Landmark-Guided Rigid Registration for Temporomandibular Joint MRI-CBCT Images with Large Field-of-View Difference

Jupeng Li[1](✉), Yinghui Wang[2], Shuai Wang[1], Kai Zhang[1], and Gang Li[2]

[1] School of Electronic and Information Engineering,
Beijing Jiaotong University, Beijing 100044, China
lijupeng@bjtu.edu.cn

[2] Department of Oral and Maxillofacial Radiology, Peking University School
and Hospital of Stomatology, Beijing 100081, China

Abstract. Fused MRI-CBCT images provide desirable complementary information of the articular disc and condyle surface for optimum diagnosis, has been shown to be accurate and reliable in Temporomandibular Disorders (TMD) assessment. But field-of-view difference between multi-modality images brings challenges to conventional registration algorithms. In this paper, we proposed a landmark-guided learning method for Temporomandibular Joint (TMJ) MRI-CBCT images registration. First, end-to-end landmark localization network was used to detect correspondence landmark pairs in the different modality images to generate the landmark guidance information. Then taking image patches centered landmarks as input, an unsupervised learning network regresses the rigid transformation matrix using mutual information as a measure of similarity between image patches. Finally combined landmarks coordinates with the rigid transformation matrix, the whole image registration can be realized. Experiment results demonstrate that our approach achieves better overall performance on registration of images from different patients and modalities with 100x speed-up in execution time.

Keywords: Multi-modality image registration · Temporomandibular joint · MRI-CBCT images · Large field-of-view difference · Landmark-guided

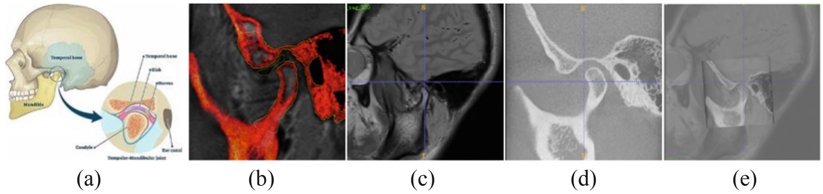
1 Introduction

TMJ is a synovial joint that contains an articular disc (shown in Fig. 1a), which allows for hinge and sliding movements [1]. TMD is an umbrella term covering pain and dysfunction of the muscles of mastication (the muscles that move the jaw) and the TMJ. TMD is common in adults; as many as one third of adults report having one or more symptoms, which include jaw or neck pain, headache, and clicking sound or grating within the joint. Although TMD is not life-threatening, it can be detrimental to quality of life; this is because the symptoms can become chronic and difficult to manage. In addition to the observer's expertise, clear image information is a substantial factor that leads to correct

C. Lian et al. (Eds.): MLMI 2021, LNCS 12966, pp. 527–536, 2021.
https://doi.org/10.1007/978-3-030-87589-3_54

diagnosis of this intractable disease [2]. The articular disc is best depicted on magnetic resonance imaging (MRI) and osseous surfaces are best seen in cone beam CT (CBCT). The fused MRI-CBCT image (see Fig. 1b) provides desirable complementary information of the articular disc and condyle surfaces for optimum diagnosis. The registration process to generate fused images has been shown to be accurate and reliable in TMD assessment [3].

Unlike registration of single-modality images, multi-modality images registration between MRI and CBCT is challenging due to significant differences in voxel size, pixel intensity, anatomical structure identification, image orientation and field-of-view (see Fig. 1c–e). Only few articles discussing MRI and CT (CBCT) image registration for TMJ visualization or assessment were published within the last 7 years [4]. Lin was the first to explore the 3D rendering of mandible from MRI and CT registered images with 12 fiducial markers attached to the facial skin-surface [5]. In a brief clinical report, Dai chose one sagittal slice of TMJ MRI and CT images from a previous study, as an example, to illustrate a hybrid image of TMJ via Photoshop® software [6]. Al-Saleh published the first study that employed MRI and CBCT registered images to assess diagnostic reliability of TMJ pathology [3]. They evaluated the quality of two techniques of image registration, extrinsic (fiducial marker-based) versus intrinsic (voxel value mutual information based) in 20 TMJ images. In a latest report, Ma, one author of this article, imported the DICOM format data of CT/CBCT and MRI into Amira® to realize automatic/semi-automatic registration of multi-modality images by adjusting the registration parameters [7].



