-uniformity, it might need adjustments to work well with heavily shaded images, such as those acquired from a single local RF coil.

Conclusions

We emphasize that our 27 primary test cases all posed difficult registration problems: it was reports of these recurring troubles that led us to examine the EPI-structural alignment problem closely. However, these datasets are representative of the quality of data often acquired at different centers. With this real data, the problem with improving alignment between EPI and structural volumes is not one of implementation or of the optimization algorithm. Rather, generic histogram-based methods are not necessarily minimized at the best

alignment. We have shown that the modality-specific LPC cost functional, which utilizes some model of the image contrast, is more accurate and robust.

While the LPC method is fully automated and highly reliable, we nonetheless recommend that users always stay close to their data as it is transformed by the diverse processing steps used in FMRI. Despite the marked improvements with LPC and the robustness of the approach, we advocate that users assess registration quality by visual inspection using the edge-enhanced method. Our results demonstrate that a quick look at whether the brain outlines overlap is wholly inadequate. The generation of composite images and their presentation for examination is fully automated in AFNI and adds only a few minutes to the data processing time for each subject. It is critically important that registration quality be assessed visually before reliance is placed on the concordance between activation maps and anatomical features. Since the execution time for LPC registration is about 30% faster than for MI, CR, or HEL registration (in the AFNI implementation, LPC alignment takes about 300 CPU s on a 2 GHz Intel CPU), there is no particular penalty in using this new cost functional. With derivative-free optimization software such as NEWUOA (Powell, 2006), it is relatively straightforward to add the LPC cost functional to image alignment programs.

Acknowledgments

This research was supported by the NIMH and NINDS Intramural Programs of the NIH. MSB was supported by NSF grant 0642532 (PI Michael Beauchamp); RD was supported by NIH/NIDCD grant 5R03DC008416-02 (PI Rutvik Desai) and NIH/NINDS grant 2R01NS33576-11 (PI Jeffery Binder).

References

Agresti, A., 2002. Categorical Data Analysis, 2nd ed. Wiley.
Argall, B.D., Saad, Z.S., Beauchamp, M.S., 2006. Simplified intersubject averaging on the cortical surface using SUMA. Hum. Brain Mapp. 27, 14–27.
Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Suetens, P., Marchal, G., 1995. Automated multi-modality image registration based on information theory. Inf. Process. Med. Imag. (IPMI) 263–274.
Collins, D.L., Neelin, P., Peters, T.M., Evans, A.C., 1994. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. J. Comput. Assist. Tomogr. 18, 192–205.
Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162–173.
Cox, R.W., Jesmanowicz, A., 1999. Real-time 3D image registration for functional MRI. Magn. Reson. Med. 42, 1014–1018.
Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. NeuroImage 9, 179–194.
Deriche, R., 1987. Optimal edge detection using recursive filtering. Int. J. Comput. Vis. 167–187.
Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. NeuroImage 9, 195–207.
Gan, R., Chung, A.C., Liao, S., 2008. Maximum distance-gradient for robust image registration. Med. Image Anal. 12, 452–468.
Hirsch, J., Ruge, M.I., Kim, K.H., Correa, D.D., Victor, J.D., Relkin, N.R., Labar, D.R., Krol, G., Bilsky, M.H., Souweidane, M.M., DeAngelis, L.M., Gutin, P.H., 2000. An integrated functional magnetic resonance imaging procedure for preoperative mapping of cortical areas associated with tactile, motor, language, and visual functions. Neurosurgery 47, 711–721 discussion 721–712.
Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5, 143–156.
Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17, 825–841.
Kepler, J. Strena seu De Nive Sexangula (Godfrey Tampach Frankfurt am Main 1611).
Likert, R., 1932. A technique for the measurement of attitudes. Arch. Psychol. 140, 1–55.
Monga, O., Deriche, R., Malandain, G., Cocquerez, J.P., 1991. Recursive filtering and edge tracking: two primary tools for 3-D edge detection. Image Vis. Comput. 4, 203–214.
O'Shea, J.P., Whalen, S., Branco, D.M., Petrovich, N.M., Knierim, K.E., Golby, A.J., 2006. Integrated image- and function-guided surgery in eloquent cortex: a technique report. Int. J. Med. Robot. + Comput. Assist. Surg.: MRCAS 2, 75–83.
Penney, G.P., Weese, J., Little, J.A., Desmedt, P., Hill, D.L., Hawkes, D.J., 1998. A comparison of similarity measures for use in 2-D-3-D medical image registration. IEEE Trans. Med. Imag. 17, 586–595.
Pluim, J.P., Maintz, J.B., Viergever, M.A., 2000. Image registration by maximization of combined mutual information and gradient information. IEEE Trans. Med. Imag. 19, 809–814.
Powell, M.J.D., 2006. The NEWUOA software for unconstrained optimization without derivatives. In: Pillo, G.D., Roma, M. (Eds.), Large-Scale Nonlinear Optimization. Springer.
R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Roche, A., Malandain, G., Pennec, X., Ayache, N., 1998. The Correlation Ratio as a New Similarity Measure for Multimodal Image Registration. MICCAI '98; Proceedings of the First International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer-Verlag.
Studholme, C., Hill, D.L., Hawkes, D.J., 1999. An overlap invariant entropy measure of 3D medical image alignment. Pattern Recogn. 32, 71–86.
Sunaert, S., 2006. Presurgical planning for tumor resectioning. J. Magn. Reson. Imaging 23, 887–905.
Taneja, I.J., 2005. Refinement inequalities among symmetric divergence measures. Aust. J. Math. Anal. Appl. 2, 1–23.
Van Essen, D.C., Drury, H.A., Joshi, S., Miller, M.I., 2000. Functional and structural mapping of human cerebral cortex: solutions are in the surfaces. Adv. Neurol. 84, 23–34.
Van Essen, D.C., Drury, H.A., Dickson, J., Harwell, J., Hanlon, D., Anderson, C.H., 2001. An integrated software suite for surface-based analyses of cerebral cortex.[comment]. J. Am. Med. Inform. Assoc. 8, 443–459.
Wells 3rd, W.M., Viola, P., Atsumi, H., Nakajima, S., Kikinis, R., 1996. Multi-modal volume registration by maximization of mutual information. Med. Image Anal. 1, 35–51.
Yetkin, O., Zerrin Yetkin, F., Haughton, V.M., Cox, R.W., 1996. Use of functional MR to map language in multilingual volunteers. Ajnr: Am. J. Neuroradiol. 17, 473–477.