Stacked Spatio-Temporal Graph Convolutional Networks for Action Segmentation

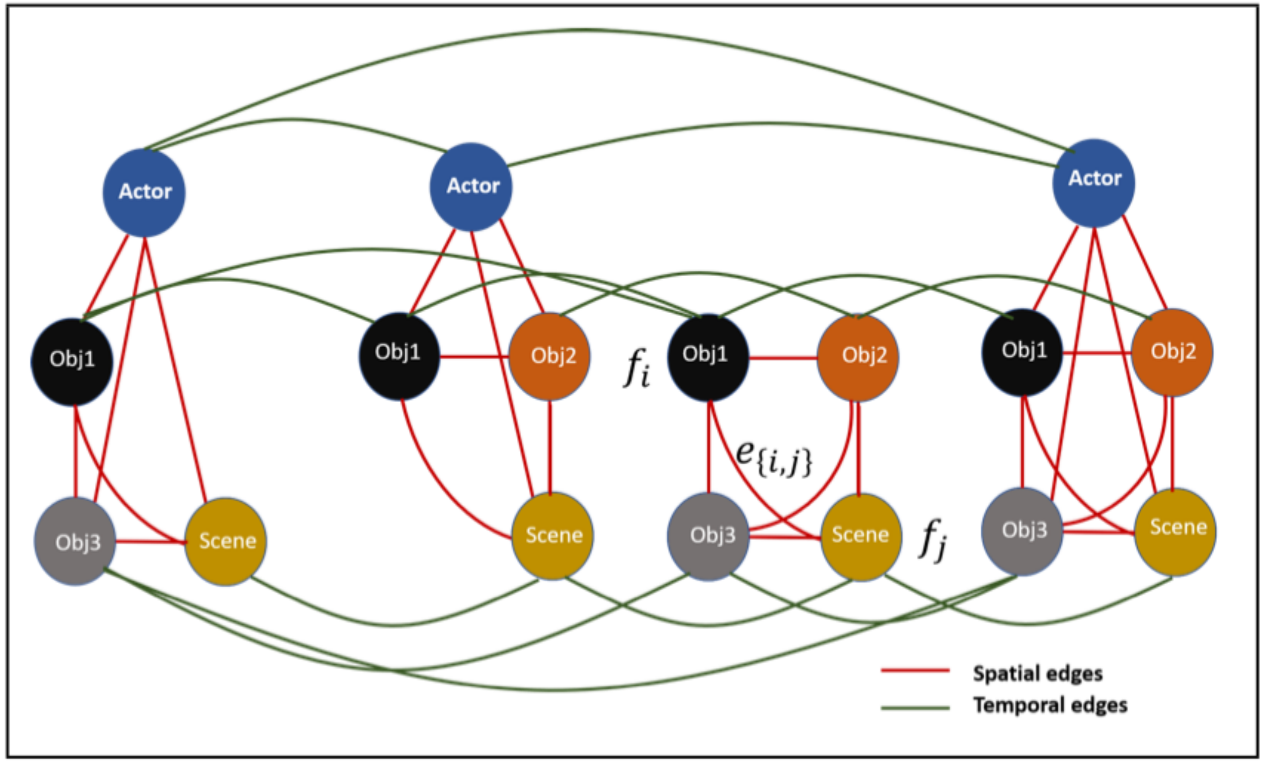
Pallabi Ghosh[1], Yi Yao[2], Larry S[1]. Davis, Ajay Divakaran[2]
[1]University of Maryland, [2]SRI International

WACV 2020

Innovations

- They proposed a stacked spatio-temporal graph convolutional network (STGCN) for action segmentation.
- The proposed network accounts for contextual cues (actors, objects, etc). However, the original STGCN accounts for skeletal joints.
- Original STGCN can only handle information across one consectitve time step. The proposed network can handle information over long video sequences.
- They introduced an extended use of stacked hourglass architecture on spatiotemporal graphs (first attempt in the field).

