Cross-modal Attention for MRI and Ultrasound Volume Registration

Xinrui Song[1], Hengtao Guo[1], Xuanang Xu[1], Hanqing Chao[1], Sheng Xu[2],
Baris Turkbey[3], Bradford J. Wood[2], Ge Wang[1], and Pingkun Yan✉[1]

[1] Department of Biomedical Engineering and Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
`yanp2@rpi.edu`

[2] Center for Interventional Oncology, Radiology & Imaging Sciences, National Institutes of Health, Bethesda, MD 20892, USA

[3] Molecular Imaging Program, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA

Abstract. Prostate cancer biopsy benefits from accurate fusion of transrectal ultrasound (TRUS) and magnetic resonance (MR) images. In the past few years, convolutional neural networks (CNNs) have been proved powerful in extracting image features crucial for image registration. However, challenging applications and recent advances in computer vision suggest that CNNs are quite limited in its ability to understand spatial correspondence between features, a task in which the self-attention mechanism excels. This paper aims to develop a self-attention mechanism specifically for cross-modal image registration. Our proposed cross-modal attention block effectively maps each of the features in one volume to all features in the corresponding volume. Our experimental results demonstrate that a CNN network designed with the cross-modal attention block embedded outperforms an advanced CNN network 10 times of its size. We also incorporated visualization techniques to improve the interpretability of our network. The source code of our work is available at `https://github.com/DIAL-RPI/Attention-Reg`.

Keywords: Self-attention · Image feature · Image registration · Multi-modal · Prostate caner

1 Introduction

Image-guided interventional procedures often require registering multi-modal images to visualize and analyze complementary information. For example, prostate cancer biopsy benefits from fusing transrectal ultrasound (TRUS) imaging with magnetic resonance imaging (MRI) to optimize targeted biopsy. However, image registration is a challenging task especially for multi-modal images. Traditional multi-modal image registration relies on maximizing the mutual information between images [9,16], which performs poorly when the input images have complex textural patterns, such as in the case of MRI and ultrasound registration. Feature

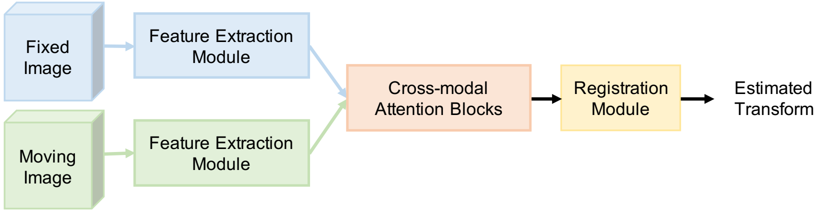
Fixed Image
Feature Extraction Module
Moving Image
Feature Extraction Module
Cross-modal Attention Blocks
Registration Module
Estimated Transform

Fig. 1. Overview of the proposed registration framework with cross-modal attention.

based methods compute the similarity between images by representing image appearances using features [6]. However, feature engineering limits the registration performance on images in different contrasts, of complicated features, and/or with strong noise.

In the past several years, deep learning has become a powerful tool for medical image registration, starting from the early works of using neural networks for similarity metric computation to direct transformation estimation [5,17]. For example, Haskins *et al.* [4] developed a deep learning metric to measure the similarity between MRI and TRUS volumes. The correspondences between the volumes is established by optimizing the similarity iteratively, which can be computationally intensive. de Vos *et al.* [14] proposed an end-to-end unsupervised image registration method to train a spatial transform network by maximizing the normalized cross correlation. Their method can directly estimate an image transformation for registration. Balakrishnan *et al.* [1] further used mean squared voxel-wise difference and local cross-correlation to train a registration network to map image features to a spatial transformation. While the way of estimating such an image transform underwent major changes, researchers also developed novel ways to supervise the network learning process. Hu *et al.* [7] trained an image registration framework in a weakly supervised fashion by minimizing the differences between segmentation labels of the fixed image and a warped moving image. Yan *et al.* [19] developed an adversarial registration framework using a discriminator to supervise the registration estimator.

