Ultrasound to CT rigid image registration using CNN for the HIFU treatment of heart arrhythmias

Batoul Dahman, Francis Bessière, Jean-Louis Dillenseger

Ultrasound to CT rigid image registration using CNN for the HIFU treatment of heart arrhythmias

Batoul Dahman[a], Francis Bessière [b], and Jean-Louis Dillenseger[a]

[a]Univ Rennes, INSERM, LTSI - UMR 1099, F-35000 Rennes, France
[b]LabTAU, INSERM, Centre Léon Bérard, Univ Lyon, F-69003 Lyon, France

ABSTRACT

Image-guided thermal ablations have become an important therapeutic option for patient with cardiac arrhythmia, it is minimally-invasive and provides better and faster patient recovery. However, to enhance the ablation guidance, the therapist needs to link by image registration the intraoperative images to the high-resolution anatomical preoperative imaging, in which the ablation path has been defined. In this work, we present a convolutional neural networks (CNNs) framework for transesophageal ultrasound/computed tomography image registration to solve the problem of high computation time of the classical iterative methods, which is not suitable for a real-time application. We propose the following process: we first pass the input moving and fixed image pairs through a siamese architecture consisting of convolutional layers, thus extracting features of moving and fixed maps analogous to dense local descriptors, then matching the feature maps, and finally pass this correspondence feature map into a registration network, which directly outputs the registration parameters set of the rigid registration. Accuracy of the registration is quantified based on the Target Registration Error (TRE) for specific anatomical landmarks. Results of the registration process show a median TRE of 2.2 mm for all the fiducial points, and the registration computation time was around 3 ms comparing to the classic iterative methods which takes around 70 seconds for one image pair. In our future work we are going to perform our approach on 2D/3D learning-based registration to refine the estimation of the transesophageal probe pose in the 3D preoperative volume.

Keywords: Rigid-body image registration, Ultrasound Imaging-cardiac, Multimodal image fusion.

1. INTRODUCTION

In the last 20 years, cardiac arrhythmias (heart rhythm problems) have become one of the most important public health problems and a significant cause of increasing health care costs in Western countries. Arrhythmias occur when the electrical impulses that coordinate the heartbeats do not propagate properly and cause rapid, uncoordinated, weak contractions of the heart 1. Moreover, in the long term, arrhythmias may increase the risk of developing other effects such as stroke and heart failures.

Treatments for arrhythmia can take many different forms, depending on the type and severity of the irregular heartbeat and its cause. Medication treatments for arrhythmia are called chemical cardioversion. The patient receives antiarrhythmic medicine orally or intravenously. But when medications don't offer a solution, an ablation treatment for arrhythmia based on thermal Radio Frequency (RF) can be considered. This ablation technique can be performed on different targets of the heart 2. For example, in atrial/ventricular fibrillation ablation, small scars are intentionally created on the cardiac wall to break up the electrical signals propagation paths that cause the irregular heartbeats. During catheter ablation, a small RF ablator is inserted into the heart, usually through a vein and then small areas of tissue that may be causing or propagating the arrhythmia are necrotized by heating 2. However, RF lesions are not always trans-mural because a good contact between the tissue and the catheter is difficult to achieve 3, and there are also many complications associated with endocardial therapy.

Recently, an ultrasound-guided transesophageal high-intensity focused ultrasound (HIFU) device has been designed to perform lesions from the esophagus into the heart walls in various locations, while preserving intervening tissues 4–6. This device is less invasive as RF and offer better targeting ability. This ablation device also

