CHAPTER

11

Scalable High Performance Image Registration Framework by Unsupervised Deep Feature Representations Learning

Shaoyu Wang[*,†], **Minjeong Kim**[*], **Guorong Wu**[*], **Dinggang Shen**[*]
University of North Carolina at Chapel Hill, Chapel Hill, NC, United States[*] *Donghua University, Shanghai, China*[†]

CHAPTER OUTLINE

Deep Learning for Medical Image Analysis. DOI: 10.1016/B978-0-12-810408-8.00015-8

11.1 INTRODUCTION

Deformable image registration is defined as the process of establishing anatomical correspondences between two medical images, which is a fundamental task in medical image processing [5–12]. To name a few among its most important applications, image registration is widely used in (a) population studies toward delineating the group difference [14,15], (b) longitudinal projects for measuring the evolution of structure changes over time [1–4,16], and (c) multi-modality fusion to facilitate the diagnosis [17,18].

Challenges of developing the state-of-the-art image registration method. In general, there are two main challenges.

1. *Lack of discriminative feature representation to determine anatomical correspondences*. Image intensity is widely used to reveal the anatomical correspondences [19–24], however, two image patches that show similar, or even the same, distribution of intensity values do not guarantee that the two points correspond from an anatomical point of view [25–28]. Handcrafted features, such as geometric moment invariants [11] or Gabor filters [29], are also widely used by many state-of-the-art image registration methods [24,26–28,30]. In general, the major pitfall of using those handcrafted features is that the developed model tends to be ad-hoc. In other words, the model is only intended to recognize image patches specific to an image modality or a certain imaging application [31]. Supervised learning-based methods have been proposed to select the best set of features from a large feature pool that may include plenty of redundant handcrafted features [31–33]. However, for this approach, the ground-truth data with known correspondences across the set of training images is required. In general, image registration methods that use supervised learning for feature selection are biased and aggravate the risk of over-fitting the learning process by either unintentionally removing many relevant features, or selecting irrelevant features due to lack of ground truth.
2. *Lack of fast development framework of image registration method for specific medical image application*. Typically, it takes months, or even years, to develop a new image registration method that has acceptable performance for a new imaging modality or a new imaging application. As shown in the top of Figs. 11.1, 7.0 T images have more texture patterns than 1.5 T images since the high signal-to-noise ratio in image acquisition [34]. Thus, it is not as straightforward to use feature representations developed for 1.5 T image in the image registration for 7.0 T images. Another example is infant image registration. Although tons of image registration methods have been developed for adult brain images, early brain development studies still lack efficient image registration methods since image appearance of infant MR brain images change dramatically from 2-week-old to 2-year-old, due to myelination [35]. Rapid development of scalable high performance image registration methods that are applicable for new modalities and new applications becomes urgent in neuroscience and clinical investigations.

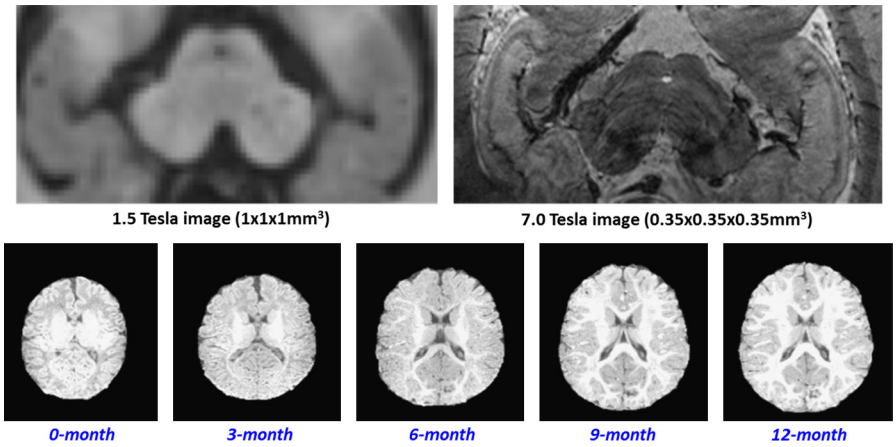

FIGURE 11.1

Challenging brain image registration cases (7.0-T MR images and infant brain images) that are in great need of distinctive feature representations.

The bottleneck, in the above challenges, is due to the insufficient number of training samples with ground truth. To overcome this difficulty, unsupervised learning-based feature selection methods are worthwhile to investigate. Because of the complexity of the data, the conventional unsupervised approaches that use simple linear models, such as PCA and ICA [36,37], are typically not suitable because they are unable to preserve the highly nonlinear relationships, when projected to low-dimensional space. More advanced methods, such as ISOMAP [38], kernel SVM [39,40], locally linear embedding (LLE) [41,42], and sparse coding [43], can learn the nonlinear embedding directly from the observed data within a single layer of projection. However, since the learned features are found using only a single layer, or a *shallow model*, the selected features may lack high-level perception knowledge (e.g., shape and context information), and may not be suitable for correspondence detection. To model complex systems with nonlinear relationships and learn high-level representations of features, unsupervised deep learning has been successfully applied to feature selection in many nontrivial computer vision problems, where deep models significantly outperform shallow models [13,40,44–52].

In this book chapter, we introduce how to learn the hierarchical feature representations directly from the observed medical images by using unsupervised deep learning paradigm. Specifically, we describe a stacked auto-encoder (SAE) [13,45,47,51] with convolutional network architecture [50,53] into our unsupervised learning framework. The inputs to train the convolutional SAE are 3D image patches. Generally speaking, our learning-based framework consists of two components, i.e., the encoder and decoder networks. On the one hand, the multi-layer encoder network is used to transfer the high-dimensional 3D image patches into the low-dimensional

feature representations, where a single auto-encoder is the building block to learn nonlinear and high-order correlations between two feature representation layers. On the other hand, the decoder network is used to recover 3D image patches from the learned low-dimensional feature representations by acting as feedback to refine inferences in the encoder network. Since the size of 3D image patches can be as large as $\sim 10^4$, it is computationally expensive to directly use an SAE to learn useful features in each layer. To overcome this problem, we use a convolutional network [50] to efficiently learn the translational invariant feature representations [50,54] such that the learned features are shared among all image points in a certain region. Finally, we present a general image registration framework with high accuracy by allowing the learned feature representations to steer the correspondence detection between two images.

To assess the performance of the proposed registration framework, we evaluate its registration performance on the conventional 1.5 T MR brain images (i.e., the elderly brains from ADNI dataset and the young brains from LONI dataset [55]), which show that the proposed registration framework achieves much better performance than exiting state-of-the-art registration methods with handcrafted features. We also demonstrate the scalability of the proposed registration framework on 7.0 T MR brain images, where we develop an image registration method for the new modality with satisfactory registration results.

The remaining sections are organized as follows: In Section 11.2, we first present the deep learning approach for extracting hierarchical feature representations and the new learning-based registration framework based on the deep learning feature. In Section 11.3, we evaluate our proposed registration framework by comparing with the conventional registration methods, and we provide a brief conclusion in Section 11.4.

11.2 PROPOSED METHOD

11.2.1 RELATED RESEARCH

Unsupervised learning is an important research topic in machine learning. For example, principle component analysis (PCA) [56] and sparse representation [57] can be used to learn intrinsic feature representations in computer vision and medical image applications. In the following, we will brief on several typical unsupervised learning methods and investigate deep learning as the reliable feature representations in medical imaging scenario.

11.2.1.1 Linear vs. Nonlinear Model

K-means [58] and Gaussian mixture model (GMM) [59] are two well-known clustering methods based upon linear learning models. In particular, given a set of training data $X_{L\times M}$, where L is the dimension of the data and M is the number of samples, the clustering methods learn K centroids such that each sample can be assigned to the closest centroid. Suppose the observed feature vectors (such as image patches and SIFT [60]) form a feature space and the appropriate K centroids in

the high-dimensional feature space are known. A typical pipeline defines a function $f : \mathcal{R}^L \rightarrow \mathcal{R}^K$ that maps the observed L-dimensional feature vector to a K-dimensional feature vector ($K < L$) [61]. For instance, we first calculate the affiliations for each observed feature vector (w.r.t. the K centroids) and then use such affiliations as morphological signatures to represent each key point in the feature space. However, the limitation of K-means and GMM is that the number of centroids is required to be larger as the input dimension grows. Thus, these clustering-based methods may not be applicable in learning the intrinsic representations for high-dimensional medical images.

