20th International Conference on Information Fusion
Xi'an, China - July 10-13, 2017

Tensor-based Descriptor for Image Registration via Unsupervised Network

Qiegen Liu[1,2], *Henry Leung*[1]

[1]Department of Electrical and Computer Engineering, University of Calgary, Calgary, T2N 1N4, Canada
[2]School of Electronic Information Engineering, Nanchang University, Nanchang, 330031, China

***Abstract*—Since the significant intensity variations existed between different modal images, the deformable registration is still very challenging. In this paper, in order to alleviate the variations deficiency and attain robust alignment, we propose a multi-dimensional tensor based modality independent neighbourhood descriptor (tMIND) to measure the similarity between the images. The tMIND compares the neighboring tensors which consisting of multi-filters induced features. In this work we learn these filters via PCA network (PCANet). We additionally describe the scheme of incorporating these filters into the tMIND. Experimental evaluations demonstrate its promise and effectiveness over the current state-of-the-art approaches.**

***Index Terms*—Non-rigid registration, tensor, unsupervised network, multi-filters, MIND.**

I. INTRODUCTION

Alignment of multi-modal images is a vital step for further analysis in the fields of medical imaging, such as magnetic resonance imaging (MRI) and computed tomography (CT), etc. Aid by the accurate and robust registration, the thereafter fusion, segmentation, etc, even the diagnostic tasks can be improved. Compared to the advances resulted in a number of successful methods for deformable registration techniques for scans of the same modality, the registration of images from different modalities is still challenging [1]. The severe intensity variations across modalities and large non-rigid motion are the primary difficulties.

Since the local intensity of different modalities is exactly different, it is essential to obtain a unified representation of the heterogeneous anatomies such that meaningful comparisons can be performed. A great number of methods have been proposed over the years to address this issue. The Mutual information (MI) and the Modality independent neighbourhood descriptor (MIND) are two representative successful approaches [1-3]. Specifically, MI is the most widely used information theoretical algorithm that globally describe the correlation between the reference and images to be registered. As a structural image representation, MIND assigns each pixels a structural vector that describes the central pixel in a local/nonlocal way. In summary, the internal image representation approaches in fact exploit the nonlocal or global information of the object itself such that the transformed coefficient under different object is similar or the same.

This work was supported in part by the National NSFC under 61362001, the international postdoctoral exchange fellowship program.

The similarity metric learning methodology via external learning also receives the attentions of researchers. The idea of using supervised learning to build a similarity metric for multimodal images has been explored in a number of works. For instance, Guetter *et al.* proposed to use Kullback-Leibler divergence to the joint-image distribution [4]. More works transfer the problem of learning a similarity metric to be a binary classification, whose goal is to discriminate between aligned and misaligned patches given pairs of aligned images. In [5], Lee *et al.* proposed a new supervised technique to learn the similarity measurement via the discriminative structured support vector machine. Within this similarity measurement, the correct correspondences are assigned high values while the wrong correspondences are assigned low ones. Michel *et al.* used a method based on Adaboost [6]. Simonovsky *et al.* relied on Convolutional neural network (CNN) learning method as the suitable set of characteristics for each type of modality combinations can be directly learned from the training data [7]. Tow shortcomings occur with the supervised learning approaches. The first is that the learning procedure is time-consuming. Like CNN networks optimize the filters by utilizing gradient method on large datasets, relying on the expertise of parameter initiation and fine tuning. Besides, supervised learning requires image labels. Yet label become scarce along with the increasing image scale. As we know, the unsupervised approaches used for registration is very few [8, 9]. Cao *et al.* learned two local dictionaries to describe the different modalities, subjecting their coefficients are the same [8]. It needs training images available for all modalities. Due to the complicated computation, the second strategy is far from satisfactory.

Inspired by the great success of deep neural networks in computer vision [10, 11, 12, 13, 14], we propose a multi-dimensional tensor-based modality independent neighbourhood descriptor (tMIND). Specifically, recent researches show that the unsupervised filter learning, incorporating with the multi-layer architecture, can achieve excellent performance in pattern recognition, deoising, etc [11, 12, 13]. In this work, the tensor-based tMIND extends the patch-based descriptor by generating multi-filters induced features. We learn these filters via PCANet, which applying Principal component analysis (PCA) to learn optimal image filters in a layer-wise way [14]. We additionally describe how incorporate these filters into the tMIND. Therefore, the present tMIND inherits the strengths of internal representation and external learning.

The rest of the paper is organized as follows: After reviewing some knowledge of the MIND, in Section II we introduce a new

978-0-9964-5270-0©2017 ISIF

tensor-based MIND descriptor, followed by introducing how to generate these multi-filters in details. Subsequently, Section III demonstrates the performance of the proposed algorithm. Finally, conclusions are given in Section IV.

II. PROPOSED TENSOR-BASED MIND

In this section, we briefly review the patch-based MIND, and then introduce a higher-dimensional tensor-based MIND descriptor: from patch to cube/volume. Subsequently, we describe how to generate the learned filters by employing unsupervised network. Finally, the detailed implementation of registration is described.

A. Proposed tMIND Descriptor

The modality independent neighborhood descriptor (MIND) [3] was proposed by Heinrich *et al.* for multi-modal image registration. The MIND was computed based on the similarities between neighboring patches. It only utilizes the image intensities for similarity computation. Thus may lead to inaccurate registration results for corners, edges and complicated textured regions.

