X-ray2Shape: Reconstruction of 3D Liver Shape from a Single 2D Projection Image

Fei Tong[1], *Student Member, IEEE,* Megumi Nakao[1], *Member, IEEE,* Shuqiong Wu[1], Mitsuhiro Nakamura[2], Tetsuya Matsuda[1], *Member, IEEE*

***Abstract*— Computed tomography (CT) and magnetic resonance imaging (MRI) scanners measure three-dimensional (3D) images of patients. However, only low-dimensional local two-dimensional (2D) images may be obtained during surgery or radiotherapy. Although computer vision techniques have shown that 3D shapes can be estimated from multiple 2D images, shape reconstruction from a single 2D image such as an endoscopic image or an X-ray image remains a challenge. In this study, we propose X-ray2Shape, which permits a deep learning-based 3D organ mesh to be reconstructed from a single 2D projection image. The method learns the mesh deformation from a mean template and deep features computed from the individual projection images. Experiments with organ meshes and digitally reconstructed radiograph (DRR) images of abdominal regions were performed to confirm the estimation performance of the methods.**

I. INTRODUCTION

Three-dimensional (3D) medical imaging, such as computed tomography (CT) and magnetic resonance imaging (MRI), can image human internal organs, and are widely used for diagnosis, intraoperative navigation, and radiotherapy. However, a large number of image slices are typically required to obtain accurate organ shapes and tumor positions from 3D medical images [1] [2]; in the case of CT, this gives rise to high exposure to ionizing radiation, which is undesirable for the patient. The measurement time and imaging radiation dose can be reduced by lowering the number or the resolution of the image slices, but the image quality may be sacrificed correspondingly. Most importantly, it is seldom possible to use CT and MRI equipment during surgery or daily radiation therapy, because they are usually located separately. When high-resolution 3D images are not available, treatment can only proceed with low-dimensional and local images, such as endoscopic images and X-ray images.

To solve this problem, some researchers have proposed a method to reconstruct 3D organ shape from a single image, such as an endoscopic or X-ray image [3] [4] [5]. For example, Wu et al. proposed a 3D shape reconstruction algorithm based on a Convolutional Neural Network (CNN) [4], and showed that the 3D shape of the lungs during a deaeration deformation process can be reconstructed from only one captured two-dimensional (2D) image. However, the shape estimated by Wu's method was represented as point clouds; this means that the surface and topological information between vertices, which is important for deformation calculation, was lost. It is difficult to accurately obtain the vertex correspondence before and after deformation for point clouds, especially for organs undergoing large deformations such as lungs and abdominal organs. Wang et al. proposed a CNN-based framework to calculate lung respiratory deformation from a digitally reconstructed radiograph (DRR) image simulating an X-ray image [5]. However, in Wang's study, the training and test experiments were performed using artificially created augmented data, and the 3D shapes were deformed from multiple 3D initial templates. Hence, the CNN-based reconstruction of organ shape for real patients has yet to be tested.

The purpose of our study was to reconstruct the 3D shape of an organ from an individual patient's single-view 2D X-ray image. Our method uses a combination of Graph Convolutional Networks (GCN) [6] and a CNN. We extend and apply the framework that is based on Pixel2Mesh [7] used on natural images to low contrast DRR images. X-ray images and DRR images are projected images, especially in the abdominal region, where clear organ contours can not be obtained and the contrast from the background is also very low. Shape reconstruction is therefore a difficult task. We set the mean shape derived from the training 3D CT data as the initial template, and aimed to calculate the deformation from the initial template to the individual organ shape using the image features of the DRR image.

II. METHODS

A. Outline of the methods

Fig. 1 shows the framework of the proposed method. The entire framework consists of a 2D image-feature CNN network and a GCN mesh deformation network. As the patient's posture is fixed relative to the X-ray irradiation position during X-ray imaging, it can be assumed that the camera position and angle at the time of imaging are known. Therefore, the 2D image pixel coordinates corresponding to each 3D shape vertex can be calculated using the camera parameters. In our proposed method, the 3D initial template is projected onto the input DRR image, and the image features corresponding to each vertex can be extracted. We then concatenate the image features and 3D vertex coordinates and incorporate them into a GCN network for deformation. We used two types of loss, MSE loss and discrete Laplacian loss, to train the network to generate an accurate mesh.

[1]F. Tong, M. Nakao, S. Wu and T. Matsuda are with Graduate School of Informatics, Kyoto University, Kyoto, Japan. (e-mail: tong.fei@sys.i.kyoto-u.ac.jp)

[2]M. Nakamura is with Graduate School of Medicine, Human Health Sciences, Kyoto University, Kyoto, Japan.