PCA [56] is one of the most commonly used methods for dimension reduction. PCA extracts a set of basis vectors from the observed data, which maximizes the data variance of the projected subspace (spanned by the basis vectors). These basis vectors are obtained by calculating the eigenvectors of the covariance matrix of the input data. Given the observed data $X = [x_1, \ldots, x_m, \ldots, x_M]$, the following steps are sequentially applied in the training stage: (i) calculate the mean by $\hat{x} = \frac{1}{M}\sum_{m=1}^{M} x_m$; (ii) compute the eigenvectors $E = [e_j]_{j=1,\ldots,L}$ for the covariance matrix $\frac{1}{M-1}\overline{X}\,\overline{X}^T$, where $\overline{X} = [x_m - \hat{x}]_{m=1,\ldots,M}$ and E are sorted in the decreasing order of eigenvalues; (iii) determine the first Q largest eigenvalues such that $\sum_{j=1}^{Q}(\lambda_j)^2 > f_Q \sum_{j=1}^{L}(\lambda_j)^2$, where f_Q defines the proportion of the remaining energy. In the end, each training data x_m can be approximately reconstructed as $x_m = \hat{x} + E_Q b$, where E_Q contains the first Q largest eigenvectors of E and $b = E_Q^T(x - \hat{x})$. In the testing stage, given the new testing data x_{new}, its low-dimensional feature representation can be obtained by $b_{new} = E_Q^T(x_{new} - \hat{x})$. This classic approach for finding low-dimensional representations has achieved many successes in medical image analysis area [62,63]. However, PCA is only an orthogonal linear transform and is not optimal for finding structures with highly non-Gaussian distributions. As shown in the experiment, such feature representations learned by PCA can hardly assist image registration to establish more accurate correspondences in feature matching.

Since there are huge variations in anatomical structures across individuals, the above linear models might have limitations in finding intrinsic feature representations [40,44,64]. A more flexible representation can be obtained by learning nonlinear embedding of features. To name a few of them, ISOMAP [38], local linear embedding [41,42], and sparse dictionary learning [49,64,65] have been extensively investigated in several medical imaging applications. In many of these applications, such nonlinear methods show much more power than their simpler linear counterpart versions [34,66–68]. However, as pointed next, the depth of inferring model is another important factor in learning optimal feature representations. Unfortunately, those learning models can only infer low-level feature representations within a single layer. Thus, the learned features may lack high-level perception knowledge, such as shape and contextual information.

11.2.1.2 Shallow vs. Deep Model

As stated above, the morphological patterns in various medical images are very complex. To describe complex anatomical structure at each image point, it is necessary to use multi-layer representations to hierarchically capture the image information from different perspectives. Therefore, many references encourage using multi-resolution frameworks, since it is simple and efficient to obtain hierarchical feature representations. For instance, multi-resolution histogram is used in [69] as the morphological signature to drive the correspondence detection in brain MR image registration. Most of these feature learning methods (including PCA, decision trees, and SVMs) use shallow models, including neural networks with only one hidden layer. Theoretical results show that the internal feature representations learned by these methods are simple and incapable of extracting intrinsic types of complex structure from medical images. Training these shallow models also requires large amounts of labeled data. Thus, the development of a learning model, which has deep architecture that can discover more abstract feature representations, is of crucial importance.

In studies on visual cortex, the brain is found to have a deep architecture consisting of several layers. In visual activities, signals flow from one brain layer to the next to obtain different levels of abstraction. By simulating the function of the deep architecture of the brain, Hinton et al. [13] introduced a deep learning model, which was composed of several layers of nonlinear transformations to stack multiple layers on top of each other to discover more abstract feature representations in higher layers. Compared with shallow models that learn feature representations in a single layer, deep learning can encode multi-level information from simple to complex. To this end, deep learning is more powerful for learning hierarchical feature representations directly from high-dimensional medical images. Among various deep learning models, we propose an intrinsic feature representation mechanism based on convolutional stacked auto-encoder [13,48,50,51,54] to learn the features from 3D image patches, as described below.

11.2.2 LEARN INTRINSIC FEATURE REPRESENTATIONS BY UNSUPERVISED DEEP LEARNING

11.2.2.1 Auto-Encoder

As one typical neural network, three sequential layers structurally define the auto-encoder (AE): the input layer, the hidden layer, and the output layer (as shown in Fig. 11.2). In existing neural network algorithms, the labeled data were required as training data that are essential to the backpropagation-based fine-tuning pass, as those labels were used to readjust the parameters. However, AE, developed by Hinton and Rumelhart [70], performs backpropagation by setting the output equal to the input, and thus is trained to minimize the discrepancy between the actual output and the expected output, where the expected output is the same as the input. Hinton et al. defined the AE as a nonlinear generalization of PCA, which uses an adaptive, multilayer "encoder" network to transform the high-dimensional input data into a low-dimensional code, while a similar "decoder" network is used to recover the data from the code

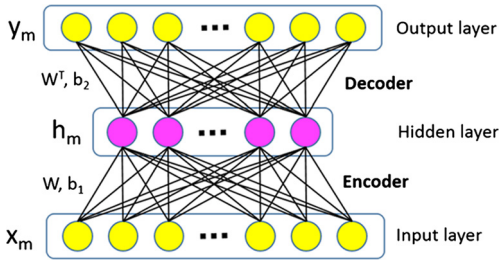

FIGURE 11.2

The architecture of an AE.

[13]. As a result, AE is able to learn compressed feature representations of the input data, without supervision by constraining the number of hidden nodes. Furthermore, sparsity is a useful constraint when the number of hidden nodes is larger than the number of input values that can discover the intrinsic features of the data. As a result of the sparse constraint, few hidden nodes can be active most of the time, and even with many hidden nodes, the network is able to learn important features of the data to reconstruct it at output layer.

Here, the goal of AE is to learn the latent feature representations from the 3D image patches collected from images. Let D and L respectively denote the dimensions of hidden representations and input patches. Given an input image patch $x_m \in \mathcal{R}^L$ $(m = 1, \ldots, M)$, AE maps an activation vector $h_m = [h_m(j)]_{j=1,\ldots,D}^T \in \mathcal{R}^D$ by $h_m = f(Wx_m + b_1)$, where the weight matrix $W \in \mathcal{R}^{D\times L}$ and the bias vector $b_1 \in \mathcal{R}^D$ are encoder parameters. Here, f is the logistic sigmoid function $f(a) = 1/(1+\exp(-a))$. It is worth noting that h_m is considered as the representation vector of the particular input training patch x_m via AE. Next, the representation h_m from the hidden layer is decoded to a vector $y_m \in \mathcal{R}^L$, which approximately reconstructs the input image patch vector x by another deterministic mapping $y_m = f(W^T h_m + b_2) \approx x_m$, where the bias vector $b_2 \in \mathcal{R}^L$ contains the decoder parameters. Therefore, the energy function in AE can be formulated as

$$\{W, b_1, b_2\} = \arg\min_{W,b_1,b_2} \sum_{m=1}^{M} \left\| f\left(W^T\left(f(Wx_m + b_1)\right)\right) + b_2 - x_m \right\|_2^2. \qquad (11.1)$$

The sparsity constraint upon the hidden nodes in the network usually leads to more interpretable features. Specifically, we regard each hidden node $h_m(j)$ as being "active," if the degree of $h_m(j)$ is close to 1, or "inactive," if the degree is close to 0. Thus, the sparsity constraint requires most of the hidden nodes to remain "inactive" for each training patch x_m. The Kullback–Leibler divergence is used to impose the sparsity constraint to each hidden node by enforcing the average activation degree over the whole training data, i.e., $\hat{\rho}_j = \sum_{m=1}^{M} h_m(j)$, to be close to a very small

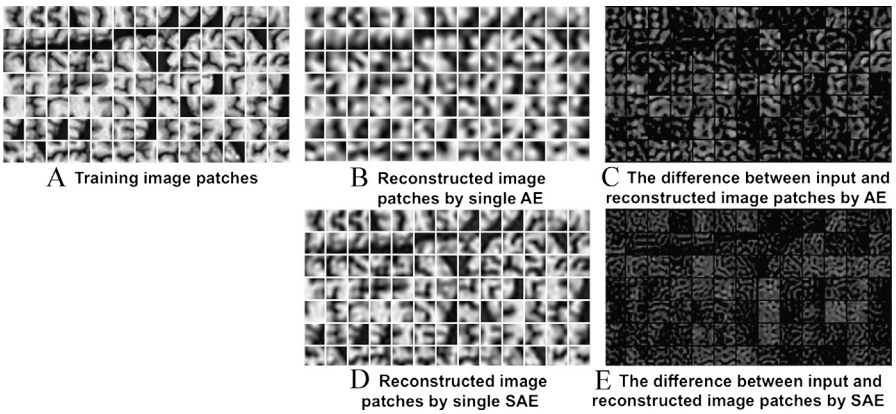

FIGURE 11.3

The reconstructed image patches by single auto-encoder (B) and stacked auto-encoder (D).

value ρ (ρ is set to 0.001 in the experiments):

$$KL\left(\rho \| \hat{\rho}_j\right) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j}. \tag{11.2}$$

Then, the overall energy function of AE with sparsity constraint is defined as

$$\begin{aligned} \{W, b_1, b_2\} = \underset{W, b_1, b_2}{\operatorname{argmin}} & \sum_{m=1}^{M} \left\| f\left(W^T\left(f(Wx_m + b_1)\right)\right) + b_2 - x_m \right\|_2^2 \\ & + \beta \sum_{j=1}^{D} KL\left(\rho \| \hat{\rho}_j\right), \end{aligned} \tag{11.3}$$

where β controls the strength of sparsity penalty term. Typical gradient based back-propagation algorithm can be used for training single AE [44,45].