Specifically, the MIND borrows the self-similarity concept introduced by Buades *et al.* [15]. It explores image self-similarities by replacing the local comparison of individual pixels with the non-local comparison of image patches. In the MIND, for two pixels at x and $x+r$ in the spatial search region R of image I, the similarity of $MIND(I,x,r)$ is calculated as:

$$MIND(I,x,r)=\frac{1}{n}\exp(-\frac{D(I,x,x+r)}{V(I,x)}) \quad r\in R \qquad (1)$$

$$D(I,x,y)=\left\|P_x-P_y\right\|^2 \qquad (2)$$

where $D(I,x,x+r)$ is the Euclidean distance between two similarity windows (i.e., image patches P_x and P_{x+r}) centered at x and $x+r$, respectively. The denominator $V(I,x)$ in Eq. (1) acts as the smoothing parameter and it is computed as the mean of the patch distances themselves within a six-neighborhood centered at x.

As can be observed in Eq. (1)(2), the MIND computes the pixel similarity based on the differences between the intensities of image patches; thus it is not robust to contrast variations. It is difficult to accurately determine self-similarity for the medical images with the complicated edge/texture features. The incorrect local image structure representation resulting from the MIND will influence the final non-rigid multi-modal registration results. To alleviate the deficiency, some researchers turn to integrate some intensity-insensitive information such as phases and gradients into the MIND principle for achieving better alignment [16-18]. In this work, we introduce a more general descriptor, which replaces the patch by tensor extracting from multi-feature images via convolutional filters. Specifically, the similarity of tMIND is computed as:

$$tMIND(I,x,r)=\frac{1}{n}\exp(-\frac{D(I,x,x+r)}{V(I,x)}) \quad r\in R \qquad (3)$$

$$D(I,x,y)=\left\|T_x-T_y\right\|^2 \qquad (4)$$

where T_x is the extracted image tensors formed by concatenating patches of multi-features centered at x. Assume there has L filters, then there exist L feature images $I_l=d_l\otimes I, l=1,\mathrm{L},L$ and form the tensor object $[I_1,\mathrm{L},I_L]$. $\otimes$ denotes the 2D convolution operation. The difference of extracted patch P_x and cube T_x is exhibited in Fig. 1.

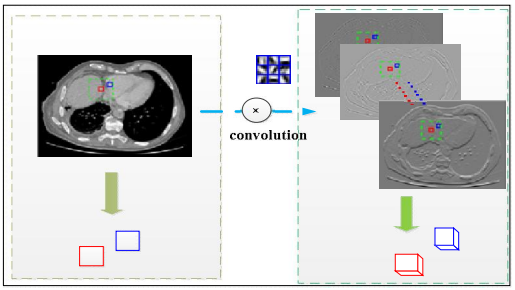


Fig. 1. Visualization of the basic elements in patch-based MIND and tensor-based tMIND.

As a generalization of MIND, tMIND involves two merits and advantages. First, from the representation viewpoint, the multi feature images are derived by the convolutional filters, which reveals the spatial relations in neighbor region and prefers to reflect the intrinsic representation of images from different modalities. Therefore, it is desirable that images from different multimodalities are transformed into a common space such that can be compared uniformly. Second, from the high-dimensional viewpoint, the new similarity metric can be seen as weight aggregation or boosting of the patch-based metric, as seen from the following derivation:

$$\begin{aligned} tMIND(I,x,r)&=\frac{1}{n}\exp(-\frac{\left\|T_x-T_{x+r}\right\|^2}{V(I,x)})\\ &=\frac{1}{n}\exp(-\frac{\sum_{l=1}^{L}\left\|[d_l\otimes I]_x-[d_l\otimes I]_{x+r}\right\|^2}{V(I,x)})\\ &=\frac{1}{n}\prod_{l=1}^{L}\exp(-\frac{\left\|[d_l\otimes I]_x-[d_l\otimes I]_{x+r}\right\|^2}{V(I,x)}) \end{aligned} \qquad (5)$$

Notice that in Eq. (3)(4), only the simple L2-norm of the tensor is employed in the similarity measure. More sophisticated property or measure (e.g., low rank) defined on the basis of tensor may prefer to better descriptor [19, 20]. This will be investigated in further study.

B. Filter Learning via Unsupervised Network

In the current work, we choice the PCANet to provide the learned filters. PCANet is a lightweight CNN in which the filters in the convolution layers are learned by the unsupervised learning method PCA. The PCANet involves two advantages: one is that PCA coding is commonly used in image processing community, like the recent PCA-LPG and patch-based PCA, both employing PCA to clustered patches [21-23]. As for PCANet, its training procedure is extremely simple and efficient because it does not involves regularized parameters or requiring numerical optimization solvers. The second is that it has strong generalization ability. As we will illustrate in the

experiment, the PCANet filters learned on well prepared databases can provide reasonably accuracy for various test images.

We use a patch of size $k_1 \times k_2$ to slide each pixel of the i th image $I_i \in R^{m \times n}$ and then reshape each patch into a column-vector, which is then concatenated to obtain a matrix $X_i \in R^{k_1 k_2 \times mn}$. Subsequently, for all the training images $\{I_i\}_{i=1}^{N}$, we can obtain the following matrix: $X = [X_1, X_2, \mathrm{L}\ , X_N] \in R^{k_1 k_2 \times Nmn}$. After subtracting patch mean from each patch (for convenience we still denote it as X), we apply PCA to the mean-removed matrix as follow:

$$\min_{U \in R^{k_1 k_2 \times L_1}} \left\| X - UU^T X \right\|_F^2 \quad s.t.\ U^T U = I_{L_1} \tag{6}$$