The aforementioned deep learning methods map the composite features from input images directly into a spatial transformation to align them. So far, the success comes from two primary sources. One is the ability of automatically learning image representations through training a properly designed network. The other is the capability of mapping complex patterns to an image transformation. The current methods mix these two components together for image registration. However, converting image features to a spatial relationship is extremely challenging and highly data-dependent, which is the bottleneck for further improvements of the registration performance.

In this paper, we propose a novel cross-modal attention mechanism to explicitly use the spatial correspondence to improve the performance of neural networks for image registration. By extending the non-local attention mechanism [15] to an attention operation between two images, we designed a cross-

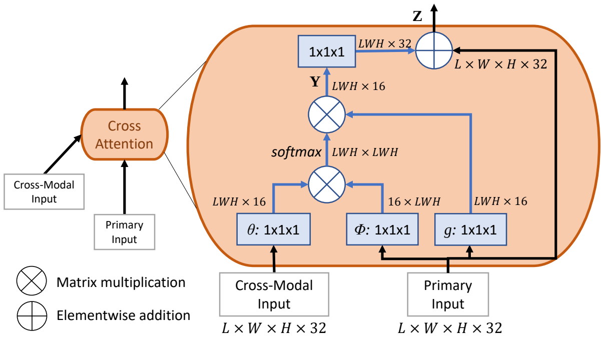

Fig. 2. The proposed cross-modal attention block.

modal attention block that is specifically oriented towards registration tasks. The attention block captures both local features and their global correspondence efficiently. Embedding this cross-modal attention block into an image registration network, as shown in Fig. 1, improves deep learning based multi-modal image registration, attaining both feature learning and correspondence establishment explicitly and synergically.

By adding the cross-modal feature correspondence, the image registration network can achieve better registration performance with a much simpler architecture. To the best of our knowledge, this is the first work to embed the non-local attention in the deep neural network for image registration.

In our experiments, we demonstrate the proposed method on the 3D MRI-TRUS fusion task, which is a very challenging cross-modality image registration problem. The proposed network was trained and tested on a dataset of 650 MRI and TRUS volume pairs. The results show that our network significantly reduced the registration error from 10.17 ± 5.75mm to 3.71 ± 1.99mm. The proposed method also outperformed state-of-the-art methods with only 1/10 to 1/5 of the number of parameters used by the competitors, as well as significantly reduced the run time.

2 Method

In this image registration application, the MRI volume is considered to be the fixed image, and the TRUS volume is the moving image. Our registration network consists of three main parts, as shown in Fig. 1. The feature extractor uses convolutional and max pooling layers to capture regional features, and down samples the input volume. Then we use the proposed cross-modal attention block to capture both local features and their global correspondence between modalities. Finally, this information is fed to the deep registrator that further fuses information from two modalities and infers the registration parameters.

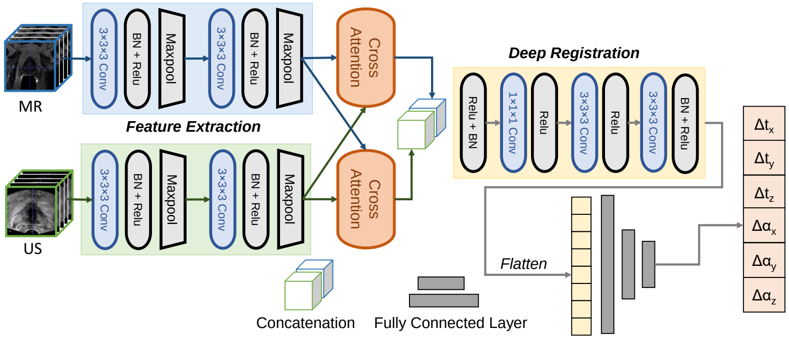

Fig. 3. Overview of the proposed network structure.