11.2.2.2 Stacked Auto-Encoder

A single AE is limited in what it can present, since it is a shallow learning model. As shown in Fig. 11.3A, a set of training image patches are sampled from brain MR images, each sized at 15×15 (for demonstration, we use 2D patches as examples). We set the number of hidden nodes to 100 in this single AE. The reconstructed image patches are shown in Fig. 11.3B. It is obvious that many details have been lost after reconstruction from low-dimensional representations, as displayed in Fig. 11.3C.

The power of deep learning emerges when several AEs are stacked to form a stacked auto-encoder (SAE), where each AE becomes a building block in the deep learning model. In order to train the SAE model, we use greedy layer-wise learning

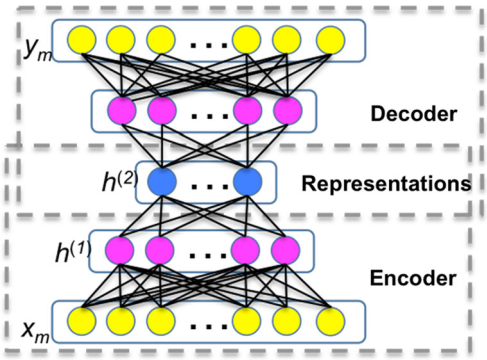

FIGURE 11.4

The hierarchical architecture of stacked auto-encoder (SAE).

[46,47] to train a single AE at each time. Specifically, there are three steps in training SAE, i.e., (i) pre-training, (ii) unrolling, and (iii) fine-tuning [13]. In the pre-training step, we train the first AE with all image patches as the input. Then, we train the second AE by using the activations $h^{(1)}$ of the first AE (pink circles in Fig. 11.4) as the input. In this way, each layer of the features captures strong, high-order correlations based on outputs from the layer below. This layer-by-layer learning can be repeated many times. After pre-training, we build a deep learning network by stacking the AE in each layer, with the higher layer AE nested within the lower layer AE. Fig. 11.4 shows an SAE consisting of 2-layer stacked AEs. Since the layer-by-layer pre-training procedure provides a good initialization for the multi-level network, we can efficiently use the gradient-based optimization method (e.g., L-BFGS or Conjugate Gradient [71]) to further refine the parameters in the fine-tuning stage. Due to the deep and hierarchical nature of the network structure, SAE can discover highly nonlinear and complex feature representations for patches in medical images. As shown in Figs. 11.3D and 11.3E, the patch reconstruction performance by SAE is much better than using a single AE, where the SAE consists of only 2 layers and the numbers of hidden nodes in each layer are 255 and 100, respectively.

11.2.2.3 Convolutional SAE Network (CSAE)

Due to the complex nature of medical images, learning the latent feature representations by employing deep learning is much more difficult than similar applications in computer vision and machine learning areas. In particular, the dimension of input training patch is often very high. For example, the intensity vector of a 3D image patch with the size of $21 \times 21 \times 21$ has 9261 elements. Thus, the training of SAE network becomes very challenging. To alleviate this issue, we resort to CSAE, which is based on efficient convolution layers to both model the locality of the pixel correlations as shared features and significantly increase the computational performance.

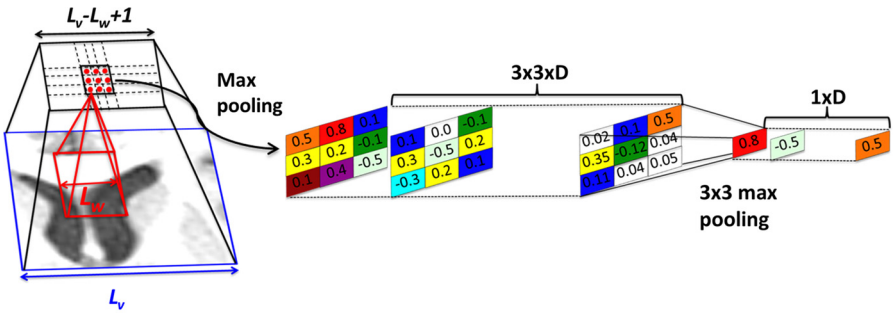

FIGURE 11.5

The 3 × 3 max pooling procedure in convolutional network.

CSAE differs from SAE as it uses convolution to take advantage of the locality of image features, and shares its weights among locations in the image patches. The convolution of an $m \times m$ matrix with an $n \times n$ window may in fact result in an $(m+n-1) \times (m+n-1)$ matrix (full convolution) or in an $(m-n+1) \times (m-n+1)$ (valid convolution). The convolution filter strides over the entire 3D image patch without overlap. The process of successive filtration finally can make a much less noisy representation of the input 3D image. To make the feature representation more spatially invariant and reduce the numbers of dimensions, we use a max pooling technique to down-sample the convolutional representation by a constant factor to take the maximum value over non-overlapping sub-regions. As shown in Fig. 11.5, the input to the convolutional SAE network is the large image patch $\mathcal{P}_v$ with patch size L_v. To make it simple, we explain the convolutional SAE network with 2D image patch as the example. Since the dimension of the image patch $\mathcal{P}_v$ is too large, we let an $L_w \times L_w$ ($L_w < L_v$) sliding window $\mathcal{P}_w$ (red box in Fig. 11.5) go through the entire large image patch $\mathcal{P}_v$, thus obtaining $(L_v-L_w+1) \times (L_v-L_w+1)$ small image patches. Eventually, we use these small image patches $\mathcal{P}_w$ to train the auto-encoder in each layer, instead of the entire image patch $\mathcal{P}_v$. Given the parameters of network (weight matrix W and bias vector b_1 and b_2), we can compute $(L_v - L_w + 1) \times (L_v - L_w + 1)$ activation vectors, where we use the red dots in Fig. 11.5 to denote the activation vectors in a 3×3 neighborhood. Then, the max pooling is used to shrink the representations by a factor of C in each direction (horizontal or vertical). The right part of Fig. 11.5 demonstrates the 3×3 max pooling procedure ($C = 3$). Specifically, we compute the representative activation vector among these 9 activation vectors in the 3×3 neighborhood by choosing the maximum absolute value for each vector element. Thus, the number of activation vector significantly reduces to $\frac{L_v-L_w+1}{C} \times \frac{L_v-L_w+1}{C}$. Since we apply the maximum operation, shrinking the representation with max pooling allows high-level representation to be invariant to small translations of the input image patches and reduces the computational burden. This translation

invariance is advantageous for establishing anatomical correspondences between two images, as is demonstrated in our experiments.

11.2.2.4 Patch Sampling

Typically, one brain MR image, with $1 \times 1 \times 1$ mm^3 spatial resolution, has over 8 million voxels in the brain volume. As a result, there are too many image patches to train the deep learning network. Therefore, adaptive sampling strategy is necessary to secure not only using enough image patches, but also selecting the most representative image patches to learn the latent feature representations for the entire training set.

To this end, there are two criteria for sampling image patches: (i) In a local view, the selected image patches should locate at distinctive regions in the image, such as sulcal roots and gyral crowns in MR brain images, since they are relatively easy to identify their correspondences; (ii) In a global view, the selected image patches should cover the entire brain volume, while the density of sampled points could be low in uniform regions and high in context-rich regions. To meet the criteria, we use the importance sampling strategy [24] to hierarchically sample the image patches. Specifically, we smooth and normalize the gradient magnitude values over the whole image domain of each training image. Then, we use the obtained values as the importance degree (or probability) of each voxel to be sampled for deep learning. Note that, although more sophisticated method [72] could be used here for guiding sample selection, we use a simple gradient guided strategy since it is computationally fast. Based on this importance (probability) map, a set of image patches can be sampled, via Monte Carlo simulation in each image. Fig. 11.6A shows the non-uniform sampling based on the importance (probability) map in learning the intrinsic feature representations for MR brain images. It can be observed from Fig. 11.6 that the sampled image patches (with the center point of each sampled patch denoted by the red dot in Fig. 11.6B) are more concentrated at the context-rich (or edge-rich) regions, where the values of importance (or probability) are high.