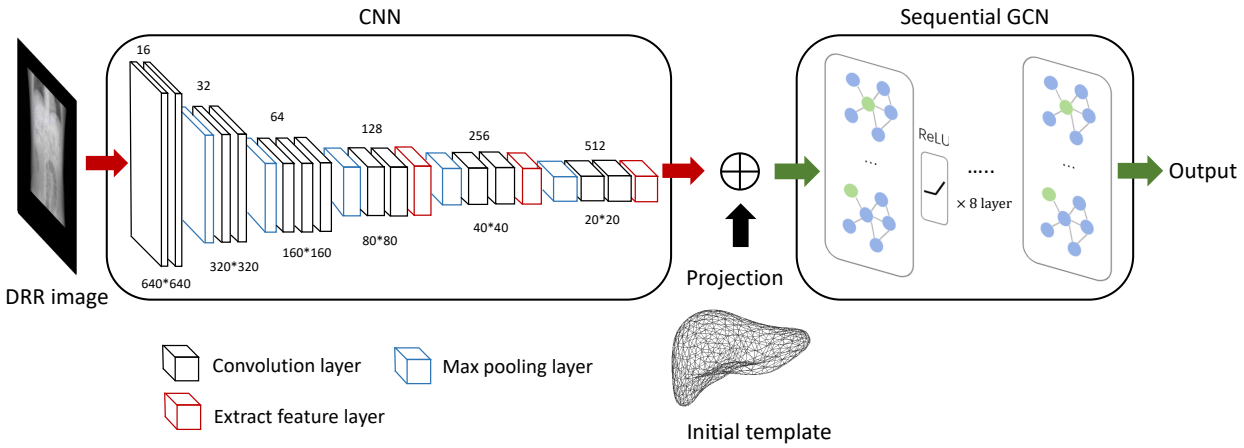

Fig. 1. The full X-ray2Shape framework contains a CNN and GCN. The CNN is used for extracting image features from the 2D image, and the GCN is for mesh deformation. $\bigoplus$ means the concatenation of image features and vertex coordinates.

B. X-ray2Shape modules

In this section, we describe the internals of each X-ray2Shape module. For the initial template, we calculated the mean shapes of the organs from the training data (Fig. 2 (a)). The initial liver template we used in our experiment contained 500 vertices and 996 faces, and all estimated shapes were deformed from this initial template.

For the image feature CNN module, the layers of the CNN shown in Fig. 1 were from an extended VGG-16 model, as it has been widely used and shown to be effective for image processing [10]. However, VGG-16 has often been used to extract features from 224×224 pixel images [8], but our input DRR images are much larger, at 640×640 pixels. To extract effective features, we changed the convolution filter size from 3×3 to 5×5, meaning that we expanded the receptive field to obtain features from a larger range of pixels. With the camera parameters, each 3D vertex can find its 2D pixel coordinates in the input DRR image. We extracted features from the latter layers of the CNN (i.e. the red layers in Fig. 1) and concatenated these high-level features to accurately learn the shape.

For the GCN module, a network that applies deep learning to graph structure data, the graph (eg. mesh) is a pair of sets $(\mathcal{V},\mathcal{E})$, where $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ is the set of edges. The graph can be deformed by updating the features of each vertex. The GCN network in our experiment consisted of eight sequential graph convolutional layers, with each convolutional layer defined as below,

$$\mathcal{F}^{(l+1)} = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}\mathcal{F}^{(l)}W^{(l)}) \tag{1}$$

where $\mathcal{F}^{(l)}$ and $\mathcal{F}^{(l+1)}$ are the feature matrix before and after convolution, $\hat{A}\in\mathbb{R}^{n\times n}$ is an adjacency matrix, $\hat{D}\in\mathbb{R}^{n\times n}$ is the degree matrix of $\hat{A}$, and n is the total number of vertices. W is the learnable parameter matrix and the feature $\mathcal{F}^{(l)}$ is the concatenation of 2D image feature from CNN and 3D vertex coordinates. The initial template can be deformed by updating $\mathcal{F}^{(l)}$.

C. Loss functions

In this section, we define two loss functions, MSE loss and discrete Laplacian loss [9], which are used to generate an accurate 3D shape. Unlike the pixel2mesh framework, whose estimated shape is deformed from an ellipsoid with fewer vertices, our estimated shape is deformed from a mean shape with the same number of vertices. Therefore, we define MSE loss to reduce the distance of the corresponding vertices between the estimated shape and ground truth. The MSE loss is defined as

$$\mathcal{L}_{MSE} = \frac{1}{n}\sum_{i=0}^{n}\|v_i-\hat{v}_i\|_2^2 \tag{2}$$

where n is the total number of vertices, $v_i\in\mathcal{V}(i=1,2,...,n)$ is the ground truth vertex, and $\hat{v}_i$ is the estimated vertex. This loss function tends to converge the estimated vertex to the correct position.