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

Fig. 1. Anatomical structure and images registration of TMJ. (a) Anatomical structure, (b) fused MRI-CBCT image, (c) MRI image, (d) CBCT image, and (e) registered image. From these images, we can see huge field-of-view difference between different modality images.

Related Works: Deep learning methods have been shown strong advantages in medical image registration [8]. In order to evaluate the posture and position of the implant during surgery, Miao proposed a hierarchical regression model to achieve six transformation parameters for real-time 2D/3D registration, in which ground truth data is synthesized by transforming aligned data [9]. Chee proposed a self-supervised affine image registration network (AIRNet) for 3D medical images that is designed to directly estimate the transformation parameters between two input images, in which the synthetic dataset was used for the training of the model [10]. The difficult nature of the acquisition of reliable ground truth has motivated research groups to explore unsupervised approaches for image registration transformation estimation [8]. Kori proposed an unsupervised image registration framework for multi-modality MRI image affine registration. Pre-trained VGG-19 was used for feature extraction followed by a key point detector. These key points were fed to the Multi-Layered Perceptron (MLP) based regression module so as to estimate the affine transformation parameters trained by generated set of random data points [11]. In order to register arbitrarily oriented reconstructed images of fetuses scanned in-utero at a wide gestational age range to a standard atlas space, Salehi proposed regression CNNs that learn to predict the angle-axis representation of 3D rotations and translations using image features. They compared mean square error and geodesic loss to train regression CNNs for 3D pose estimation in slice-to-volume registration and volume-to-volume registration [12]. Combined with unsupervised network, coarse-to-fine multi-scale iterative framework and image deformation, Shu proposed an unsupervised network for microscopic image rigid registration. The network optimizes its parameters directly by minimizing the mean square error loss between registered image and reference image without ground truth [13]. As far as we know, no research work involves the problem of different field-of-views in multi-modality medical image registration, which bring difficulties to the current learning-based registration methods.

Contribution: The main contributions of this work are summarized as follows: (1) *Landmark-guided mechanism* was introduced to effectively register MRI-CBCT images of TMJ with large different field-of-views, without any prior assumption on the image pairs. (2) Compared with affine matrix learning methods for rigid images registration, our image spatial transform regression network predicts *real rigid transformation* for multi-modality images.

2 Method

2.1 Overall Framework

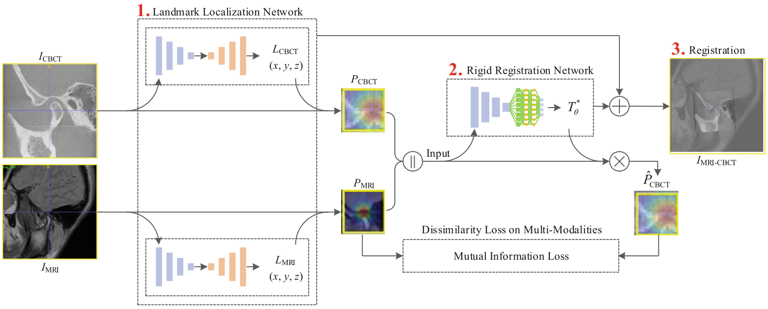


Fig. 2. The workflow of our landmark-guided rigid registration framework applied for TMJ MRI-CBCT images with different field-of-views. We highlight that all registrations are done in 3D throughout this paper. For clarity and simplicity, we depict the 2D formulation of our method in this paper.