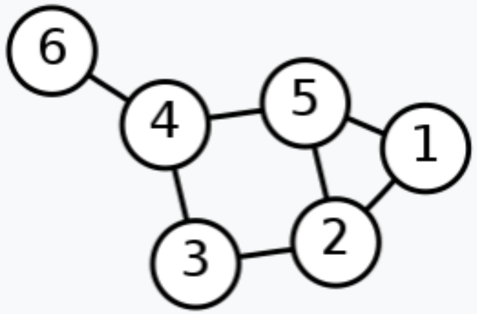
GCN



W: weight matrix
H: input matrix
$\hat{A} = I + A, A = [e_{i,j}]$
$e_{i.j}$: edge weights
$\widehat{D}$: node degree matrix of $\hat{A}$

$$H^{l+1} = g(H^l, A) = \sigma(\hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}H^l W^l) \quad (1)$$

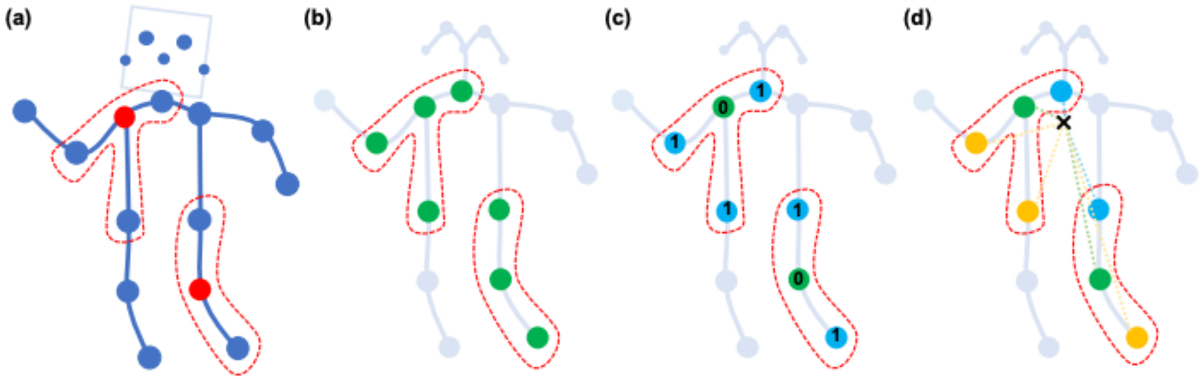
Laplacian matrix

Labelled graph	Degree matrix	Adjacency matrix	Laplacian matrix
	$\begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}$

Degree matrix: a diagonal matrix which contains information about the degree of each vertex.

