Structural-RNN: Deep Learning on Spatio-Temporal Graphs

Authors: Ashesh Jain, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena

Speaker: Shuijing Liu

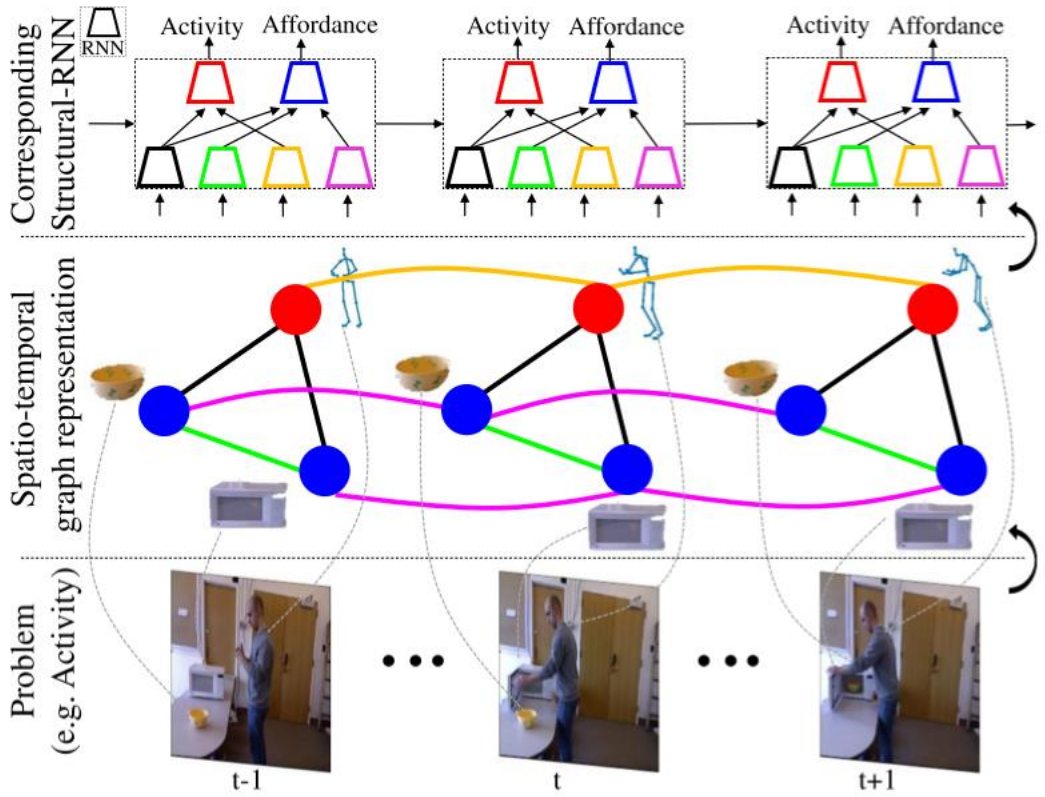
7/2/2020

Introduction

- Deep recurrent neural networks (RNN) are remarkably capable at modeling sequences, but lack an intuitive spatio-temporal structure
- Spatio-temporal graphs (st-graphs) are good at representing such spatio-temporal structures
  - The nodes represent the problem components, and the edges represent their spatio-temporal interactions
- This paper develop a general method to transform an arbitrary st-graph into a mixture of RNNs called structural-RNN(S-RNN) [1]
  - It can model the problems comprised of components that interact with each other in space and time, and is end-to-end trainable

Example problem

- Bottom: an example activity (human microwaving food) that requires spatial and temporal reasoning to model
- Middle: St-graph capturing spatial and temporal interactions between the human and 2 objects
- Top: structural-RNN architecture derived from st-graph