11.2.3 REGISTRATION USING LEARNED FEATURE REPRESENTATIONS

11.2.3.1 Advantages of Feature Representations Learned by Deep Learning Network

The advantage of using deep learning-based features in registration is demonstrated in Fig. 11.7. A typical image registration result for the elderly brain images is shown in the top of Fig. 11.7, where the deformed subject image (Fig. 11.7C) is way of being well registered with the template image (Fig. 11.7A), especially for ventricles. Obviously, it is very difficult to learn meaningful features given the inaccurate correspondences derived from imperfect image registration, as suffered by many supervised learning methods.

The discriminative power of our learned features is shown in Fig. 11.7F. For a template point (indicated by the red cross in Fig. 11.7A), we can successfully find its corresponding point in the subject image, whose ventricle is significantly larger.

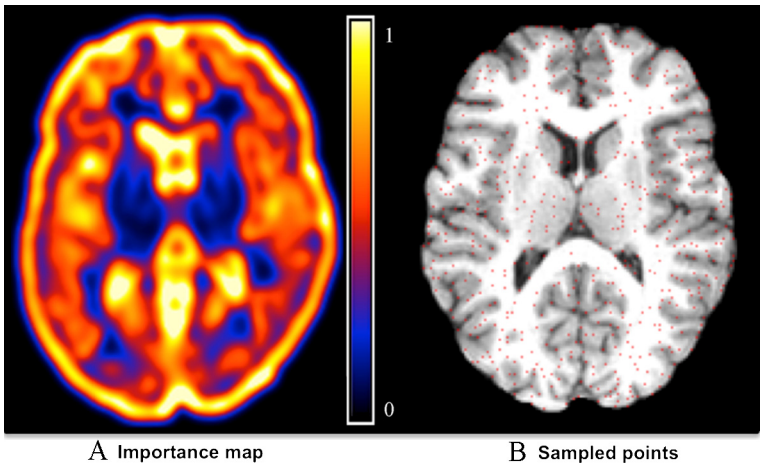

FIGURE 11.6

The importance map and the sampled image patches (denoted by the red dots) for deep learning. The color bar indicates the varying importance values for individual voxels.

Other handcrafted features either detect too many non-corresponding points (when using the entire intensity patch as the feature vector as shown in Fig. 11.7D) or have too low responses, and thus miss the correspondence (when using SIFT features as shown in Fig. 11.7E). In general, our method reveals the least confusing correspondence information for the subject point under consideration, and implies the best correspondence detection performance.

11.2.3.2 *Learning-Based Image Registration Framework*

After training the convolutional SAE on a large amount of 3D image patches, it is efficient to obtain the low-dimensional feature representations (blue circles in Fig. 11.4) by simple matrix multiplication and addition in each encoder layer. Such low-dimensional feature representation, regarded as the morphological signature, allows each point to accurately identify the correspondence during image registration, as demonstrated above. Here, we use normalized cross correlation as the similarity measurement between the feature representation vectors of the two different points under comparison.

Since the convolutional SAE can directly learn feature representations from the observed data, the learning procedure can be free of the limitation of requiring ground-truth data. Thus, it is straightforward to learn optimal feature representations for specific dataset, with little or even no human intervention. Then, we can incorporate the state-of-the-art deformation mechanisms developed in many registration methods in a learning-based registration framework by replacing with the learned feature representations and still inheriting the existing deformation mechanism to derive the deformation pathway.

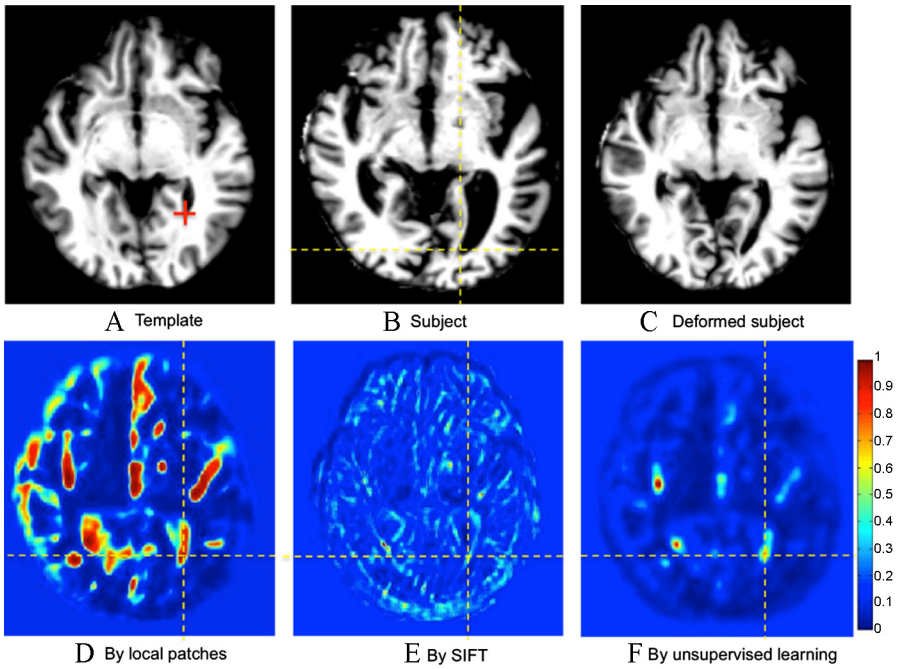

FIGURE 11.7

The similarity maps of identifying the correspondence of the red-crossed point in the template (A) w.r.t. the subject (B) by handcraft features (D)–(E) and the learned features by unsupervised deep learning (F).

Without loss of generalization, we show two examples by integrating the feature representations learned via deep learning using state-of-the-art registration methods. First, many image registration methods, e.g., diffeomorphic Demons [23], use the gradient-based optimization approach to iteratively estimate the deformation fields. To utilize multiple image information, such as multi-modality images, a multi-channel Demons algorithm was proposed by allowing one channel carrying one information source [73–75]. Therefore, it is straightforward to deploy multi-channel Demons, by regarding all elements of the learned feature representations as multiple channels.

Second, we can replace the handcrafted attribute vectors (i.e., the low-order statistics in multi-resolution histograms) in the feature-based registration method, e.g., symmetric HAMMER [27], with the learned feature representations by the convolutional SAE. Specifically, after training the convolution SAE upon large amount of 3D image patches, it is efficient to obtain the low-dimensional feature representations (blue circles in Fig. 11.2 and Fig. 11.4) by simple matrix multiplication and addition in each encoder layer. Such low-dimensional feature representation, regarded as the morphological signature, allows each point to accurately identify the correspondence

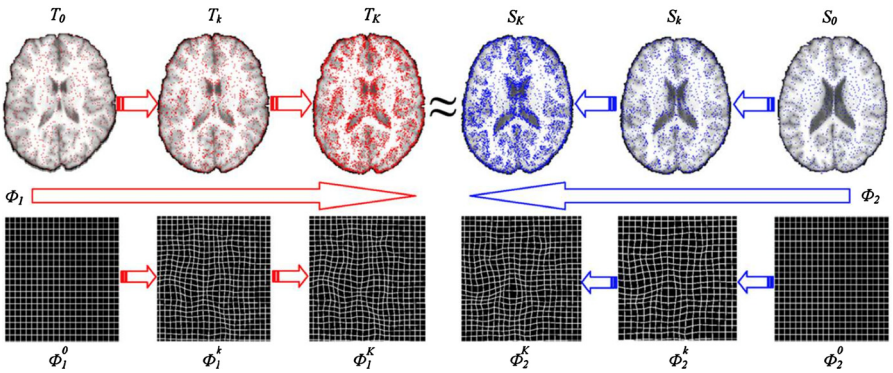

FIGURE 11.8

The demonstration of hierarchical feature matching mechanism with symmetric estimation of deformation pathway.

by robust feature matching [28,76]. Next, we simultaneously deform template and subject images toward each other until they meet at the middle point. Thus, we can obtain two deformation pathways, ϕ_1 from template T, and ϕ_2 from subject S, as denoted by red and blue arrows in Fig. 11.8, respectively. In the end, the deformation field from template to subject can be calculated by $F = \phi_1 \circ \phi_2^{-1}$, where $\circ$ denotes the composition of two deformation pathways [20]. To further improve the correspondence detection, a small number of key points (denoted in red and blue for template and subject in Fig. 11.8, respectively) with more distinctive features than other image points will be selected to drive the entire deformation field by requiring deformation on less distinctive points to follow the correspondences of nearby key points. The key points are hierarchically selected during registration by using non-uniform sampling strategy, which ensures that most key points are located at salient regions and also cover the whole image. Since ϕ_1 and ϕ_2 are iteratively refined during registration, we use k to denote the iteration. Thus, in the beginning of registration ($k = 0$), $T_0 = T$ and $S_0 = S$, along with identity deformation pathway ϕ_1^0 and ϕ_2^0. With the progress of registration, the template image T gradually deforms to $T^k = T(\phi_1^k)$. Similarly, subject image S deforms to $S^k = S(\phi_2^k)$. In the meantime, more key points are selected to refine the deformation pathways ϕ_1^k and ϕ_2^k w.r.t. T^k and S^k, which is repeated until the deformed template T^k and deformed subject S^k become very similar at the end of registration.

