Deformable Registration of High-Resolution and Cine MR Tongue Images

Jonghye Woo[1,2], Maureen Stone[1], and Jerry L. Prince[2]

[1] Department of Neural and Pain Science, University of Maryland Dental School, Baltimore MD, USA
[2] Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore MD, USA
{jwoo,mstone}@umaryland.edu, prince@jhu.edu

Abstract. This work investigates a novel 3D multimodal deformable registration method to align high-resolution magnetic resonance imaging (MRI) with cine MRI of the tongue for better visual and motion analysis. Both modalities have different strengths to characterize and analyze the tongue structure or motion. Visual and motion analysis of combined anatomical and temporal information can synergistically improve the utility of each modality. An automated multimodal registration method is presented utilizing structural information computed from the 3D Harris operator to encode spatial and geometric cues into the computation of mutual information. The robustness and accuracy of the proposed method have been demonstrated using experiments on clinical datasets and yielded better performance compared to the conventional method and an average error comparable to the inter-observer variability.

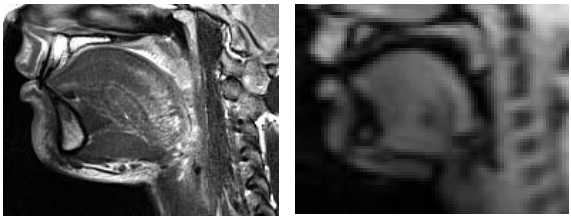
1 Introduction

Assessment of tongue motion can help the early diagnosis of disease, the evaluation of speech quality before and after surgery, and the functional analysis of the tongue [1]. Tongue anatomy is unusual; the tongue has three orthogonal fiber directions and extensive fiber inter-digitation, no bones or joints. This architecture makes the motion pattern of the tongue difficult to measure and quantify.

Assessment, diagnosis, and treatment of tongue disorders and understanding the tongue's motor control can be improved through a combinatorial analysis of tongue muscle anatomy and related tissue motion observed in magnetic resonance (MR) images [2,3]. For example, high-resolution magnetic resonance imaging (hMRI) provides muscle anatomy as shown in Figure 1(a) and cine MRI provides tongue surface motion as shown in Figure 1(b). The combination of hMRI and cine MRI offers complementary information in the study of tongue motion. However, each modality has its limits. hMRI is restricted to a static position and cine MRI does not have sufficient spatial resolution to provide high-quality tongue anatomy. To enhance the advantages of both modalities, it is necessary to combine them through registration.

In this work, we develop a fully automated and accurate 3D deformable registration method to align hMRI with cine MRI. Anatomical (hMRI) and temporal

G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part I, LNCS 6891, pp. 556–563, 2011.



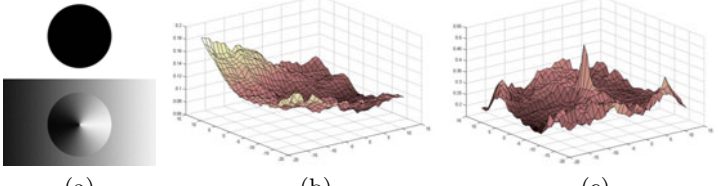
(a) (b)

Fig. 1. Example of the high-resolution MRI (hMRI) that was acquired at rest (a) and the first time frame of cine MRI that was acquired during speech task of "a geese" (b)

(cine MRI) data can be registered to provide correspondences between muscle anatomy identified in hMRI and tongue surface motions in cine, thus mapping changes in muscle pattern with surface motion. To our knowledge, this is the first study to perform registration between these two modalities.

Although mutual information (MI) is considered as a gold standard similarity measure for multimodal image registration, there are two problems in the conventional MI-based registration method. First, MI cannot handle the local intensity variations, which affects the estimation of joint histogram in MI computation [4,5]. Second, the statistics that are computed from overlap regions only considers corresponding intensity information and thus cannot encode spatial information [6]. Figure 2 illustrates a simple yet demonstrative example of this problem. Aligned synthetic circles as in [7] are used to show the limitation of the conventional MI. We compute the cost values with respect to different translations where no local maximum is found in conventional MI as shown in Figure 2(b) whereas the proposed method coincides with a local maximum corresponding to correct alignment as illustrated in Figure 2(c).

