Unsupervised Deep Feature Learning for Deformable Registration of MR Brain Images

Guorong Wu, Minjeong Kim, Qian Wang, Yaozong Gao,
Shu Liao, and Dinggang Shen

Department of Radiology and BRIC, University of North Carolina at Chapel Hill, USA

Abstract. Establishing accurate anatomical correspondences is critical for medical image registration. Although many hand-engineered features have been proposed for correspondence detection in various registration applications, no features are general enough to work well for all image data. Although many learning-based methods have been developed to help selection of best features for guiding correspondence detection across subjects with large anatomical variations, they are often limited by requiring the known correspondences (often presumably estimated by certain registration methods) as the ground truth for training. To address this limitation, we propose using an unsupervised deep learning approach to directly learn the basis filters that can effectively represent all observed image patches. Then, the coefficients by these learnt basis filters in representing the particular image patch can be regarded as the morphological signature for correspondence detection during image registration. Specifically, a stacked two-layer convolutional network is constructed to seek for the hierarchical representations for each image patch, where the high-level features are inferred from the responses of the low-level network. By replacing the hand-engineered features with our learnt data-adaptive features for image registration, we achieve promising registration results, which demonstrates that a general approach can be built to improve image registration by using data-adaptive features through unsupervised deep learning.

1 Introduction

Deformable image registration is very important in many neuroscience and clinical studies to normalize the individual subjects to the reference space [1, 2]. The principle behind image registration is to reveal the anatomical correspondences by maximizing the feature similarities between two images. Thus, image registration often relies on hand-engineered features, e.g., Gabor filters, to drive deformable registration [3].

However, the pitfall of these hand-engineered image features is that they are not guaranteed to work well for all image data, especially for correspondence detection. For example, it is not effective to use the responses of Gabor filters as the image features to help identify the point in the uniform white matter region of MR (Magnetic Resonance) brain images. Accordingly, learning-based methods have been proposed recently to select a set of best features from a large feature pool for characterizing each image point [4, 5]. The criterion is usually set to require the feature vectors on

K. Mori et al. (Eds.): MICCAI 2013, Part II, LNCS 8150, pp. 649–656, 2013.

the corresponding points to be **(1)** discriminative against other non-corresponding points and **(2)** consistent with the corresponding points across training samples [4, 6]. Then the learnt best features can often improve the registration accuracy and robustness. However, the current learning-based methods require many known correspondences in the training data, which have to be approximated by certain registration methods. Thus, besides being stuck in this chicken-and-egg causality dilemma, these supervised learning-based methods could also be affected by the quality of provided correspondences due to limited accuracy of employed registration methods.

To address these limitations, we aim to seek for the independent bases derived directly from the image data by unsupervised learning. Specifically, we first consider a feature space consisting of all possible image patches. Then, we aim to learn a set of independent bases that are able to well represent all image patches. Next, the coefficients derived from the patch representation by these learnt bases can be regarded as the morphological signature to characterize each point for correspondence detection.

Inspired by the recent progress in machine learning, we adopt a stacked convolutional ISA (Independent Subspace Analysis) method [7] to learn the hierarchical representations for patches from MR brain images. Generally speaking, ISA is an extension of ICA (Independent Component Analysis) to derive image bases for image recognition and pattern classification [8]. To overcome the limitation of high dimensional data in video processing, Le *et al.* [7] introduced the deep learning techniques such as stacking and convolution [9] to build a layered convolutional neural network that progressively performs ISA in each layer for unsupervised learning. In our application, we deploy the stacked convolutional ISA method to learn the hierarchical representations for the high-dimensional 3D image patches, thus allowing us to establish accurate anatomical correspondences by using hierarchical feature representations (which include not only the low-level image features, but also the high-level features inferred from large-scale image patches).

To show the advantage of unsupervised feature learning in image registration, we integrate our learnt features, from 60 MR brain images, into multi-channel demons [10] and also a feature-based registration method [11]. Through the evaluation on IXI dataset with 83 manually labeled ROIs and also the ADNI dataset, the performances of both state-of-the-art registration methods have been improved substantially, compared with their counterpart methods using hand-engineered image features. These results also show a general way of improving image registration by using hierarchical feature representations through unsupervised deep learning.

2 Method

2.1 Motivations

Since there are no universal image features that can work well with all image data, learning-based methods were recently developed to learn the best features for all image points to guide registration. Specifically, in the training stage, all sample images are first registered to the particular template by a certain state-of-the-art registration algorithm. Then, the correspondences from the estimated deformation fields

are regarded as ground truth. Next, the procedure of feature selection is performed on each point to pick up the best features, so that the similarities between corresponding points can be preferably increased [4]. In the application stage, each subject point has to be pre-calculated with all kinds of features used in the training stage, and then its correspondence is established by using the learnt best features for each target template point under registration.