The solution of this minimization is the L_1 principal eigenvectors of XX^T. The resulting learned PCA filters are termed as

$$d_l^1 = mat_{k_1,k_2}(q_l(XX^T)) \in R^{k_1 \times k_2}, \quad l = 1, 2, \mathrm{L}\ , L_1 \tag{7}$$

where the operator $mat_{k_1,k_2}(u_l)$ maps the vector $u_l \in R^{k_1 k_2}$ to its matrix formulation $d_l^1 \in R^{k_1 \times k_2}$. Given the first layer's convolution filter bank $d^1 = \{d_1^1, d_2^1, \mathrm{L}\ , d_{L_1}^1\}$, we convolve each training image I_i with the L_1 filters:

$$I_i^1 = I_i \otimes d^1, i = 1, 2, \mathrm{L}\ , N \tag{8}$$

where $I_i^1 = \{I_i \otimes d_1^1, I_i \otimes d_2^1, \mathrm{L}\ , I_i \otimes d_{L_1}^1\}$ is the first layer's feature map set of image I_i convolved by the filters.

Almost repeating the same process as in the first stage, we collect overlapping patches from the feature maps $\{I_{i,l}^1\}_{l=1,\mathrm{L}\ ,L_1}^{i=1,\mathrm{L}\ ,N}$, employ PCA to the mean-removed patches induced matrix, and obtain the PCA filter bank $d^2 = \{d_1^2, d_2^2, \mathrm{L}\ , d_{L_2}^2\}$. The whole procedure of the unsupervised network is shown in Fig. 2.

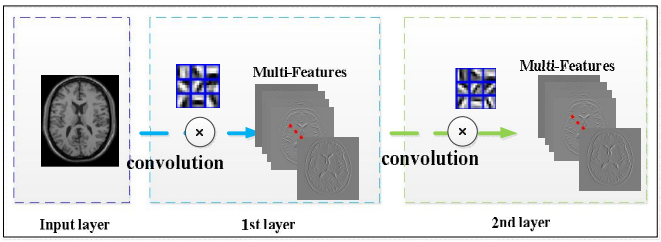


Fig. 2. Illustration of the two-layer unsupervised network.

After the two-layer filters are learned, we adopt a new fashion to collect them into the registration procedure. i.e., at first, borrowing the ideas from the GoogLeNet [24], we use these multi-layer filters in a parallel way to convolute the images simultaneously. Secondly, due to the redundancy, some filters at different layers tend to appear repeated. Motivated by the work in dictionary learning [25] that enforcing incoherence between the different dictionaries, we propose to remove some strongly correlated filter from the filter set, thereby improve the discriminatory power of the filter set. A cross correlation criteria is employed in this work: $\{(d_i^1)^T d_j^2\}_{j=1,\mathrm{L}\ ,L_2}^{i=1,\mathrm{L}\ ,L_1}$. The filter in 2nd layer whose correlation value above 0.95 with any filter in the 1st layer filter set will be removed. An example is shown in Fig. 3.

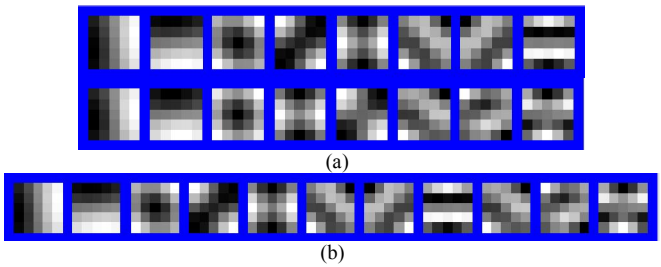


Fig. 3. Visualization of learned two-layer PCANet filters (a) and the final selected filters (b).

C. Multi-modal Registration Procedure

One motivation of using convolutional operator as a basic tool is that it allows to align multi-modal images via a simple similarity metric across modalities. Once the descriptors are extracted for both images, yielding a vector for each voxel, the similarity metric between two images is defined as the SSD between their corresponding descriptors. Therefore efficient optimization algorithms, which converge rapidly can be used without further modification.

The registration procedure is implemented same as in [3]. Concretely, the registration with tMIND descriptor is derived by the following objective function:

$$\min_u \sum_x S(I_1(x), I_2(x+u))^2 + \alpha tr(\nabla u(x)^T * \nabla u(x)) \tag{9}$$

where $S(I_1(x), I_2(x)) = \frac{1}{|R|} \sum_{r \in R} \left| tMIND(I_1, x, r) - tMIND(I_2, x, r) \right|$. The weighting parameter α balances the similarity fidelity and the local deformation regularization.

Eq. (9) can be rewritten as

$$\min_u f^T f, \quad f = [S(I_1(x), I_2(x+u)), \sqrt{\alpha} \nabla u(x)] \tag{10}$$

Since it has $S_{I_2(x+u)} \approx S_{I_2(x)} + \nabla S_{I_2(x)} u$ and $\sqrt{\alpha} \nabla u \equiv \sqrt{\alpha} \nabla u$, Its derivative is $J = [\nabla S_{I_2(x)}, \sqrt{\alpha} \nabla]$. By applying the Gauss-Newton, at each iteration the update rule is $J^T J u = -J^T f$, J is the derivative of f with respect to variable u. Hence the resulting deformation field at iteration u^{j+1} is

$$(\nabla S_{I_2(x)}^T \nabla S_{I_2(x)} - \alpha \Delta) u^{j+1} = -(\nabla S_{I_2(x)}^T S_{I_2(x)} - \alpha \Delta u^j) \tag{11}$$

At each step, a symmetric deformable registration is followed such as to obtain diffeomorphic transformations, avoiding physically implausible folding of volume occurs [26]. Additionally, a multi-resolution scheme is used to represent coarse-to-fine details of both volumes for fast and robust registration.

