Estimation of Registration Accuracy
Applied to Multi-Atlas Segmentation
R. D. Datteri, A. J. Asman, B. A. Landman, and B. M. Dawant

Department of Electrical Engineering and Computer Science,
Vanderbilt University, Nashville, TN 37235, USA

Abstract. Multi-atlas registration-based segmentation has recently become a popular technique in medical imaging. Since the quality of individual atlas segmentations affect the quality of the results, atlas selection and atlas fusion have become important areas of research for multi-atlas segmentation. In this paper, we present an automatic technique that approximately calculates the quality of registration. We applied our method to multi-atlas segmentation and find that our measure correlates strongly ($R^2 = 0.79$) with the ground truth DICE similarity index. When applied to atlas fusion using a majority vote technique weighted by our measure of registration quality, our algorithm performs statistically better than both an un-weighted majority vote technique and a voting technique weighted by residual normalized mutual information.

Keywords: Image registration, registration circuits, atlas selection.

1 Introduction

Estimation of registration accuracy is a critical problem in the medical image processing and analysis community. In the rigid-body case the problem has been studied extensively in [1], [2], and [3]. In the non-rigid case, the estimation of error does not currently have an analytical solution. Non-rigid registration is often used to perform atlas-based segmentation, and, in this context, registration accuracy has been measured by comparing manual and automatic segmentation results [4].

A recent active theme of investigation in this area is multi-atlas segmentation. This concept requires the registration of multiple atlases and then the selection or combination of segmentation results provided by each atlas. Atlas combination techniques range from a straightforward majority vote to more sophisticated statistical fusion [5] and weighted voting [6] techniques. Atlas selection can be achieved using a range of methods that include intensity-based similarity measures (e.g., mutual information, correlation coefficient, etc.) between the volumes after registration or properties of the deformation fields that register the images [6] [7] [8].

In this work, we propose a new method to assess registration accuracy. The next section describes a generalized version of our technique, which we have called AQUIRC for Assessing Quality Using Image Registration Circuits. Following this, we compare the results of our algorithm to the ground truth DICE values. Finally, we apply the AQUIRC framework to atlas fusion and compare our results to a current popular method.
2 Methods

2.1 General AQUIRC algorithm
The AQUIRC algorithm builds on the idea of registration circuits which was proposed as a consistency measure by [9], [10], and [11]. A registration circuit involves three images A, B, and C and three transformations $T_{AB}$, $T_{BC}$, and $T_{CA}$. As discussed by Fitzpatrick [12], using only one registration circuit can lead to an underestimation of registration error because the error made along one edge in the circuit could correct error introduced from another one.

![Figure 1: Example complete graph with one circuit shown with red arrows](image)

In this work, we expand upon the idea of registration circuits to multiple circuits. We start with a set of images and compute pair-wise registrations between all elements in the set, creating a complete graph as shown in Figure 1. The complete graph of registrations is similar to what is done in [13] where it is used as an overall measure of the quality of a registration algorithm, rather than as a method to determine the quality of individual registrations as we have done here. If our initial set contains $N$ images, the graph contains $\binom{N}{2}$ edges. With each edge in this graph, we associate an initially unknown measure of registration quality, called $\varepsilon$ that we wish to solve for (we use the term quality because, so far, our method does not result in the true registration accuracy but rather an approximate calculation). There are $\binom{N}{3}$ unique registration circuits that can be formed from a complete graph (this is using registration circuits of size 3, the circuit size can be increased to form more registration circuits but this was not explored here).

Next, we define a measure of registration error that can be computed across a circuit. To compute this error, we select a certain set of points in image A, say $X$. We
then compute the transformed points $X'$ as $X' = T_{AB}(T_{BC}(T_{CA}(X)))$. The quality of registrations across circuit A, B, C, is then defined as $E_c = \text{dissimilarity}(X, X')$. There are many possible ways to define the dissimilarity function, and $X$ can be defined over the entire image or on specific regions of interest. The value, $E_c$, is affected by the error of three registrations, i.e., the registration error between A and B, the registration error between B and C, and the registration error between C and A. With only one circuit the contribution of each component cannot be computed. It can, however, be estimated with more than one circuit. To achieve this we make the assumption that each registration affects the quality measure additively, i.e., $\epsilon_{ABC} = \epsilon_A + \epsilon_B + \epsilon_C$. Computing this expression for all possible circuits and rearranging them in matrix form, we obtain

