DYNAMIC SPATIO-TEMPORAL GRAPH CONVOLUTIONAL NETWORKS FOR CARDIAC MOTION ANALYSIS

Ping Lu[1], *Wenjia Bai*[2,3], *Daniel Rueckert*[2], *J.Alison Noble*[1]

[1]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK
[2]Department of Computing, Imperial College London, London, UK
[3]Department of Brain Sciences, Imperial College London, London, UK

ABSTRACT

We propose a dynamic spatio-temporal graph convolutional network (DST-GCN) approach to learn the left ventricular (LV) motion patterns from cardiac MR cine images. We represent the myocardial geometry using a graph that is constructed from sample nodes on endo- and epicardial contours. The DST-GCN follows an encoder-decoder framework. The encoder accepts a given cardiac motion represented by a sequence of ST-GCN. The decoder employs a graph-based gated recurrent unit (G-GRU) to predict future cardiac motion. We show that the DST-GCN can automatically quantify the spatio-temporal patterns in cardiac MR that characterise cardiac motion. Experiments are performed on the UK Biobank dataset. We compare four methods from two architecture variances. Experiments show that the proposed method inputting node velocities with residual connection in the decoder outperform others, and achieves a mean squared error of 0.135 pixel between the ground truth node locations and our prediction.

Index Terms— Graph convolutional networks, gated recurrent unit, cardiac MR, myocardium, cardiac motion.

1. INTRODUCTION

Cardiac motion analysis plays a significant role in the diagnosis of heart conditions [1] [2]. Quantitative image-derived phenotypes, such as displacement, strain or strain rates, are important biomarkers for diagnosis. They are sensitive to subtle changes in myocardial function and often indicate early onset of cardiac disease [1]. Cardiac motion can be evaluated by the sampling nodes of the endocardium and the epicardium from magnetic resonance imaging (MRI). The aim of cardiac motion analysis is to perform accurate estimation of the motion trajectories for the myocardial contours or meshes.

In recent years, geometric deep learning-based methods have achieved promising results in medical image analysis for disease classification, image segmentation and landmark detection. These methods are also popular in computer vision. For instance, the spatial-temporal graph neural network (ST-GCN) is widely used for human action recognition, which predicts graph-structured data in times series based on skeleton and joint trajectories of human bodies [3] [4]. Inspired by these applications of GCNs [5], we propose to employ graph convolutional neural networks (GCN) to model cardiac motion in the geometry space of the GCN instead of convolutional neural networks (CNNs) on a regular grid.

In this paper, we propose a geometric deep learning-based architecture, named dynamic spatio-temporal graph convolutional networks (DST-GCN), with a self-supervised training strategy. This method predicts the future 2D left ventricular (LV) cardiac motion given the previous observed motion trajectories. The cardiac motion is represented on a graph constructed from sample nodes on myocardial contours. The spatio-temporal graph convolutional network (ST-GCN) and a graph-based gated recurrent unit (G-GRU) are connected using a encoder-decoder framework. We introduce difference operators to describe cardiac motion dynamics. We investigate different difference operators as inputs and two architecture variances for modelling the spatio-temporal patterns of cardiac motion.

The contributions of this work are as follows. (1) We propose a geometric deep learning-based architecture for LV cardiac motion estimation. To our knowledge, this is the first method to exploit spatio-temporal patterns with a graph-based gates recurrent unit (G-GRU) for cardiac motion estimation. (2) We evaluate the impact of different inputs and the residual connection on the accuracy of the 2D LV cardiac motion prediction. (3) We demonstrate that a DST-GCN working at the velocity space improves the performance.

