IEEE Xplore® Digital Library

ISCAS 2020

2020 IEEE International Symposium on Circuits and Systems
Virtual, October 10-21, 2020

Bidirectional Independently Recurrent Neural Network for Skeleton-based Hand Gesture Recognition

Shuai Li[*^], Longfei Zheng[*], Ce Zhu[*], Yanbo Gao[*^]

[*]University of Electronic Science and Technology of China (UESTC)
[^]Shandong University (SDU)

2020 IEEE International Symposium on Circuits and Systems
Virtual, October 10-21, 2020

Outline

- Introduction
  - Skeleton-based hand gesture recognition
  - Recurrent Neural Network
- Proposed bidirectional independently recurrent neural network (Bi-IndRNN) for Skeleton-based Hand Gesture Recognition
- Experimental results
- Conclusion

Skeleton-based Hand Gesture Recognition

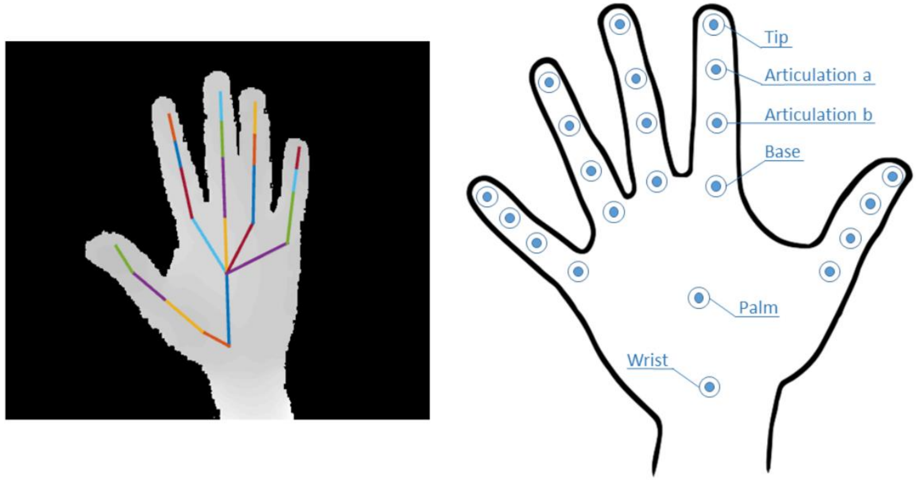


Fig. 1. Definition of the hand skeleton used in the DHG 14/28 Dataset (22 joints).

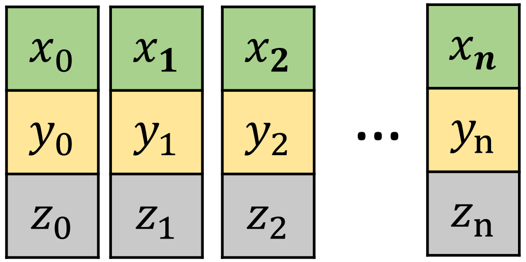


Fig. 2. Sequence Of Each Joint World Coordinate

Gesture categories:

Grab (G),
Tap (T),
Expand(E),
Pinch (P),
Rotation clockwise (RC),
Rotation counterclockwise (RCC),
Swipe right (SR),
Swipe left (SL),
Swipe up (SU),
Swipe down (SD),
Swipe x (SX),
Swipe + (S+),
Swipe v (SV),
Shake (SH)

Related Work

- Handcrafted features
  - Shape of connected joints or hand orientation
  - Dynamic time warping or hidden Markov model
- CNN based methods
  - 3DCNN on joint coordinates
- RNN based methods
  - RNN and LSTM for temporal processing

Motivation

Recurrent Neural Network:

The equation of RNN:

$$h_t = \sigma(Wx_t + Uh_{t-1} + b)$$

RNN appropriate for sequence modelling

Limit of RNN:

- Gradient vanishing and explosion problem
- Difficult to construct deep networks