$$
\begin{bmatrix}
1 & 1 & 0 & \ldots & 0 \\
1 & 0 & 1 & \ldots & 0 \\
1 & 1 & 0 & \ldots & 0 \\
0 & 1 & 1 & \ldots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 1 & 1 & \ldots & 0 \\
\end{bmatrix}
\begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\epsilon_3 \\
\vdots \\
\vdots \\
\epsilon_{(N/2)} \\
\end{bmatrix}
= 
\begin{bmatrix}
E_{c_1} \\
E_{c_2} \\
E_{c_3} \\
\vdots \\
\vdots \\
E_{(N/2)} \\
\end{bmatrix}
$$

(1)

in which $E_{ci}$ is defined as the dissimilarity($X, X'$) value around circuit $i$. This expression can be rewritten as $P \bar{\epsilon} = \bar{E}_c$ and, as a result of the additive assumption, $\bar{\epsilon}$ can be solved for using a linear least squares solution

$$
\bar{\epsilon} = (P^T P)^{-1} P^T \bar{E}_c
$$

(2)

We are currently working on a proof of conditions on the registration circuits for when $P$ is full rank and therefore $(P^T P)$ is invertible. Empirically $P$ has been observed to be full rank in the general case when $N \geq 5$. We define $P$ to be all possible circuits in the graph, although it may be possible to utilize fewer circuits to eliminate redundancy.

There are two known instances where the AQUIRC algorithm is not able to identify the error in a registration between two images.

- First, if each of the registrations in a complete graph resulted in identical transformations then the metric $E_c$ would be the same for each registration circuit and AQUIRC would be unable to identify a relevant error value for each registration.
- Secondly, if there is an error that is the same across all registrations into one image, the AQUIRC algorithm is unable to account for this error. This is because for every possible circuit, the error is first added to, and then subtracted from the resulting combination of transformations. No circuit is thus able to account for that error since it is always removed from the final error value of a circuit.

We do not believe these two instances are likely to appear often in practice, but their impacts are under continuing evaluation.
2.2 Applied AQUIRC

In this applied example, the dissimilarity function we have chosen is the DICE [15] metric. Thus, $X$ is a set of manual segmentations and $X'$ is the projection of the manual segmentations across the circuit. DICE is calculated as

$$DICE = \frac{|X \cap X'|}{\frac{1}{2}(|X| + |X'|)} \quad (3)$$

We could have also utilized other methods such as a grid of points instead of manual segmentations, but this was not explored here. In the context of atlas selection, we do not have manual segmentations on the target image. Therefore we utilize all possible circuits except those circuits that begin on the target image. For example, in Figure 1, if image $I_1$ is the target image, we would not utilize any circuit that begins on image $I_1$. Thus, a circuit such as $T_{1i_2}(T_{2i_3}(T_{3i_4}(X)))$ would not be utilized because there is no manual segmentation $X$, to transform. We do, however, consider all other possible paths. This mimics the situation where there are 9 atlases with segmentations and one target image to segment. In practice, the set of atlases can be registered together ahead of time. Thus, only registrations between the multiple atlases and target image is required to produce a segmentation with this method. Empirically $P$ has been observed to be full rank in this case when $N \geq 10$ (i.e. 9 atlases and 1 target image). $N$ is larger here than in the general case because there are circuits that are not utilized.

2.3 Image Information and Registration Method

In this experiment, we used 10 T1-weighted head MRI atalases with dimensions of 256x256x256 and voxels of 1 mm in each direction from the OASIS data set [15], each containing manual delineations of 40 structures. The images were registered utilizing the Adaptive Bases Algorithm (ABA) [16], which uses Normalized Mutual Information (NMI) [17] as its similarity measure. Briefly, ABA computes a deformation field that is modeled as a linear combination of radial basis functions with finite support. This results in a transformation with thousands of degrees of freedom. Two transformations (one from the atlas to the subject and the other from the subject to the atlas) are computed simultaneously and constrained to be inverses of each other. Once the deformation field is computed between two MR images, the manual labels are transferred using nearest neighbor interpolation. When transferring labels across a circuit, $T_{AB}(T_{BC}(T_{CA}(X)))$, the deformation fields are first interpolated together using tri-linear interpolation. The interpolated deformation field is then applied to the labels which are transferred using nearest neighbor interpolation. DICE is calculated for each individual label and averaged over the 40 regions of interest.