However, these learning-based image registration methods have the following limitations:

1) The correspondences provided for training may be inaccurate. A typical example for elderly brain images is shown in the top row of **Fig. 1**, where the deformed subject image (**Fig. 1**(c)) is far from well-registered with template (**Fig. 1**(a)), especially for the ventricles. Thus, it is difficult to learn meaningful features for allowing accurate correspondence detection.
2) The best features are often learnt only at the template space. Once the template image is changed, the whole learning procedure needs to be redone, which is time consuming.
3) Current learning-based methods are not straightforward to include new image features for training, unless repeating the whole training procedure again.
4) Considering the computational cost, the best features are learnt only from a few types of image features (e.g., only 3 types of image features, each with 4 scales, as used in [6]), which limits the discriminative power of the learnt features.

To overcome the above limitations, we propose the following unsupervised learning approach to learn the hierarchical representations for image patches.

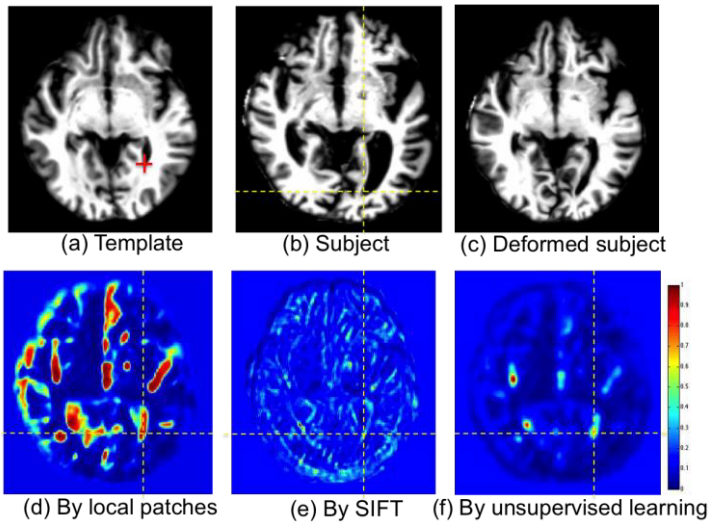

Fig. 1. (a-c) Comparison of template, subject, and deformed subject. (d-f) Similarity between a template point (indicated by red cross) and all subject points, measured with hand-engineered features (d, e) and the features learnt by unsupervised learning, respectively.

2.2 Unsupervised Learning by Independent Subspace Analysis

Here, we use x^t to denote a particular image patch, which is arranged into the column vector with length L, i.e., $x^t = [x_1^t, x_2^t, \ldots, x_L^t]'$. The superscript $t = \{1, \ldots, T\}$ denotes the index of all T image patches from the training MR brain images. In the classic feature extraction, a set of filters $W = \{w^i\}_{i=1,\ldots,N}$ are hand-designed to extract features from x^t, where each w^i is a column vector ($w^i = [w_1^i, w_2^i, \ldots, w_L^i]'$) and totally N filters are used. A certain feature can be computed by the dot product of x^t and w^i, i.e., $s^{t,i} = x^t \odot w^i$.

ISA is an unsupervised learning algorithm that automatically learns the basis filters $\{w^i\}$ from image patches $\{x^t\}$. As an extension of ICA, the responses $s^{t,i}$ are not required to be all mutually independent in ISA. Instead, these responses can be divided into several groups, each of which is called independent subspace [8]. Then, the responses are dependent inside each group, but dependencies among different groups are not allowed. Thereby, similar features can be grouped into the same subspace to achieve invariance. We use matrix $V = [v_{i,j}]_{i=1,\ldots,L, j=1,\ldots,N}$ to represent the subspace structure of all observed responses $s^{t,i}$, where each entry $v_{i,j}$ indicates whether basis vector w^i is associated with the j^{th} subspace. Here, N denotes for the dimensionality of subspace of response $s^{t,i}$. It is worth noting that the matrix V is fixed when training ISA [7].