III. EXPERIMENTAL RESULTS

In this section, we perform a number of challenging registration experiments to evaluate the accuracy and robustness of the present method. We evaluate our findings based on qualitative and quantitative measures. The Root Mean Square Error (RMSE) and Target Registration Error (TRE) are used for quantitative comparisons of registered results. For simulated

data, the RMSE is calculated between the source and aligned images. The target registration error (TRE) of anatomical landmarks (The TRE for a given transformation $u(x)$ and an anatomical landmark pair (x, x')) is defined by:

$$TRE = \|x + u(x) - x'\|_2 / |G| \quad (12)$$

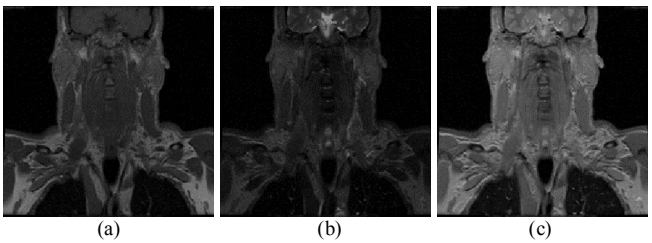
where G and $|G|$ are the set of anatomical landmarks and the number of landmarks in the reference image, respectively. We ran all the tests on an Intel Core i7-4700MQ CPU 2.4 GHz Windows 64-bit operating system with 8 GB RAM.

We first investigate the parameter of tMIND. Then we apply the method to some simulated data and perform deformable registrations ten CT datasets of lung cancer patients. In all the experiments, we manually tuned the weighting parameter α for both MIND and tMIND methods such that the best RMSE/TRE is achieved. In the MI and Residual Complexity (RC) method, the regularization parameter is also optimized manually.

A. Parameter Setting

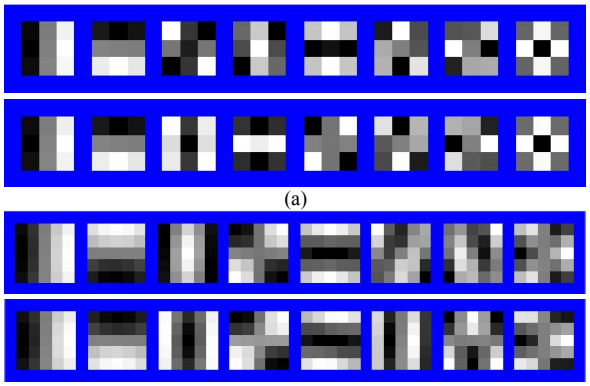
In the patch-based MIND, its search region is usually in the range of 3×3, the Gaussian patch parameter sigma is below 0.5, which roughly equals to 3×3 [3]. As an extension to tensor formulation, we argue that a good tensor formulation should has near or equal length on each mode. Therefore, we restraint our learned filter number (i.e., L_1 and L_2) no more than 9 and the filter size (i.e., k_1 and k_2) no more than 5.

We use the eighteen T1, T2 and PD weighted MRI scans with size 256×256 from the Visible Human dataset [27] as samples to train the PCANet filters. One sample scan is depicted in Fig. 4 and the learned filters with filter number $L_1 = L_2 = 8$ are shown in Fig. 5. The cases of $k_1 = k_2 = 3$ and $k_1 = k_2 = 5$ are depicted. As can be seen, the first and second filters look like the gradient filters. Additionally, the 8th filter in Fig. 5 is very similar to the LapLacian filter. In order to investigate the registration performance of the tMIND with regard to filter number and filter size. We apply these two filter set with filter number ranging from 1 to 8 to register a T1-PD pair as shown in Fig. 6. In the synthetic data, a geometric distortion with a thin-plate spline (TPS) model is applied to the corresponding source image to get the fixed and moving images [28]. The RMSE result is shown in Fig. 7. It can be observed that the case of $k_1 = k_2 = 3$ is generally better than that of $k_1 = k_2 = 5$. Furthermore, varying L_1 from 1 to 8, the result improves and more robust. Hence in our experiment, we set $k_1 = k_2 = 3$ and $L_1 = L_2 = 8$.



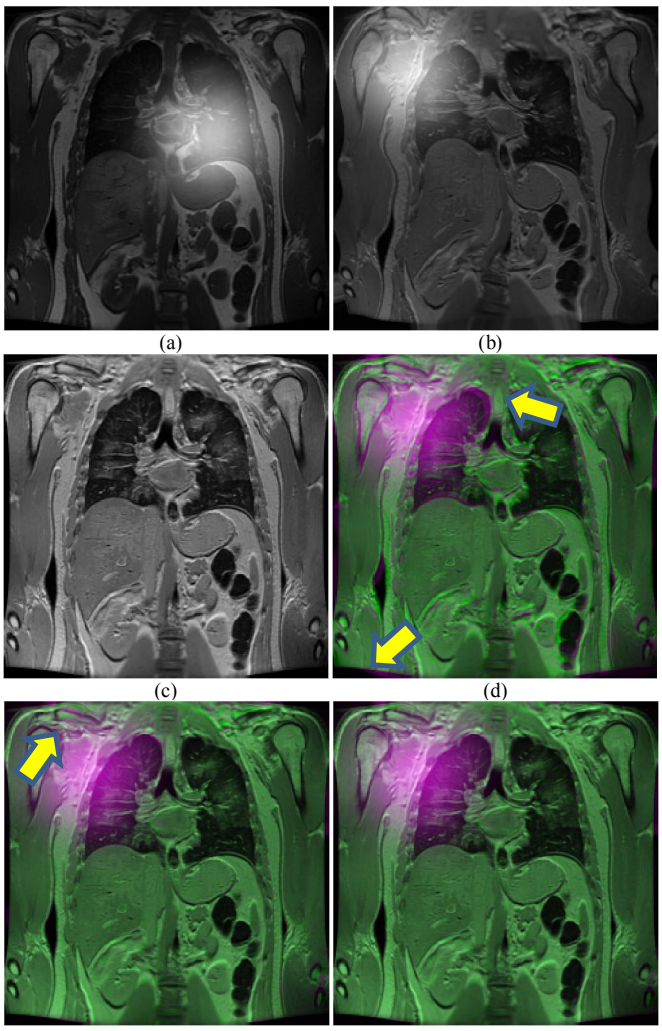
(a) (b) (c)