To preserve the smoothness of the surface, we used a discrete Laplacian loss function. The discrete Laplacian at vertex v_i is defined as

$$L(v_i) = \frac{1}{N(v_i)}\sum_{j\in N(v_i)}(v_i-v_j) \tag{3}$$

where $N(v_i)$ is the number of adjacent vertices of one ring connected by vertex v_i and the edge, v_j is the neighboring vertex of v_i, and the discrete Laplacian loss can be defined as

$$\mathcal{L}_{laplacian} = \frac{1}{n}\sum_{i=0}^{n}\|L(v_i)-L(\hat{v}_i)\|_2^2 \tag{4}$$

where $L(v_i)$ and $L(\hat{v}_i)$ are the discrete Laplacian before and after deformation. MSE loss makes vertices move too freely, while discrete Laplacian loss properly limits the freedom of vertex movements.

The total loss is the weighted sum of two loss functions which is expressed as

$$\mathcal{L}_{total} = \lambda_{MSE}\mathcal{L}_{MSE} + \lambda_{laplacian}\mathcal{L}_{laplacian} \tag{5}$$

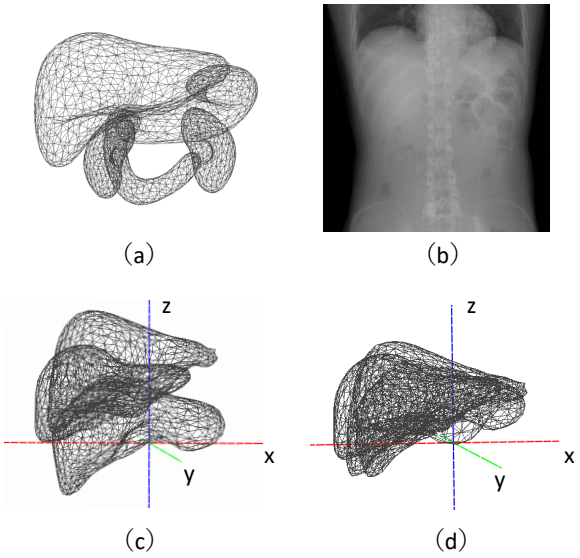

Fig. 2. Experimental data, (a) Initial templates of five organs (liver, stomach, duodenum, left kidney, right kidney) with mean shapes, (b) input DRR image, (c) liver coordinates of different cases before preprocessing, (d) liver coordinates of different cases after preprocessing

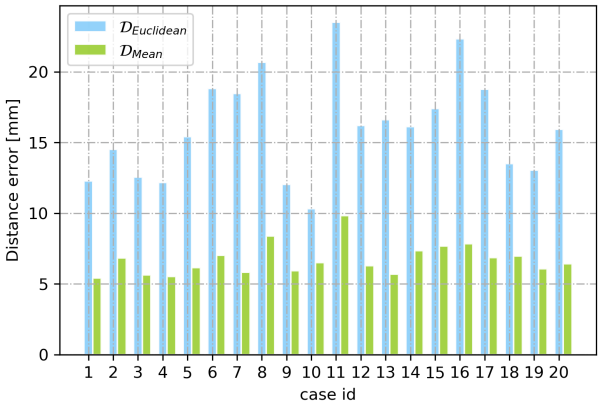

Fig. 3. Mean distance error and Euclidean distance error of 20 test cases

TABLE I
COMPARISON BETWEEN OUR LOSSES AND PIXEL2MESH LOSSES

Our losses		pixel2mesh losses	
$\bar{\mathcal{D}}_{Mean}$	$\bar{\mathcal{D}}_{Euclidean}$	$\bar{\mathcal{D}}_{Mean}$	$\bar{\mathcal{D}}_{Euclidean}$
6.71mm	16.0mm	9.1mm	19.3mm

In our experiment, we set the hyper-parameter $\lambda_{MSE} = 1$ and $\lambda_{laplacian} = 100$ to balance the two loss functions.

III. EXPERIMENTS

A. Dataset and preprocess

The 3D organ data we used in this experiment were generated from 3D CT datasets from 124 patients. The 3D organ data used a surface triangle mesh structure, with the liver mesh having 500 vertices and 996 faces. Once we have the 3D organ meshes, we can generate the corresponding front-view 2D projections that we refer to as the DRR image (Fig. 2 (b)), with the size of these DRR images being 640×640 pixels.