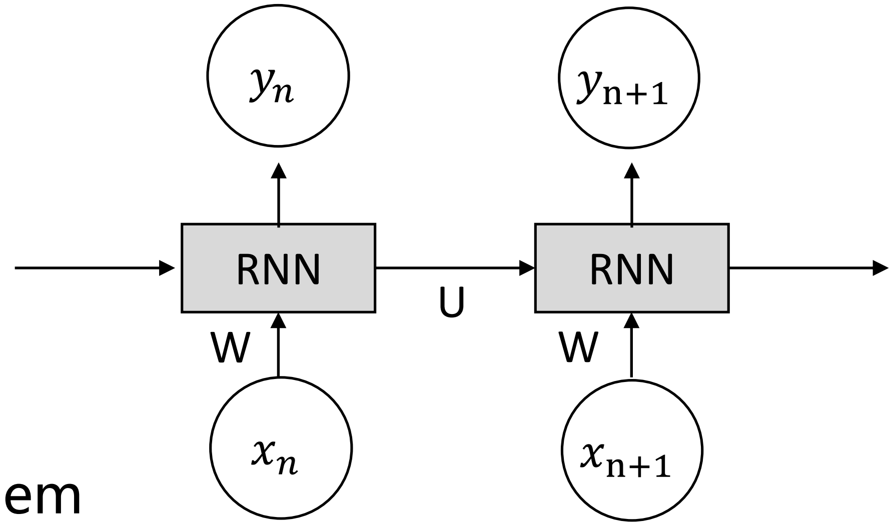


Illustration of RNN.

Motivation

Independently Recurrent Neural Network :

$$h_t = \sigma(Wx_t + u \odot h_{t-1} + b)$$

Gradient backpropagation through time:

$$\frac{\partial J_n}{\partial h_{n,t}} = \frac{\partial J_n}{\partial h_{n,T}} u_n^T \prod_{k=t}^{T-1} \sigma'_{n,k+1}$$

- The gradient vanishing and exploding problems can be effectively solved by regulating the recurrent weights.
- IndRNN with ReLU can be robustly trained.
- Multiple layers of IndRNNs can be efficiently stacked to explore deep features.

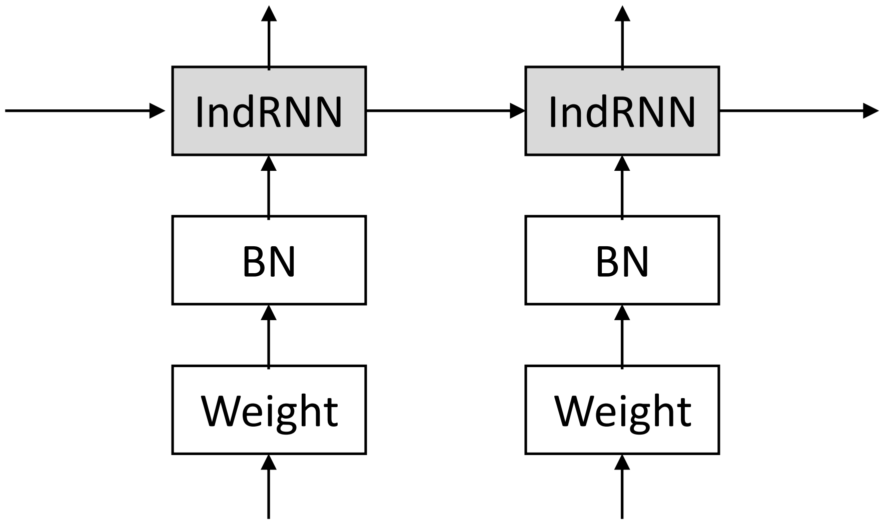


Illustration of IndRNN

Motivation

Similar hand gestures with different directions

Swipe up (SU)
Swipe down (SD)

Similar gesture features (Swipe Vertically)

Swipe left (SL)
Swipe right (SR)

Similar gesture features (Swipe Horizontally)

Rotation clockwise (RC)
Rotation counterclockwise (RCC)

Similar gesture features (Rotational gesture)

Similar gesture, different directions

Similar spatial features, different temporal features

Idea:

Using the same network structure to extract similar structural features.

Using reverse structure to extract features of different directions.

Proposed Bidirectional IndRNN Architecture.