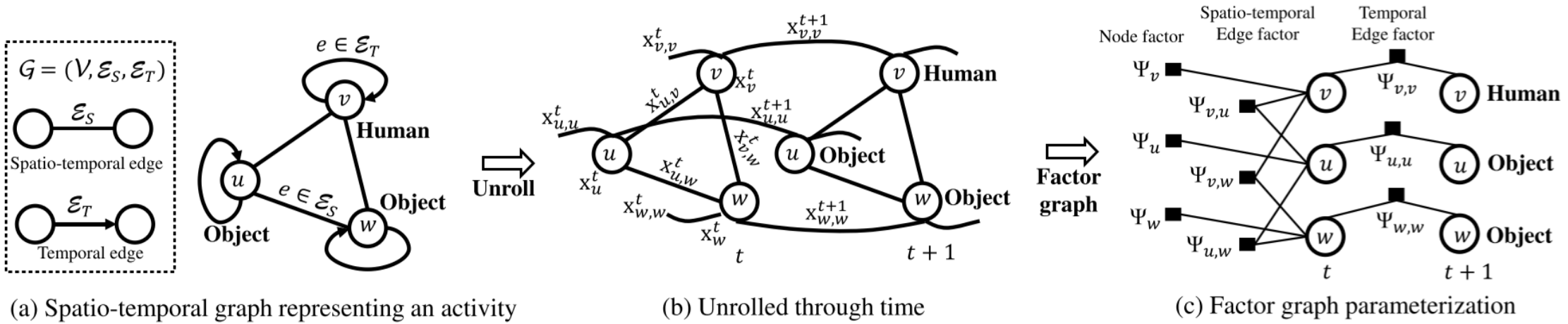
Representation of st-graphs

Formulating problems as st-graphs

(a) We can represent the previous example problem as a st-graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E}_S, \mathcal{E}_T)$, which consists of:

- Nodes v $\in \mathcal{V}$, spatial edges e $\in \mathcal{E}_S$, and temporal edges e $\in \mathcal{E}_T$

(b) In the unrolled st-graph, different nodes at the same time t are connected with *spatial* edges, and same nodes at adjacent time steps (e.x. t and $t+1$) are connected with *temporal* edges



(a) Spatio-temporal graph representing an activity

(b) Unrolled through time

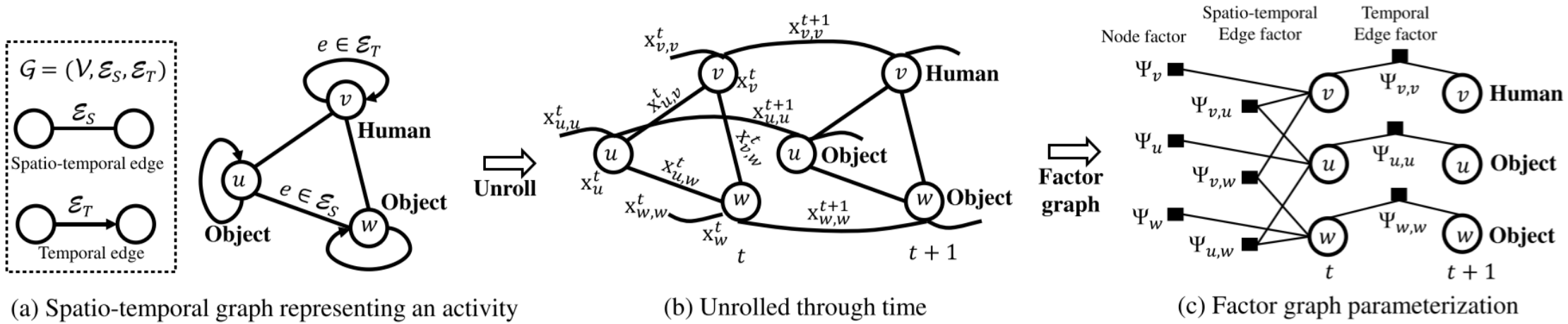
(c) Factor graph parameterization

Formulating problems as st-graphs

(c) Given an st-graph and the features of nodes x_v^t and edges x_e^t, the goal is to predict the node labels y_v^t, which is affected by both its node and its interactions with other nodes and edges. In the example problem,

- Node features = human and object poses
- Edge features = the relative positions of human and objects
- Node labels = the human activity and object affordance

Our factor graph has a factor function $\psi(y_v, x_v)$ for each node, and a pairwise factor $\psi_e(y_{e1}, y_{e2}, x_e)$ for each edge. In this way, st-graph **factorizes** a complicated function into many simpler functions.