However, the CT imaging range varies from patient to patient, which means that the variation in the 3D organ coordinates is very large. For example, in Fig. 2 (c), the liver coordinates vary greatly between the three different patients, affecting the accuracy of the estimated organ shapes. To solve this problem, we preprocess the 3D organ mesh. We calculate the center of gravity of five organs (liver, stomach, duodenum, left kidney, right kidney) and use this as the origin to translate the 3D organs. The translated 3D liver meshes are shown in Fig. 2 (d), in which we can see the livers of different patients translated into the same coordinate range. In our experiment, we randomly split the dataset into 104 cases for training and 20 cases for testing. The initial liver template in Fig. 1 is the mean shape of the translated training cases.

B. Evaluation

In this section, the mean distance $\mathcal{D}_{Mean}$ [11] and the Euclidean distance $\mathcal{D}_{Euclidean}$ are used to evaluate the difference between ground truth and estimated shapes. The mean distance $\mathcal{D}_{Mean}$ is the mean value of the nearest bidirectional point-to-surface distance. Considering that the ground truth and estimated shapes require point-to-point correspondence, we define the Euclidean distance below to calculate the distance between corresponding points,

$$\mathcal{D}_{Euclidean} = \frac{1}{n}\sum_{i=0}^{n}\sqrt{(v_i - \hat{v}_i)^2} \tag{6}$$

However, while these two metrics can evaluate the distance error between ground truth and estimated shapes, they do not reflect the smoothness and estimated shape quality. We visualize the estimated shape in Fig. 4 to better understand these aspects.

C. Training and Results

Our experiment aimed to generate liver shape from a single DRR image with low background contrast, and to do so using limited datasets. We used an extended VGG-16 model without pre-training. The width (channel number) of each layer is marked at the top of the layers in Fig. 1, and the size of each layer is marked below. The whole network was implemented in Tensorflow-GPU, and the network was trained using an Adam optimizer with a learning rate of 1×10^{-4} and weight decay of 5×10^{-6}. The batch size was 1 and the total number of training epochs was 1000. The network took 4.5 hours to train on a single NVIDIA GeForce RTX 2070.

The distance error of each test case is shown in Fig. 3. The blue bars show the Euclidean distance error of 20 test cases, with the mean value of the Euclidean distance error being 16.0 mm. The green bars show the mean distance error, with the mean value of these 20 test cases being 6.71 mm. We compared training the network with our proposed loss functions with training it with the loss functions proposed in pixel2mesh. It should be noted that we already knew the

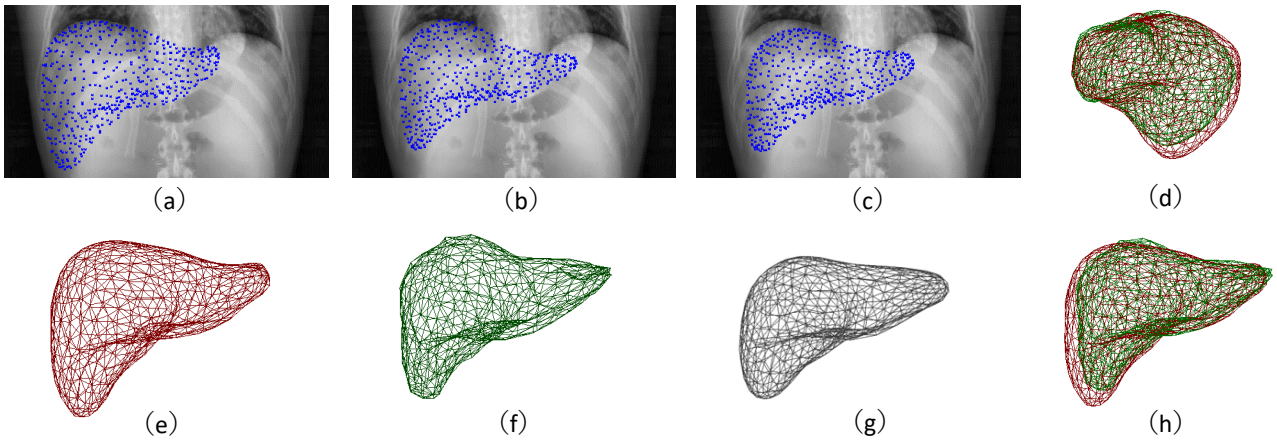

Fig. 4. An example of 3D shape reconstruction using the proposed method, (a) projection of the ground truth liver shape (the pixels corresponding to each 3D vertex are colored blue), (b) projection of the estimated liver shape, (c) projection of the initial template of the liver, (d) sideview of the overlap between estimated and ground truth shapes, (e) ground truth liver shape, (f) estimated liver shape, (g) initial template of the liver, (h) frontview of the overlap between estimated and ground truth shapes