These problems were addressed partly by incorporating spatial information into calculation of the MI. Pluim *et al.* [6] combined spatial information by multiplying the MI with an external local gradient. Rueckert *et al.* [8] proposed higher-order mutual information. Russakoff *et al.* [9] proposed regional mutual information to take neighboring information of corresponding pixels into account. Yi *et al.* [7] proposed to include spatial variability via a weighted



(a) (b) (c)

Fig. 2. Comparison between conventional mutual information (b) and the proposed method (c). With aligned synthetic circles (a), we plot the cost values with respect to the different translations along x and y axes.

combination of normalized mutual information and local matching statistics. Zhuang *et al.* [5] proposed to unify spatial information into the computation of the joint histogram. Loeckx *et al.* [10] investigated the conditional MI by incorporating both intensity dimensions and spatial dimension to express the location of the joint intensity pair.

To mitigate the limitations of conventional MI, we propose a novel mechanism to incorporate both spatial and geometric information into the calculation of MI using the Harris operator.

2 Method

Our method is based on an iterative framework of computing MI incorporating spatial information and geometric cues. The underlying idea is to split the image into a set of non-overlapping regions using a 3D Harris operator and to perform registration on spatially meaningful regions. Additionally, we exploit structural information describing gradient of the local neighborhood of each pixel to define structural saliency to compute MI.

2.1 Maximization of Mutual Information

We denote the images $I_1 : \Omega_1 \subset \mathbb{R}^n \to \mathbb{R}$ and $I_2 : \Omega_2 \subset \mathbb{R}^n \to \mathbb{R}$, defined on the open and bounded domains Ω_1 and Ω, as the template and target images, respectively. Given two images, a deformation field is defined by the mapping $u : \Omega_2 \mapsto \Omega_1$. The goal of registration is to find a deformation field at each pixel location $\mathbf{x}$ such that the deformed template $I_1(u(\mathbf{x}))$ is as close as possible to $I_2(\mathbf{x})$ satisfying the given criterion. Since I_1 and I_2 are considered to be different modalities, we focus on the MI criterion for registration [4]. The main idea is to find the deformation field $\hat{u}$ by maximizing the statistical dependency between the intensity distributions of the two images, i.e.,

$$\hat{u} = \arg\max_{u}(\mathcal{M}(I_1(u(\mathbf{x})), I_2(\mathbf{x}))), \tag{1}$$

where $\mathcal{M}$ denotes the mutual information of the two distributions. $\mathcal{M}$ can be computed using joint entropy $\mathcal{H}$ as

$$\begin{aligned} \mathcal{H}(I_1(u(\mathbf{x})), I_2(\mathbf{x})) &= -\iint p(i_1, i_2) \log p(i_1, i_2) di_1 i_2 \\ \mathcal{M}(I_1(u(\mathbf{x})), I_2(\mathbf{x})) &= \mathcal{H}(I_1(u(\mathbf{x}))) + \mathcal{H}(I_2(\mathbf{x})) - \mathcal{H}(I_1(u(\mathbf{x})), I_2(\mathbf{x})) \\ &= \int_{\mathbb{R}^3} p_u(i_1, i_2) \log \frac{p_u(i_1, i_2)}{p_{I_1}(i_1) p_{I_2}(i_2)} di_1 di_2, \end{aligned} \tag{2}$$

where $i_1 = I_1(u(\mathbf{x})), i_2 = I_2(\mathbf{x})$, and $p_{I_1}(i_1)$ and $p_{I_2}(i_2)$ are marginal distributions. $p_u(i_1, i_2)$ denotes the joint distribution of $I_1(u(\mathbf{x}))$ and $I_2(\mathbf{x})$ in the overlap region $V = u^{-1}(\Omega_1) \cap \Omega_2$ which can be computed using the Parzen window given by

$$p_u(i_1, i_2) = \frac{1}{|V|} \int_V \varphi\left(\frac{i_1 - I_1(u(\mathbf{x}))}{\rho}\right) \varphi\left(\frac{i_2 - I_2(\mathbf{x})}{\rho}\right) d\mathbf{x}\,, \tag{3}$$

where φ is a Gaussian kernel and ρ controls the width of window.