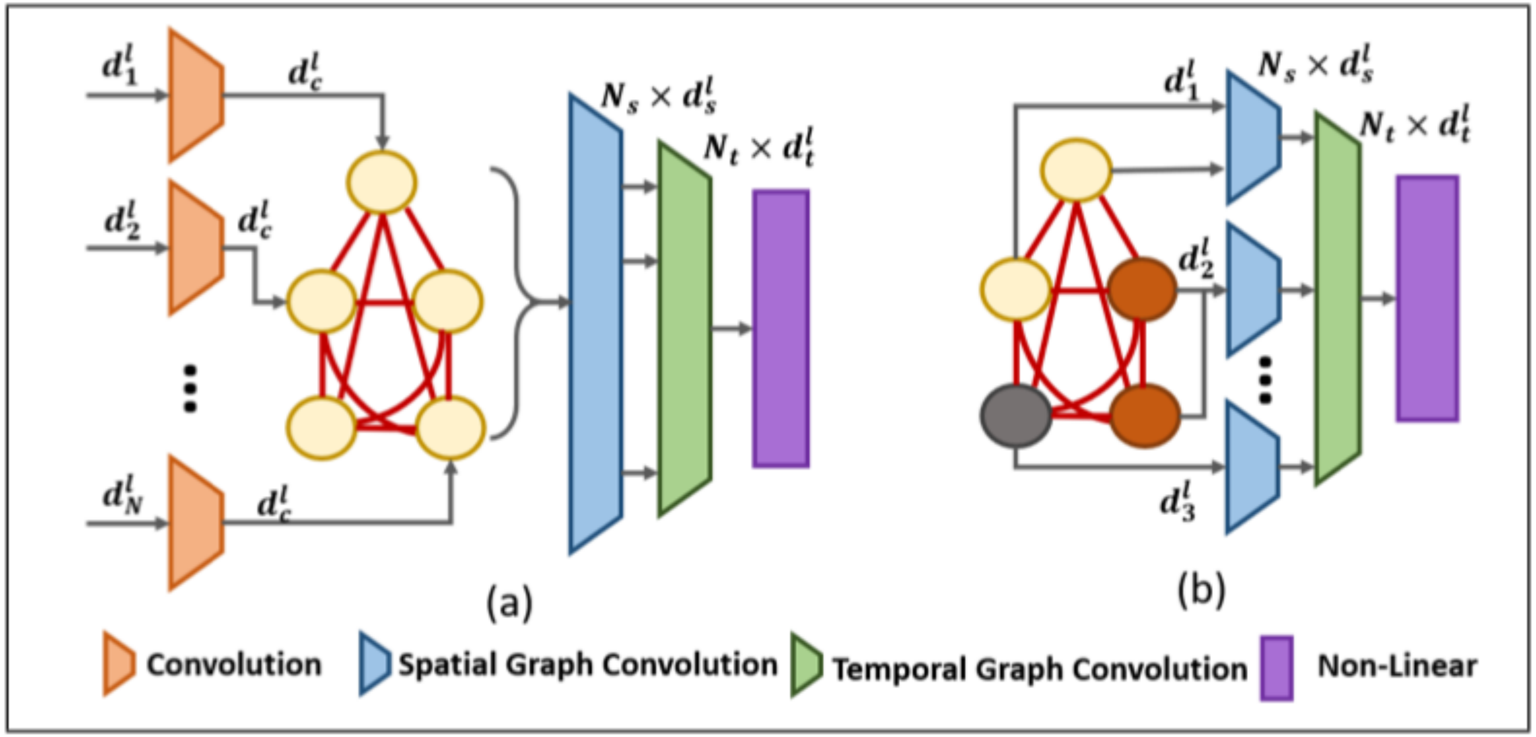
Adjacency matrix: a matrix indicate whether pairs of vertices are adjacent or not.

