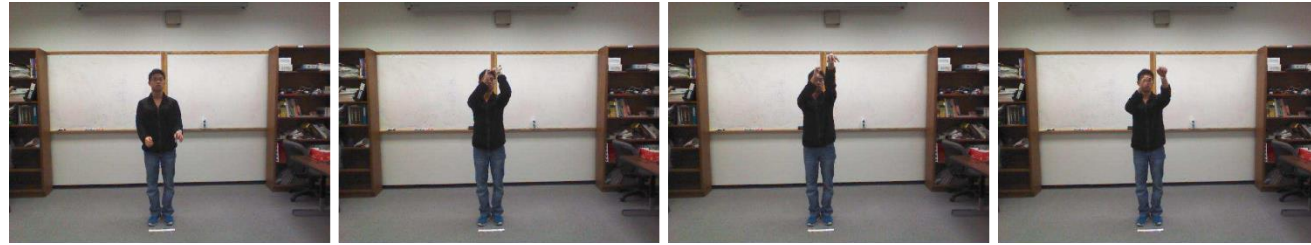# Human Action Recognition by Recurrent Neural Network (RNN)
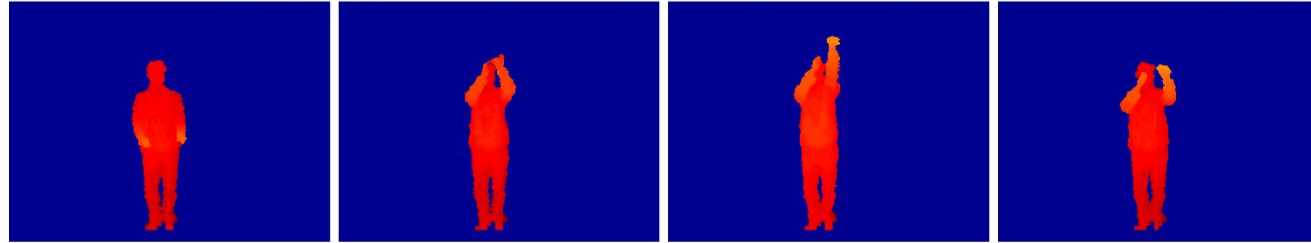
Zhifei Zhang
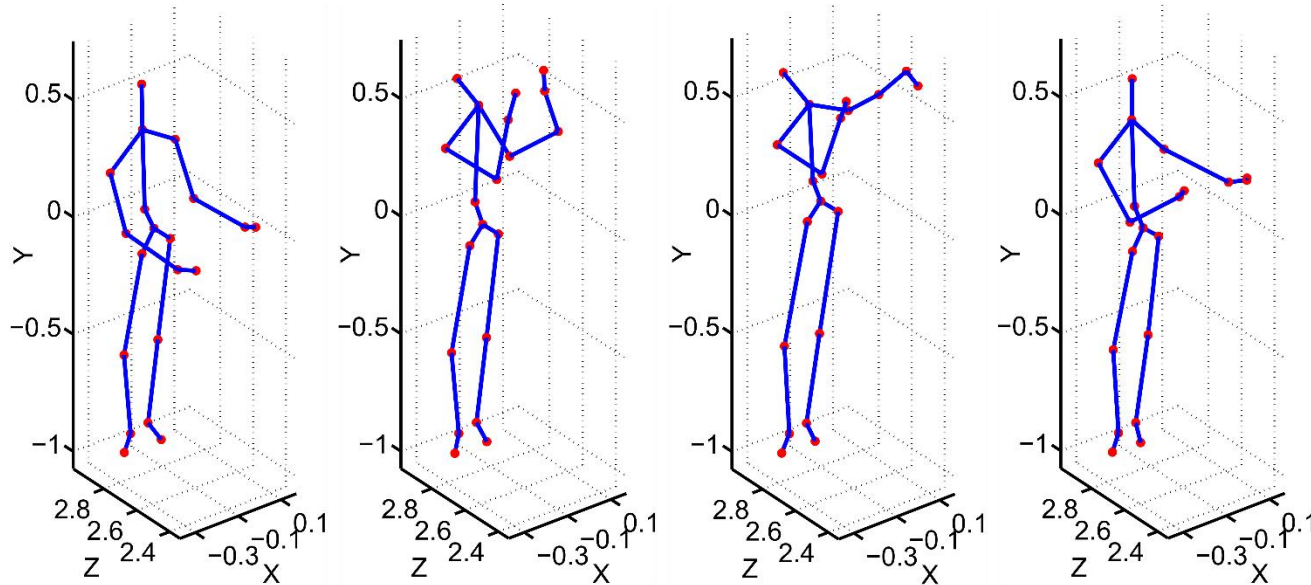
Human action recognition