The equation of IndRNN:

$$h_t = \sigma(Wx_t + u \odot h_{t-1} + b)$$

The equation of bidirectional IndRNN:

$$h_{f,t} = \sigma(W_f x_t + u_f \odot h_{f,t-1} + b_f)$$

$$h_{b,t} = \sigma(W_b x_t + u_b \odot h_{b,t-1} + b_b)$$

$$h_t = concat(h_{f,t}, h_{b,t})$$

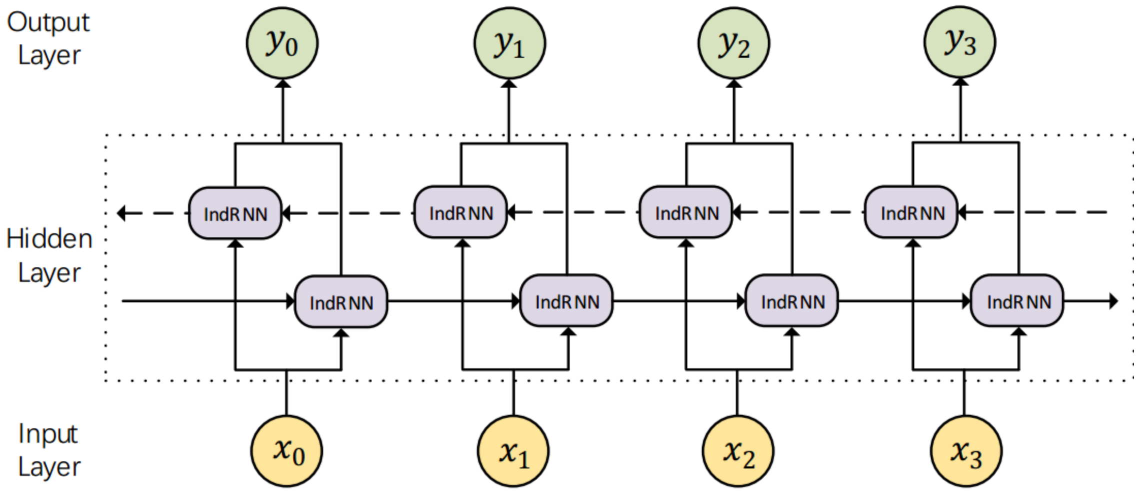


Illustration of the proposed bidirectional IndRNN architecture.