As shown in Fig. 2, our overall framework for MRI-CBCT images registration includes three stages. Firstly, landmark numerical coordinate regression network takes I_{MRI} and

I_{CBCT} as input and estimates landmarks' coordinate $L_{\mathrm{MRI}}(x, y, z)$ and $L_{\mathrm{CBCT}}(x, y, z)$ respectively. Then the spatial transform network regress the rigid transformation matrix T_θ between two image patches P_{MRI} and P_{CBCT} centered landmarks. Finally, combined the rigid transformation matrix T_θ^* with the landmark-guided information, the rigid registration of MRI-CBCT images is achieved.

2.2 Landmark Localization Network

Inspired by landmark localization in human pose estimation [14], we proposed an end-to-end landmark localization network for 3D medical images (L^{2}Net) by converting heat-map regression into coordinate regression task. L^{2}Net consists of feature extraction network (U-Net) and coordinate regression layer (Fig. 3). For more technical details you can read our previous work in reference [15].

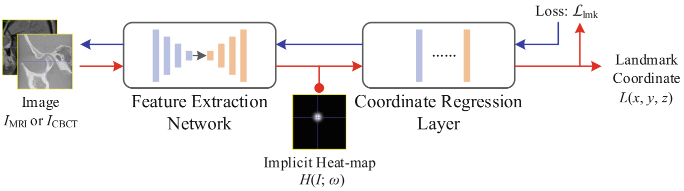

Loss: $\mathcal{L}_{\mathrm{lmk}}$
Image
I_{MRI} or I_{CBCT}
Feature Extraction Network
......
Coordinate Regression Layer
Landmark Coordinate
$L(x, y, z)$
Implicit Heat-map
$H(I;\omega)$

Fig. 3. Architecture of L^{2}Net for 3D medical images: a feature extraction network that extracts modality independent feature as implicit heat-map $H(I;\omega)$; and a coordinate regression layer that map the feature $H(I;\omega)$ to landmark coordinate $L(x, y, z)$.

Given a MRI/CBCT image I with size of $v = m \times n \times k$, U-Net learning the image feature to output the same size implicit normalized heat-map $H(I;\omega)$. By taking the probabilistic interpretation of $H(I;\omega)$, we can represent the landmark coordinates, $L(x, y, z)$ as center of mass (centroid) function defined as

$$L = (x, y, z) \sum_{(x,y,z)\in v} (x, y, z) * H(I;\omega) \Big/ \sum_{v} H(I;\omega) \tag{1}$$

Loss Function: Since the coordinate regression layer outputs numerical coordinates, it is possible to directly calculate Euclidean distance between the predicted coordinate $L_{\mathrm{inf}}(x, y, z)$ and ground truth $L_{\mathrm{gt}}(x, y, z)$. We take advantage of this fact to formulate the main term of landmark localization loss function (Eq. 2).

$$\mathcal{L}_{\mathrm{euc}}(L_{\mathrm{gt}}, L_{\mathrm{inf}}) = \left\| L_{\mathrm{inf}} - L_{\mathrm{gt}} \right\|_2 \tag{2}$$

The shape of the implicit heat-map also affects the regression accuracy of landmark coordinate [16]. More specifically, to force the implicit heat-map to resemble a spherical Gaussian distribution, we can minimize the divergence between the heat-map $H(I;\omega)$ and an appropriate target normal distribution $N(L_{\mathrm{inf}}, \sigma_t^2)$. Equation 3 defines distribution regularization, where $D(\cdot\|\cdot)$ is the Jensen-Shannon divergence.

$$\mathcal{L}_{\mathrm{reg}}(H(I;\omega), L_{\mathrm{inf}}, \sigma_t) = D(H(I;\omega)||N(L_{\mathrm{inf}}, \sigma_t^2)) \tag{3}$$

Equation 4 shows how regularization is incorporated into the Euclidean distance function. A regularization coefficient, λ, is used to set strength of the regularizer, $\mathcal{L}_{\text{reg}}$.

$$\mathcal{L}_{\text{lmk}} = \mathcal{L}_{\text{euc}}(L_{\text{gt}}, L_{\text{inf}}) + \lambda \cdot \mathcal{L}_{\text{reg}}(H(I;\omega), L_{\text{inf}}, \sigma_t) \tag{4}$$