2.1 Cross-modal Attention

The proposed cross-modal attention block takes image features extracted from MRI and TRUS volumes by the preceding convolutional layers. Unlike the non-local block [15] computing self-attention on a single image, the proposed cross-modal attention block aims to establish spatial correspondences between features from two images in different modalities. Fig. 2 shows the structure of the proposed cross-modal attention block. The two input feature maps of the block are denoted as primary input $P \in \mathbb{R}^{LWH \times 32}$ and cross-modal input $C \in \mathbb{R}^{LWH \times 32}$, respectively. LWH indicates the size of each 3D feature channel after flattening. The block computes the cross-modal feature attention as

$$\mathbf{y}_i = \frac{\sum_{\forall j} f(\theta(\mathbf{c}_i)^T \phi(\mathbf{p}_j)) g(\mathbf{p}_j)}{\sum_{\forall j} f(\theta(\mathbf{c}_i)^T \phi(\mathbf{p}_j))}, \tag{1}$$

where $\mathbf{c}_i$ and $\mathbf{p}_j$ are features from $\mathbf{C}$ and $\mathbf{P}$ at location i and j, $\theta(\cdot)$, $\phi(\cdot)$ and $g(\cdot)$ are all linear embeddings, and $f(\cdot) = \exp(\cdot)$. In Eq. 1, $f(\cdot)$ computes a scalar representing correlations between the features of these two locations, $\mathbf{c}_i$ and $\mathbf{p}_j$. The result $\mathbf{y}_i$ is a normalized summary of features on all locations of $\mathbf{P}$ weighted by their correlations with the cross-modal feature on location i. Thus, the matrix $\mathbf{Y}$ composed by $\mathbf{y}_i$ integrated non-local information from $\mathbf{P}$ to every position in $\mathbf{C}$. Finally, the block's output $\mathbf{Z}$ is the sum of $\mathbf{Y}$ and $\mathbf{P}$ to allow efficient back-propagation. Therefore, the feature of a location k in $\mathbf{Z}$ summarizes non-local correlation between the entire primary feature map and location k of the cross modality feature map, as well as the information from the original primary feature map at k.

2.2 Feature Extraction and Deep Registration Modules

In the proposed network shown in Fig. 3, feature extraction modules precede the cross-modal attention block to efficiently represent the input volumes. Each feature extraction module consists of two sets of convolutional and maxpooling layers. The deep registration module fuses the concatenated outputs of the cross-modal attention blocks, and predicts the transformation for registration. Other works have used very deep neural networks to automatically learn the complex features of inputs [18]. However, since the cross-modal attention blocks help determine the spatial correspondence between the two sets of volumes, our registration module can afford to be light weighted. Thus, only three convolutional layers are used to fuse the two feature maps. The final fully connected layers convert the learnt spatial information into an estimated transformation.

2.3 Implementation Details

Due to the difficulty in representing the complex image appearances of MRI and TRUS images, surface-based and surface to volume registration methods have been investigated with considerable success [2, 13, 20]. That inspired us to replace the MRI volume with the prostate segmentation label volume in our work. The network remains the same and we only need to set the fixed image input as either MRI volume or segmentation label. The corresponding networks are named as Attention-Reg (image) and Attention-Reg (label), respectively. One advantage of using MRI prostate segmentation is that the binary representation is much more tolerant to image quality and device specificity than MRI volume. Moreover, using segmentation as input can readily extend the proposed method to other imaging modalities, like computed tomography. This implies that while we trained our segmentation guided model on MRI and ultrasound, it may potentially be used on any two modalities.

In this work, we focus on rigid transformation based registration. This decision is determined by the better accessibility of ground truth labels for rigid transformation, and the idea of focusing on network structure comparison only. Rigid transformations in this work are performed with 4×4 matrices generated from 6 degrees of freedom $\theta = \{\Delta t_x, \Delta t_y, \Delta t_z, \Delta a_x, \Delta a_y, \Delta a_z\}$. These 6 transformation parameters represent translations and rotations along the x, y, and z directions, respectively. We supervise the network by calculating the MSE (Mean Square Error) between the prediction and the ground truth parameters.