Extract features from more time intervals

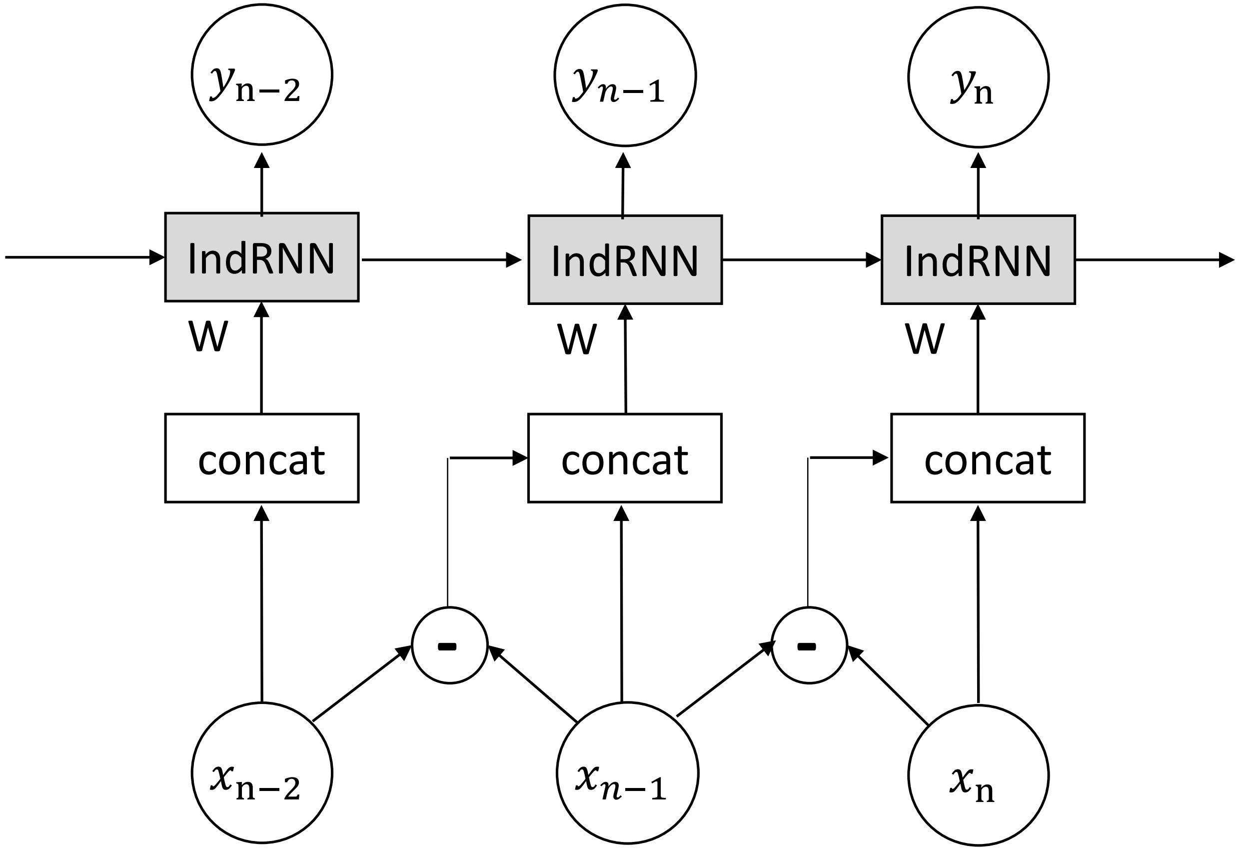
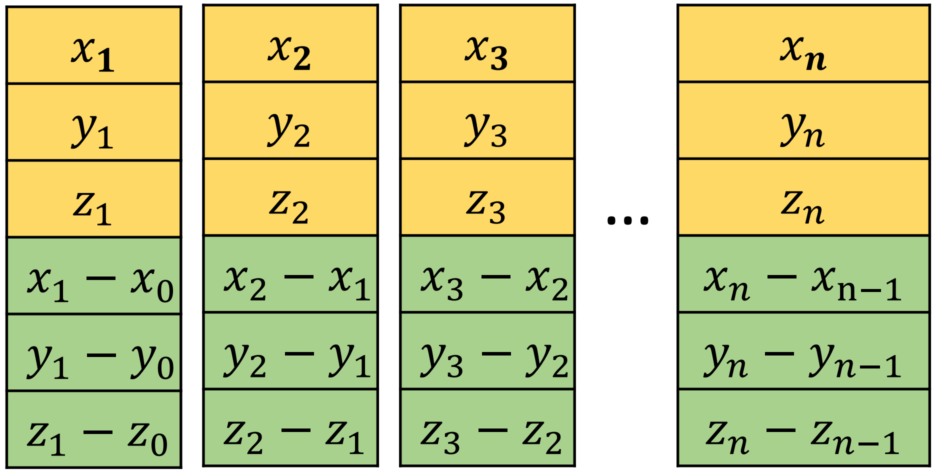


Illustration of concatenating the temporal displacement.

Features in the temporal domain (temporal displacement)