RGB image
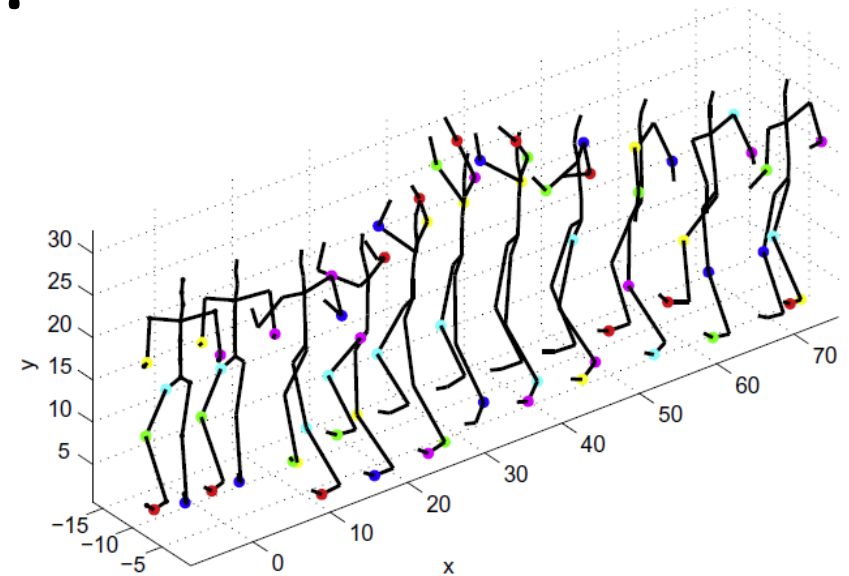

Depth map

Skeleton

From Single Image to Sequential Images

A traditional neural network assumes the inputs are **independent**, e.g., NN, AE, RBM, and CNN
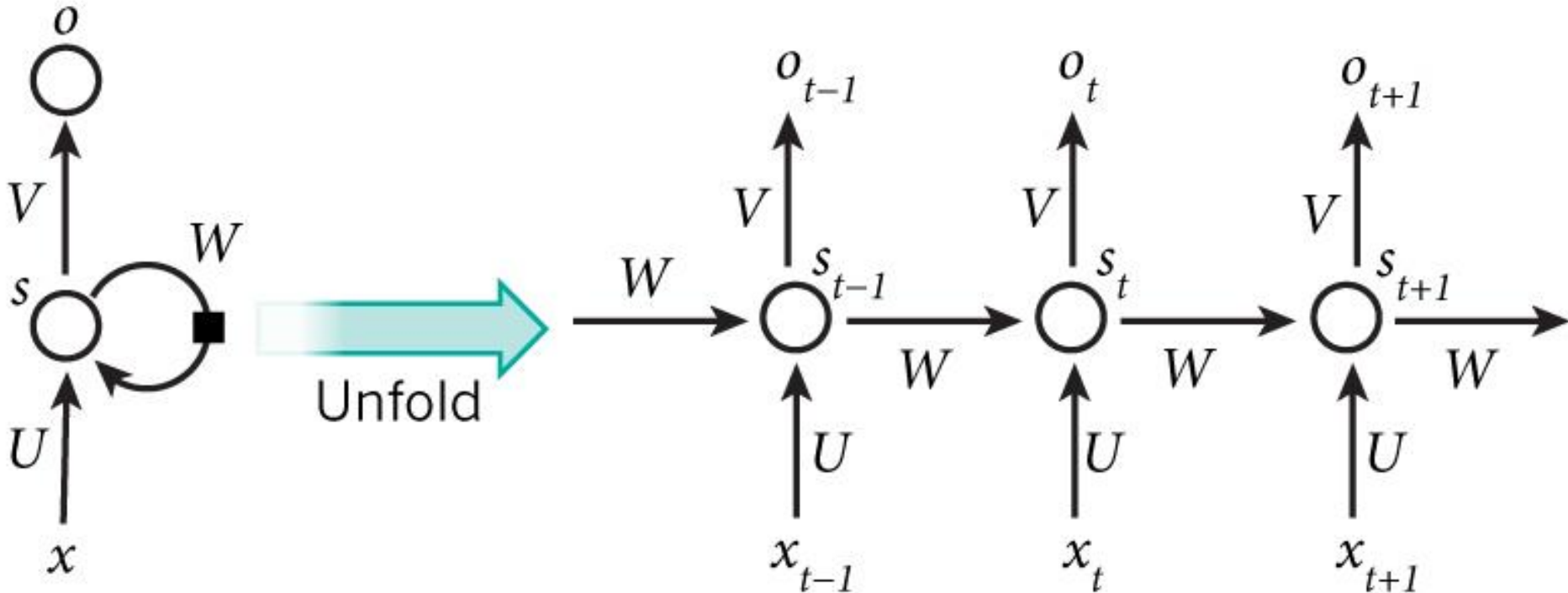
However, the video frames are **correlated**.