In our experiments, we included the recent methods of MSReg by Guo *et al.* [3] and DVNet by Sun *et al.* [12] as benchmarks. We used Adam optimizer [8] with maximum of 300 epochs to train all the networks including our proposed Attention-Reg approach. We used a step learning rate scheduler for MSReg training, with initial learning rate 5×10^{-5} which then decays to 0.9 every 5 epochs, as suggested in [3]. For DVNet, we used the same scheduler but with initial learning rate adjusted to 1×10^{-3}. The models were trained on a NVIDIA DGX-1 deep learning server with batch size of 16 for MSReg, and 8 for our proposed network. The testing phase and runtime benchmark were performed on a

Table 1. Performance comparison between Attention-Reg and similarity-based iterative registration methods.

Method	Initialization	Result SRE (mm)
Mutual Information [9]	8mm	8.96±1.28
SSD MIND [6]		6.42±2.86
Attention-Reg (img)		**3.63±1.86**
Attention-Reg (label)		**3.54±1.91**
Mutual Information [9]	16mm	10.07±1.40
SSD MIND [6]		6.62±2.96
Attention-Reg (img)		**4.17±2.14**
Attention-Reg (label)		**4.06±2.10**

work station equipped with NVIDIA GeForce RTX 2080 Ti and AMD Ryzen 9 3900X. Both the proposed and the MSReg methods were implemented in Python using the open source PyTorch library [10]. Our implementation of the proposed Attention-Reg is available at: `https://github.com/DIAL-RPI/Attention-Reg`.

3 Experiments and Results

3.1 Dataset and Preprocessing

In this work, we used 528 cases of MRI-TRUS volume pair for training, 66 cases for validation, and 68 cases for testing. Each case contains a T2-weighted MRI volume and a 3D ultrasound volume. Each MRI volume has 512×512×26 voxels with 0.3mm resolution in all directions. The ultrasound is reconstructed from an electro-magnetic tracked freehand 2D sweep of the prostate. The training set was generated afresh for every training epoch to boost model robustness. On the contrary, the validation set consists of 5 pre-generated initialization matrices for each case, resulting in 330 total samples. The reason for not regenerating new validation sets every epoch is to monitor the epoch-to-epoch performance in a more stable manner. For testing, we generated 40 random initialization matrices for each case. The same test set is used for all experiments.

We measured the image registration performance using surface registration error (SRE). To accurately generate a dataset of with known SRE for training and validation, we perturbed each ground truth transformation parameter randomly within the range of 5mm of translation or 6 degrees of rotation, and then scale the perturbation to a random SRE within the desired range.

3.2 Experimental Results

We first compared our approach to classical iterative registration methods. Table 1 summarizes the comparison of our method and traditional iterative registration approaches, including mutual information [9] and MIND [6] based registration as in [5]. We tested our result on two sets of initial registrations. One set is initialized at SRE=8mm, and the other set is initialized at SRE=16mm.

Table 2. Performance comparison between Attention-Reg, MSReg [3], and DVNet [12]. Both parameter count and runtime were measured per stage. SRE values are in mm.

Method	Initial.	Stage 1	Stage 2	#Parameters	Runtime
DVNet [12]	[0,20mm]	4.77±3.17	-	5,275,832	3ms
MSReg [3]		4.75±2.63	4.04±2.30	16,106,076	6ms
Attention-Reg (image)		4.50±2.58	3.71±1.99	1,248,777	3ms
Attention-Reg (label)		**4.44±2.32**	**3.60±2.01**	1,248,777	3ms
Feature-Reg (image)	[0,20mm]	5.14±2.58	-	1,244,393	3ms
Feature-Reg (label)		5.22±2.81	-	1,244,393	3ms

The results of our proposed models are averaged from 6,800 test samples, with 68 cases of MRI-TRUS volume pair and 100 initialization matrices each. We used this large test set to improve the robustness of the evaluation. Attention-Reg (img) stands for the the registration result of our proposed network with MRI volume as the input of fixed image, whereas Attention-Reg (label) uses MRI prostate segmentation label as the fixed image. In both test scenarios, our methods outperformed the traditional approaches significantly (p <0.001 under t-test). It is also worth noting that when using MRI prostate segmentation as the fixed image, the performance of our network is slightly improved with statistical significance (p <0.001 under t-test).