Proposed 6-layer Bi-IndRNN Network

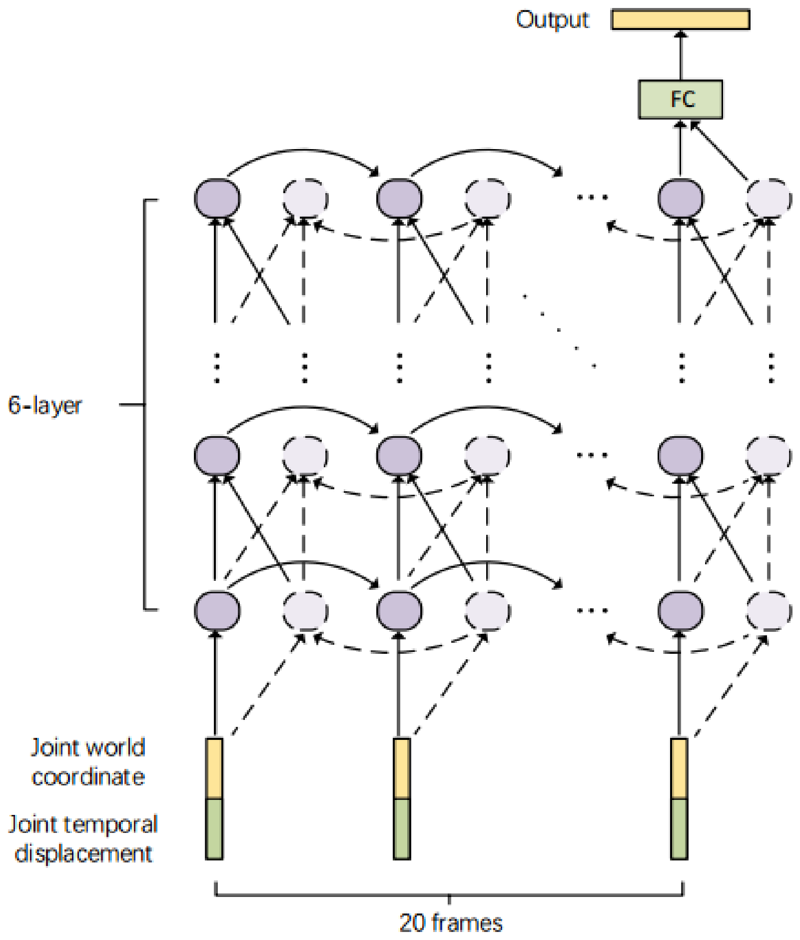
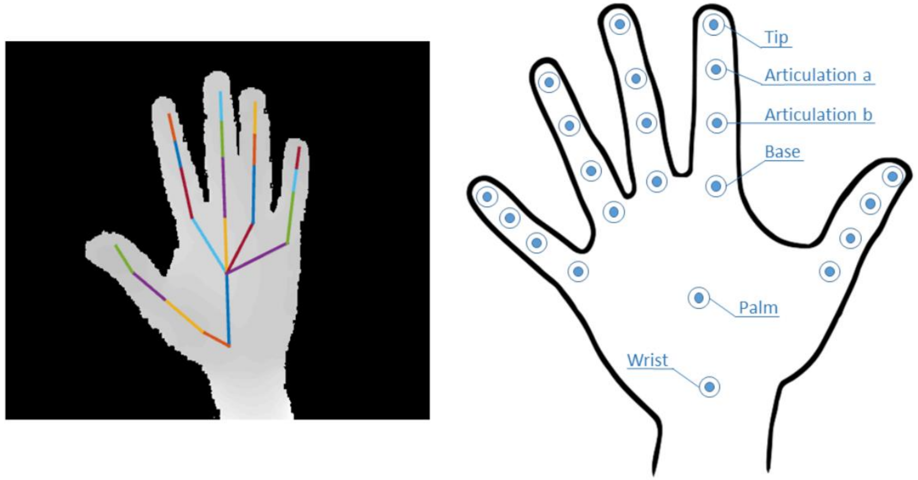


Illustration of the proposed 6-layer Bi-IndRNN network for skeletal gesture recognition.

6 layers of Bi-IndRNN
FC for classification

Experiments On DHG 14/28 Dataset

DHG(Dynamic Hand Gesture) Dataset：



Definition of the hand skeleton used in the DHG 14/28 Dataset.

2800 sequences in total
1960 sequences for training
837 sequences for testing

Gesture categories:

Grab (G),
Tap (T),
Expand(E),
Pinch (P),
Rotation clockwise (RC),
Rotation counterclockwise (RCC),
Swipe right (SR),
Swipe left (SL),
Swipe up (SU),
Swipe down (SD),
Swipe x (SX),
Swipe + (S+),
Swipe v (SV),
Shake (SH)

Experiments On DHG 14/28 Dataset

Setup:

Layers: 6

Number of Neurons of each layer: 512

Optimizer: Adam

Batch Size: 128

Dropout probability : 0.2

Learning rate: initial to 2 $*10^{-4}$ and is decayed by 10 when the accuracy of the validation set has not improved with patience 20.