2.2 Volume Labeling Using 3D Harris Operator

The Harris corner detector [11] was first introduced to detect corner features that contain high intensity changes in the horizontal and vertical directions. In this work, we extend the 2D Harris detector used for images or video sequences to localize meaningful features in 3D images. The Harris operator is based on the local autocorrelation function of the intensity, which measures the local changes of the intensity with patches shifted in different directions. We first define the autocorrelation function as

$$c(x,y,z) = \sum_{x_i,y_i,z_i} W(x_i,y_i,z_i)\left[I(x_i,y_i,z_i) - I(x_i+\Delta x, y_i+\Delta y, z_i+\Delta z)\right]^2, \tag{4}$$

where $I(\cdot,\cdot,\cdot)$ denotes the image function, (x_i, y_i, z_i) are the points in the Gaussian function $W(\cdot,\cdot,\cdot)$ centered on (x,y,z) and $(\Delta x, \Delta y, \Delta z)$ represents a shift to define the neighborhood area. Using a first-order Taylor expansion, we can write

$$\begin{aligned} c(x,y,z) &= \sum_{x_i,y_i,z_i} \left[W \cdot I(x_i+\Delta x, y_i+\Delta y, z_i+\Delta z) \begin{bmatrix}\Delta x\\ \Delta y\\ \Delta z\end{bmatrix} \right]^2 \\ &= [\Delta x\ \Delta y\ \Delta z] \begin{bmatrix} \sum W \cdot I_x^2 & \sum W \cdot I_x I_y & \sum W \cdot I_x I_z \\ \sum W \cdot I_x I_y & \sum W \cdot I_y^2 & \sum W \cdot I_y I_z \\ \sum W \cdot I_x I_z & \sum W \cdot I_y I_z & \sum W \cdot I_z^2 \end{bmatrix} \begin{bmatrix}\Delta x\\ \Delta y\\ \Delta z\end{bmatrix} \\ &= [\Delta x\ \Delta y\ \Delta z]\, \mathcal{C}(x,y,z) \begin{bmatrix}\Delta x\\ \Delta y\\ \Delta z\end{bmatrix}, \end{aligned} \tag{5}$$

where I_x, I_y, and I_z denote the partial derivatives in the x, y, and z axes, respectively, and the *local structure matrix* $\mathcal{C}(x,y,z)$ captures the intensity structure of the local neighborhood. Let $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3$ denote the eigenvalues of the matrix $\mathcal{C}(x,y,z)$ and let the 3D Harris operator be defined as

$$H = \det(\mathcal{C}) - k(\mathrm{trace}(\mathcal{C}))^3, \tag{6}$$

where k is an arbitrary constant. Then each pixel can be classified as one of three types using a threshold T and the following definitions

- Type 1: $H \geq T$, Location having significant local variation
- Type 2: $H \leq -T$, Location having moderate local variation
- Type 3: $-T \leq H \leq T$, Location having small local variation

We assume that Type 1 and Type 2 regions have more structural and characteristic information compared to Type 3 (homogeneous) region to calculate local statistics. Thus we consider Type 1 and Type 2 regions to calculate MI. One example result of the voxel labeling is shown in Figure 3(b). The white, gray and black color represent the Type 1, Type 2, and Type 3, respectively.