Recurrent Neural Network (RNN) is good at modeling dependencies, e.g., temporal or spatial dependencies.
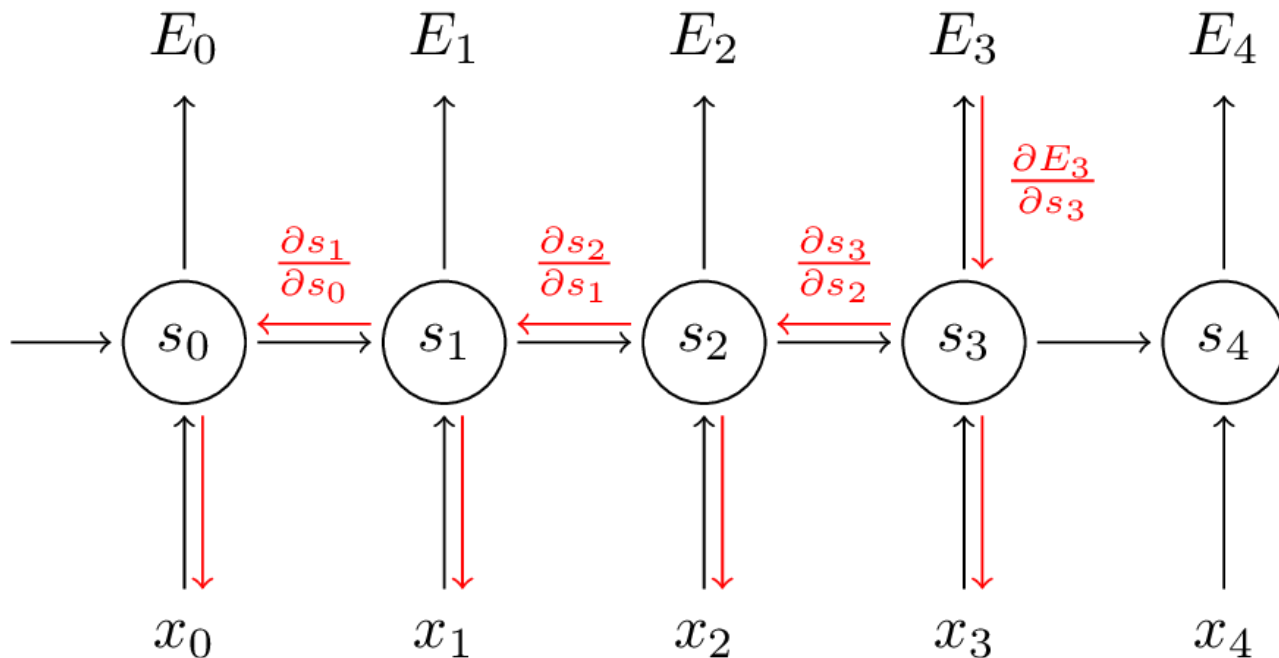
Recurrent Neural Network (RNN)
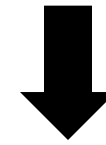
Feed forward: $o_t = \text{softmax}(V \times s_t)$

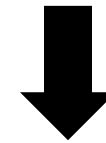$$s_t = \tanh(Ux_t + Ws_{t-1})$$

Backpropagation Through Time (BPTT)