Corresponding author: batoul.dahman@univ-rennes1.fr

integrates in its middle a 2D US imaging transducers for intraoperative guidance purpose. This imaging device produces an US image perpendicular to the device axis. Classically, the planning of the intervention (ablation path) is defined on a high-resolution preoperative anatomical imaging (CT/MRI). The goal of the cardiologist is then to follow the pre-defined path with the help of the 2D US image alone. But this imaging device used for guidance offers only a very limited view of the anatomy. It seems the important to merge this two information (3D preoperative CT and 2D intraoperative US) by registration in order to improve the guidance of the gesture 7. In a previous work of our team, Sandoval *et al.* proposed a solution to locate an intraoperative 2D US slice within the 3D CT volume 8. Their method starts from the anatomical hypothesis that the intraoperative 2D US slices are perpendicular to the esophagus axis. So, they reformat the preoperative 3D CT volume into 2D CT slices perpendicular to the esophagus axis. For each 2D CT slice within a candidate zone, they perform a 2D US to 2D CT image-based registration. They use the classical scheme in which an optimizer iteratively searches the transformation (rigid in this case) which gives the highest similarity (Normalized Mutual Information in their case). At the end, the 2D CT slice with the highest similarity after registration is considered as the final result. One of the main drawbacks of this method is that the iterative 2D/2D registration process is too time consuming to be used in the operating room. Recent studies have demonstrated the potential of deep learning methods in directly solving the registration problem without any iterative process. but the main problem of deep learning registration methods is the lack of ground truth for learning. This is why many methods use unsupervised learning 9. However, if the ground truth is known, supervised methods can be considered. Among the class of rigid registration supervised methods, Miao *et al.* 10 are the first to use deep learning to predict rigid transformation parameters via hierarchical learning. They use a CNN to predict the transformation matrix associated with the rigid registration of 2D/3D X-ray attenuation maps and 2D X-ray images. This approach outperformed the classical image and optimization-based registration approaches in terms of both accuracy and computational efficiency. Recently, Chee *et al.* 11 use a CNN to predict the transformation parameters used to rigidly register 3D brain MR volumes. In their framework called Affine Image Registration network (AIRNet), the Mean Square Error (MSE) between the predicted and ground truth affine transforms is used to train the network. Salehi *et al.* 12 use a deep residual regression network, a correction network, and a bivariate geodesic distance-based loss function to rigidly register T1 and T2 weighted 3D fetal brain MRs for atlas construction. However, these methods mainly concern monomodal images.

In this work, we propose a CNN framework for CT/US image registration, using Deep Features representation for a supervised rigid transformation estimation.

2. METHOD

The heart is a moving organ. However, some characteristics of the cardiac movement allowed us to consider a rigid registration scheme. First, we had at our disposal a Cine CT from a patient's heart composed of 20 volumes at each 5% phase of the RR interval. Second, we are interested in ventricular fibrillation. During its diastolic phase, the ventricle is relatively stationary. The HIFU treatment will be shot in this phase to have a fixed focal point in relation to the organ and thus avoid a dispersion of heat prejudicial to the necrosis of the tissues. So, a quasi-static ventricle pose can be considered. Moreover, on our US system, the acquisition is synchronized with the ECG. Thus, it is relatively easy to create pairs of US/CT images at the same phase and so to consider rigid registration.

In this section, we will introduce the proposed framework for estimating the transformation parameters of a rigid image registration between a preoperative CT slice and an intraoperative US image. In our case and following the approach described in 8, only a 2D rigid transform with three Degrees of Freedom (DOF) – one rotation and 2 translations- has to be estimated.

The main idea is to estimate the registration that best aligns some common characteristics of the images. The information contained in the two images is of a very different nature (gray levels proportional to the X-ray absorption coefficient of the tissues for the CT and information formed by the reflection of waves on surfaces and speckle for US). We must therefore first extract from the two imaging modalities a common information, in our case the shapes of the organs, before performing registration. Therefore, we propose the following process (see Figure 1):

(i) Descriptors are extracted from the moving I_M and the fixed I_F images using Deep Learning. For this, I_M and I_F are passed through a siamese CNN architecture (ResNet18) consisting of convolutional layers, thus extracting two feature maps f_M, and f_F which are analogous to dense local descriptors;

(ii) These feature maps are combined in a concatenating layer;

(iii) This corresponding feature maps are the set as input into a convolutional network which directly outputs the parameters set T (two translations, one rotation) of the rigid registration. This framework should be trainable end-to-end for rigid registration task.