2.3 Mutual Information Using Local Structure Matrix

MI represents the statistical relationship between the template and target images. As shown in Eq. (2), MI is calculated using the marginal and joint distributions of the two images. To address limitations stated before, we compute a weighted joint distribution in order to encode both spatial and geometric information in the objective function. The local structure matrix $\mathcal{C}(x, y, z)$ derived in Eq. (5) exhibits local intensity structure that implies gradient directions within a local neighborhood of each pixel. We can rewrite the joint distribution weighted by the distance between two matrices defined in corresponding pixels as:

$$p_u^{\mathcal{C}}(i_1, i_2) = \frac{1}{|V|} \int_V \gamma(\mathbf{x}) \cdot \varphi\left(\frac{i_1 - I_1(u(\mathbf{x}))}{\rho}\right) \varphi\left(\frac{i_2 - I_2(\mathrm{x})}{\rho}\right) d\mathbf{x} \tag{7}$$

where $\gamma(\mathbf{x})$ is a weighting function that incorporates the distance between local structure matrices between corresponding pixels given by

$$\gamma(\mathbf{x}) = \exp(-\frac{\Delta(\mathcal{C}_{i_1}(\mathbf{x}), \mathcal{C}_{i_2}(\mathbf{x}))}{m}). \tag{8}$$

Here, $\Delta(\mathcal{C}_{i_1}(\mathbf{x}), \mathcal{C}_{i_2}(\mathbf{x}))$ is a distance between two matrices, m is a normalization constant, and $\mathcal{C}_{i_1}(\mathbf{x})$ and $\mathcal{C}_{i_2}(\mathbf{x})$ are the local structure matrices of the corresponding pixels in $I_1(u(\mathbf{x}))$ and $I_2(\mathbf{x})$, respectively. The local structure matrices do not reside in a vector space and therefore the Euclidean metric does not hold. However, local structure matrices are symmetric and positive semidefinite (like covariance matrices), and therefore belong to a connected Riemannian manifold that is locally Euclidean [12]. Accordingly, we can define the distance between two structure matrices as

$$\Delta(\mathcal{C}_{i_1}(\mathbf{x}), \mathcal{C}_{i_2}(\mathbf{x})) = \sqrt{\sum_{n=1}^{N} \ln^2 \lambda_n(\mathcal{C}_{i_1}(\mathbf{x}), \mathcal{C}_{i_2}(\mathbf{x}))}, \tag{9}$$

where λ_n are the generalized eigenvalues of $\mathcal{C}_{i_1}(\mathbf{x})$ and $\mathcal{C}_{i_2}(\mathbf{x})$ and N is the number of rows and columns in each matrix. This definition of distance satisfies the metric properties including symmetry, positivity, and the triangle inequality.

We can rewrite MI based on the above weighting scheme as follows:

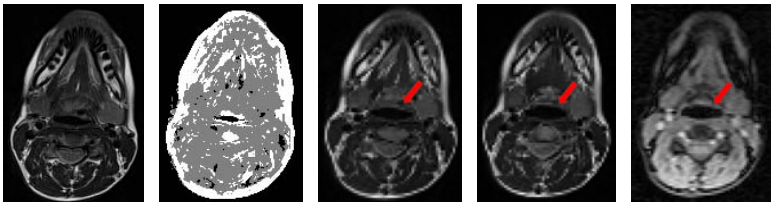
$$\mathcal{M}^{\mathcal{C}}(I_1(u(\mathbf{x})), I_2(\mathbf{x})) = \int_{R^3} p_u^{\mathcal{C}}(i_1, i_2) \log \frac{p_u^{\mathcal{C}}(i_1, i_2)}{p_{I_1}(i_1) p_{I_2}(i_2)} di_1 di_2. \tag{10}$$

Using this (modified) MI, the local structure matrices provide a geometric similarity measure while the image intensities continue to provide an appearance measure, thereby allowing us to find correspondence more reliably and address the limitation of the conventional MI-based registration.