The graphical depiction of ISA is shown in **Fig. 2**(a). Given image patches $\{x^t\}$ (in the bottom of **Fig. 2**(a)), ISA learns optimal W (in the middle of **Fig. 2**(a)) via finding independent subspaces (indicated by the pink dots in **Fig. 2**(a)) by solving:

$$\widehat{W} = \arg\min_W \sum_{t=1}^{T} \sum_{j=1}^{N} p_j(x^t; W, V), \quad s.t. WW' = I, \tag{1}$$

where $p_j(x^t; W, V) = \sqrt{\sum_{i=1}^{L} v_{i,j} (x^t \odot w^i)^2}$ is the activation of particular x^t in ISA. The orthonormal constraint is used to ensure the diversity of the basis filters $\{w^i\}$. Batch projected gradient descent is used to solve Eq. 1, which is free of tweaking with learning rate and convergence criterion [7]. Given the optimized W and any image patch y, it is straightforward to obtain the activation of y in each subspace, i.e., $\xi(y) = [p_j(y; W, V)]_{j=1,\ldots,N}$. Note that $\xi(y)$ is regarded as the representation coefficient vector of a particular patch y with the learnt basis filters W, which will be used as the morphological signature for the patch y during the registration.

To ensure the accurate correspondence detection, multi-scale image features are necessary to use, especially for the ventricle example shown in **Fig. 1**. However, it also raises a problem of high-dimensionality in learning features from the large-scale image patches. To this end, we follow the approach used in video data analysis [7] by constructing a two-layer network, as show in **Fig. 2**(b), for scaling up the ISA to the large-scale image patches. Specifically, we first train the ISA in the first layer based on the image patches with smaller scale. After that, a sliding window (with the same scale in the first layer) convolutes with each large-scale patch to get a sequence of overlapped small-scale patches (shown in **Fig. 2**(c)). The combined responses of these overlapped patches through the first layer ISA (a sequence of blue triangles in **Fig. 2**(b)) are whitened by PCA and then used as the input (pink triangles in **Fig. 2**(b))

to the second layer that is further trained by another ISA. In this way, high-level understanding of large-scale image patch can be perceived from the low-level image features detected by the basis filters in the first layer. It is apparent that this hierarchical patch representation is fully data-adaptive, thus free of requirement on known correspondences.

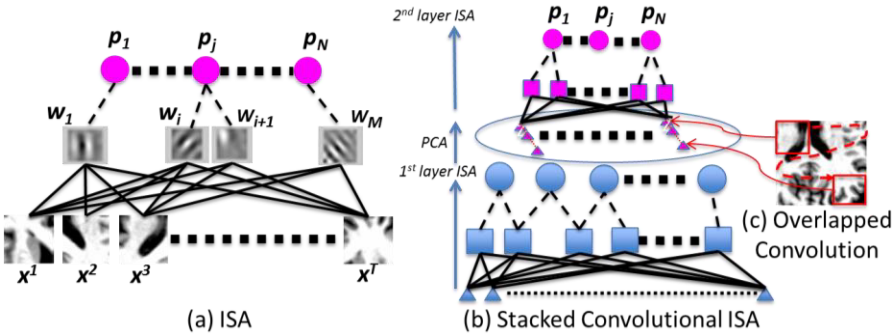

Fig. 2. Graphical depiction of ISA and the stacked convolutional ISA network

The basis filters learnt in the first layer, from 60 MR brains, are shown in **Fig. 3**, where we show only a 2D slice to represent each 3D filter. Most of them look like Gabor filters that can detect edges in different orientations. Given the stacked 2-layer convolutional ISA network, the input image patch y is now extracted in a large scale. The hierarchical representation coefficient $\xi(y)$ is calculated as follows: (1) extract a set of overlapped small-scale patches from y by sliding window (**Fig. 2**(c)); (2) calculate the response of each small-scale patch in the 1st layer ISA; (3) combine the responses in step (2) and further reduce the dimension by learnt PCA; (4) calculate the response in the 2nd layer ISA as the hierarchical representation coefficients $\xi(y)$ for patch y. In registration, we extract image patch in the large scale for each underlying point and use $\xi(y)$ as the morphological signature to detect correspondence. Here, we use normalized cross correlation as the similarity measurement between two representation coefficient vectors. The performance of our learnt features is shown in **Fig. 1**(f), where, for a template point (indicated by red cross in **Fig. 1**(a)), we can successfully find its corresponding point in the subject image even with large ventricle. Other hand-engineered features either detect too many non-corresponding points (when using entire intensity patch as the feature vector in **Fig. 1**(d)) or have too low responses and thus miss the correspondence (when using SIFT features in **Fig. 1**(e)).