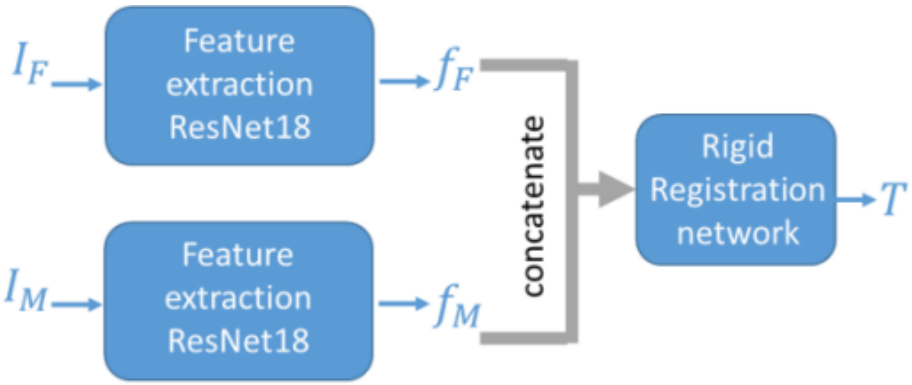

Figure 1. The overall of the proposed framework.

2.1 Feature extraction

The first step of the framework is feature extraction. Features extraction is a classical tool in deep learning. One common feature extraction technique is to feed the image to a conventional pre-trained neural network and use the representation for that particular image in the intermediate layers of the neural network. We used the ResNet18, which is one of the most efficient standard feature extraction model that can be used in many medical applications 13. This model handles the vanishing or exploding gradient problem when the CNN goes deeper.

Resnet18 can be found implemented in PyTorch. This implementation offers a version with the weights pre-trained for feature extraction on ImageNet, the large benchmark database, Each of our input modalities has its own image characteristics. Thus we passed each of the two images to be registered in its own network. Each of these networks produces a feature map f composed of image descriptors.

2.2 Matching

These two feature maps should be combined across images as a single tensor to input it to the rigid transformation parameters estimation network. To achieve this, a concatenation of descriptors along the channel dimensions is performed in the concatenation layer.

2.3 Registration network

We will present the network architecture which consists of three blocks of convolutional layers using a kernel size of 5, each followed by batch normalization layers, and a rectified linear unit (ReLU). The last layer is a fully connected layer to estimate the rigid registration parameters. The network expects the concatenated map extracted from moving and fixed images as its input, and directly estimates the parameters (t_x, t_y, and θ) of the rigid transformation that connects these images. The idea behind this architecture is that the estimation is performed in a bottom-up manner where early convolutional layers vote for candidate transformations, and these are then processed by the later layers to aggregate the votes. The first convolutional layers can also enforce local neighborhood consensus by learning filters which only fire if nearby descriptors in image I_M are matched to nearby descriptors in image I_F.

2.4 Training

We consider a supervised learning scheme, the training datasets include pairwise images of moving CT and fixed US image pairs, and their associated rigid geometric transformation parameters as ground truth GT. The training loss function can be formulated as the $L2$ norm of the error between the GT (T_{GT}) and the predicted transformation parameter T_{Est}.

$$L = \alpha \|t_{GT} - t_{Est}\|^2 + \beta \|\vartheta_{GT} - \vartheta_{Est}\|^2 \tag{1}$$

With t_{GT} and t_{Est} the translation vector of respectively the ground truth transformation and the estimated one expressed in mm, ϑ_{GT} and ϑ_{Est} the rotation angle expressed in degrees and α and β are weights controlling the balance between the translation and the rotation losses. The choice to use mm for translation and degrees for rotation gave a certain coherence and balance between these parameters. Indeed, an error of 1 degree in rotation leads to a displacement of 1 mm at 60 mm from the center of rotation. Because of this consistency, α and β could be set to 1.

The network is trained by the gradient of the loss function with respect to the estimated rigid parameters (t_x, t_y, and ϑ). This gradient is then used to minimize the loss function by using backpropagation and Stochastic Gradient Descent.

After training, the network can be applied for registration of unseen image pairs. We implemented the network using PyTorch and we trained it on a NVIDIA TitanX GPU with 10000 iterations.