2. METHODS

In this paper, we formulate cardiac motion estimation as a prediction problem, which estimates future motion trajectories in the next T time frames, given the observations in the past. We define the historical cardiac structure as $[X_{t-n}, ..., X_{t-1}, X_t] \in \mathbb{R}^{m\times(n+1)\times d}$, where $X_t \in \mathbb{R}^{m\times d}$ with m nodes and $d = 2$ feature dimensions at time t. We propose a model ρ to predict future cardiac motion structure. The model is described as $[X_{t+1}, ..., X_{t+T}] =$

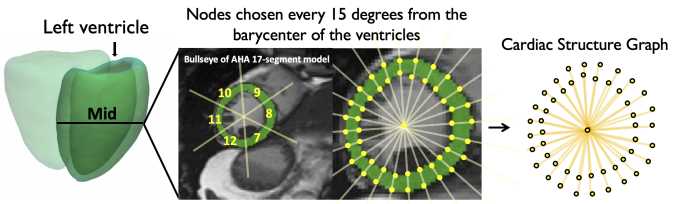

Fig. 1: *Overview of the proposed framework for the cardiac structure graph construction in the mid-ventricular of short-axis view cardiac MR image sequences. The barycenter of the left ventricle (LV) and the 48 node locations from both the endocardium and epicardium form a cardiac structure graph.*

$\rho([X_{t-n}, ...X_{t-1}, X_t])$ where n is the length of historical time series and T is the length of the predicted time series.

To design the dynamic spatio-temporal graph convolutional networks (DST-GCN), we consider two components: a Spatio-Temporal Graph Convolutional Neural Network (ST-GCN) and a graph-based gated recurrent unit (G-GRU).

2.1. Cardiac Structure Graph Construction

Fig.1 illustrates the construction of a cardiac structural graph. The cardiac structure sequence is represented by 2D coordinates of nodes on both the endocardium and epicardium in each cardiac MR frame. These nodes are chosen by the left and right ventricle geometry, based on the mid-slice 6-segments model of the 17-Segment AHA model. The detail of how to sample nodes is described in our previous work [6]. Moreover, these selected node locations are the ground truth in our work.

We construct one undirected spatio-temporal graph $G = (V, E)$ on the cardiac structure with N nodes and T frames. Nodes on both the endocardium and epicardium connect to the barycenter of the LV respectively. The detail of how to define the spatio-temporal graph is described in our previous work [6]. The cardiac structure graph will be used as the $0th$ order difference input to the proposed architecture.

2.2. Dynamic Spatio-Temporal Graph Convolutional Neural Network

Difference operator. We investigate two different inputs to the proposed architecture, node velocities and node locations. We define the node locations as the $0th$ order difference and the node velocities as the $1st$ order difference. We define node velocities as the differences between the current and the immediately previous node locations, without the division by the constant scanning time between two consecutive cardiac MR frames.

Spatio-temporal GCN. Let $A \in \{0, 1\}^{N \times N}$ be the adjacency matrix of the graph. If the i-th and the j-th nodes are

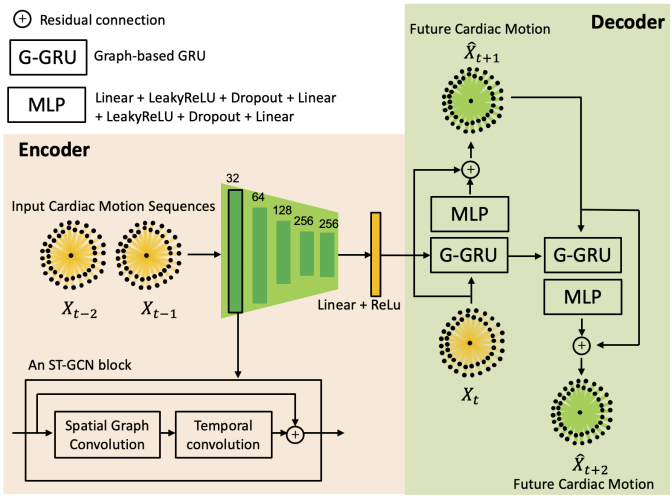

Fig. 2: *Network overview. The cardiac motion sequence is given as the input to the DST-GCN encoder-decoder framework. The output of the encoder and the previous motion status are fed into the decoder. The sum of the output of the decoder and the previous motion status represents the output future cardiac motion trajectory (predicted node locations), which can used in the left ventricular function evaluation.*

connected, $A_{i,j} = 1$. Otherwise, $A_{i,j} = 0$. Let $D \in \mathbb{R}^{N \times N}$ represent the diagonal degree matrix where $D_{i,i} = \sum_j A_{i,j}$.