2.3 Improving Deformable Image Registration with Learnt Features

Without loss of generalization, we show two examples of integrating the learnt features by ISA into the state-of-the-art registration methods. First, it is straightforward to deploy multi-channel demons [12] by regarding each channel with the element in $\xi(y)$. Second, we replace the hand-engineered attribute vector (i.e., local intensity

histogram) in a feature-based registration method, i.e., HAMMER [11][1], with our learnt features, while keeping the same optimization mechanism for deformable registration. In the following, we will show the registration improvement for these two methods after equipping with our learnt features.

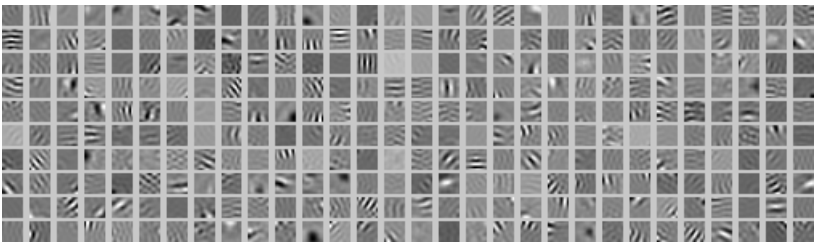

Fig. 3. The learnt bases filters (13×13) in the first layer from 60 MR brain images

3 Experiments

In this section, we demonstrate the performance of unsupervised learning method by evaluating the registration accuracy w/o learnt features. Specifically, we use 60 MR images from ADNI dataset (http://adni.loni.ucla.edu/) for training. For each image, we randomly sample ~15,000 image patches, where the patch size for the 1st and 2nd layers are $13 \times 13 \times 13$ and $21 \times 21 \times 21$, respectively. The number of learnt basis filters in the 1st layer and the dimensionality of subspace in the 2nd layer are $N = 300$ and $N = 150$, respectively. Therefore, the overall dimensionality of morphological signature $\xi(y)$ used for registration is 150 for each image patch y.

For comparison, we set diffeomorphic demons [2] and the HAMMER registration method [11] as the baseline methods. Next, we integrate the learnt image features into each channel of multi-channel demons and also replace the hand-engineered image features in HAMMER, which are denoted below as M+ISA, and H+ISA, respectively. Since PCA is also an unsupervised dimensionality reduction method by calculating the Eigen vectors, we further integrate the PCA-based dimension-reduced image patches with multi-channel demons (M+PCA) and HAMMER (H+PCA), in order to show the better performance by deep learning. To keep similar number of features as ISA, we deploy PCA on $7 \times 7 \times 7$ patches and keep >70% energy of image patches.

3.1 Experiment on IXI Dataset

IXI dataset[2] consists of 30 subjects, each with 83 manually delineated ROIs. FLIRT in FSL software package (http://fsl.fmrib.ox.ac.uk) is used to perform affine registration for all subjects to the template space. Then, these images are further registered with 6 above methods, respectively. In this way, we can calculate the Dice ratio for each of

[1] Source code can be downloaded at `http://www.nitrc.org/projects/hammerwml`

[2] `http://biomedic.doc.ic.ac.uk/brain-development/index.php?n=Main.Datasets`

83 ROIs, and also their overall Dice ratio. Specifically, the overall Dice ratios are 78.5% by Demons, 75.2% by M+PCA, 79.0 by M+ISA, 78.9% by HAMMER, 75.4% by H+PCA, and 80.1% by H+ISA, respectively. The detailed Dice ratios in 10 typical brain structures are also shown in **Fig. 4**. It is clear that the registration accuracy is improved by the learnt image features, compared to the baseline methods. The performances by M+PCA and H+PCA are worse than the baseline methods, because PCA assumes Gaussian distribution of image patches and may be not able to model the actual complicated patch distribution. Furthermore, we apply the paired t-test on Dice ratios by the above 6 registration methods. We found that M+ISA and H+ISA have significant improvements over their respective baseline methods ($p<0.05$) in 37 and 68 out of 83 ROIs in IXI dataset.