3. RESULTS

3.1 dataset

Because ground truth cannot be obtained on real data, *i.e.* we can never ensure that a real US image is perfectly associated to a real CT plane, we decided to simulate US images from CT data.

Our study has been conducted on a public available dataset of twenty contrast enhanced cardiac CT volumes 14,15. All the data were obtained from two state-of-the-art 64-slice CT scanners (Philips Medical Systems, Netherlands) using a standard coronary CT angiography protocol at two sites affiliated to Shanghai Shuguang Hospital. Images were acquired in the axial view, covering the whole heart from the upper abdominal to the aortic arch. The in-plane resolution was about 0.44×0.44 mm and the average slice thickness was 0.60 mm.

In these volumes, the esophagus was roughly segmented manually.

From these volumes, we create a set of corresponding pair of CT and US images with known transformation (Figure 2) First we extracted randomly 4000 2D CT oblique cut planes from the 20 CT volumes (200 images per volume). For this we:

1. We choose randomly 4000 initial poses along the esophagus axes within the 20 CT volumes.
2. For each pose we create a new referential by setting some randomly transformations near these initial poses with some translations within ± 10 mm and rotations within ± 15 degree around each coordinate axis.
3. The $x - y$ plane of this referential will serves as the fixed CT image I_{CT}. The origin of the $x - y$ plane served also as origin of I_{CT}.
4. For each I_{CT}, we simulated the corresponding US images with the method described in 16 that predict the appearance and properties of a B-scan ultrasound image from a probe origin pose, the point spread function of the US device, the acoustical impedance of the tissues and some tissue-adapted distribution of point scatterers. We randomly define the pose of the simulated US origin probe within a range of ± 10 mm from the I_{CT} origin and we randomly rotate the probe in a range of ± 15 degree around this origin. These two translation and rotation defines the ground truth transformation T_{GT} between the CT and US image.
5. So, for each I_{CT} we get an US image I_{US} and a transformation ground truth T_{GT}.

From this dataset, the network was trained by selecting the 3600 pairs of corresponding I_CT and I_{US} slices from 18 of the 20 cardiac CT scans. For validation, we used 400 image pairs from the 2 remaining volumes

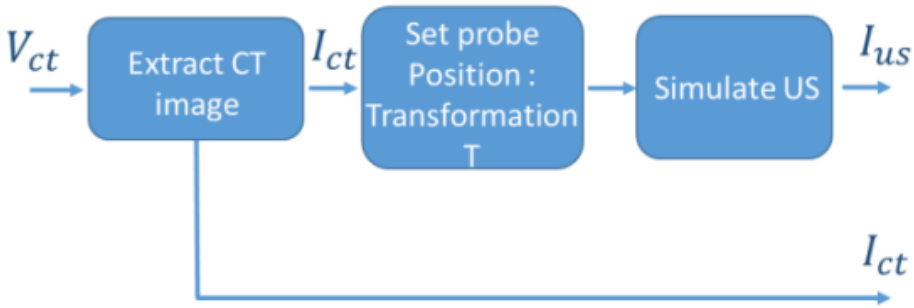

Figure 2. Datasets creation workflow

3.2 Evaluation

We compared the registration results obtained by the proposed methods to these obtained by the classical iterative rigid registration method implemented in the SimpleITK Library 17. We used Normalized Mutual Information as similarity measure because it has been found as one of the most suited for our CT/US registration problem 18, We also used 2D Euler transform to presents the spatial mapping of points from the fixed image space to points in the moving image space. We evaluate the performance of the method in terms of computation time and registration accuracy.

3.2.1 Computation time

The mean registration computation time for all the 400 image pairs is now lower than 3 ms for each image pair. which is suitable for a real-time application. For comparison, the classic iterative method takes around 70 seconds to register one image pair.

3.2.2 Transformation estimation error

We compared the parameters of the transformation obtained by our proposed methods with the ground truth (GT). A transformation is composed by a translation vector t (t_x, t_y) and a rotation ($vartheta$). We evaluate separately the translation errors and the rotation errors between the estimated pose of each of the 400 validation image pairs and their associated GT.