Vanishing Gradient Problem:
gradient values shrinks exponentially

⬇

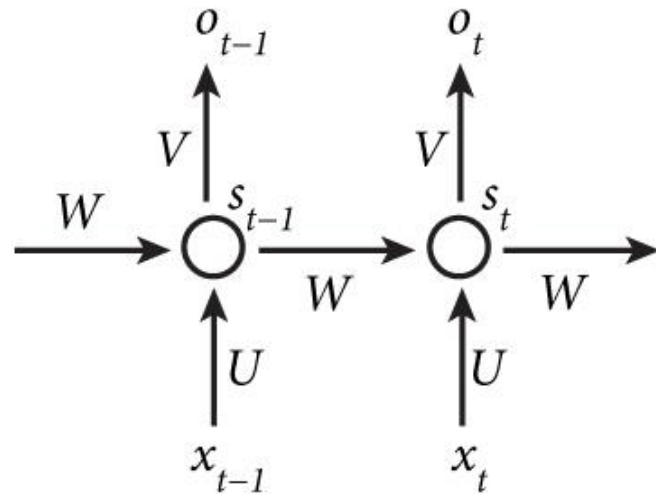**States far away do not contribute
to current learning**

⬇

**Fail to learn long-range
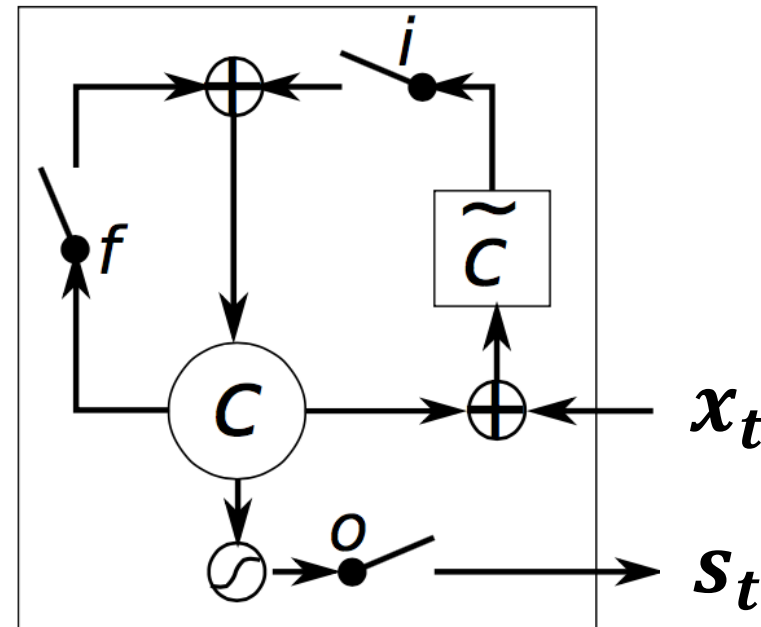dependencies**

Long-Short Term Memory (LSTM)

[Hochreiter al et., 1997]

LSTM is a way of updating S_t to prevent vanishing problem

Classic RNN: $\quad s_t = \tanh(Ux_t + Ws_{t-1})$

RNN with LSTM:

Related Works in Recent Years

(RNN refers to RNN with LSTM in later slides)

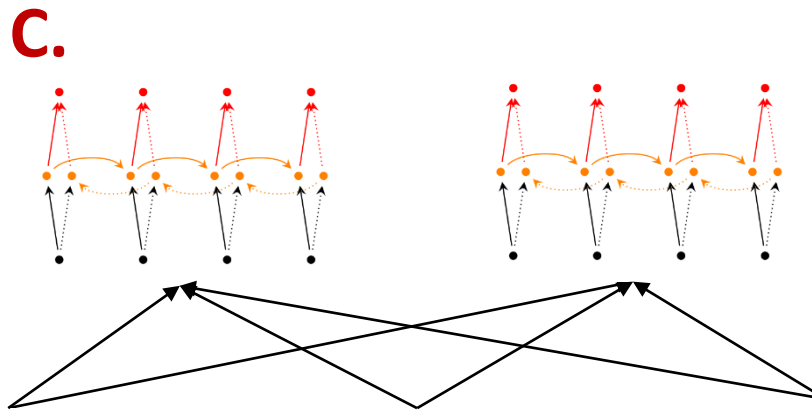| Conference	Description
PAMI 2013	3DCNN (CNN on video cube)
ICCV 2015	Factorized Spatiotemporal CNN
ICCV 2015	dRNN (modified LSTM)
ICCV 2015	Encoder + RNN + Decoder
CVPR 2015	BRNN (Hierarchical Bidirectional RNN)
CVPR 2015	CNN + RNN
CVPR 2016	CNN + BRNN
CVPR 2016	Regularizing LSTM (Encoder + LSTM)