11.3 EXPERIMENTS

Here, we evaluate the performance of deformable image registration algorithm that uses deep learning for feature selection. For comparison, we set diffeomorphic

Demons and HAMMER as baselines for intensity-based and feature-based registration methods, respectively. Then, we extend the diffeomorphic Demons from a single channel (i.e., image intensity) to multi-channel Demons by adapting the learned feature representations via deep learning to multiple channels (M + DP). Similarly, we modify HAMMER to use the feature representations learned via deep learning (H + DP). Since PCA is widely used for unsupervised learning, we apply PCA to infer the latent low-dimensional feature representations. After integrating such low-dimensional feature representations by PCA into multi-channel Demons and HAMMER, we can obtain two other new registration methods, denoted as M + PCA and H + PCA, respectively.

11.3.1 EXPERIMENTAL RESULT ON ADNI DATASET

We randomly select 66 MR images from the ADNI dataset (http://adni.loni.ucla.edu/), where 40 images are used to learn feature representations and the other 26 images are used to test image registration. The preprocessing steps include skull removal [77], bias correction [78], and intensity normalization [79]. For each training image, we sample around 7000 image patches, where the patch size is set to $21 \times 21 \times 21$. The convolutional SAE consists of 8 layers (stacked with 4 AEs). We only apply the max pooling in the lowest layer with the pooling factor $C = 3$. From the lowest to the highest level, the numbers of hidden nodes in each stacked AE are 512, 512, 256, and 128, respectively. Thus, the dimension of final feature representations after deep learning is 128. To keep the similar dimension of learned features by PCA, we set the portion of remaining energy f_Q to 0.7 in this experiment.

In image registration, one image is selected as the template and the other 25 images are considered as subject images. Before deploying deformable image registration, FLIRT in FSL package (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/) is used to linearly align all subject images to the template image. After that, we apply 6 registration methods, i.e., diffeomorphic Demons (simply named as Demons below), M + PCA, M + DP, HAMMER, H + PCA, and H + DP, to normalize those 25 subject images to the template image space, respectively. To quantitatively evaluate the registration accuracy, we first use FAST in FSL package to segment each image into white matter (WM), gray matter (GM), and cerebral-spinal fluid (CSF). After that, we further separate the ventricle (VN) from the CSF segmentation. Here, we use these segmentation results to evaluate the registration accuracy by comparing the Dice ratio of tissue overlap degrees between the template and each registered subject image. Specifically, the Dice ratio is defined as

$$D(R_A, R_B) = \frac{2|R_A \cap R_B|}{|R_A| + |R_B|}, \tag{11.4}$$

where R_A and R_B denote two ROIs and $|\cdot|$ stands for the volume of the region. The Dice ratios on WM, GM, and VN by 6 registration methods are shown in Table 11.1. It is clear that (i) the registration methods, integrated with the feature representations by deep learning, consistently outperform the counterpart baseline methods and also

Table 11.1 The Dice ratios of WN, GM, and VN on ADNI dataset (in %)

Method	WM	GM	VN	Overall
Demons	85.7	76.0	90.2	84.0
M + PCA	85.5	76.6	90.2	84.1
M + DP	85.8	76.5	90.9	84.4
HAMMER	85.4	75.5	91.5	84.1
H + PCA	86.5	76.9	91.7	85.0
H + DP	88.1	78.6	93.0	86.6

Table 11.2 The Dice ratio of hippocampus on ADNI dataset (in %)

Method	Demons	M + PCA	M + DP	HAMMER	H + PCA	H + DP
Dice ratio	72.2 ± 3.1	72.3 ± 2.9	72.5 ± 2.8	75.5 ± 2.9	75.6 ± 2.5	76.8 ± 2.2

use PCA-based feature representations only; (ii) H + DP achieves the highest registration accuracy with almost 2.5% improvement in overall Dice ratio, compared to the baseline HAMMER registration method. Since ADNI provides hippocampus labeling for the template and all 25 subject images, we can further evaluate the Dice overlap ratio on hippocampus. The mean and the standard deviation of the Dice ratios on hippocampus by 6 registration methods are shown in Table 11.2. Compared to the baseline methods, M + DP and H + DP obtain 0.3% and 1.3% improvements in terms of Dice ratios, respectively. Particularly, the reason of the less improvement by M + DP compared to H + DP might be related with the high number of channels (128 channels) used in M + DP, compared with only less than 10 channels used in [73–75].

11.3.2 EXPERIMENTAL RESULT ON LONI DATASET

In this experiment, we use the LONI LPBA40 dataset [80], which consists of 40 brain images, each with 56 manually labeled ROIs. We randomly use 20 images to learn the latent feature representations, and the other 20 images to test registration performance. The preprocessing procedures include bias correction, intensity normalization, and linear registration by FLIRT, which are the same as in Section 11.3.1. For each training image, we sample around 9000 image patches, where the patch size is again set to $21 \times 21 \times 21$. Other parameters in training convolutional SAE are also the same as in Section 11.3.1. Therefore, the dimension of feature representations after deep learning is 128. To keep the similar dimension of learned features by PCA, we set f_Q to 0.65 in this experiment.

One of the 20 testing images is selected as the template, and then we apply 6 registration methods to register the rest of 19 testing images to the selected template. The averaged Dice ratio in each ROI by the 6 registration methods is shown in Fig. 11.9. The overall Dice ratios across all 56 ROIs by the 6 registration methods are provided