We estimated the translation errors by equation (2), where $t_{GT,i}$ and $t_{Est,i}$ are the translation parameters of respectively the GT and the estimated one.

The rotation error has been estimated using the angle difference in degrees estimated between the orientation parameter of the pose of respectively the GT and the estimated rotation angle (equation (3)), where $\vartheta_{GT,i}$ and $\vartheta_{EST,i}$ are the angles that encode the orientation parameter of the pose of respectively the GT and the estimated rotation.

$$\|t_{GT,i} - t_{EST,i}\|^2 \tag{2}$$

$$|\vartheta_{GT,i} - \vartheta_{EST,i}| \tag{3}$$

Figure 3.a shows the boxplots of the 400 translation errors from our CNN-based registration and the classical iterative one. The median translation errors are 1.1 mm using CNN and 1.2 mm using the classical approach.

The boxplots of the rotation errors in Figure 3.b show that the median rotation errors are 2.1 degree using CNN and 2.4 degree when using the iterative classical method.

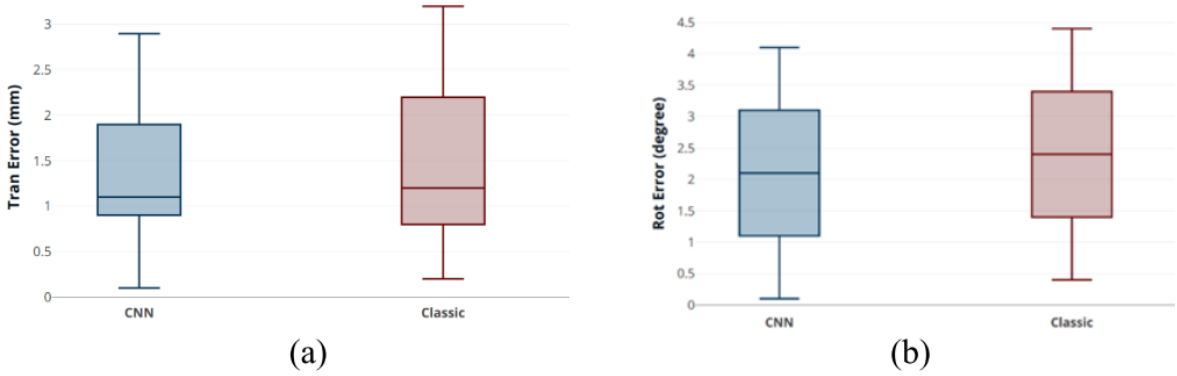

Figure 3. Box plots of a) the Translation Estimation Errors, b) the Rotation Estimation Errors.

3.2.3 Target Registration Error (TRE)

The validation can also be done by estimating the registration errors on some fiducial markers. To quantify the error, we defined eight specific feature points (or landmarks) P_j in the US fixed images, and we used the two transformations matrices, the estimated T_{Est} and the GT T_{GT} one to project these points in the corresponding CT images: $P_{Est,j} = T_{Est}P_j$ and $P_{GT,j} = T_{GT}P_j$. The Euclidean distance between the corresponding projected points $P_{Est,j}$ and $P_{GT,j}$ gives the TRE. Figure 4 shows the boxplot of the TREs for all the 8 fiducial points of all the 400 test images. Quantitative results show a median TRE of 2.2 mm for all the fiducial points of all the 400 test images using CNN, and 2.7 mm using the classical method.