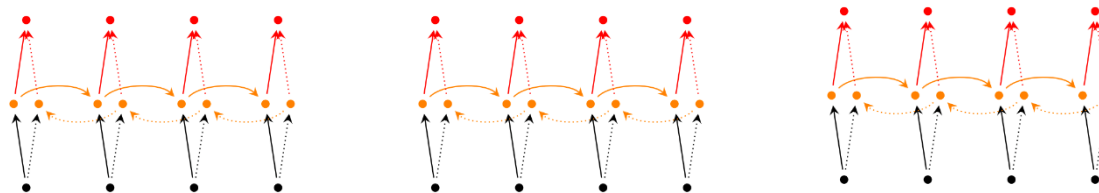# Five Directions on RNN-based Methods

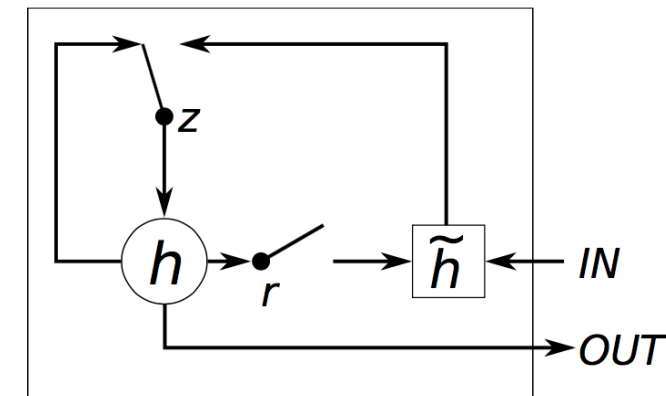A. **Bidirectional RNN**
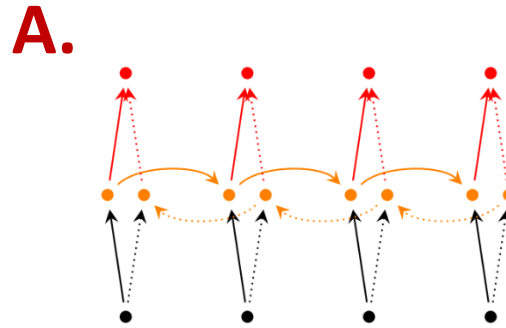
B. **Deep RNN**
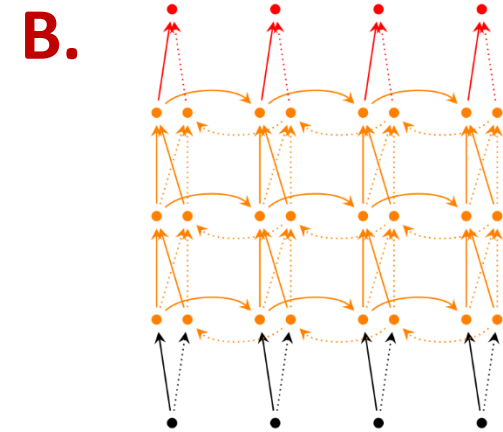
C. **Hierarchical RNN**

D. **LSTM Modification**

E. **CNN (AE) + RNN**

A.

B.

C.

D. GRU

Performance of RNN as Compared to Other Methods

Experimental results on MSR Action3D dataset
(*on subsets that is easier)

Conference	Description	Accuracy (%)
ECCV 2014	HOPC	91.64
CVPR 2014	Lie Group	92.46
CVPR 2015	HON4D	88.89
ICCV 2015	RNN	87.78
ICCV 2015	dRNN (modified LSTM)	92.03
CVPR 2015	BRNN (Hierarchical Bidirectional RNN)	94.49*
	Delay Embedding	*93.77 / 94.52**

THANK YOU
for your
ATTENTION!