Fig. 4. Example of multimodal images in Visible Human dataset. (a) T1 MR, (b) T2 MR, and (c) PD MR.



(a) (b)

Fig. 5. Visualization of two-layer filters learned from Visible Human dataset. (a) size $k_1 = k_2 = 3$, (b) size $k_1 = k_2 = 5$.



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

Fig. 6. Registration comparison of T1-PD pair. (a) fixed image, (b) moving image, (c) source/reference of the moving image, (d)(e)(f) registration display between the reference and moving image obtained by MIND, tMIND with 2 and 8 filters.

It is worth noting that in the case of $k_1 = k_2 = 3$ as shown in Fig. 5(a), the filter set includes the gradient and LapLacian filters. This indicates that our proposed model can be seen as a generalization of several recent works for improving the MIND descriptor [16-18].

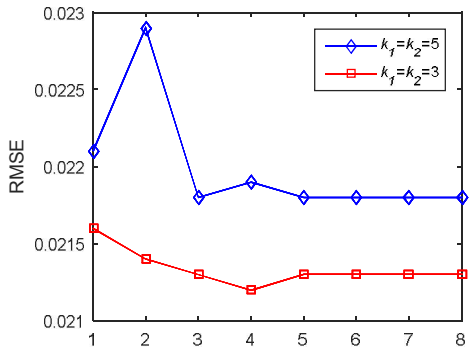
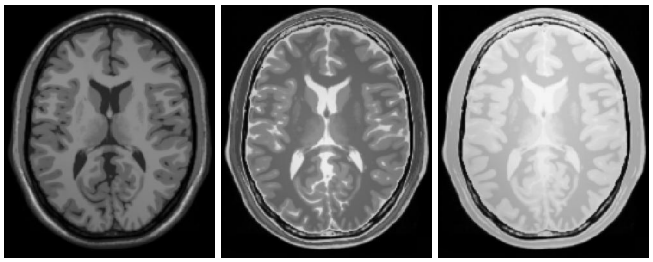


Fig. 7. Registration error RMSE of tMIND vs learned filter number.

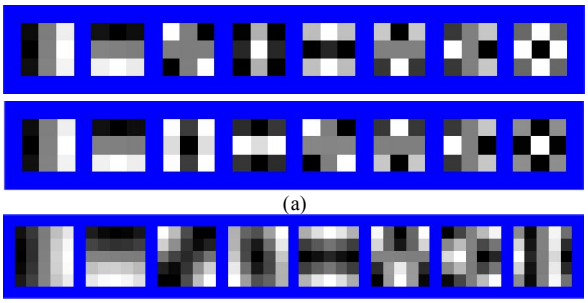
B. Deformable registration of simulated data

We also extract thirty T1, T2 and PD weighted MRI scans with size 181×217 from the BrainWeb dataset as samples to train the PCANet filters. As shown in Fig. 8, T1 MR image shows better anatomical detail and T2 MR reflects pathological changes better. Fig. 9 depicts the learned PCANet filters. It can be observed that the filters learned from the simulated dataset is very similar to that of Visible Human dataset, in terms of visual inspection. This implies that it is desirable to construct a generic training dataset for medical images specifically for MR images. We use the learned filters in Fig. 5(a) in the following two experiments, conducted in Brainweb dataset [29] and the Brain Tumor Segmentation (BRATS) challenge [30], respectively.

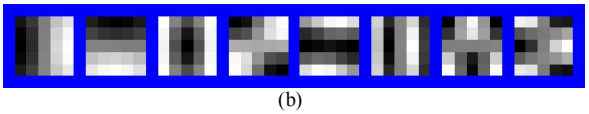


(a) (b) (c)

Fig. 8. Example of multimodal images in BrainWeb dataset. (a) T1 MR, (b) T2 MR, and (c) PD MR.



(a)

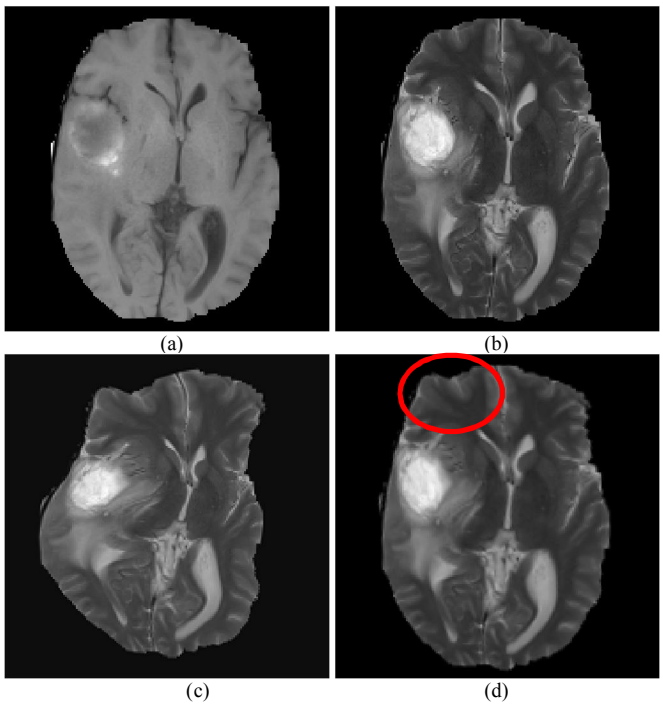


(b)