Table 2 lists the results of our method and other end-to-end rigid registration techniques, including MSReg by Guo *et al.* [3] and DVNet by Sun *et al.* [12]. The ResNeXt structure that Guo *et al.* adopted is one of the more advanced variations of CNN [18], adding more weight to this comparison. The 2D CNN network in DVNet [12] treats 3D volumes as patches of 2D images, a lighter approach in handling 3D volume registration. We tested these networks on 2,720 testing samples, which consists of 68 cases with 40 initialization positions for each case. To better compare our Attention-Reg with MSReg [3], which used two consecutive networks to boost performance, we also trained our network twice on two differently distributed training sets. The model for the 1^{st} stage was trained and tested on a generated dataset with initial SRE uniformly distributed within the range of $[0, 20mm]$, and the range for the 2^{nd} stage was set to be $SRE \in [0, 8mm]$. The trained networks were concatenated together to form a two-stage registration network.

As shown in the top part of Table 2, our cross-modal attention network outperformed MSReg in both registration stages. Furthermore, the better result was achieved with only 1/10 the number of parameters, and half the runtime. The significantly smaller model and simpler calculation demonstrate that the proposed cross-modal attention block can efficiently capture key features of the image registration task. Again, we observed that the performance of our network with segmentation label as input was consistently better, with significantly reduced SRE when compared to MSReg (p <0.001) in both stages.

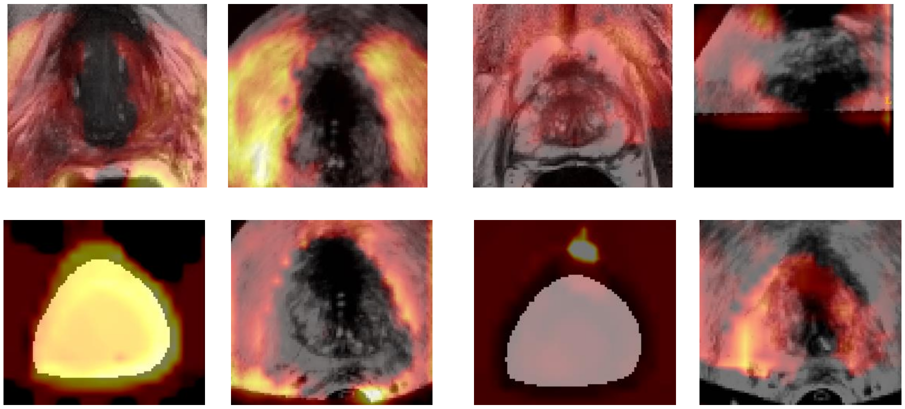

Fig. 4. Grad-CAM visualization of four pairs of feature maps resulting from the multi-modal attention blocks of **(top)** Attention-Reg (image) and **(bottom)** Attention-Reg (label). The image on the left and right in each pair are from the fixed and moving images, respectively.

To demonstrate the contribution of the proposed cross-modal attention block, we trained our Attention-Reg network without the attention block, *i.e.*, directly concatenating the outputs of feature extraction modules and feeding to the deep registration module. The results are shown in the bottom half of Table 2, which prove the importance of the proposed cross-modal attention block. Without the attention module, the registration performance under both settings was significantly reduced (p <0.001 with paired t-test). Also, note that without the attention block, using segmentation label as fixed image no longer has an advantage over MRI volume. We speculate that this is also caused by the loss of attention block, which establishes a sensible spatial correlation between the MRI segmentation and the ultrasound volume, as shown in Fig. 4.

To help understand the function of cross-modal attention blocks, we employed Grad-CAM [11] to visualize the output of the two multi-modal attention blocks. Similar with Grad-CAM, we used the preceding CNN layer's weight gradient to scale the importance of each feature map channel, and thereby acquired a single volume that represents the output of the multi-modal attention block. Fig. 4 shows the visualization result. It is apparent that both MRI and ultrasound features are roughly the shape and location of the corresponding ultrasound frame. This means that the network is focusing on the same region of information in both volumes.