2.4 Label Fusion
To combine the labels for multi-atlas segmentation, we used a weighted voting technique described as

$$\psi_i = \arg\max_s \sum_j I(D_{ij} = s) * w_j$$

(4)

where $\psi_i$ is the estimated label for voxel $i$, $s$ is the list of labels, $D_{ij}$ is the decision of rater (in this case, atlas segmentation) $j$, $I$ is the indicator function, and $w_j$ is the assigned weight of rater $j$. When we utilized an un-weighted voting scheme, $w_j$ was set equal to one for every rater.

3 Experiments and results

Two experiments have been performed to illustrate the performance of the AQUIRC algorithm. The first experiment correlates our measure of registration quality with a ground truth DICE metric; the second experiment applies the method to the problem of label fusion, we compare our results to residual NMI of the registered images, a method that is often used for atlas selection and atlas fusion.

3.1 Comparison to ground truth

To validate the AQUIRC algorithm we compared our results to the DICE value of each atlas-based segmentation since DICE is often utilized to verify the quality of a registration. We first computed a pair-wise registration between each of the 10 images in the dataset using the ABA registration algorithm. This resulted in a registration between each of the 10 images, with 45 unique registrations in total. The DICE similarity value was computed between the projected structures and the original structures and then correlated to the $\epsilon$ values. The AQUIRC algorithm was then applied in a leave-one-out framework. Each of the 10 images were treated as a target image to be segmented, using the other 9 images as atlases. All registration circuits were used except for those that began with the target image to be segmented. This is done to simulate an applicable situation, in which there will be no manual segmentation on a new target image.

Repeating the method for each image in the set resulted in 90 estimations of the quality of a registration because we estimate each registration’s $\epsilon$ value twice. For example, in Figure 1, $\epsilon_{11}$ is calculated when $I_1$ is the target image and again when $I_2$ is the target image. This tests if AQUIRC is consistent across different circuit configurations, which we found to be the case. When comparing the two $\epsilon$ values calculated for the same registration, we found that there was no statistical difference between the two sets of 45 $\epsilon$ values ($p > 0.05$). The mean difference between the $\epsilon$ values calculated for the same registration, is -0.0012 with a standard deviation of 0.0075.
As a comparison, we utilized a popular \cite{6} \cite{18} method often used in multi-atlas based segmentation to rank atlases, residual NMI. This method is calculated by taking the transformed atlas after registration, and calculating the residual NMI between the atlas and target MR images. To calculate the NMI, we utilized a bin size

**Figure 2:** Scatter plot of the $\epsilon$ quality of registration value compared to the ground truth DICE of 90 pair-wise registrations

**Figure 3:** Scatter plot of the residual NMI compared to the ground truth DICE of the 45 pair-wise registrations
of 16 for the histogram, and we calculated this for each of the 45 unique registrations. The ABA algorithm which was utilized in the registrations is bijective, therefore we did not calculate the residual NMI in both directions (i.e. atlas to target and target to atlas). Figure 4 shows the DICE and $\varepsilon$ values in a scatterplot. There is a strong correlation between the ground truth error metric (DICE) and the $\varepsilon$ values, with an $R^2 = 0.79$. The correlation for residual NMI on the other hand, is moderate, with an $R^2 = 0.44$.

### 3.2 Label Fusion Results

The quantitative measure of registration quality that results from AQUIRC lends itself to a label fusion technique. Therefore, we utilized an implementation of a weighted majority vote technique to fuse the labels of the multiple atlas segmentations. This was done by weighting the labels of each registered atlas using the $\varepsilon$ values calculated in the AQUIRC algorithm. Utilizing Equation (4), we assigned $w_j = \varepsilon_i$ where $\varepsilon_i$ is the estimated quality of the registration between atlas $j$ and the target image. This process was then repeated for each target atlas. For example, in Figure 1, if $I_1$ is the target image, $I_2$ through $I_5$ are each registered to $I_1$ and their labels are projected to $I_1$ resulting in four observed segmentations. Each segmentation is then combined using Equation (4), weighting their labels as $w_2 = \varepsilon_1$, $w_3 = \varepsilon_2$, $w_4 = \varepsilon_3$, $w_5 = \varepsilon_4$.