2.4 Registration Model

Data fidelity Term. With the modified MI defined in Eq. (10) and labeled regional information, we can define data fidelity term given by

$$\mathcal{D}(I_1(u(\mathbf{x})), I_2(\mathbf{x})) = \sum_{k=1}^{K} w_k \chi_{D_k}(u(\mathbf{x})) \mathcal{M}^{\mathcal{C}}(I_1(u(\mathbf{x})), I_2(\mathbf{x})), \tag{11}$$



(a) (b) (c) (d) (e)

Fig. 3. One example of the results: (a) template image (hMRI) (b) volume labeling of the template image using the Harris operator, (c) a resulting image using conventional MI-based registration, (d) a resulting image using the proposed method and (e) the target image (cine MRI). The red arrows show that (d) and (e) are better aligned than (c) and (e) in terms of vocal tract edge.

where $w_k \in \mathbb{R}^+$ is the weight assigned to kth region and χ_{D_k} is kth characteristic function defined by

$$\chi_{D_k}(x) = \begin{cases} 1, & x \in D_k \\ 0, & x \notin D_k \end{cases} \tag{12}$$

Transformation Model. We use free-form deformations (FFD) based on uniform cubic B-splines to model the deformable registration as in [13]. Additionally, a multi-resolution scheme is used to represent coarse-to-fine details of both volumes for fast and robust registration. The energy functional is minimized using a Simultaneous Perturbation Stochastic Approximation (SPSA) [14] method.

3 Experiments and Results

3.1 Subjects and Task

Nine normal native American English speakers were subjects in this experiment. The speech task was "a geese". Both types of MRI datasets were recorded in the same session using a head and neck coil. Cine MRI datasets were collected with a 6mm slice thickness and had an in-plane resolution of 1.875mm/pixel resolution. hMRI datasets were 3mm thick with an in-plane resolution of 0.94mm/pixel. The subjects were required to remain still from 1.5 to 3 minutes for each plane.

3.2 Evaluation of the Registration Method

To evaluate the accuracy and robustness of the proposed method, we have performed two registration experiments on nine pairs of 3D axial MRI volumes described above. Both registrations were performed on the same two static volumes: (1) the first time frame of axial cine MRI that was acquired during speech task of "a geese" and (2) the axial hMRI volume that was acquired at rest. The registration methods used affine registration as an initialization, followed by the deformable registration using the proposed and conventional MI-based method using FFD. In our experiments, we set the number of histograms to 50, and used

Table 1. Registration errors and observer variability

TRE (voxel)	Before	Affine	Conventional	Proposed	Observer Variability
Tongue Tip	6.2±3.7	3.8±1.4	3.6±1.8	2.5±1.2	1.8±1.3
Lower Tip	3.9±1.8	2.8±1.1	2.6±1.4	1.8±1.2	2.7±1.7
Posterior pharynx	3.9±5.8	1.9±0.9	1.5±0.7	1.5±0.9	1.4±1.3
Average	4.7±4.0	3.1±2.8	2.7±2.6	2.1±1.2	2.0±1.4

Table 2. Registration errors in different non-uniformity fields

TRE (voxel)	Affine	Conventional method	Proposed method
Small bias field (20%)	3.8±1.6	3.5±2.6	2.3±1.2
Medium bias field (40%)	3.7±1.3	3.6±2.6	2.4±1.2
Large bias field (60%)	3.8±1.5	3.8±2.5	2.7±1.5

the entire volume as the sample size. We used control point spacings of 8 mm in each axis. For the 3D Harris operator, we set k=0.001 and T=50,000,000. The method stops when the movement is less than 0.001 mm or iteration reaches the predefined iteration number 100 in both methods.

The first experiment assessed the accuracy of the registration method using target registration error (TRE) [15]. Two expert observers independently selected three corresponding anatomical landmarks from each volume including tongue tip, lower lip, and posterior pharynx. Table 1 lists the mean and standard deviation of TRE and inter-observer variability using both methods. The TRE results show that the proposed method provides accurate results compared to the conventional MI-based method and is comparable to the observer-variability. Figure 3 shows one result of the first experiment. It is apparent in the figure that the proposed method has better alignment. Of note, selecting anatomical landmarks is of great importance, and a challenging task even for humans, in assessing the accuracy of the registration method. There is no true gold standard other than visual judgment, which is marred by inter-observer variability.