4 Conclusion

This paper introduced a novel attention mechanism for the task of medical image registration. By comparing the proposed network with other classical methods and purely CNN-based networks up to ten times of its size, we demonstrated the effectiveness of the new cross-modal attention block. To emphasize the importance of prostate boundary, we also quantitatively evaluated the effect of replacing an MRI volume with its segmentation mask as network input. Our proposed methods have led to significant improvements in image registration accuracy over the previous registration methods. Through feature map visualization, we observed that the network indeed extracted meaningful features to guide image registration. We expect to see our methods tested out in other medical image registration settings in the future with such improvement in accuracy and efficiency, and interpretability.

References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging **38**(8), 1788–1800 (2019)
2. Bashkanov, O., Meyer, A., Schindele, D., Schostak, M., Tönnies, K., Hansen, C., Rak, M.: Learning multi-modal volumetric prostate registration with weak intersubject spatial correspondence (2021)
3. Guo, H., Kruger, M., Xu, S., Wood, B.J., Yan, P.: Deep adaptive registration of multi-modal prostate images. Computerized Medical Imaging and Graphics **84**, 101769 (2020)
4. Haskins, G., Kruecker, J., Kruger, U., Xu, S., Pinto, P.A., Wood, B.J., Yan, P.: Learning deep similarity metric for 3d mr–trus image registration. International journal of computer assisted radiology and surgery **14**(3), 417–425 (2019)
5. Haskins, G., Kruger, U., Yan, P.: Deep learning in medical image registration: a survey. Machine Vision and Applications **31**(1), 8 (2020)
6. Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, S.M., Schnabel, J.A.: MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. Medical Image Analysis **16**(7), 1423 – 1435 (2012)
7. Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C.M., Emberton, M., et al.: Weakly-supervised convolutional neural networks for multimodal image registration. Medical image analysis **49**, 1–13 (2018)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P.: Multimodality image registration by maximization of mutual information. IEEE TMI **16**(2), 187–198 (1997)
10. Paszke, A., Gross, S., Chintala, S., et al.: Automatic differentiation in pytorch. In: NIPS 2017 Workshop Autodiff. pp. 1–4 (2017)
11. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
12. Sun, Y., Moelker, A., Niessen, W.J., van Walsum, T.: Towards robust ct-ultrasound registration using deep learning methods. In: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, pp. 43–51. Springer (2018)

13. Thomson, B.R., Smit, J.N., Ivashchenko, O.V., Kok, N.F., Kuhlmann, K.F., Ruers, T.J., Fusaglia, M.: MR-to-US registration using multiclass segmentation of hepatic vasculature with a reduced 3d u-net. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 275–284. Springer (2020)
14. de Vos, B.D., Berendsen, F.F., Viergever, M.A., Staring, M., Išgum, I.: End-to-end unsupervised deformable image registration with a convolutional neural network. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 204–212. Springer (2017)
15. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
16. Wells, W.M., Viola, P., Atsumi, H., Nakajima, S., Kikinis, R.: Multi-modal volume registration by maximization of mutual information. Medical image analysis **1**(1), 35–51 (1996)
17. Wu, G., Kim, M., Wang, Q., Gao, Y., Liao, S., Shen, D.: Unsupervised deep feature learning for deformable registration of mr brain images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 649–656. Springer (2013)
18. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on CVPR. pp. 1492–1500 (2017)
19. Yan, P., Xu, S., Rastinehad, A.R., Wood, B.J.: Adversarial image registration with application for MR and TRUS image fusion. In: International Workshop on Machine Learning in Medical Imaging. pp. 197–204. Springer (2018)
20. Zhang, Y., Bi, J., Zhang, W., Du, H., Xu, Y.: Recent advances in registration methods for mri-trus fusion image-guided interventions of prostate. Recent Patents on Engineering **11**(2), 115–124 (2017)