Original STGCN

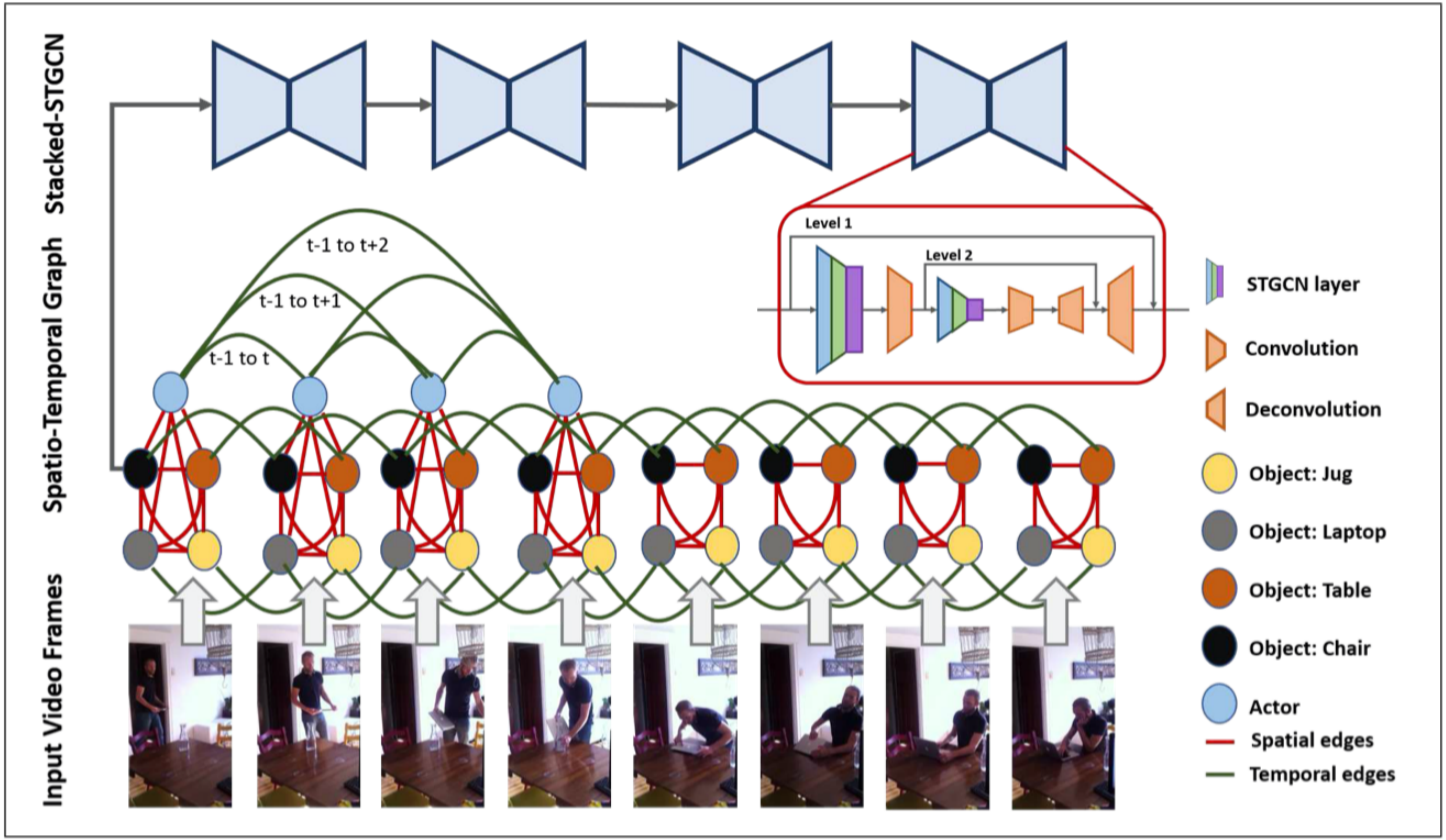


- Nodes are skeletal joints
- Spatial connections depend on physical adjacency of these joints.
- It is not directly applicable to their task since their network needs to handle action segmentation with contextual cues.

Spatio-temporal GCN



$$\boldsymbol{H}^{l+1} = g_t(\boldsymbol{H}_s^l, \boldsymbol{A}_t) = \sigma(\hat{\boldsymbol{D}}_t^{-1/2} \hat{\boldsymbol{A}}_t \hat{\boldsymbol{D}}_t^{-1/2} \boldsymbol{H}_s^l \boldsymbol{W}_t^l)$$
$$\boldsymbol{H}_s^l = g_s(\boldsymbol{H}^l, \boldsymbol{A}_s) = \hat{\boldsymbol{D}}_s^{-1/2} \hat{\boldsymbol{A}}_s \hat{\boldsymbol{D}}_s^{-1/2} \boldsymbol{H}^l \boldsymbol{W}_s^l \tag{2}$$



Stacked-STGCN
Spatio-Temporal Graph
Input Video Frames
t-1 to t+2
t-1 to t+1
t-1 to t
Level 1
Level 2
STGCN layer
Convolution
Deconvolution
Object: Jug
Object: Laptop
Object: Table
Object: Chair
Actor
Spatial edges
Temporal edges