2.3 Rigid Registration Network

The architecture of spatial transform regression network used for rigid registration is shown in Fig. 4. The input is the concatenated patch pairs ($P_{\text{CBCT}} \parallel P_{\text{MRI}}$) that centered landmarks and the output is the transform matrix T_θ which indicates the spatial relationship between two image patches. Each convolution layer is zero-padded and is followed by ReLU activations. After max pooling for two times, two fully connected (FC) layers with ReLU activation function are used to gather information from the entire images to give the *rigid transform matrix*: $M = [\theta x, \theta y, \theta z, \Delta x, \Delta y, \Delta z,]$.

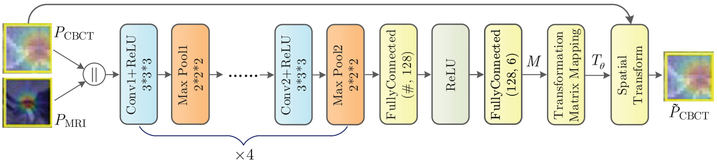


Fig. 4. Architecture of multi-modality image rigid transformation matrix regression network.

Transformation Matrix Mapping: This layer converts the transform M into exact rigid transformation matrix, instead of affine matrix used in reference [10, 11, 13]. Therefore, the shearing transformation caused by affine transformation can be eliminated, so as to improve the accuracy of rigid registration. To make the entire registration network training process back-propagating, this parameters mapping process must meet the derivability requirements. For rotation we can get rotation matrix R_x as following,

$$\begin{aligned} R_x = \cos(\theta_x) * \begin{bmatrix} 0\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,0 \end{bmatrix} + \sin(\theta_x) * \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 1\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,0\,1 \end{bmatrix} \\ = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\theta_x) & -\sin(\theta_x) & 0 \\ 0 & \sin(\theta_x) & \cos(\theta_x) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{aligned}$$

The rotations matrices R_y and R_z (similar definitions with R_x), specified in this way, determine an amount of rotation about each of the individual axes of the coordinate system. And for translation matrix,

$$D = \Delta x * \begin{bmatrix} 0\,0\,0\,1 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \end{bmatrix} + \Delta y * \begin{bmatrix} 0\,0\,0\,0 \\ 0\,0\,0\,1 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \end{bmatrix} + \Delta z * \begin{bmatrix} 0\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,0\,1 \\ 0\,0\,0\,0 \end{bmatrix} + \begin{bmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \end{bmatrix} = \begin{bmatrix} 1\,0\,0\,\Delta x \\ 0\,1\,0\,\Delta y \\ 0\,0\,1\,\Delta z \\ 0\,0\,0\,\ 1 \end{bmatrix}$$

Combined these rotation and translation matrices together in certain order, we can get the rigid transformation matrix T_θ,

$$T_\theta = R_x * R_y * R_z + D \tag{5}$$

Once the transformation matrix T_θ is obtained, a spatial transformation layer [14] is used to warp the moving image using the deformation field T_θ. Each voxel in the warped image $\tilde{P}_{\text{CBCT}}$ is calculated by bi-linear interpolation in the corresponding location, as given by the displacement vector, in the subject image P_{CBCT}:

$$\tilde{P}_{\text{CBCT}}(p) = \sum_{q \in N(p+\omega)} P_{\text{CBCT}}(q)(1 - |p + \omega - q|_2^2) \tag{6}$$

Where p and q are coordinates on the image, and ω is the displacement of p, $N(p + \omega)$ is the set of 8-pixel cubic neighbors of $p + \omega$.