The spatio-temporal GCN (ST-GCN) includes a range of the ST-GCN blocks [3]. Each block has a spatial graph convolution and then a temporal convolution, which extracts spatial and temporal features respectively. The key point of ST-GCN is the spatial graph convolution, which provides a weighted average of neighboring features for each node.

Let $F_{in} \in \mathbb{R}^{C \times T \times N}$ represents the input features, where C is the number of channels, T is the temporal length and N is the number of nodes in one frame. Let $F_{out} \in \mathbb{R}^{C \times T \times N}$ represents the output features obtained from the spatial graph convolution. The parameters m is the edge weight matrix, w is the feature importance matrix, $\widetilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ is the normalized adjacent matrix, $\circ$ represents the Hadamard product. The spatial graph convolution can be represented as

$$\mathrm{F}_{out} = \sum m \circ \widetilde{A} F_{in} w.$$

Let K_t be the kernel size of the temporal dimension. A $1D$ temporal convolution ($K_t \times 1$ convolution) can be applied to extract features in the temporal dimension.

Graph-based gated recurrent unit. Following graph guidance, hidden cardiac motion states can be learned and updated from the graph-based GRU (G-GRU). The graph is trained to generate a future cardiac structure by regularising the motion states. Let $A_{hid} \in \mathbb{R}^{m \times m}$ be the adjacent matrix of the graph from time $t-1, ..., t-n$. The G-GRU needs two inputs at time t: the initial state H_0 and the 2D cardiac structure-based feature $F_t \in \mathbb{R}^{m \times d}$. A graph convolution is used on the hid-

den states H_t and generates the state for the next frame. The G-GRU(F_t, H_t) can be denoted as

$$\begin{aligned}
\mathrm{u}_t &= \sigma(u_{in}(F_t) + u_{hid}(A_{hid}H_tW_{hid})),\\
r_t &= \sigma(r_{in}(F_t) + r_{hid}(A_{hid}H_tW_{hid})),\\
s_t &= tanh(s_{in}(F_t) + u_t \otimes s_{hid}(A_{hid}H_tW_{hid})),\\
H_{t+1} &= r_t \otimes H_t + (1 - r_t) \otimes s_t,
\end{aligned}$$

where W_{hid} indicates the trainable weights and the functions u_{in}, u_{hid}, r_{in}, r_{hid}, s_{in}, s_{hid} are linear transformations.

Encoder-decoder architecture. The encoder extracts the spatio-temporal features from the observed cardiac motion and provides the motion states as a hidden variable to the decoder. Firstly, we normalize input ($0th$ order difference or $1st$ order difference) with a batch normalization layer. The ST-GCN model contains 5 layers of spatial-temporal graph convolution blocks (ST-GCN blocks). The output channel for these layers are 32, 64, 128, 256, 256. Each layer has a temporal kernel size of 5. The ResNet mechanism is applied to each ST-GCN block. After that, we add a linear layer (the input dimensions are 256, the output dimensions are 49) and a ReLu layer.

The decoder is to predict future cardiac structure sequences, which are represented by node locations on the myocardial geometry. We model cardiac displacement between two consecutive frames using the proposed graph-based GRU (G-GRU) and a multilayer perceptron (MLP). To get the estimated cardiac motion (see Fig.2), we add a residual connection between the input and the output of each G-GRU cell.