Hourglass STGCN

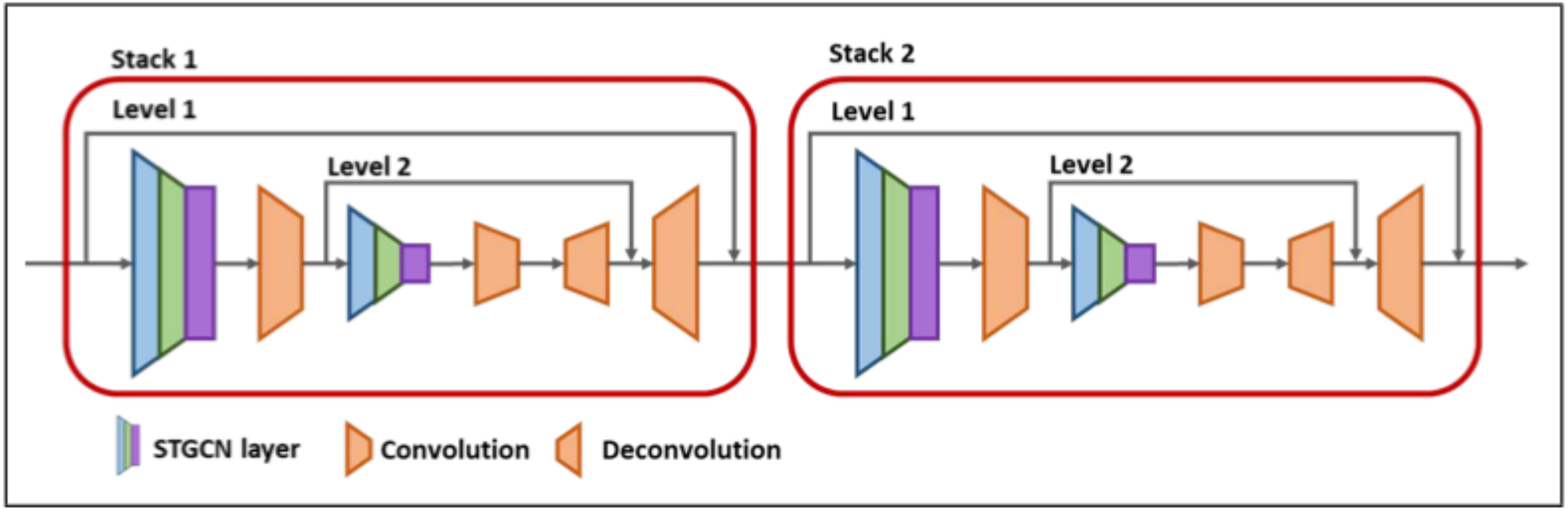


Figure 4. Illustration of stacked hourglass STGCN with two levels.

CAD120 experiment

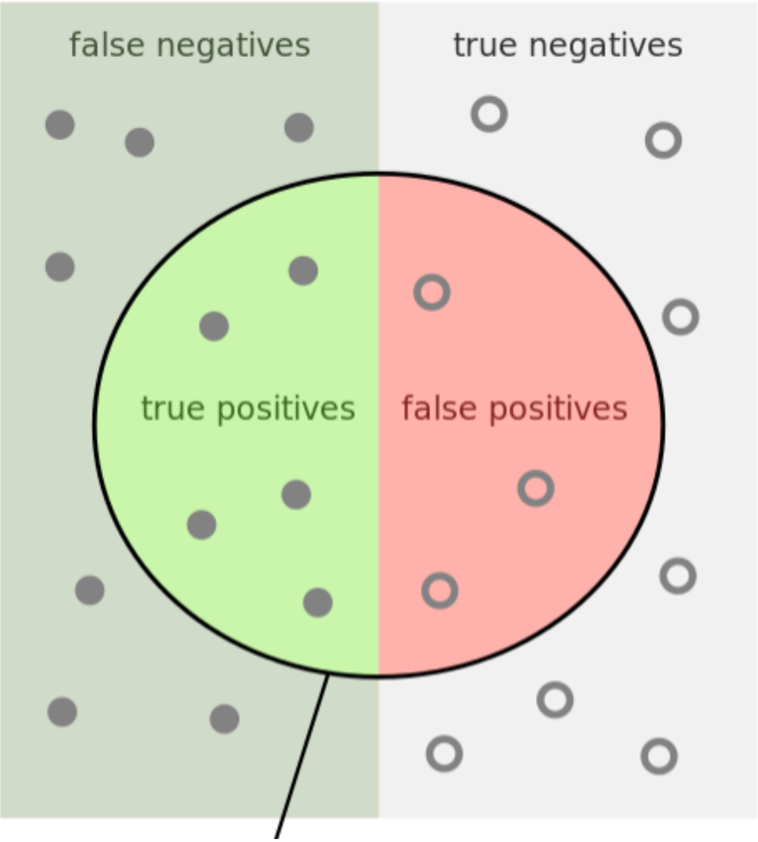
- 120 videos on 4 subjects as well as skeletal data.
- Actor nodes have length 630.
- Object nodes have length 180.