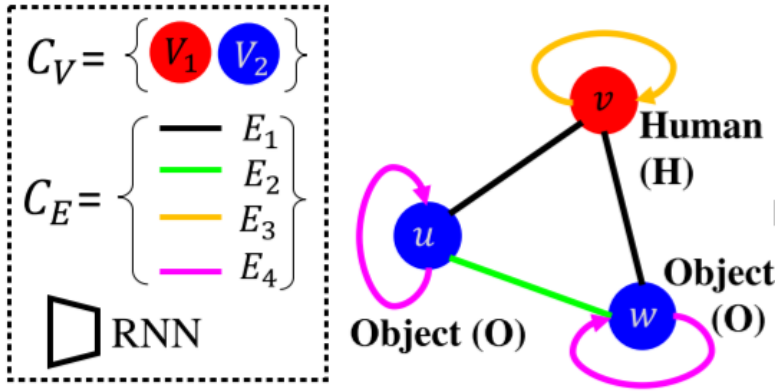
(a) Spatio-temporal graph representing an activity

(b) Unrolled through time

(c) Factor graph parameterization

Sharing factors between nodes

- Learning a different set of parameters for each factor is not scalable if the size of st-graph becomes larger
- To solve this, semantically similar nodes or edges can share factors. For example,
  - All “object nodes” (blue) in the example can share the same node factors and parameters
  - All human-object edges (black) can share the same edge factors and parameters
- Thus, we divide all nodes and edges into partitions, C_V and C_E, where each partition consists of nodes/edges sharing the same parameters

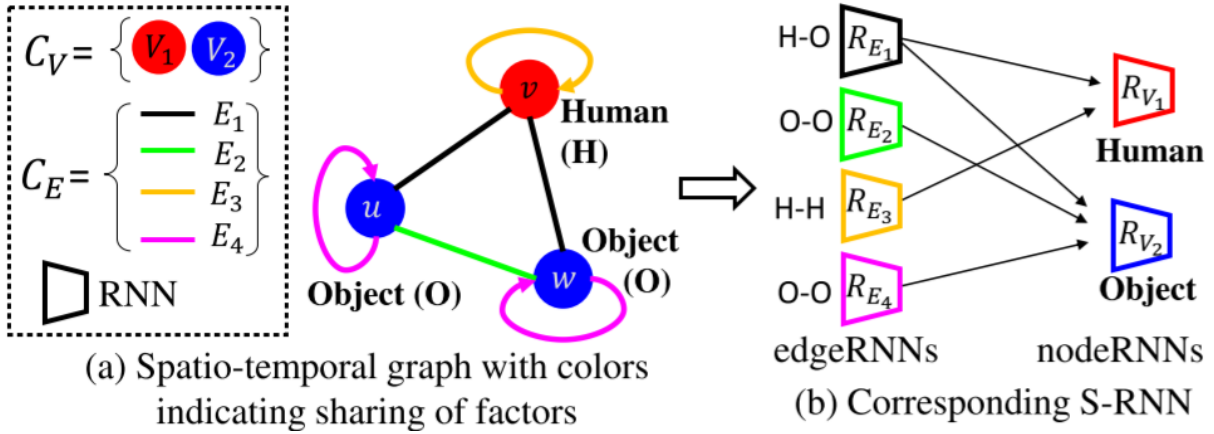


(a) Spatio-temporal graph with colors indicating sharing of factors

From st graphs to structural-RNN

Deriving S-RNN from st-graphs

- We represent each factor with an RNN:
  - the RNNs obtained from the node factors are referred as nodeRNNs or R_V
  - the RNNs obtained from edge factors are referred as edgeRNNs or R_E
- The interactions in st-graph are captured through connections between the nodeRNNs and the edgeRNNs
  - An edgeRNN is connected to a nodeRNN iff the edge and the node are neighbors in the st-graph
  - The nodeRNNs combine the outputs of the edgeRNNs they are connected to
  - The predictions of nodeRNNs interact through the edgeRNNs