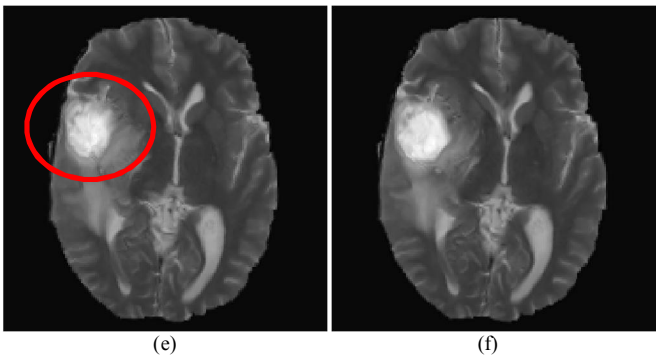
Fig. 9. Visualization of two-layer filters learned from BrainWeb dataset. (a) size $k_1 = k_2 = 3$, (b) size $k_1 = k_2 = 5$.

In the first experiment of aligning the T2-T1 pair, the spatially-varying intensity distortion is added to both fixed and moving images, as done in Fig. 6. In the result, the reference image is displayed in green and the registered version of moving image in magenta. As expected, a well-done registration yields a gray-scale image and larger color differs imply higher misregistration. The registration comparison is shown in Fig. 10. It can be observed that the RC method cannot yield good result and the MI measure also fails to the heavy intensity distortion. tMIND produces better registration results in the region with simulated intensity distortions and large spatial deformations than MIND, as indicated by the yellow arrows. The superior result indicates that the high-dimensional and deep representation enabled tMIND to better avoid the uncertainty existed in the registration procedure.

In the second experiment of aligning the T1-T2 pair, both the fixed and source image have tumor. We apply a spatially-varying intensity distortion as described by Myronenko *et al.* [28] to the source image to obtain the moving image. As shown in Fig. 11(d), the MI method cannot accurately correct the deformations in the left boundary area. In Fig. 11(e), the MIND fails to align the region around the tumor, even though different combinations of parameters were tried. There still exist some excessive localized deformations, which are highlighted by the red circles. While it can be observed that the tMIND descriptor less dependent on the intensity differences and the associated result keeps more consistent with the reference image.



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



(e) (f)

Fig. 11. Registration comparison of T1-T2 pair. (a) fixed image, (b) source image, (c) moving image, (d) MI method, (e) MIND and (f) tMIND.

C. Deformable registration of real data

We further test the presented algorithm on DIR-Lab dataset [31], which consists of ten 4D thoracic CT scan pairs between inhale and exhale phase of the breathing cycle. Each scan pair is acquired on the thorax and upper abdomen of patients treated for esophagian cancer, between inhale and exhale phase of the breathing cycle. The slice thickness is 2.5 mm, and in-plane resolution varies in the interval of 0.97 and 1.16 mm. Major challenges arise from possible contrast variations between tissue and air induced by lung compression, motion discontinuities at the lung/rib cage interface, as well as large deformations of small features such as lung vessels, airways.

In the experiment, we apply registration directly between each pair of the original CT scans. For each scan, 300 anatomical landmark pairs within the lungs have been carefully annotated by thoracic imaging experts. We evaluate the registration accuracy based on these landmark point sets. Table 1 summarizes the average TRE comparison results. The mean landmark distance and corresponding standard deviations are recorded. It can be observed that tMIND achieves lower TRE value than the MIND method. The visual comparison of case 7 is shown in Fig. 12. In this visualization, the source image is shown in magenta while the reference image is shown in green. Gray scale image will emerge in the regions where the images are fully aligned. In the unregistered case, magenta and green areas can clearly be observed indicating that the morphology is not aligned. In the registered cases, these colored areas almost disappear indicating that the images are successfully registered. Particularly, tMIND diminishes larger regions than that of MIND.