	Demons	M+PCA	M+DP		HAMMER	H+PCA	H+DP
L sup. frontal gyrus	80.2	79.5	79.1		77.3	77.1	78.5
R sup. frontal gyrus	79.7	79.8	80.0		77.5	77.2	78.4
* L middle frontal gyrus	77.4	78.1	78.2		79.5	78.5	82.3
* R middle frontal gyrus	77.0	76.5	77.1		78.2	78.6	80.4
L inf. frontal gyrus	72.2	72.8	72.6		72.8	73.0	74.6
R inf. frontal gyrus	72.0	72.1	72.3		72.6	72.4	74.1
L precentral gyrus	67.9	68.5	68.4		68.6	69.1	71.5
R precentral gyrus	68.6	68.5	69.0		65.0	65.2	65.1
* L middle orbitofrontal gyrus	66.9	66.1	67.1		69.8	69.9	73.5
* R middle orbitofrontal gyrus	66.8	67.5	67.7		69.4	69.5	73.9
* L lateral orbitofrontal gyrus	58.1	58.1	58.5		60.5	61.5	64.9
R lateral orbitofrontal gyrus	55.4	55.6	55.7		64.7	64.1	68.9
L gyrus rectus	66.7	66.1	66.9		67.9	67.2	69.8
R gyrus rectus	68.1	67.7	68.1		65.5	65.0	65.0
L postcentral gyrus	60.5	60.7	61.4		60.5	61.2	63.0
* R postcentral gyrus	62.9	62.4	62.5		63.5	63.1	65.2
* L sup. parietal gyrus	70.7	70.9	70.6		72.6	72.7	74.7
* R sup. parietal gyrus	70.9	70.6	71.2		71.8	71.9	73.5
L supramarginal gyrus	63.8	63.4	64.1		65.1	65.5	67.8
R supramarginal gyrus	63.3	63.7	63.8		65.1	66.4	68.5
* L angular gyrus	63.2	62.8	63.5		66.9	66.8	70.0
R angular gyrus	65.0	65.1	65.7		65.4	65.8	67.5
* L precuneus	65.9	65.7	66.4		70.6	70.9	74.0
* R precuneus	67.3	67.2	67.8		70.8	71.5	76.5
L sup. occipital gyrus	58.1	58.0	58.2		61.2	62.4	65.5
R sup. occipital gyrus	55.4	55.9	56.2		64.5	65.4	67.7
* L middle occipital gyrus	68.7	68.5	68.4		72.6	74.9	80.6
R middle occipital gyrus	67.9	67.4	67.8		71.1	72.0	73.1
L inf. occipital gyrus	67.2	67.8	67.9		65.8	65.5	66.4
R inf. occipital gyrus	66.1	66.5	67.1		62.0	62.0	62.1
L cuneus	63.4	63.4	63.9		64.1	64.5	64.7
* R cuneus	62.2	62.4	62.5		66.0	67.2	70.1
L sup. temporal gyrus	72.5	72.5	72.7		69.5	69.5	70.7
* R sup. temporal gyrus	72.6	73.1	73.4		74.1	74.5	76.4
* L middle temporal gyrus	66.4	66.8	66.8		67.1	66.9	69.5
R middle temporal gyrus	67.9	67.5	67.9		68.3	68.4	69.7
L inf. temporal gyrus	65.6	65.2	65.9		65.8	65.1	66.3
* R inf. temporal gyrus	66.4	66.4	66.5		66.9	67.8	70.0
L parahippocampal gyrus	68.1	68.2	68.5		68.0	69.1	70.1
R parahippocampal gyrus	66.9	66.7	67.2		67.5	69.0	71.0
* L lingual gyrus	69.7	69.8	68.9		69.4	69.2	71.8
* R lingual gyrus	70.6	70.5	70.6		73.6	74.3	77.5
L fusiform gyrus	68.9	68.8	69.1		66.5	66.1	66.2
R fusiform gyrus	68.3	68.3	68.5		67.5	67.5	67.8
L insular cortex	76.4	76.1	76.5		77.5	77.9	79.7
R insular cortex	74.2	74.6	74.7		75.1	76.0	76.1
* L cingulate gyrus	68.1	68.2	68.8		69.5	69.9	71.4
* R cingulate gyrus	67.5	67.4	67.2		69.2	70.5	72.2
* L caudate	73.4	73.4	73.8		74.5	75.0	77.8
* R caudate	73.1	73.0	73.5		76.2	76.4	78.3
* L putamen	76.3	76.5	76.7		77.0	77.7	80.0
* R putamen	76.5	76.3	76.4		76.5	78.6	80.6
* L hippocampus	72.7	72.6	72.8		74.7	75.8	77.7
* R hippocampus	72.8	72.6	73.1		75.9	77.1	81.3
* cerebellum	84.9	85.1	85.9		86.0	87.8	90.3
* brainstem	80.6	81.4	83.1		85.5	86.6	89.1

FIGURE 11.9

The Dice ratios of 56 ROIs on LONI dataset by 6 registration methods.

Table 11.3 The overall Dice ratio of 54 ROIs on LONI dataset (in %)

Method	Demons	M + PCA	M + DP	HAMMER	H + PCA	H + DP
Dice ratio	68.9	68.9	69.2	70.2	70.6	72.7

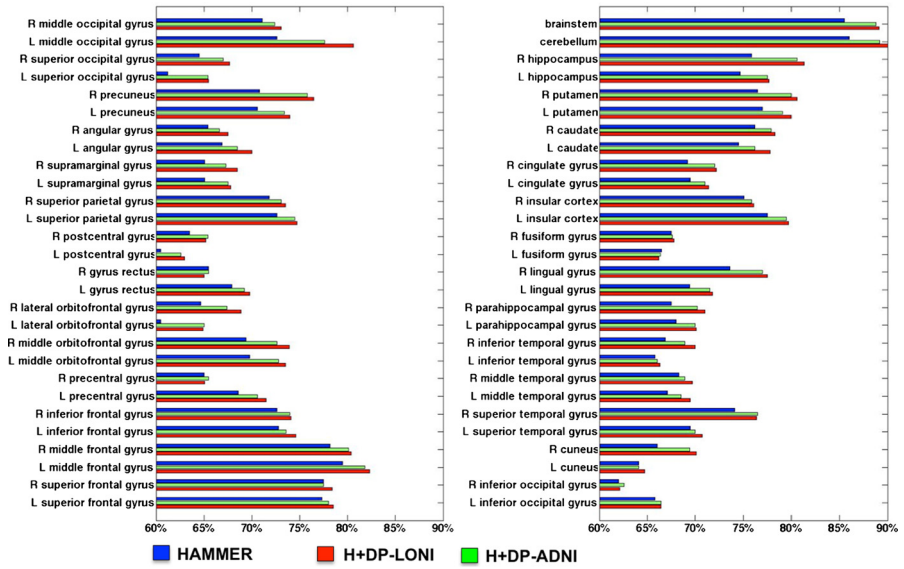

FIGURE 11.10

The Dice ratios of 56 ROIs in LONI dataset by HAMMER (blue), H + DP-LONI (red), and H + DP-ADNI (green), respectively. Note, H + DP-LONI denotes for the HAMMER registration integrating with the feature representations learned directly from LONI dataset, while H + DP-ADNI stands for applying HAMMER registration on LONI dataset but using the feature representations learned from ADNI dataset, respectively.

in Table 11.3. Again, H+DP achieves the largest improvement (2.5%) in comparison to the baseline HAMMER registration method. Specifically, we perform the paired t-test between H + DP and all other 5 registration methods, respectively. The results indicate that H + CP has the statistically significant improvement over all other 5 registration methods in 28 out of 54 ROIs (designated by the red stars in Fig. 11.9).

Recall that we have obtained the feature representations by deep learning on the ADNI dataset in Section 11.3.1. It is interesting to evaluate the generality of deep learning by applying the learned feature representations from the ADNI dataset (which mostly contains elderly brains) to register the images in the LONI dataset (i.e., young brains). Fig. 11.10 shows the Dice ratio in each ROI in the LONI dataset by (i) the baseline HAMMER registration method (in blue), (ii) H + DP-LONI (in red), which denotes we learned the feature representations from LONI dataset and integrated the learned feature representations with HAMMER, and (iii) H+DP-ADNI (in

green) where we apply the learned feature representations from ADNI dataset as the morphological signature to drive the registration on the LONI dataset. It is apparent that the registration performance by H + DP-ADNI is comparable to H + DP-LONI, where the average improvements over the baseline HAMMER registration method are 1.99% by H + DP-ADNI, and 2.55% by H + DP-LONI, respectively. It indicates that the learned feature representations by convolutional SAE network are general, although the appearances of two dataset can be quite different (i.e., due to aging).

11.3.3 EXPERIMENTAL RESULT ON 7.0-T MR IMAGE DATASET

In previous experiments, we demonstrated the power of learned feature representations by deep learning in terms of the improved registration accuracy, which is represented by an overlap ratio of structures. As mentioned earlier, another attractive advantage of deep learning is that it can rapidly learn intrinsic feature representations for a new imaging modality. In this section, we apply the convolutional SAE to 7.0 T MR brain images. The learned feature representations are then integrated with HAMMER, which allows us to develop a specific registration method for 7.0 T MR brain images.

The advent of 7.0-T MR imaging technology [81] achieves high signal-to-noise ratio (SNR), as well as the dramatically increased tissue contrast, compared to the 1.5 or 3.0 T MR image. A typical 7.0-T MR brain image (with the image spatial resolution of $0.35 \times 0.35 \times 0.35$ mm^3) is shown in Fig. 11.11B, along with a similar slice from a 1.5 T scanner (with the resolution of $1 \times 1 \times 1$ mm^3) in Fig. 11.11A for comparison. As demonstrated in [82], 7.0 T MR image can reveal brain structures with resolution equivalent to that obtained from thin slices *in vitro*. Thus, researchers are able to clearly observe the fine brain structures on the μm scale, which was only possible with *in vitro* imaging previously. Without a doubt, 7.0 T MR imaging technique has the potential to become the standard method in discovering morphological patterns of human brain in the near future.

Unfortunately, all existing state-of-the-art deformable registration methods, developed for 1.5 or 3.0 T MR images, do not work well for the 7.0 T MR images, mainly because (i) severe intensity inhomogeneity issue in 7.0 T MR images, and because (ii) much richer texture information than that in 1.5 or 3.0 T MR images, as displayed in Fig. 11.1.