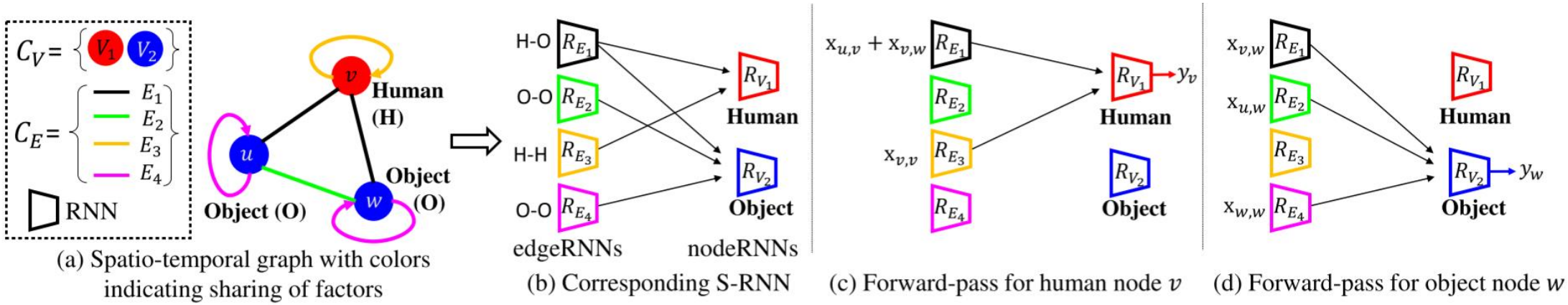
(a) Spatio-temporal graph with colors indicating sharing of factors

(b) Corresponding S-RNN

Training structural-RNN

Training the S-RNN

- In the forward pass of a node $v \in \mathcal{V}$,
  - The input into edgeRNN R_{E_m} is the sequence of edge features x_e^t on the edges $e \in E_m$ connected to v
  - The nodeRNN concatenates the node feature x_v^t and the outputs of edgeRNN it is connected to, and outputs the prediction label



(a) Spatio-temporal graph with colors indicating sharing of factors

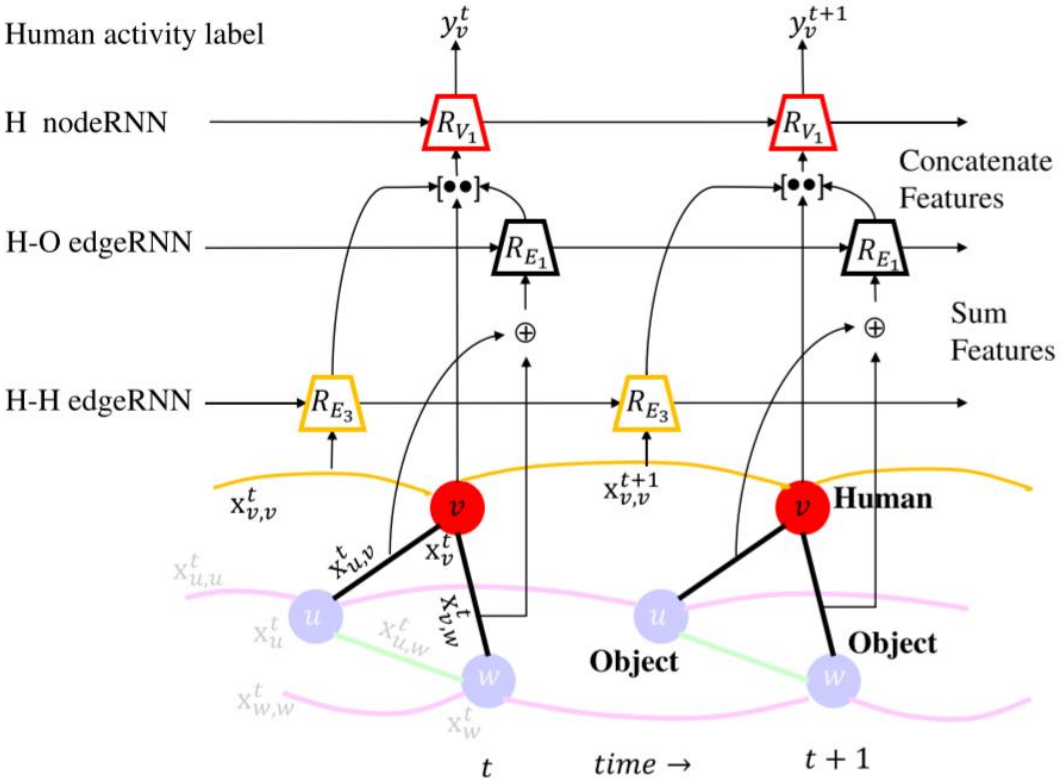
(b) Corresponding S-RNN

(c) Forward-pass for human node v

(d) Forward-pass for object node w

The forward pass

- In the forward pass, we feed the sum of spatial edge features to the human-object edgeRNN, R_{E_1} (black)
  - The summation instead of concatenation allows us to handle variable number of object nodes without changing the architecture
- We feed the temporal edge features to the human-human edgeRNN, R_{E_3} (yellow)
- The nodeRNN R_{V_1} (red) concatenates the human node features with the outputs from R_{E_1} and R_{E_3}, and predicts the human's activity at each time step