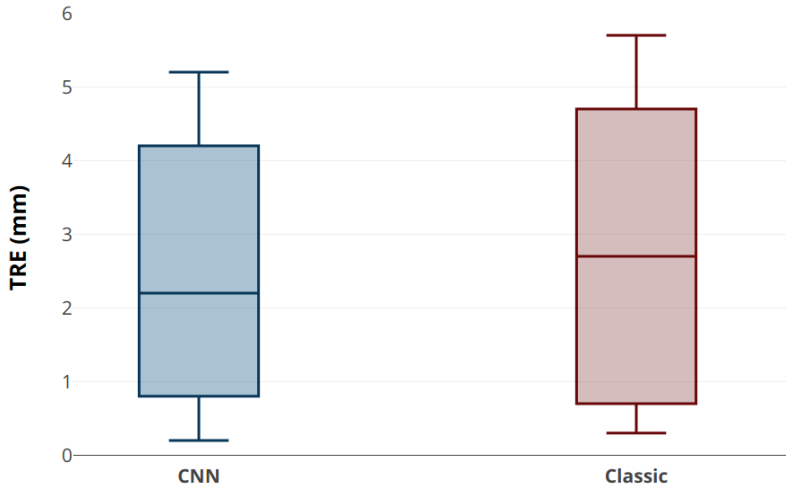

Figure 4. Boxplots of the mean Target Registration Errors

3.2.4 Visual validation

Figure 5 shows a visual comparison between the simulated fixed US image, and the corresponding moving CT image pair. It shows the overlap between the moving CT image and the fixed US image: a) before registration b) after registration with the proposed method. Visually, the results obtained by the proposed method seems to provide a good alignment, this can be seen for example at the probe center, and the bottom of the image on the thoracic chest.

3.2.5 Discussion

From the previous quantitative results, we can conclude that on the one hand, the registration accuracy obtained by CNN is of the same order as that obtained by the classical iterative method. The results obtained by CNN

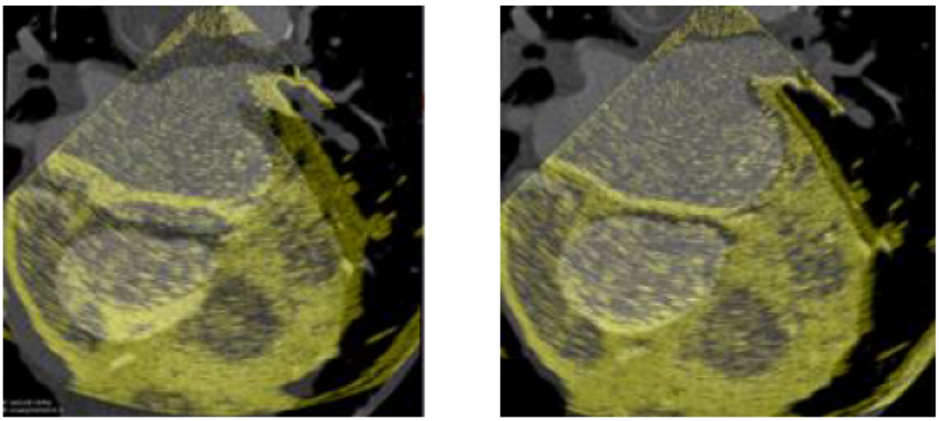

Figure 5. An example of the registration of an image pair. The overlap between the moving CT image and the fixed US image a) before registration, and b) after registration with the proposed method

are even slightly better, even if statistically this improvement is not significant. Compared to other methods in the literature, the global target registration error (TRE) of 2.2 mm is on the same range of magnitude as those reported in 8, 10.

On the other hand, CNN allows to strongly accelerate the processing time. The registration between two images takes only 3 ms (instead of 70 s for the classical iterative method). This gain in computation time allows us to consider implementing the 3D CT/2D US registration technique proposed by 8 in clinical practice.

These results were obtained from simulated US images. We are fully aware that there are differences between simulated and real US images (signal attenuation compensation, acoustic shadowing, post processing of real US images...). However, we found in a previous study that a method developed on simulated data performed well on real data 8, 10. We are therefore confident that our method will also work on real data.

4. CONCLUSION

In this paper, we presented a deep feature learning-based approach for the registration of transesophageal US/CT cardiac images. The results showed a strong improvement in terms of computation time with a promising result in terms of registration accuracy. In future work, we will integrate the features learning approach to a minimally-invasive HIFU procedure to improve the therapy planning and guidance. We will apply our approach our approach on 2D/3D learning-based registration to refine the estimation of the transesophageal probe pose placement in the 3D preoperative volume. Finally, we will include data from physical phantom and real-patients to evaluate the contribution of our registration scheme to the therapy guidance.

