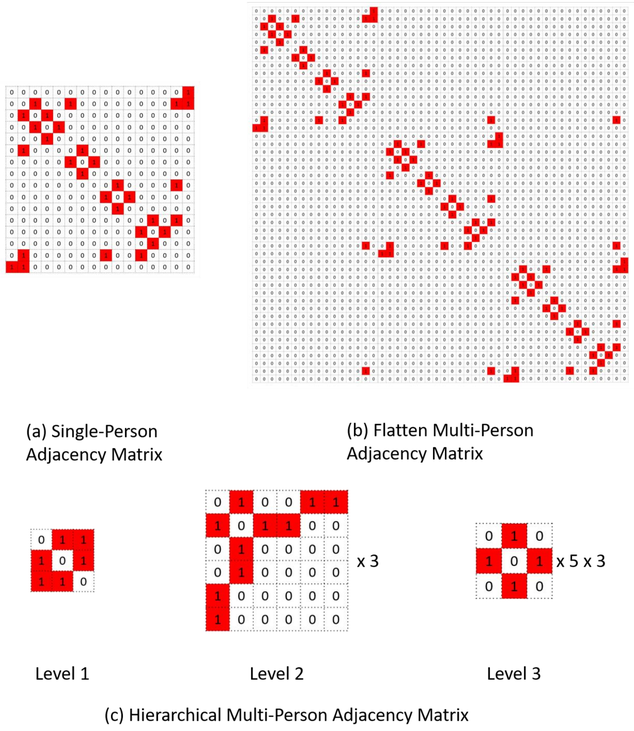
HDG-ODE: A Hierarchical Continuous-Time Model for Human Pose Forecasting

Yucheng Xing, Xin Wang

Department of Electrical and Computer Engineering, Stony Brook University

Motivation

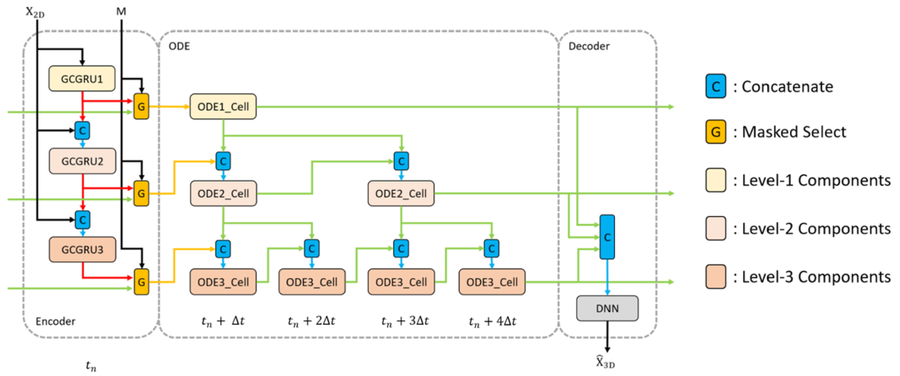
- Accurate and timely estimation and forecasting of human poses are important for further analyzing complex activities;
- Occlusion problem is severe in monocular images and the 2D poses extracted from them, which will influence the performance of 3D pose forecasting, but the temporal information can ease it in some degree;
- In multi-person scenarios, the relations among people are helpful for the pose forecasting, which are not well-used in most current methods;
- When using graphs to model human poses, there's a sparsity problem, which is more obvious in multi-person scenarios, as shown below. It is harmful to model efficiency;



Objective

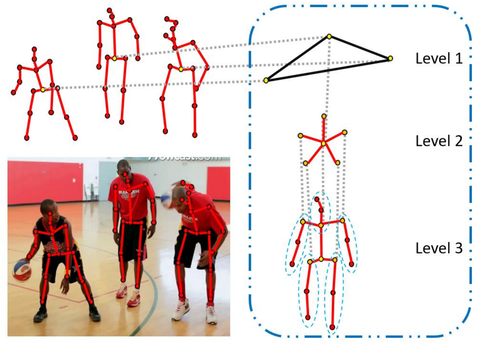
We design a new continuous model with the hierarchical structure, focusing on the restoration from a sequence of monocular multi-person 2D skeletons to predict the corresponding 3D poses at a future time point.