correct position of each vertex. Hence, in this comparison we replaced the pixel2mesh's Chamfer loss with MSE loss, while the rest of the losses in pixel2mesh were unchanged. The results are shown in Table 1, where $\bar{\mathcal{D}}$ means the mean value of the distance errors. They show that the mean value of the Euclidean distance errors and the mean distance errors calculated with our losses are totally smaller than those obtained with the pixel2mesh losses. As our deformation starts from the mean shape, the variation in the shape is smaller than that with pixel2mesh. On the other hand as the ground truth coordinates of each vertex are already known, the two loss functions proposed in this paper are sufficient.

The visualization result is shown in Fig. 4. To understand how the estimated 3D shape is deformed from the initial template, we projected the 3D shape onto the DRR image with the pixels corresponding to each 3D vertex colored in blue (Fig. 4 (a–c)). It can be observed that the projection of the initial template is completely misplaced (Fig. 4 (c)). However, the projection of the estimated 3D shape (Fig. 4 (b)) has been improved, especially the diaphragm part, where an obvious contour can be seen. The overlap between the estimated shape and the ground truth shape is shown in Fig. 4 (d) and (h).

IV. CONCLUSIONS

This paper proposed X-ray2Shape, a deep neural network that combines a GCN and CNN to reconstruct 3D liver shape from only one low-contrast DRR image. To generate an accurate and smooth 3D shape, we trained our network using MSE loss and discrete Laplacian loss. Our experimental dataset contains 3D CT from 124 patients without data augmentation. Distinct from the natural and captured images, our input DRR images were low-contrast, without clear organ contours. However, as shown by the evaluation error and visualization results, our trained X-ray2Shape network was still effective. In future work, we will use multi-view DRR images for training, and design other losses to more accurately reconstruct the shape.

ACKNOWLEDGMENT

This research was supported by JSPS Grant-in-Aid for Scientific Research (B) (grant number 18H02766) and for challenging Exploratory Research (grant number 18K19918).

REFERENCES

[1] M. Islam, T. Purdie, B. Norrlinger, H. Alasti, D. Moseley, M. Sharpe, J. Siewerdsen, and D. Jaffray, "Patient dose from kilovoltage cone beam computed tomography imaging in radiation therapy", *Medical Physics*, vol. 33, pp. 1573-1582, 2006.

[2] M. Kan, L. Leung, W. Wong, and N. Lam, "Radiation dose from cone beam computed tomography for image-guided radiation therapy", *International Journal of Radiation Oncology Biology Physics*, vol.70, pp. 272-279, 2008.

[3] A. Saito, M. Nakao, Y. Uranishi, and T. Matsuda, "Deformation Estimation of Elastic Bodies Using Multiple Silhouette Images for Endoscopic Image Augmentation", *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 170-171, Sep 2015.

[4] S. Wu, M. Nakao, J. Tokuno, T. Chen-Yoshikawa, and T. Matsuda, "Reconstructing 3d lung shape from a single 2D image during the deaeration deformation process using model-based data augmentation", *IEEE Int. Conf. on Biomed. Health Info.(BHI)*, pp. 1-4, 2019.

[5] Y. Wang, Z. Zhong, and J. Hua, "DeepOrganNet: On-the-fly reconstruction and visualization of 3D/4D lung models from single-view projections by deep deformation network", *arXiv preprint*, Art. no. arXiv:1907.09375, 2019.

[6] T. N. Kipf, and M. Welling, "Semi-supervised classification with graph convolutional networks", in *International Conference on Learning Representations (ICLR)*, 2016.

[7] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2mesh: Generating 3d mesh models from single rgb images", in *Proc. of the European Conference on Computer Vision (ECCV)*, 2018

[8] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", in *International Conference on Learning Representations (ICLR)*, 2015.

[9] M. Nakao, M. Nakamura, T. Mizowaki, and T. Matsuda, "Statistical deformation reconstruction using multi-organ shape features for pancreatic cancer localization", *arXiv preprint*, Art. no. 1911.05439, 2019.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge", *Int. J. Comput. Vis. (IJCV)*, vol. 115, pp. 211-252, 2015.

[11] J. Kim, C. Valdes-Hernandez Mdel, N. A. Royle, and J. Park, "Hippocampal shape modeling based on a progressive template surface deformation and its verification", *IEEE Trans. Med. Imag.*, vol. 34, pp. 1242-1261, 2015.