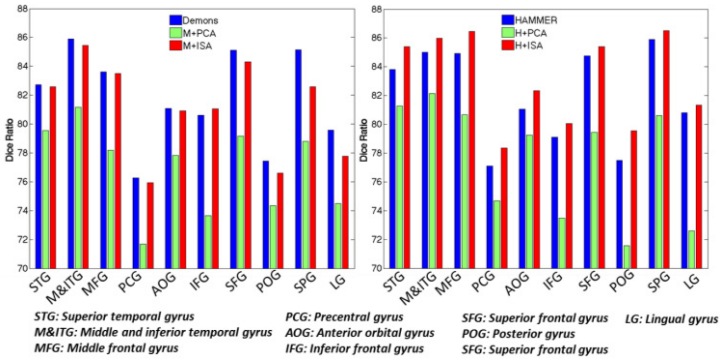

Fig. 4. Dice ratios for 10 typical ROIs in IXI dataset

3.2 Experiment on ADNI Dataset

In this experiment, we randomly select 20 MR images from ADNI dataset (different with the training images). The preprocessing steps include skull removal, bias correction, and intensity normalization. All subject images are linearly registered with template image by FLIRT. After that, we deploy 6 deformable registration methods to further normalize all subjects onto the template space. First, we show the Dice ratios on 3 tissue types (ventricle, gray mater, and white matter) by 6 registration methods in Table 1, where H+ISA method achieves the highest Dice ratio as in IXI dataset. It is worth noting that Demons achieves very high overlap (although still lower than H+ISA), because its registration is guided by intensities which are also used for tissue segmentation in our experiment (by FAST in FSL software package). Thus, it is not very fair for the feature-based registration method, although H+ISA still gets best.

Table 1. The Dice ratios of VN, GM, and WM on ADNI dataset. (unit: %)

Methods	Ventricle	Gray Matter	White Matter	Overall
Demons	93.2	78.0	89.7	86.9
M+PCA	84.5	71.6	80.5	78.9
M+ISA	88.9	76.5	87.8	84.4
HAMMER	91.5	72.5	82.4	82.1
H+PCA	90.6	71.9	83.5	82.0
H+ISA	95.0	78.6	88.1	87.3

Second, since ADNI also provides the labeled hippocampi, we further compare the overlap ratio of hippocampus by the registration methods using different features, i.e., our learnt ISA features and PCA features. Taking HAMMER as example, H+ISA achieves overall 2.74% improvement, compared to both H+PCA and HAMMER which have similar performance. On the other hand, M+ISA achieves overall 0.19% 0.24% improvements compared to M+PCA and the original Demons, respectively. We also apply the paired *t*-test upon M+ISA v.s. Demons and H+ISA v.s. HAMMER, and found that only H+ISA has significant improvement ($p < 0.05$) over the baseline method (HAMMER).

4 Conclusion

We have presented using the unsupervised learning method to explore the optimal image features for deformable image registration. In particular, a stacked convolutional ISA network is built to learn the hierarchical basis filters from a number of image patches in the MR brain images, thus the learnt basis filters are fully adaptive to both global and local image appearances. After incorporating these learnt image features into the existing state-of-the-art registration methods, we achieved promising registration results, showing that a general registration approach could be built by using hierarchical and data-adaptive features through unsupervised deep learning.

References

1. Zitová, B., Flusser, J.: Image registration methods: A survey. Image and Vision Computing 21(11), 977–1000 (2003)
2. Vercauteren, T., et al.: Diffeomorphic demons: efficient non-parametric image registration. NeuroImage 45(1, suppl. 1), S61–S72 (2009)
3. Liu, J., Vermuri, B.C., Marroquin, J.L.: Local frequency representations for robust multimodal image registration. IEEE Trans. Med. Imaging 21(5), 462–469 (2002)
4. Ou, Y., et al.: DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting. Med. Image Anal. 15(4), 622–639 (2011)
5. Wu, G., Qi, F., Shen, D.: Learning-based deformable registration of MR brain images. IEEE Trans. Med. Imaging, 25(9), 1145–1157 (2006)
6. Wu, G., Qi, F., Shen, D.: Learning best features and deformation statistics for hierarchical registration of MR brain images. In: Karssemeijer, N., Lelieveldt, B. (eds.) IPMI 2007. LNCS, vol. 4584, pp. 160–171. Springer, Heidelberg (2007)
7. Le, Q.V., 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)
8. Hyvarinen, A., Hoyer, P.: Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. Neural Comput. 12(7), 1705–1720 (2000)
9. LeCun, Y., Bengio, Y.: Convolutional network for images, speech, and time series. In: The Handbook of Brain Theory and Neural Networks (1995)
10. Peyrat, J., et al.: Registration of 4D cardiac CT sequences under trajectory constraints with multichannel diffeomorphic demons. IEEE Trans. Med. Imaging 29(7), 1351–1368 (2010)
11. Shen, D.G.: Image registration by local histogram matching. Pattern Recognition 40(4), 1161–1172 (2007)
12. Peyrat, J.-M., Delingette, H., Sermesant, M., Pennec, X., Xu, C., Ayache, N.: Registration of 4D time-series of cardiac images with multichannel Diffeomorphic Demons. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part II. LNCS, vol. 5242, pp. 972–979. Springer, Heidelberg (2008)