ACKNOWLEDGMENTS

This work was part of the CHORUS (ANR-17-CE19-0017) project which have been supported by the French National Research Agency (ANR).

REFERENCES

[1] Sinclair-Smith, B. C., "Electrical reversion of cardiac arrhythmias.," *South Med J* **65**(3), 289–293 (1972).
[2] Huang, S. K. S. and Wood, M. A., [*Catheter ablation of cardiac arrhythmias e-book*], Elsevier Health Sciences (2014).

[3] Reddy, V. Y., Shah, D., Kautzner, J., Schmidt, B., et al., "The relationship between contact force and clinical outcome during radiofrequency catheter ablation of atrial fibrillation in the TOCCATA study," *Heart Rhythm* **9**(11), 1789–1795 (2012).

[4] Pichardo, S. and Hynynen, K., "New design for an endoesophageal sector-based array for the treatment of atrial fibrillation: a parametric simulation study," *IEEE Trans Ultrason Ferroelectr Freq Control* **56**(3), 600–612 (2009).

[5] Constanciel, E., N'Djin, A., Bessière, F., et al., "Design and evaluation of a transesophageal hifu probe for ultrasound-guided cardiac ablation: simulation of a hifu mini-maze procedure and preliminary ex vivo trials.," *IEEE Trans Ultrason Ferroelectr Freq Control* **60**, 1868–1883 (2013).

[6] Bessière, F., N'djin, W. A., Constanciel-Colas, E., et al., "Ultrasound-guided transesophageal high-intensity focused ultrasound cardiac ablation in a beating heart: a pilot feasibility study in pigs," *Ultrasound Med Biol* **42**(8), 1848–1861 (2016).

[7] Markelj, P., Tomaževič, D., Likar, B., and Pernuš, F., "A review of 3d/2d registration methods for image-guided interventions.," *Med Image Anal* **16**, 642–661 (2012).

[8] Sandoval, Z., Castro, M., Alirezaie, J., Lafon, C., Bessière, F., and Dillenseger, J.-L., "Transesophageal 2D ultrasound to 3D computed tomography registration for the guidance of a cardiac arrhythmia therapy," *Phys Med Biol* **63**(15), 155007 (2018).

[9] de Vos, B. D., Berendsen, F. F., Viergever, M. A., et al., "End-to-end unsupervised deformable image registration with a convolutional neural network," in [*Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017*], 204–212 (2017).

[10] Miao, S., Wang, Z. J., and Liao, R., "A CNN regression approach for real-time 2D/3D registration," *IEEE T Med Imaging* **35**(5), 1352–1363 (2016).

[11] Chee, E. and Wu, Z., "AIRNet: Self-supervised affine registration for 3D medical images using neural networks," *CoRR* **abs/1810.02583** (2018).

[12] Mohseni Salehi, S. S., Khan, S., Erdogmus, D., and Gholipour, A., "Real-time deep pose estimation with geodesic loss for image-to-template rigid registration," *IEEE T Med Imaging* **38**(2), 470–481 (2019).

[13] Lopes, U. K. and Valiati, J. F., "Pre-trained convolutional neural networks as feature extractors for tuberculosis detection.," *Computers in biology and medicine* **89**, 135–143 (Oct. 2017).

[14] Zhuang, X., "Challenges and methodologies of fully automatic whole heart segmentation: a review," *J Healthc Eng* **4**(3), 371–408 (2013).

[15] Zhuang, X. and Shen, J., "Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI," *Med Image Anal* **31**, 77–87 (2016).

[16] Dillenseger, J.-L., Laguitton, S., and Delabrousse, E., "Fast simulation of ultrasound images from a CT volume.," *Comput Biol Med* **39**(2), 180–186 (2009).

[17] Lowekamp, B. C., Chen, D. T., Ibáñez, L., and Blezek, D., "The design of SimpleITK," *Front Neuroinf* **7** (2013).

[18] Sandoval, Z. and Dillenseger, J.-L., "Evaluation of computed tomography to ultrasound 2D image registration for atrial fibrillation treatment," in [*Computing in Cardiology*], 245–248 (Sept. 2013).