Loss Function: Mutual information (MI) has become a common loss function for (especially multi-modality) image registration [17]. Formally, the mutual information between our image patches P_{MRI} and $\tilde{P}$CBCT is defined as the following,

$$\text{MI}(P_{\text{MRI}}, \tilde{P}_{\text{CBCT}}) = \sum_x \sum_y p_{\text{MRI,CBCT}}(x, y) \log \frac{p_{\text{MRI,CBCT}}(x, y)}{p_{\text{MRI}}(x) p_{\text{CBCT}}(y)} \tag{7}$$

In order to realize the training of end-to-end image registration network, here we use *Parzen* windowing [18] to calculate differentiable MI for loss function,

$$\mathcal{L}_{\text{sim}}(P_{\text{MRI}}, \tilde{P}_{\text{CBCT}}) = -\text{MI}(P_{\text{MRI}}, \tilde{P}_{\text{CBCT}}) \tag{8}$$

The optimal rigid transformation matrix T_θ^* is finally obtained through the network training. Finally, the coordinate offset between landmarks $L_{\text{MRI}}(x, y, z)$ and $L_{\text{CBCT}}(x, y, z)$ is mapped to the transformation matrix T_θ^*, accordingly the spatial transformation matrix between the I_{MRI} and I_{CBCT} is obtained.

3 Experiments

3.1 Data

The TMJ dataset consists of 204 CBCT and paired MRI images from 102 patients in Peking University School and Hospital of Stomatology. CBCT images are of size 481 × 481 × 481, where each voxel is of size 0.125 × 0.125 × 0.125 mm^3. MRI images are 256 × 256 × (7–11) with voxel size of 0.546875 × 0.546875 × 3.3 mm^3. We group the intensity of each image according to the modality and perform histogram matching.

3.2 Training Details

Network Training. Localization networks were trained using mini-batches of 1-sample each. The implicit heat-map is the same size as the resampled input image with $\sigma = 5$. In our experiments, we picked $\lambda = 1$ using cross validation. The models were optimized with RMSProp using an initial learning rate of 2.5×10^{-4}. Each model was trained for 120 epochs, with the learning rate reduced by a factor of 10 at epochs 60 and 90 (an epoch is one complete pass over the training set). In order to increase the ability of registration network to capture displacement of the input image patches, along the *x*, *y*, and *z* directions the landmark coordinates are randomly added by an offset [−60 ~ + 60] as the center position of image patches. We set the learning rate to 1.0e-2, and use exponential decay to adjust the learning rate parameters (ExponentialLR) method, where the basis coefficient *gamma* is set to 0.95. We implemented our method using Pytorch, and used a workstation equipped with single Maxwell-architecture NVIDIA Titan X GPU.

Landmark Localization Results. The mean radial error (MRE, in mm) is the commonly used evaluation index for medical image landmark detection task [15]. Compared with heat-map based method, the MRE result of our networks is lower obviously and modality independent (Fig. 5).

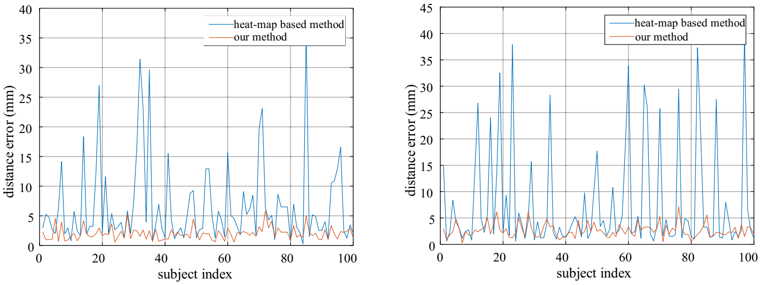


Fig. 5. Distance error of the landmark localization. MRI (left): 6.6803 ± 7.0876 mm (heat-map based method), 2.0244 ± 1.0635 mm (our method); CBCT (right): 7.6375 ± 10.0229 mm (heat-map based method), 2.6371 ± 1.2982 mm (our method).

Rigid Registration Results. In order to evaluate the performance of our proposed unsupervised learning rigid registration network, here we choose to compare the methods including SimpleElastix[1], ANTs[2], two software packages based on traditional iterative optimization, and an affine transformation matrix regression based on convolutional neural networks such as reference [10].