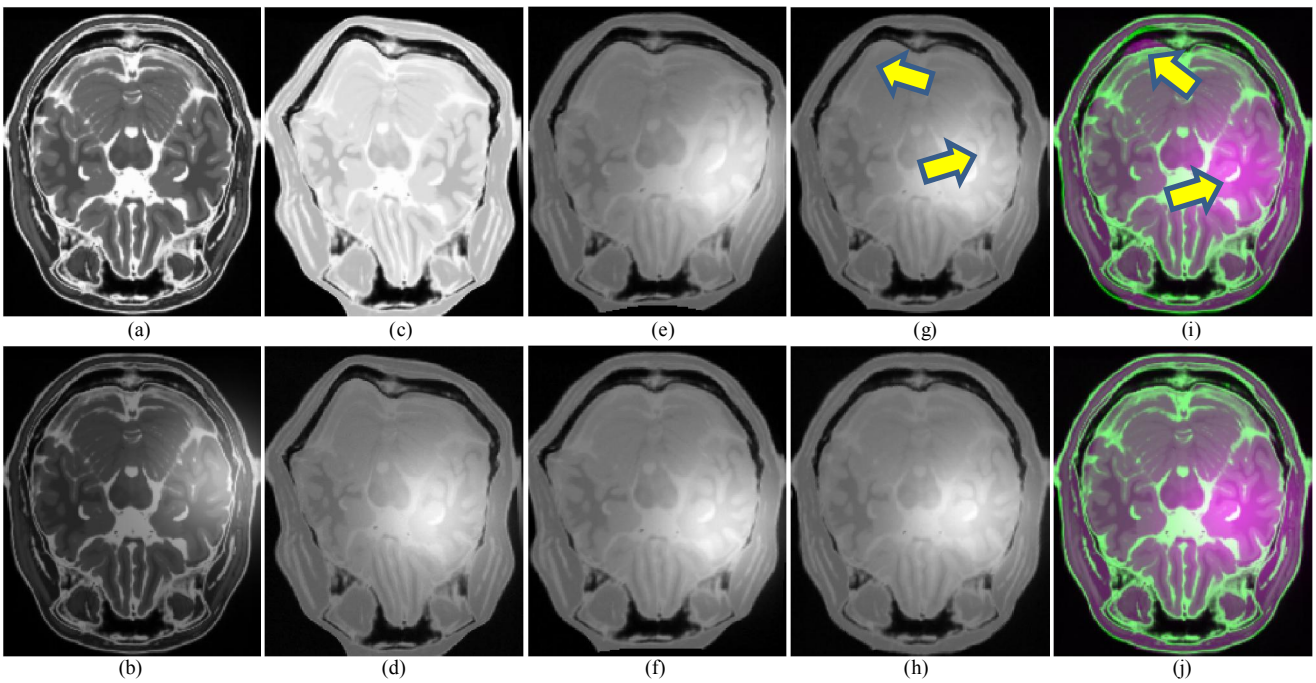
Table 1. Target registration error (in millimeters) obtained over the 10 cases of thorax CT-scans for all tested experimental conditions.

	BEFORE	MIND	tMIND
TRE	8.46(6.58)	1.64(3.04)	1.21(1.15)

IV. CONCLUSIONS

Aid by multi-filters learned from the unsupervised network, this paper presented a new tensor-based descriptor tMIND for structural representation of images to be registered. tMIND concentrates multi-features as a whole object to be compared. We employed the PCANet to learn filters so as to produce multi-features. Experiments were conducted on simulated and real data to validate the strengths of the proposed method.

The primary contribution of this paper is to building a general framework of constructing high-dimensional descriptor for deformable registration by means of unsupervised deep learning. Forthcoming study will focus on developing new unsupervised learning methods to enhance the effectiveness and efficiency. Besides, more registration experiments evaluated by clinical experts will be investigated.



(a) (c) (e) (g) (i)

(b) (d) (f) (h) (j)

Fig. 10. Registration comparison of T2-T1 pair. (a) Source image of the fixed image (b), (c)(d) moving image, (e) MI method, (f) RC method, (g) MIND, (h) tMIND, (i)(j) registration display between the reference and moving image obtained by MIND and tMIND.

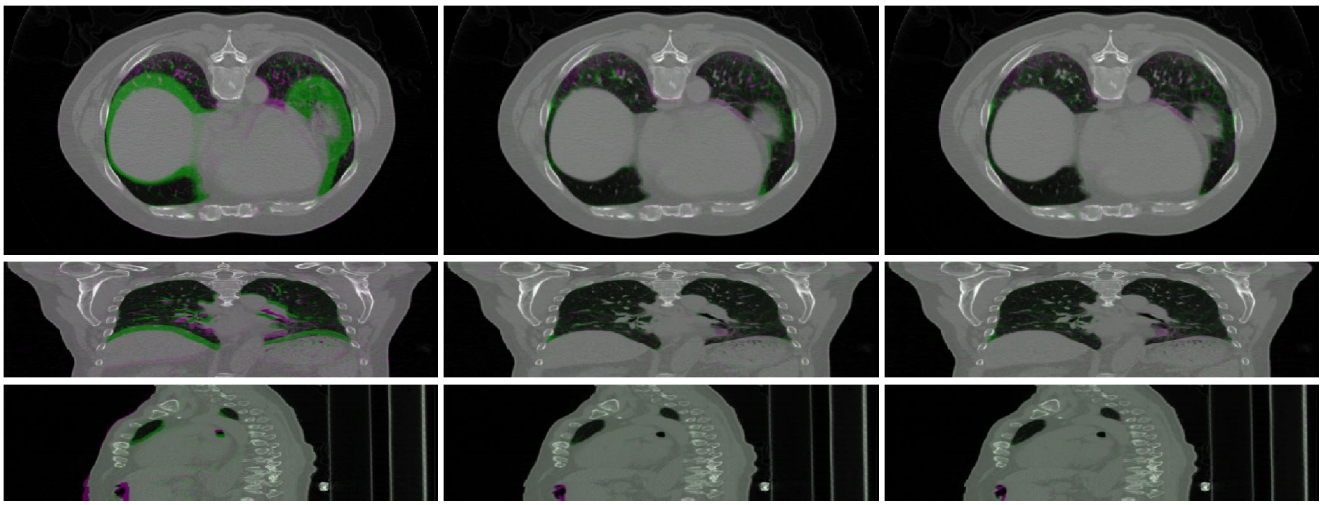


Fig. 12. Deformable registration result for Case 7 of the CT dataset. Top: axial, middle: sagittal and bottom: coronal plane. Left row: before registration, center and right row: registration by MIND and the proposed tMIND technique. The target image is displayed in magenta and the source image in green (complementary color).

V. REFERENCES

[1] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, 16, 187–198, 1997.