Methods



Hierarchical Dynamic Graph Ordinary Differential Equation (HDG-ODE)

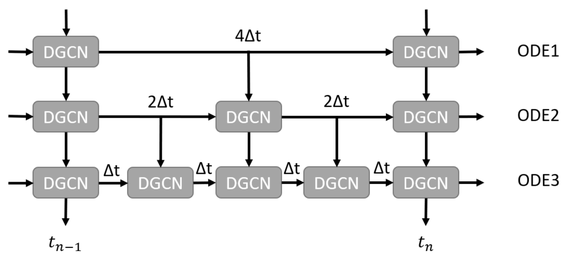
- Hierarchical Graph Structure
  - Reducing the sparsity of the adjacency matrix to reduce the meaningless calculation within graph convolutions;
  - Distinguishing different types of connections and process them separately;
  - Guiding the forecasting of lower level with the result of higher level in a cascade manner;



- Time-Varying Dynamic Graph Convolution

$$Z_t^{(l+1)} = \sigma(W Z_t^{(l)} \rho(W_{\tilde{A}} \odot \tilde{A}_t))$$

  - Learning an additional weight matrix $W_{\tilde{A}}$ to indicate the importance of different joints when doing the forecasting;
  - Adaptive to different joint missing cases caused by occlusions with dynamic matrix formed by $W_{\tilde{A}}$ and the normalized time-varying adjacency matrix $\tilde{A}_t$;
- Parallel Ordinary Differential Equation
  - The model is continuous, given 2D skeleton input of time t_n, the 3D prediction is not limited for the next observation time point t_{n+1}, but can be any time $\tau \in (t_n, t_{n+1}]$;
  - Hierarchical structure and parallel processing allow different propagation steps and model complexity at different levels;



Results

- Multi-Person Scenario

  According to the given occlusion masks in dataset, only observed joints are fed into the model as input

	MPJPE(↓)	PCKh@α (↑)		
		α=0.1	α=0.5	α=1.0
Discrete				
STGCN	0.6004	13.03%	60.35%	77.76%
Graph-GRU	0.5573	14.86%	62.39%	80.04%
SemGCGRU	0.4375	17.09%	69.61%	85.96%
Continuous				
ODE-RNN	0.4396	20.21%	71.78%	85.45%
Graph-ODE	0.4066	19.80%	71.57%	87.00%
Ours				
HDG-ODE	**0.3038**	**29.72%**	**78.67%**	**92.18%**

Table 1. Testing Performance of different models on MuPoTS-3D.

- Single-Person Scenario

  During experiments, we assume only partial of the 2D joints are available due to occlusions:
  - p_t: ratio of the frames within the 2D sequence are observable;
  - p_s: ratio of the joints available within each observable frame;

	$p_t = 0.8, p_s = 0.5$		$p_t = 0.6, p_s = 0.4$	
	MPJPE	PCKh@1.0	MPJPE	PCKh@1.0
STGCN	0.3966	80.19%	0.4376	75.82%
Graph-GRU	0.3723	83.34%	0.4188	77.39%
ODE-RNN	0.3527	84.33%	0.3826	80.66%
Graph-ODE	0.3473	83.98%	0.4019	78.54%
HDG-ODE	**0.3052**	**88.29%**	**0.3543**	**82.36%**

Table 2. Performance of different models on Human3.6M with different observable ratios.

Conclusion

- We propose a new continuous-time model called HDG-ODE to forecast the future 3D human pose representations in multi-person videos given the 2D pose sequence as input. Through experiments, our model performs better than those literature works;
- Currently, the maximum number of people in the scenes of the video is limited and pre-defined. In the future, we will make our model more adaptive;