Overall, 20 7.0 T MR images acquired by the method in [81] were used in this experiment, where 10 are used to train deep learning, and the other 10 images are used to test registration performance. We randomly select one image as the template. For the 7.0 T scanner (Magnetom, Siemens), an optimized multichannel radiofrequency (RF) coil and a 3D fast low-angle shot (Spoiled FLASH) sequence were utilized, with TR = 50 ms, TE = 25 ms, flip angle 10°, pixel band width 30 Hz/pixel, field of view (FOV) 200 mm, matrix size $512 \times 576 \times 60$, 3/4 partial Fourier, and number of average (NEX) 1. The image resolution of the acquired images is isotropic, e.g., $0.35 \times 0.35 \times 0.35$ mm^3. The hippocampi were manually segmented by a neurologist [81]. All images were pre-processed by the following steps: (i) inhomogeneity cor-

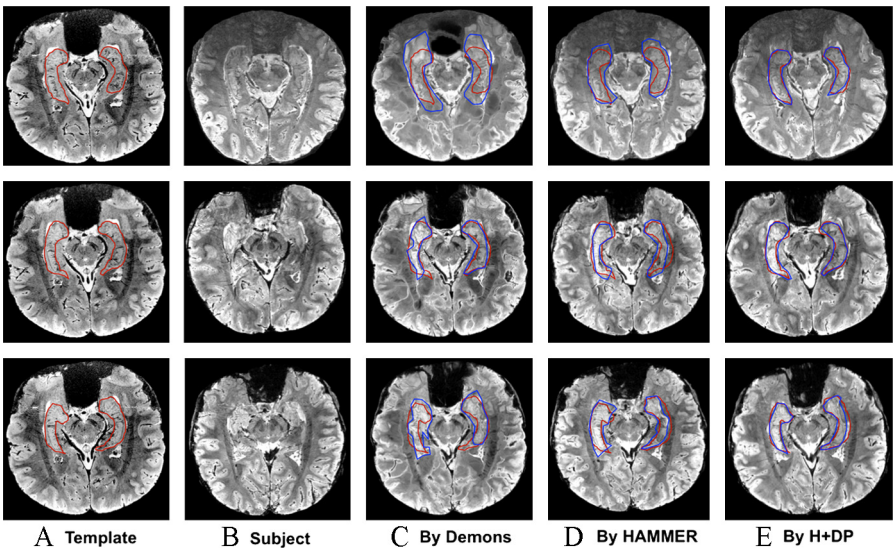

FIGURE 11.11

Typical registration results on 7.0-T MR brain images by Demons, HAMMER, and H + DP, respectively. Three rows represent three different slices in the template, subject, and registered subjects.

rection using N4 bias correction [78]; (ii) intensity normalization for making image contrast and luminance consistent across all subjects [79]; (iii) affine registration to the selected template by FSL.

For each training image, we sample around 9000 image patches, where the patch size is set to $27 \times 27 \times 27$. The convolutional SAE consists of 10 layers (stacked with 5 AEs). We only apply the max pooling in the lowest layer, with the pooling factor $C = 3$. From low level to high level, the numbers of hidden nodes in each stacked AE are 1024, 512, 512, 256, and 128, respectively. Thus, the dimension of deep learning-based feature representations after deep learning is 128. In order to achieve the best registration performance, we integrate the learned feature representations trained from 7.0 T MR images with the HAMMER registration method.

Several typical registration results on 7.0 T MR images are displayed in Fig. 11.11, where the template and subject images are shown in Figs. 11.11A and 11.11B, respectively. Here, we compare the registration results with diffeomorphic Demons (Fig. 11.11C) and HAMMER (Fig. 11.11D). The registration results by H + DP, i.e., integrating the learned feature representations by deep learning with HAMMER, are display in Fig. 11.11E, where the manually labeled hippocampus on the template image, and the deformed subject's hippocampus, by different registration methods, are shown by red and blue contours, respectively. Through visual inspection (the overlap of red and blue contours), the registration result by

H + DP is much better than both diffeomorphic Demons and HAMMER. Since we have manually labeled hippocampus for the template and all subject images, we can further quantitatively measure the registration accuracy. The mean and standard deviation of Dice ratios on hippocampus are $(53.5 \pm 4.9)\%$ by diffeomorphic Demons, $(64.9 \pm 3.1)\%$ by HAMMER, and $(75.3 \pm 1.2)\%$ by H + DP, respectively. Obviously, H + DP achieves significant improvement in registration accuracy. This experiment demonstrates that (i) the latent feature representations inferred by deep learning can well describe the local image characteristics; (ii) we can rapidly develop image registration for new medical imaging modalities by using deep learning framework to learn the intrinsic feature representations; and (iii) the whole learning-based framework is fully adaptive to learn the image data and reusable to various medical imaging applications.

11.4 CONCLUSION

In this book chapter, a new deformable image registration approach that uses deep learning for feature selection was introduced. Specifically, we proposed an unsupervised deep learning feature selection framework that implements a convolutional-stacked auto-encoder network (SAE) to identify the intrinsic features in the 3D image patches. Using the LONI and ADNI datasets, the image registration performance was compared to two existing state-of-the-art deformable image registration frameworks that use handcrafted features. The results show that the new image registration framework consistently demonstrated better Dice ratio scores, when compared to state-of-the-art methods. In short, because the trained deep learning network selected features more accurately capture the complex morphological patterns in the image patches, it is allowed to produce better anatomical correspondences, which ultimately resulted in better image registration performance.

To demonstrate the scalability of the proposed registration framework, image registration experiments were also conducted on 7.0 T brain MR images. Likewise, the results showed that the new image registration framework consistently demonstrated better Dice ratio scores, when compared to state-of-the-art methods. Unlike those existing image registration frameworks, the deep learning architecture can be quickly developed, and trained using no ground-truth data with superior performance, which allows deep learning architecture to be eventually applied to new imaging modalities with the least effort.

REFERENCES

1. T. Paus, et al., Structural maturation of neural pathways in children and adolescents: in vivo study, Science 283 (5409) (1999) 1908–1911.

2. E.R. Sowell, et al., Localizing age-related changes in brain structure between childhood and adolescence using statistical parametric mapping, NeuroImage 9 (6 (Part 1)) (1999) 587–597.
3. P.M. Thompson, et al., Growth patterns in the developing brain detected by using continuum mechanical tensor maps, Nature 404 (2000) 190–193.
4. S. Resnick, P. Maki, Effects of hormone replacement therapy on cognitive and brain aging, Ann. N.Y. Acad. Sci. 949 (2001) 203–214.
5. P.M. Thompson, et al., Tracking Alzheimer's Disease, Ann. N.Y. Acad. Sci. 1097 (1) (2007) 183–214.
6. J. Lerch, et al., Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls, Neurobiol. Aging 29 (1) (2008) 23–30.
7. N. Schuff, et al., MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers, Brain (2009) 1067–1077.
8. G. Frisoni, et al., In vivo neuropathology of the hippocampal formation in AD: a radial mapping MR-based study, NeuroImage 32 (1) (2006) 104–110.
9. L. Apostolova, et al., Conversion of mild cognitive impairment to Alzheimer disease predicted by hippocampal atrophy maps, Arch. Neurol. 64 (9) (2007) 1360–1361.
10. A.D. Leow, et al., Alzheimer's Disease Neuroimaging Initiative: a one-year follow up study using tensor-based morphometry correlating degenerative rates, biomarkers and cognition, NeuroImage 45 (3) (2009) 645–655.
11. D. Shen, C. Davatzikos, HAMMER: hierarchical attribute matching mechanism for elastic registration, IEEE Trans. Med. Imaging 21 (11) (2002) 1421–1439.
12. J. Maintz, M. Viergever, A survey of medical image registration, Med. Image Anal. 2 (1) (1998) 1–36.
13. G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
14. P. Thompson, A. Toga, A framework for computational anatomy, Comput. Vis. Sci. 5 (2002) 13–34.
15. C. Davatzikos, et al., Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy, NeuroImage 14 (6) (2001) 1361–1369.
16. Y. Li, et al., Discriminant analysis of longitudinal cortical thickness changes in Alzheimer's disease using dynamic and network features, Neurobiol. Aging 33 (2012) 415–427.
17. C. Saw, et al., Multimodality image fusion and planning and dose delivery for radiation therapy, Med. Dosim. 33 (2008) 149–155.
18. J. Cízek, et al., Fast and robust registration of PET and MR images of human brain, NeuroImage 22 (1) (2004) 434–442.
19. V. Arsigny, et al., A log-euclidean framework for statistics on diffeomorphisms, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006, 2006, pp. 924–931.
20. J. Ashburner, A fast diffeomorphic image registration algorithm, NeuroImage 38 (1) (2007) 95–113.
21. B. Avants, M. Grossman, J. Gee, Symmetric diffeomorphic image registration: evaluating automated labeling of elderly and neurodegenerative cortex and frontal lobe, in: Biomedical Image Registration, 2006.
22. D. Rueckert, et al., Nonrigid registration using free-form deformations: application to breast MR images, IEEE Trans. Med. Imaging 18 (8) (1999) 712–721.