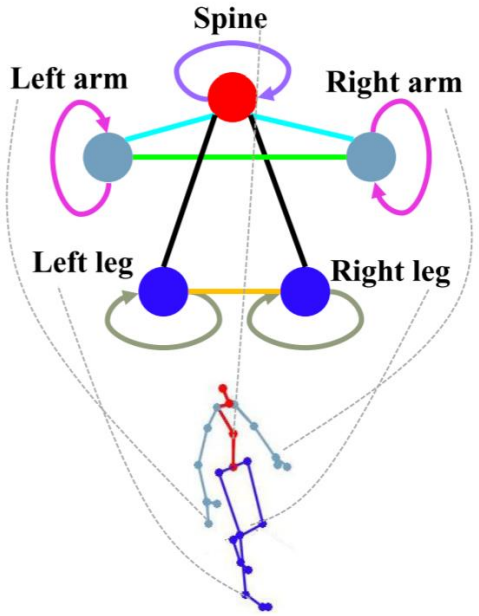


Forward-pass for human node v

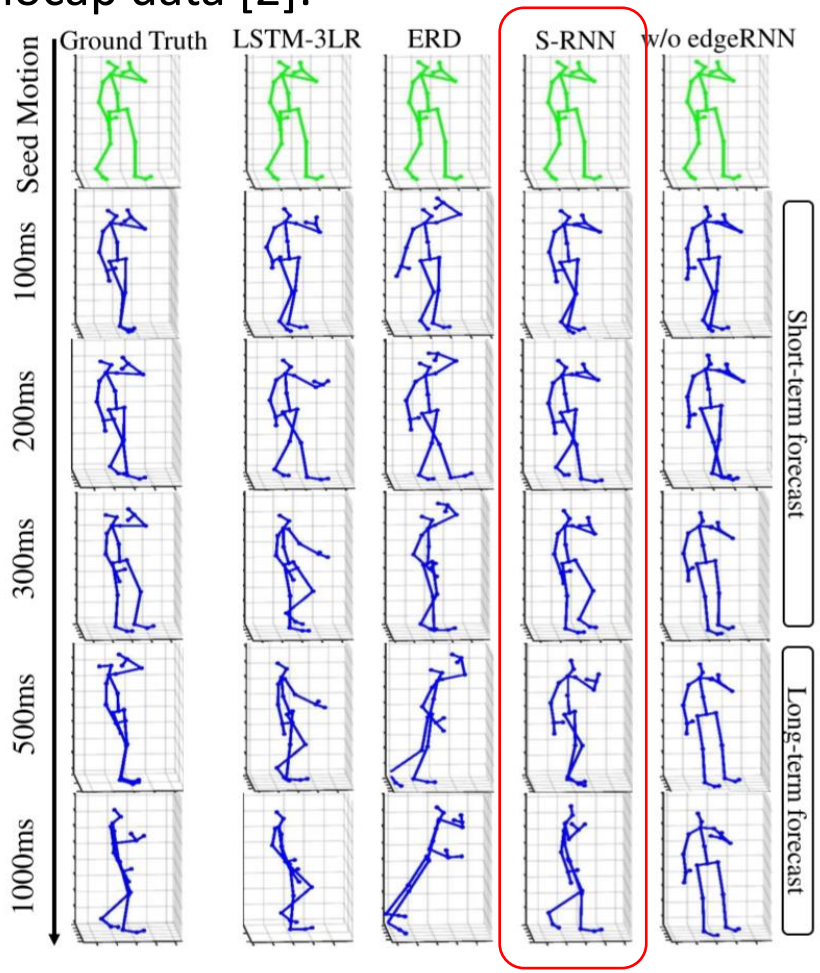
Applications of structural-RNN

Human motion modeling and forecasting

- Human body is a good example of separate but well related components
- To formulate a human body as a st-graph, the spine interacts with all body parts, and the arms and legs interact with each other