As a comparison to the AQUIRC method, we utilized residual NMI with the same implementation of a weighted voting scheme, weighting the observed segmentations by the residual NMI between the registered atlas and the target image. That is, $w_j = NMI$ (registered Atlas, target image), with $j$ being the atlas that was used. Going back to Figure 1, if $I_1$ is again the target image, with $I_2$ through $I_5$ registered to $I_1$, the weighting of each of the segmentations using Equation (4) would be $w_2 = NMI$ ($T_{I_2}, I_1$), $w_3 = NMI$ ($T_{I_3}, I_1$), $w_4 = NMI$ ($T_{I_4}, I_1$), $w_5 = NMI$ ($T_{I_5}, I_1$).

As a base line, we used a majority vote technique with equal weights. Each of the 10 images in the dataset was treated as a target image and the number of fused atlas-based segmentations was varied from 3 to 9. The order in which the atlases were fused was chosen randomly for each target image, but kept the same between each of the three methods. Figure 4 shows the average DICE score of all three methods, fusing 3 to 9 atlases, averaged over each of the 10 target images.

AQUIRC performed consistently better than both the un-weighted voting method and residual NMI weighted voting method. To show this, the DICE value for each individual fused atlas using both the residual NMI and un-weighted techniques were subtracted from the DICE value of the fused atlas of the AQUIRC method. Figure 5 contains the box plot of the DICE of the AQUIRC method subtracted by the DICE of the NMI weighted voting method and the pure weighted voting method. The increase in DICE score is statistically significant compared to both the majority vote method as well as compared to the residual NMI weighted vote method, with $p < 0.001$. 
Figure 4: Average and standard deviation of the DICE values of the AQUIRC, Majority Vote and Residual NMI atlas fusion methods, starting with 3 atlases and increasing the number of atlases fused to 9.

Figure 5: Box plot of the DICE value of the atlas-fused segmentation of AQUIRC subtracted by the DICE value of the atlas-fused segmentation of the Majority Vote and Residual NMI weighted techniques. The whiskers represent the minimum and maximum values. The red segment is the 2nd quartile and the blue segment is the 3rd quartile. The median is the line between the red and blue segments.
4 Discussion/Future Work

We have presented a new algorithm, AQUIRC, to measure the registration quality. AQUIRC correlates highly with the DICE ground truth, much more so than the popular technique of residual NMI. This suggests that AQUIRC is able to estimate the quality of a registration and that the algorithm could be easily applied to atlas selection. We intend to apply AQUIRC to a larger data set of images and use it for atlas selection.

Used as an atlas fusion method, the AQUIRC algorithm outperformed residual NMI as well as an un-weighted majority vote method. Although the AQUIRC algorithm resulted in a small improvement, this improvement was statistically significant and consistent over both the majority vote method and residual NMI method - two popular methods for multi-atlas segmentation. There is no agreement in the community over which methods are the strongest to use for atlas fusion; therefore, in the future we intend to compare our algorithm to a wider range of atlas fusion techniques.

The results we have obtained so far are very encouraging but represent only the beginning of our investigations in this area and we are pursuing a number of avenues. For instance, from our results we have not yet identified a direct correspondence between the value of $\varepsilon$ and the exact measure of TRE in a registration. It is possible that we may be able to “calibrate” our graph and determine a relation between $\varepsilon$ and TRE for a particular graph configuration and graph size. It is also possible to generalize this problem to several different error metrics and to look at ROI level and voxel level error metrics. As others have proposed [9], it may be better to choose some atlases over certain regions of interest in a target image and different atlases to segment other regions of interest for that same image. This can be readily achieved by computing $\varepsilon$ values over subsections of the images.

Acknowledgements

This work was partially supported by NSF grant DMS-0334769

References