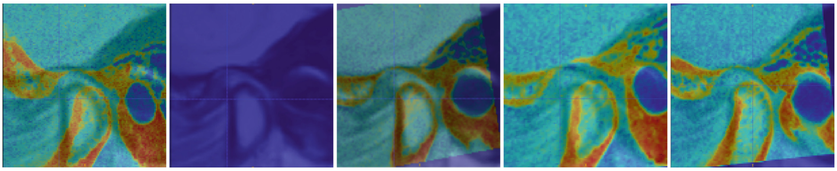


Fig. 6. Comparison of image patch registration results of various methods. (a) Inital image, (b) ANTs-Affine, (c) Simple Elastix, (d) Learning Affine, (e) Our method.

Figure 6 shows the registration results of multiple methods on the same image patches. In these figures, there are two layers: the bottom layer is an MRI image and the top layer is CBCT image with a color overlay. The first column shows the whole TMJs at the center of the image patch. The two initial images have obvious position and angle misalignment in Fig. 6a. ANTs software package cannot effectively complete the registration task (Fig. 6b). The SimpleElastix obtained good alignment (Fig. 6c) after iterative optimization. The shearing transformation in the affine registration makes the spatial position relationship tilted (Fig. 6d). The mutual information and structural similarity (SSIM) [18] between the registered patches is the most commonly used index to measure the alignment. Table 1 gives quantitative evaluation results of the registration methods.

Table 1. Comparison MI and SSIM of the registration results of different methods.

Methods	Mutual information		Structural similarity		Time
	Mean	Std	Mean	Std	Mean
Intial image	0.36	0.037	0.012	0.0087	–
Manual method	–	–	–	–	>2.0 h
Ants-rigid [2]	0.43	0.202	0.560	0.3390	1.465 s
Elastix-rigid [1]	0.59	0.061	0.128	0.0451	13.765 s
Affine learning [10]	0.48	0.051	0.044	0.0223	0.016 s
Our method	0.57	0.076	0.068	0.0209	0.016 s

[1] Medical Image Registration Library – SimpleElastix: https://simpleelastix.github.io/.
[2] Advanced Normalization Tools – ANTs: https://stnava.github.io/ANTs/.

Combined landmark localization with unsupervised image registration stages together, the registration result for the whole MRI-CBCT images is shown in Fig. 7. Conversely, SimpleElastix and Ants software packages, neither of them can achieve registration processing result successfully.

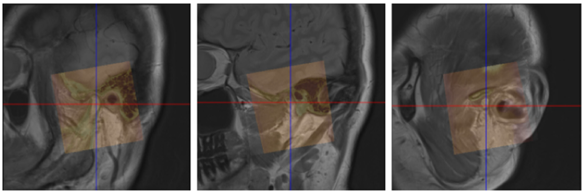


Fig. 7. Rigid registration result for the whole TMJ MRI-CBCT images. The center square area is the result of MRI and CBCT superposition.

4 Conclusion

We proposed an rigid registration network guided by landmarks for the common clinical application of multi-modality medical image registration problems. End-to-end landmark localization network effectively solves the influence of field-of-view difference between different modality images, and rigid transformation regression improves the registration accuracy and speed. We conclude that our method can effectively solve similar image registration applications.

Acknowledgments. This work was supported by the Fundamental Research Funds for the Central Universities (2021JBM003) and the National Natural Science Foundation of China with Project (81671034). Computations used the Department of Radiology and Biomedical Research Imaging Center (BRIC), University of North Carolina at Chapel Hill facility.