[2] P. Viola, W. Wells III, "Alignment by maximization of mutual information," *Int J. Comput. Vis.*, 24, 137–154, 1997.

[3] M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, S. M. Brady, and J. A. Schnabel, "MIND: modality independent neighbourhood descriptor for multi-modal deformable registration," *Med. Image Anal.*, 16(7), 1423-1435, 2012.

[4] C. Guetter, C. Xu, F. Sauer, J. Hornegger, "Learning based non-rigid multi-modal image registration using kullback-leibler divergence," In: Duncan, J.S., Gerig, G. (eds.) *MICCAI* 2005. LNCS, vol. 3750, pp. 255–262. Springer, Heidelberg, 2005.

[5] D. Lee, M. Hofmann, F. Steinke, Y. Altun, N.D. Cahill, B. Schölkopf, "Learning similarity measure for multi-modal 3d image registration," in: *Proc. IEEE CVPR*, 2009, pp. 186–193.

[6] F. Michel, M. Bronstein, A. Bronstein, N. Paragios, "Boosted metric learning for 3D multi-modal deformable registration," In: *ISBI*, pp. 1209–1214, 2011.

[7] M. Simonovsky, B. Gutiérrez-Becker, D. Mateus, N. Navab, N. Komodakis, "A deep metric for multimodal registration," *MICCAI* 2016, pp 10-18.

[8] T. Cao, C. Zach, S. Modla, D. Powell, K. Czymmek, "Multi-modal registration for correlative microscopy using image analogies," *Med. Image Anal.*, vol. 18, no. 6, pp. 914-926, 2014.

[9] Z. Yi, S. Soatto, "Multimodal registration via spatial-context mutual information," *Information Processing in Medical Imaging*, vol. 6801. Springer, Berlin/Heidelberg, pp. 424–435, 2011.

[10] K. Kavukcuoglu, P. Sermanet, Y.L. Boureau, K. Gregor, M. Mathieu and Y. LeCun, "Learning convolutional feature hierarchies for visual recognition," in *Proc. NIPS*, vol. 23, pp. 1090-1098, 2010.

[11] M.D. Zeiler, D. Krishnan, G.W. Taylor, and R. Fergus, "Deconvolutional networks," in: *Proc. IEEE CVPR*, pp. 2528-2535, 2010.

[12] Z. Lei, D. Yi, S. Li, "Learning stacked image desripttor for face recognition," *IEEE Transactions on Circuits and System for Video Technology*, vol. 26, no. 9, Sep 2016.

[13] S. Zhang, J. Wang, X. Tao, Y. Gong, N. Zheng, "Constructing deep sparse coding network for image classification," *Pattern Recognition*, 64:130-140, 2017.

[14] T. H. Chan, K. Jia, S. H. Gao, J. W. Lu, Z. N. Zeng, and Y. Ma, "PCANet: A simple deep learning baseline for image classification?" In arXiv: 1404.3606v2, 2014.

[15] A. Buades, B. Coll, J.M. Morel, "A non-local algorithm for image denoising," in: *Proc. IEEE CVPR*, June 2005, pp. 60–65.

[16] S. Thiruvenkadam, "Dense multi-modal registration with structural integrity using non-local gradients," *VISAPP*, (1) 2013:258-263.

[17] Z. Li, L. J. van Vliet, and F. M. Vos, "Image registration based on autocorrelation of local structure," *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 63-75, Jan. 2016.

[18] B Denis de Senneville, C Zachiu, M Ries and C Moonen, "EVolution: an edge-based variational method for non-rigid multi-modal image registration," *Phys. Med. Biol.*, pp. 7377-7396, 2016.

[19] A. Ghaffari, E. Fatemizadeh, "RISM: single-modal image registration via rank-induced similarity measure," *IEEE Trans. Image Process.*, 24 (12) (2015), pp. 5567–5580.

[20] L. R. Tucker, "Implications of factor analysis of three-way matrices for measurement of change," *In Problems in Measuring Change*, pp. 122-137. University of Wisconsin Press, 1963.

[21] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1838-1857, July 2011.

[22] L. Zhang, W. Dong, D. Zhang, G. Shi, "Two-stage image denoising by principle component analysis with local pixel grouping," *Pattern Recognition*, vol. 43, pp. 1531-1549, Apr. 2010.

[23] C.A. Deledalle, J. Salmon, A. S. Dalalyan, "Image denoising with patch based PCA: local versus global," *BMVC* 2011.

[24] C. Szegedy,W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in: *Proc. IEEE CVPR*, 2015.

[25] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in Proc. CVPR, pp. 3501-3508, 2010.

[26] B. Avants, C. Epstein, M. Grossman, J. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Med. Image Anal.*, 12, 26–41, 2008.

[27] M. Ackerman, "The visible human project," *Proceedings of the IEEE*, 86, 504–511, 1998.

[28] A. Myronenko, X. Song, "Intensity-based image registration by minimizing residual complexity," *IEEE Trans Med. Imag.,* 29(11), 1882–1891, 2010.

[29] C.A. Cocosco, V. Kollokian, R.K.-S. Kwan, A.C. Evans, "BrainWeb: Online interface to a 3D MRI simulated brain database," *NeuroImage*, vol.5, no.4, part 2/4, S425, 1997.

[30] MICCAI 2012 Challenge on Multimodal Brain Tumor Segmentation. Available online: http://www.imm.dtu.dk/projects/BRATS2012.

[31] R. Castillo *et al.*, "A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets," *Phys. Med. Biol.*, vol. 54, no. 7, pp. 1849–1870, Apr. 2009.