Experiments On DHG 14/28 Dataset

Method	14 gestures	28 gestures
Histogram of Oriented 4D Normals	78.53	74.03
Shape Analysis of Motion Trajectories on Riemannian Manifold	79.61	62.00
Convolutional neural network for key frames	82.90	71.90
Joint angles similarities and HOG2	83.85	76.53
Motion feature augmented recurrent neural network	84.68	80.32
SoCJ + HoHD + HoWR	88.24	81.90
Parallel convolutional neural network	91.28	84.35
IndRNN(joint coordinate)	**92.07**	**85.82**
IndRNN(joint coordinate + displacement)	**92.19**	**88.87**
Bi-directional IndRNN(joint coordinate + displacement)	**93.15**	**91.13**

Table 1. Result On The DHG Dataset in Terms Of The Accuracy.

Experiments On DHG 14/28 Dataset

	G	T	E	P	RC	RCC	SR	SL	SU	SD	SX	S+	SV	SH
G	78.7	5.5	0.09	14.46	0.89	0.09				0.27				
T	0.76	95.8		0.29	2.01			0.67			0.38			0.1
E	0.66	2.78	94.02			0.08				2.46				
P	9.82	3.33	0.09	84.54	1.02	1.2								
RC	2.37	0.2		1.97	92.4		1.15	0.2						1.7
RCC	1.06			1.95		96.37		0.09		0.53				
SR					1.65		97.76				0.46	0.13		
SL		0.08		0.08	2.51	2.18		94.9		0.24				
SU	0.55		5.26	1.52		0.14			92.53					
SD	2.37	1.54	0.96	1.92				0.32		91.55			1.43	
SX								0.15			95.74		4.11	
S+							0.16	0.33		0.08		99.42		
SV							0.19		0.48		0.58	0.48	98.27	
SH	1.32	1.54				1.83		0.8	0.95		0.88	1.02	0.15	91.51

Table 2. Confusion matrix for DHG-14 using the proposed bi-IndRNN network.

Experiments On DHG 14/28 Dataset

	G(1)	G(2)	T(1)	T(2)	E(1)	E(2)	P(1)	P(2)	RC(1)	RC(2)	RCC(1)	RCC(2)	SR(1)	SR(2)	SL(1)	SL(2)	SU(1)	SU(2)	SD(1)	SD(2)	SX(1)	SX(2)	S+(1)	S+(2)	SV(1)	SV(2)	SH(1)	SH(2)
G(1)	65.15	4.55	1.52				24.24		4.55																			
G(2)		79.01						19.75				1.23																
T(1)	3.33	1.11	87.78				1.11		2.22	3.33																	1.11	
T(2)	1.09			95.65				3.26																				
E(1)					90.72	6.19											3.09											
E(2)						100																						
P(1)	6.73	0.96	1.92				78.85	1.92	1.92		4.81								2.88									
P(2)		15.38						78.02				2.2							1.1	3.3								
RC(1)			1.3				3.9		87.01	5.19																	2.6	
RC(2)									5.75	93.1										1.15								
RCC(1)	2.15						3.23		1.08		88.17	5.38																
RCC(2)											1.18	96.47															2.35	
SR(1)										3.19			85.11	10.64											1.06			
SR(2)													5.94	94.06														
SL(1)									5.56	2.22	1.11				85.56	5.56												
SL(2)										3.3					1.1	95.6												
SU(1)					2.2	3.3	2.2						2.2				86.81	3.3										
SU(2)																		100										
SD(1)																			96.15	3.85								
SD(2)																			0.96	99.04								
SX(1)			1.3														1.3				89.61				6.49		1.3	
SX(2)														2.47							1.23	91.36				4.94		
S+(1)																							97.12	2.88				
S+(2)																							1.14	98.86				
SV(1)						0.9			0.9				0.9						1.8		2.7		0.9		90.99	0.9		
SV(2)														3.45											2.3	94.25		
SH(1)									1.87			0.93															97.2	
SH(2)				0.89																								99.11

Table 3. Confusion matrix for DHG-28 using the proposed bi-IndRNN network.

Conclusion

- Propose a Bidirectional IndRNN (Bi-IndRNN)
- Combine temporal displacement to enhance the input features
- State-of-the-art performance on DHG 14/28 dataset(93.15% for the 14 gesture classes case and 91.13% for the 28 gesture classes case).

Thanks