References

1. Asim, K.B., Santhosh, G., Aparna, S., et al.: Imaging of the temporomandibular joint: an update. World J. Radiol. **6**(8), 567–582 (2014). https://doi.org/10.4329/wjr.v6.i8.567
2. Al-Saleh M.A, Punithakumar K., Lagravere M., et al.: Three-dimensional assessment of temporomandibular joint using MRI-CBCT image registration, PLoS One **12(1)**, e0169555 (2017). https://doi.org/10.1371/journal.pone.0169555
3. Al-Saleh M.A., Jaremko J.L., Alsufyani N., et al.: Assessing the reliability of MRI-CBCT image registration to visualize temporomandibular joints. Dentomaxillofac. Radiol. **44(6)**, 20140244 (2015). https://doi.org/10.1259/dmfr.2014024
4. Al-Saleh, M.A., Punithakumar, K., Jaremko, J.L., et al.: Accuracy of magnetic resonance imaging-cone beam computed tomography rigid registration of the head: an in-vitro study. Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod. **121**(3), 316–321 (2016). https://doi.org/10.1016/j.oooo.2015.10.029

5. Lin, Y., Liu, Y., Wang, D., et al.: Three-dimensional reconstruction of temporomandibular joint with CT and MRI medical image fusion technology. Hua Xi Kou Qiang Yi Xue Za Zhi **26**(2), 140–143 (2008)
6. Dai, J., Dong, Y., Shen, S.: Merging the computed tomography and magnetic resonance imaging images for the visualization of temporomandibular joint disk. J. Craniofac. Surg. **23**(6), e647–e648 (2012). https://doi.org/10.1097/SCS.0b013e3182710517
7. Ma, R., Li, G., Sun, Y., et al.: Application of fused image in detecting abnormalities of temporomandibular joint. Dentomaxillofac. Radiol. **48**(3), 20180129 (2019). https://doi.org/10.1259/dmfr.20180129
8. Haskins, G., Kruger, U., Yan, P.: Deep learning in medical image registration: a survey. Mach. Vis. Appl. **31**(1–2), 1–18 (2020). https://doi.org/10.1007/s00138-020-01060-x
9. Miao, S., Wang, Z., Liao, R.: A CNN regression approach for real-time 2D/3D registration. IEEE Trans. Med. Imaging **35**(5), 1352–1363 (2016). https://doi.org/10.1109/TMI.2016.2521800
10. Chee, E., Wu, J.: AIRNet: self-supervised affine registration for 3D medical images using neural networks. arXiv:1810.02583 (2018)
11. Kori, A., Krishnamurthi, G.: Zero shot learning for multi-modal real time image registration. arXiv:1908.06213 (2019)
12. Salehi, S.S.M., Khan, S., Erdogmus, D.: Real-time deep pose estimation with geodesic loss for image-to-template rigid registration. IEEE Trans. Med. Imaging **38**(2), 470–481 (2019). https://doi.org/10.1109/TMI.2018.2866442
13. Shu, C., Chen, X., Xie, Q., et al.: An unsupervised network for fast microscopic image registration, In: Tomaszewski, J.E., Gurcan, M.N. (eds.) Medical Imaging 2018: Digital Pathology, vol. 10581, 105811D. International Society for Optics and Photonics (2018). https://doi.org/10.1117/12.2293264
14. Nibali, A., He, Z., Morgan, S., et al.: Numerical coordinate regression with convolutional neural networks. arXiv:1801.07372 (2018)
15. Li, J., Wang, Y., Mao, J., Li, G., Ma, R.: End-to-end coordinate regression model with attention-guided mechanism for landmark localization in 3D medical images. In: Liu, M., Yan, P., Lian, C., Cao, X. (eds.) MLMI 2020. LNCS, vol. 12436, pp. 624–633. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59861-7_63
16. Payer, C., Štern, D., Bischof, H., et al.: Integrating spatial configuration into heatmap regression based CNNs for landmark localization. Med. Image Anal. **54**, 207–219 (2019). https://doi.org/10.1016/j.media.2019.03.007
17. Huang, Y., Song, T., Xu, J., et al.: KLDivNet: an unsupervised neural network for multimodality image registration. arXiv:1908.08767 (2019)
18. Pluim, J.P.W., Maintz, J.B.A., Viergever, M.A.: Mutual-information-based registration of medical images: a survey. IEEE Trans. Med. Imaging **22**(8), 986–1004 (2003). https://doi.org/10.1109/TMI.2003.815867