The second experiment further demonstrated the performance of the registration method. Three different levels of intensity non-uniformity (bias) were generated including small (20%), medium (40%) and large (60%) bias fields. In these experiments, we also used TRE to measure the performance of the methods. As shown in Table 2, the results of the proposed method were superior to the conventional method and were also robust against the bias fields.

4 Conclusion

In this work, we propose a novel registration algorithm to align hMRI with cine MRI. We utilize structural information computed from the 3D Harris operator to encode spatial and geometric cues into the computation of MI. Fully automated 3D deformable registration of hMRI with cine MRI of tongue can be performed

accurately with average error of TRE comparable to inter-observer variability. The proposed approach can be applied to the mapping of muscle anatomy in hMRI to tongue surface motions in cine MRI.

References

1. Stone, M., Davis, E., Douglas, A., NessAiver, M., Gullapalli, R., Levine, W., Lundberg, A.: Modeling the motion of the internal tongue from tagged cine-MRI images. The Journal of the Acoustical Society of America 109(6), 2974–2982 (2001)
2. Narayanan, S., Nayak, K., Lee, S., Sethy, A., Byrd, D.: An approach to real-time magnetic resonance imaging for speech production. The Journal of the Acoustical Society of America 115(4), 1771–1776 (2004)
3. Parthasarathy, V., Prince, J., Stone, M., Murano, E., NessAiver, M.: Measuring tongue motion from tagged cine-MRI using harmonic phase (HARP) processing. The Journal of the Acoustical Society of America 121(1)
4. Pluim, J., Maintz, J., Viergever, M.: Mutual-information-based registration of medical images: a survey. IEEE Trans on Medical Imaging 22(8), 986–1004 (2003)
5. Zhuang, X., Hawkes, D.J., Ourselin, S.: Unifying encoding of spatial information in mutual information for nonrigid registration. In: Prince, J.L., Pham, D.L., Myers, K.J. (eds.) IPMI 2009. LNCS, vol. 5636, pp. 491–502. Springer, Heidelberg (2009)
6. Pluim, J.P.W., Maintz, J.B.A., Viergever, M.A.: Image registration by maximization of combined mutual information and gradient information. In: Delp, S.L., DiGoia, A.M., Jaramaz, B. (eds.) MICCAI 2000. LNCS, vol. 1935, pp. 452–461. Springer, Heidelberg (2000)
7. Yi, Z., Soatto, S.: Nonrigid registration combining global and local statistics. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2200–2207 (2009)
8. Rueckert, D., Clarkson, M., Hill, D., Hawkes, D.: Non-rigid registration using higher-order mutual information. In: Proceedings of SPIE, vol. 3979, p. 438 (2000)
9. Russakoff, D.B., Tomasi, C., Rohlfing, T., Maurer Jr., C.R.: Image similarity using mutual information of regions. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3023, pp. 596–607. Springer, Heidelberg (2004)
10. Loeckx, D., Slagmolen, P., Maes, F., Vandermeulen, D., Suetens, P.: Nonrigid image registration using conditional mutual information. IEEE Trans on Medical Imaging 29(1), 19–29 (2009)
11. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference, Manchester, UK, pp. 147–151 (1988)
12. Donoser, M., Urschler, M., Hirzer, M., Bischof, H.: Saliency driven total variation segmentation. In: CVPR, pp. 817–824 (2010)
13. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J.: Nonrigid registration using free-form deformations: Application to breast mr images. IEEE Trans. Medical Imaging 18(8), 712–721 (1999)
14. Spall, J.: An overview of the simultaneous perturbation method for efficient optimization. Johns Hopkins APL Technical Digest 19(4), 482–492 (1998)
15. Fitzpatrick, J.M., West, J.B., Maurer, C.R.: Predicting error in rigid-body point-based registration. IEEE Trans. Medical Imaging 17, 694–702 (1998)