For the $0th$ order difference input, at the time t, the decoder is described as

$$\begin{aligned}
\mathrm{H}_{\,t+1} &= \textit{G-GRU}(\hat{X}_t, H_t),\\
\hat{X}_{t+1} &= \hat{X}_t + \tau(H_{t+1}).
\end{aligned}$$

For the $1st$ order difference input, at the time t, the decoder is described as

$$\begin{aligned}
\mathrm{H}_{\,t+1} &= \textit{G-GRU}(\triangle\hat{X}_t, H_t),\\
\triangle\hat{X}_{t+1} &= \triangle\hat{X}_t + \tau(H_{t+1}).
\end{aligned}$$

Here the function τ is implemented by MLP. As shown in Fig.2, we add a dropout layer and a LeakyReLU layer (the activation function) between two fully-connected layers. The dropout probability is set to 0.3. The initial state H_0 is the final output of the encoder. For the $0th$ order difference input, at time t, $\hat{X}_t$ means X_t (the current observed cardiac structure). For the $1st$ order difference input, at time t, $\triangle\hat{X}_t$ means $X_t - X_{t-1}$ (the current observed node velocities of cardiac structure). Then the predicted node location $\hat{X}_{t+1} = X_t + \triangle\hat{X}_{t+1}$. The proposed model always has a sustained encoder hidden variable, which can avoid the problem of the encoder information vanishing. The input hidden dimensions are 256 and the output hidden dimensions are 49.

3. EXPERIMENTS

3.1. Data Acquisition

In this study, we use short-axis view cardiac MR image sequences from the UK BioBank[1]. The cardiac MR is obtained from 1.5 Tesla scanner (MAGNETOM Aera, Syngo Platform VD13A, Siemens Healthcare, Erlangen, Germany). A stack of short-axis images, around 12 slices, cover the entire left and right ventricles. We perform motion tracking on the 3 mid-ventricular slices. In-plane resolution is $1.8 \times 1.8mm^2$, while the slice gap is $2.0mm$ and the slice thickness is $8.0mm$. Each sequence contains 50 consecutive time frames per cardiac cycle. We randomly selected image sequences of 1071 subjects for training, 270 subjects for validation and 270 subjects for testing.

3.2. Implementation Details

Pre-processing. The segmentation of the LV endocardial and epicardial borders and the RV was generated from using a FCN method [7] and used for node extraction. For training and testing, we obtained 1 barycenter node location of the LV and 48 node locations from both the endocardial and epicardial borders from the sampling frames shown in Fig.1. The nodes extraction are described in section 2.1 and Fig.1. The input features are described with tensors (C, T, N). Here N denotes 49 nodes in each time frame. T denotes 3 consecutive sampling time frames. For the $1st$ order difference input (inputting node velocities), C denotes 2 channels for node locations' differences between two consecutive MR frame. For the $0th$ order difference input (inputting node locations), C denotes 2 channels for the 2D pixel coordinate (x, y).

Training. The model is trained over 200 epochs via gradient descent optimization - Adaptive Moment Estimation (Adam) with learning rate 0.001 and a batch size of 1. In each training sample, we set the input difference operators length to 3 frames, and we predict future cardiac structure in 2 frames. As shown in Fig.2, 3 consecutive sampling time frames are selected for each sample. The mean squared error (MSE) using the node locations is chosen as the evaluation metrics. The proposed network was implemented using Python 3.7 with Pytorch. Experiments are run with computational hardware GeForce GTX 1080 Ti GPU 10 GB.

3.3. Results

Quantitative results. We compared the following four methods: two methods inputting node velocities - with and without residual connection in the decoder respectively – and two methods inputting node locations – with and without residual connection in the decoder. We measured the predicted node location error from these four methods with the ground truth.

[1]UK BioBank. https://www.ukbiobank.ac.uk/

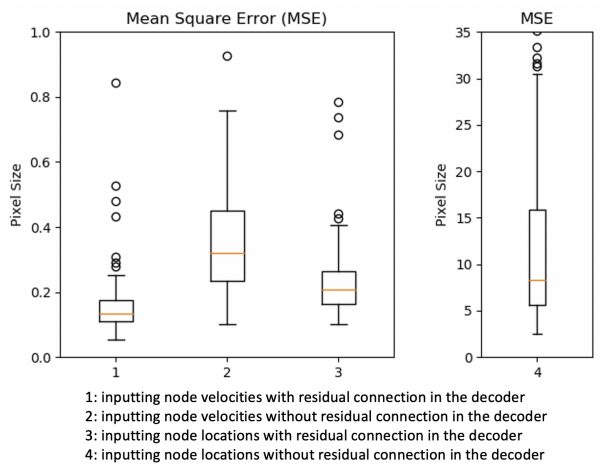

Fig. 3: *Box plot of the mean square error (MSE) for four different methods using the proposed architecture: two methods inputting node velocities - with and without residual connection in the decoder respectively – and two methods inputting node locations – with and without residual connection in the decoder.*