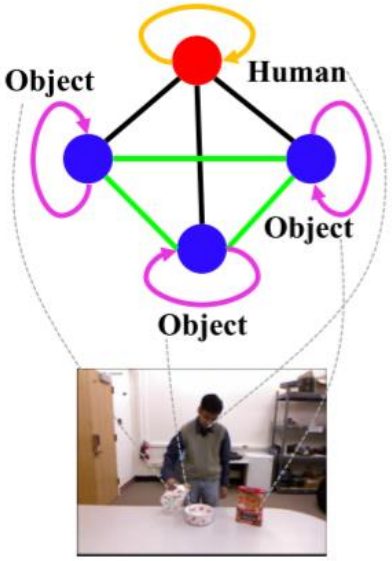


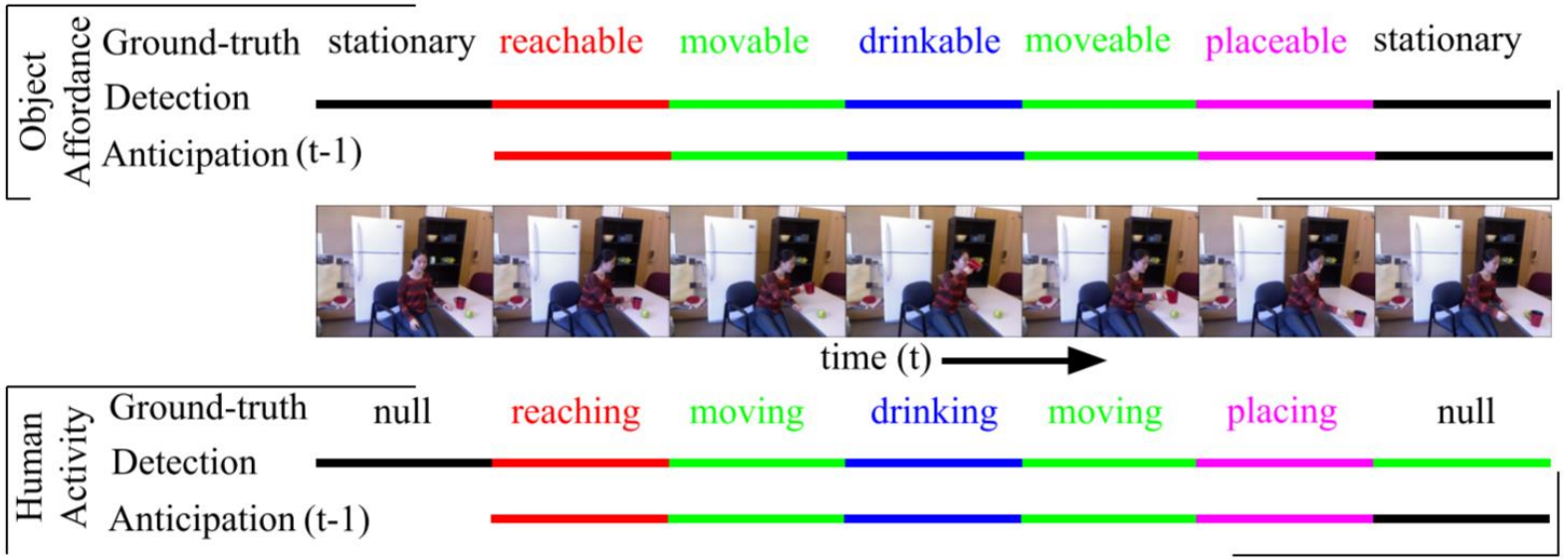
- Here's the result of forecasting eating activity on mocap data [2]:



Human activity detection and anticipation

- CAD-120 dataset [3] contains activities involve rich human-object interactions. Each activity consist of a sequence of sub-activities (e.x. moving, drinking) and objects' affordance (e.x. reachable, drinkable).
- The authors use S-RNN to detect and anticipate sub-activities and affordance (left figure). Qualitative results are in right figure.





References

[1] Jain, A., Zamir, A. R., Savarese, S., & Saxena, A. (2016). Structural-rnn: Deep learning on spatio-temporal graphs. In Proceedings of the ieee conference on computer vision and pattern recognition (pp. 5308-5317).

[2] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE PAMI, 36(7), 2014.

[3] H. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. IJRR, 32(8), 2013.