Method	**F1-score (%)**
Koppula et al. [20, 21]	80.4
S-RNN w/o edge-RNN [17]	82.2
S-RNN [17]	83.2
S-RNN(multitask) [17]	82.4
Ours (STGCN)	**88.5**

Table 1. Performance comparison based on the F1 score using the CAD120 dataset. Our STGCN improves the F1 score over the best reported result (i.e., S-RNN) by approximately 5.3%.

CAD 120 experiment

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$



CAD 120 experiment

Figure 5. Action segmentation results of our Stacked-STGCN on CAD120. Green/red: correct/erroneous detection.

Charades experiment

- 9848 videos, 157 action classes, 38 object classes, 33 verb class
- At each time step there can be more than one action label.
- Explored 2 types of features, one based on. VGG, and the other based on I3D.

Description
Scene Features
N1. FC7 layer output of VGG network trained on RGB frames
Motion Features
N2. FC7 layer output of VGG network trained on flow frames
Segment Features
N3. I3D pre-final layer output trained on RGB frames
N4. I3D pre-final layer output trained on flow frames
Actor Features
N5.GNN-based Situation Recognition trained on the ImSitu dataset
Object Features
N6. Top 5 object detection features from Faster-RCNN

Table 2. Features for the Charades dataset.

Charades experiment

(A1)	All Features; Baseline	8.13
(A2)	All Features; STGCN	10.26
(A3)	VGG-RGB; STGCN; 1 time step	6.77
(A4)	VGG-RGB; STGCN	7.06
(A5)	All Features; Stacked-STGCN; 1 time step	11.29
(A6)	VGG-RGB; Stacked-STGCN;	8.66
(A6)	VGG-RGB+VGG-Flow; Stacked-STGCN	10.94
(A7)	**All Features; Stacked-STGCN**	**11.73**

Table 3. Comparison of our Stacked-STGCN (A7) with baseline (A1), STGCN without hourglass (A2), different temporal connections (A3-A5), and different input features (A6). Input features include VGG-RGB for scene, VGG-Flow for motion, Situation Recognition for action, and Faster RCNN for object.

Charades experiment

Method	VGG mAP	I3D mAP
Baseline [30]	6.56	17.22
LSTM [30]	7.85	18.12
Super-Events [30]	8.53	**19.41**
Stacked-STGCN (VGG only)	**10.94**	
Stacked-STGCN (all features)	**11.73**	
Stacked-STGCN (I3D)		19.09

Table 4. Performance comparison based on mAP between our Stacked-STGCN and the best reported results published in [30] using the Charades dataset. Our Stacked-STGCN yields an approximate 2.41% and 3.20% improvement in mAP using VGG features only and all four types of features, respectively.

Charades experiment

Method	mAP
Random [44]	2.42
RGB [44]	7.89
Predictive-corrective [7]	8.90
Two-Stream [44]	8.94
Two-Stream + LSTM [44]	9.60
Sigurdsson *et al*. standard [44]	9.69
Sigurdsson *et al*. post-processing [44]	12.80
R-C3D [55]	12.70
I3D [5]	17.22
I3D +LSTM [30]	18.10
I3D+Temporal Pyramid [30]	18.20
I3D + Super-events [30]	**19.41**
I3D +Stacked-STGCN (ours)	**19.09**

Table 5. Performance comparison based on mAP with previous works using the Charades dataset.