With inputting node velocities, the predicted node locations are the sum of the predicted node velocities and the node locations in the immediately previous frame. We found that the method inputting node velocities outperforms the method inputting node locations. Moreover, the residual connection in the decoder improves the results of our model. As shown in Fig.3, the method inputting node velocities with residual connection has the least MSE and achieved best-predicted performance. The method inputting node locations without residual connection achieved the least performance.

Representative examples. There are 44 predicted frames of node locations in a cardiac cycle for each sequence. Fig.4 shows an example of cardiac motion estimation on frames 4, 12, 20, 28, 36, 44 of the MRI sequence between the proposed method. We can see unusually high error on the $20th$ frame. The predicted nodes are positioned nearby the ground truth nodes, but not as close as in the other frames. The rest of the node locations are predicted very accurately, especially on the $36th$ and $44th$ frame.

4. DISCUSSION AND CONCLUSION

In this work, we propose a dynamic spatio-temporal graph convolutional network to characterise cardiac motion. We investigated the factors which can improve the accuracy of cardiac motion estimation. We found that the proposed method inputting node velocities with residual connection in the decoder outperformed other, achieved a mean squared error of 0.135 pixel. Using the $1st$ order difference (node velocities) as both inputs and outputs boosts prediction because cardiac

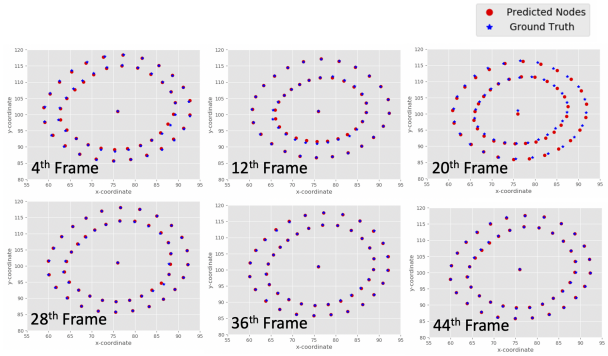

Fig. 4: Cardiac motion estimation comparison between the proposed method and the ground truth. The predictions and the ground truth on the $4th$, $12th$, $20th$, $28th$, $36th$ and $44th$ frame of the MRI sequence are shown there.

structure changes slightly in consecutive frames. We also found the traditional motion estimation method (optical flow) is very sensitive to image intensities and cannot catch cardiac motion in a meaningful way.

5. COMPLIANCE WITH ETHICAL STANDARDS

This study used data from the UK Biobank Resource under Application Number 40119. The authors declare no conflicts of interest related to this research study. **Acknowledgements.** This work is supported by SmartHeart. EPSRC grant EP/P001009/1.

References

[1] Lu et al., "Going deeper into cardiac motion analysis to model fine spatio-temporal features," in *MIUA 2020*.

[2] Zheng et al., "Explainable cardiac pathology classification on cine MRI with motion characterization by semi-supervised learning of apparent flow," *Med Image Anal 2019*.

[3] Yan et al., "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI-18*.

[4] Li et al., "Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction," in *CVPR*, 2020.

[5] Kipf et al., "Semi-supervised classification with graph convolutional networks," *arXiv:1609.02907 2016*.

[6] Lu et al., "Modelling cardiac motion via spatio-temporal graph convolutional networks to boost the diagnosis of heart conditions," in *STACOM Workshop, MICCAI 2020*.

[7] Bai et al., "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks," *JCMR*, vol. 20, no. 1, pp. 65, 2018.