23. T. Vercauteren, et al., Diffeomorphic demons: efficient non-parametric image registration, NeuroImage 45 (1, Suppl. 1) (2009) S61–S72.
24. Q. Wang, et al., Attribute vector guided groupwise registration, NeuroImage 50 (4) (2010) 1485–1496.
25. T. Rohlfing, Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable, IEEE Trans. Med. Imaging 31 (2) (2012) 153–160.
26. D. Shen, C. Davatzikos, HAMMER: hierarchical attribute matching mechanism for elastic registration, IEEE Trans. Med. Imaging 21 (11) (2002) 1421–1439.
27. G. Wu, et al., S-HAMMER: hierarchical attribute-guided, symmetric diffeomorphic registration for MR brain images, Hum. Brain Mapp. 35 (3) (2014) 1044–1060.
28. G. Wu, et al., TPS-HAMMER: improving HAMMER registration algorithm by soft correspondence matching and thin-plate splines based deformation interpolation, NeuroImage 49 (3) (2010) 2225–2233.
29. Y. Zhan, D. Shen, Deformable segmentation of 3-D ultrasound prostate images using statistical texture matching method, IEEE Trans. Image Process. 25 (3) (2006) 256–272.
30. G. Wu, et al., Feature-based groupwise registration by hierarchical anatomical correspondence detection, Hum. Brain Mapp. 33 (2) (2012) 253–271.
31. G. Wu, F. Qi, D. Shen, Learning-based deformable registration of MR brain images, IEEE Trans. Med. Imaging 25 (9) (2006) 1145–1157.
32. G. Wu, F. Qi, D. Shen, Learning best features and deformation statistics for hierarchical registration of MR brain images, in: Information Processing in Medical Imaging, 2007, pp. 160–171.
33. Y. Ou, et al., DRAMMS: deformable registration via attribute matching and mutual-saliency weighting, Med. Image Anal. 15 (4) (2011) 622–639.
34. M. Kim, et al., Automatic hippocampus segmentation of 7.0 tesla MR images by combining multiple atlases and auto-context models, NeuroImage 83 (2013) 335–345.
35. R.C. Knickmeyer, et al., A structural MRI study of human brain development from birth to 2 years, J. Neurosci. 28 (47) (2008) 12176–12182.
36. P. Comon, Independent component analysis, a new concept?, Signal Process. 36 (3) (1994) 287–314.
37. A. Hyvarinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.
38. J.B. Tenenbaum, V. Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 640–646.
39. E.-N. I, et al., A support vector machine approach for detection of microcalcifications, IEEE Trans. Med. Imaging 21 (12) (2002) 1552–1563.
40. H. Larochelle, et al., An empirical evaluation of deep architectures on problems with many factors of variation, in: Proceedings of the 24th International Conference on Machine Learning, Corvalis, Oregon, ACM, 2007, pp. 473–480.
41. S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323.
42. M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.
43. B.A. Olshausen, D.J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature 381 (1996) 607–609.
44. L. Arnold, et al., An introduction to deep-learning, in: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN, 2011.

45. Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, arXiv:1206.5538v3 [cs.LG], 23 Apr 2014.
46. Y. Bengio, et al., Greedy layer-wise training of deep networks, in: Advances in Neural Information Processing Systems, NIPS, 2006.
47. G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
48. Q.V. Le, et al., Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2011.
49. H. Lee, C. Ekanadham, A.Y. Ng, Sparse deep belief net model for visual area V2, in: Advances in Neural Information Processing Systems, NIPS, 2008.
50. H. Lee, et al., Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, in: Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, Quebec, Canada, ACM, 2009, pp. 609–616.
51. H. Lee, et al., Unsupervised learning of hierarchical representations with convolutional deep belief networks, Commun. ACM 54 (10) (2011) 95–103.
52. Y.F. Li, C.Y. Chen, W.W. Wasserman, Deep feature selection: theory and application to identify enhancers and promoters, in: Recomb 2015, Res. Comput. Mol. Biol. 9029 (2015) 205–217.
53. Y. LeCun, Y. Bengio, Convolutional network for images, speech, and time series, in: The Handbook of Brain Theory and Neural Networks, 1995.
54. H.-C. Shin, et al., Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1930–1943.
55. D.W. Shattuck, et al., Construction of a 3D probabilistic atlas of human cortical structures, NeuroImage 39 (3) (2008) 1064–1080.
56. S. Mika, et al., Kernel PCA and denoising in feature space, Adv. Neural Inf. Process. Syst. 11 (1) (1999) 536–542.
57. A. Hyvarinen, P. Hoyer, Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces, Neural Comput. 12 (7) (2000) 1705–1720.
58. J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceeding of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1967, pp. 281–297.
59. D.M. Titterington, A.F.M. Smith, U.E. Makov, Statistical Analysis of Finite Mixture Distributions, John Wiley & Sons, 1985.
60. K. Mikolajczyk, C. Schmid, A performance evaluation of local descriptors, IEEE Trans. Pattern Anal. Mach. Intell. 27 (10) (2005) 1615–1630.
61. A. Coates, A.Y. Ng, Learning feature representations with K-means, in: Neural Networks: Tricks of the Trade, in: Lect. Notes Comput. Sci., Springer, 2012.
62. T.F. Cootes, et al., Active shape models-their training and application, Comput. Vis. Image Underst. 16 (1) (1995) 38–159.
63. T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 681–685.
64. M.A. Ranzato, et al., Efficient learning of sparse representations with an energy-based model, in: Advances in Neural Information Processing Systems, NIPS, 2006.
65. R. Raina, et al., Self-taught learning: transfer learning from unlabeled data, in: 24th International Conference on Machine Learning, Corvalis, OR, 2007.

66. J. Hamm, et al., GRAM: a framework for geodesic registration on anatomical manifolds, Med. Image Anal. 14 (5) (2010) 633–642.
67. R. Wolz, et al., LEAP: learning embeddings for atlas propagation, NeuroImage 49 (2) (2010) 1316–1325.
68. R. Wolz, et al., Manifold learning for biomarker discovery, in: MICCAI Workshop on Machine Learning in Medical Imaging, Beijing, China, 2010.
69. D. Shen, Image registration by local histogram matching, Pattern Recognit. 40 (2007) 1161–1171.
70. D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (6088) (1986) 533–536.
71. Q.V. Le, et al., On optimization methods for deep learning, in: ICML, 2011.
72. T. Kadir, M. Brady, Saliency, scale and image description, Int. J. Comput. Vis. 45 (2) (2001) 83–105.
73. J.M. Peyrat, et al., Registration of 4D time-series of cardiac images with multichannel Diffeomorphic Demons, Med. Image Comput. Comput. Assist. Interv. 11 (Pt 2) (2008) 972–979.
74. J.M. Peyrat, et al., Registration of 4D cardiac CT sequences under trajectory constraints with multichannel diffeomorphic demons, IEEE Trans. Med. Imaging 29 (7) (2010) 1351–1368.
75. D. Forsberg, et al., Improving registration using multi-channel diffeomorphic demons combined with certainty maps, in: Multimodal Brain Image Registration, in: Lect. Notes Comput. Sci., vol. 7012, 2011.
76. H. Chui, A. Rangarajan, A new point matching algorithm for non-rigid registration, Comput. Vis. Image Underst. 89 (2–3) (2003) 114–141.
77. F. Shi, et al., LABEL: pediatric brain extraction using learning-based meta-algorithm, NeuroImage 62 (2012) 1975–1986.
78. N. Tustison, et al., N4ITK: improved N3 bias correction, IEEE Trans. Med. Imaging 29 (6) (2010) 1310–1320.
79. A. Madabhushi, J. Udupa, New methods of MR image intensity standardization via generalized scale, Med. Phys. 33 (9) (2006) 3426–3434.
80. D.W. Shattuck, M.M., V. Adisetiyo, C. Hojatkashani, G. Salamon, K.L. Narr, R.A. Poldrack, R.M. Bilder, A.W. Toga, Construction of a 3D probabilistic atlas of human cortical structures, NeuroImage 39 (3) (2008) 1064–1080.
81. Z.-H. Cho, et al., Quantitative analysis of the hippocampus using images obtained from 7.0 T MRI, NeuroImage 49 (3) (2010) 2134–2140.
82. Z.-H. Cho, et al., New brain atlas – mapping the human brain in vivo with 7.0 T MRI and comparison with postmortem histology: will these images change modern medicine?, Int. J. Imaging Syst. Technol. 18 (1) (2008) 2–8.