DeepGRU: Deep Gesture Recognition Utility

University of Central Florida

Interactive Computing Experiences
Research Cluster

Mehran Maghoumi[1,2]
Joseph J. LaViola Jr.[1]

[1]University of Central Florida
[2]NVIDIA

October 7, 2019

https://github.com/Maghoumi/DeepGRU

Overview

- Motivation & Contribution
- DeepGRU
- Experiments and Results
- Analysis
- Future Outlook
- References

Motivation

Gesture interactions are as popular as ever...

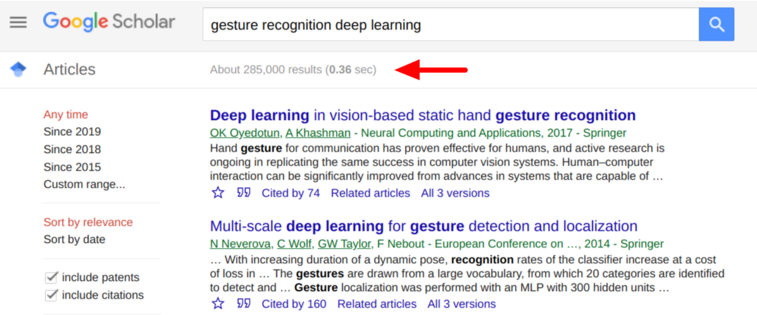
- Novel interactions techniques
- Fast and (mostly) reliable
- Sensors are getting better

Challenges:

- Many devices
- Many modalities
- And most importantly...

Motivation (cont'd)

The Tyranny of Choice!

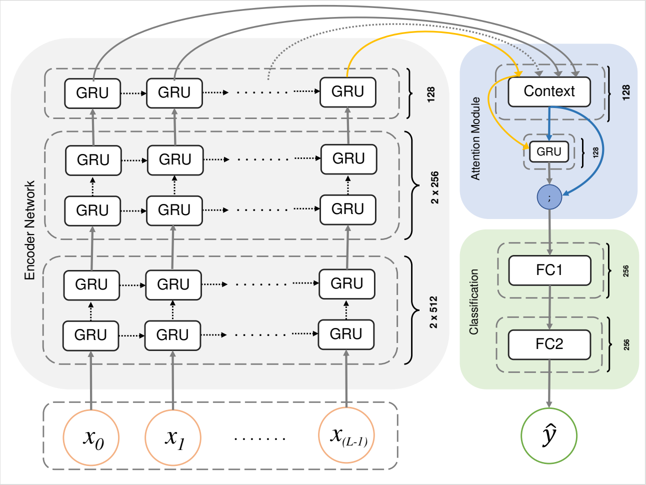


Contributions

Our method puts focus on application:

- Easy to understand
- Easy to implement and use
- Ease to train, not much parameter tuning
  - Various datasets (small, large)
  - Various modalities
- Quick training, even without powerful hardware
- High recognition accuracy

DeepGRU



Encoder Network
GRU
128
2 x 256
2 x 512
x_0
x_1
$x_{(L-1)}$
Attention Module
Context
128
;
Classification
FC1
256
FC2
256
$\hat{y}$

DeepGRU

Encoder Network

- Standard gated recurrent units (GRUs)
- We used GRUs because they are faster and simpler than LSTMs!

$$
\begin{aligned}
h_t &= \Gamma\big(x_t, h_{(t-1)}\big) \\
r_t &= \ \sigma \ \Big(\big(W_x^r \, x_t + b_x^r\big) \ + \ \big(W_h^r \, h_{(t-1)} + b_h^r\big)\Big) \\
u_t &= \ \sigma \ \Big(\big(W_x^u \, x_t + b_x^u\big) \ + \ \big(W_h^u \, h_{(t-1)} + b_h^u\big)\Big) \\
c_t &= \tanh\Big(\big(W_x^c \, x_t + b_x^c\big) \ + \ r_t\big(W_h^c \, h_{(t-1)} + b_h^c\big)\Big) \\
h_t &= u_t \circ h_{(t-1)} \ + \ \Big(1 - u_t\Big) \circ c_t
\end{aligned}
$$

- We zero-pad all inputs to the same length

DeepGRU

Attention Module

- Learn the most important subsequences
- Compute the context vector c with trainable parameters W_c
  - h_{L-1}: last hidden state
  - $\bar{h}$: all hidden states from $t = 0$ to $t = L - 1$

$$c = \operatorname{softmax}\left(h^{\intercal}_{(L-1)} W_c \bar{h} \right) \bar{h}$$
$$= \left(\frac{\exp\left(h^{\intercal}_{(L-1)} W_c \bar{h} \right)}{\sum_{t=0}^{L-1} \exp\left(h^{\intercal}_{(L-1)} W_c h_t \right)} \right) \bar{h}$$

- Inspired by Luong [20] *et al*.

DeepGRU

Attention Module (cont'd)

- Typically $\left[c\ ;\ h_{(L-1)}\right]$ is used, however...
  - Susceptible to sequence length variation
- Use an additional GRU to decide what to do

$$c = \text{softmax}\left(h^{\intercal}_{(L-1)} W_c \bar{h}\right) \bar{h}$$

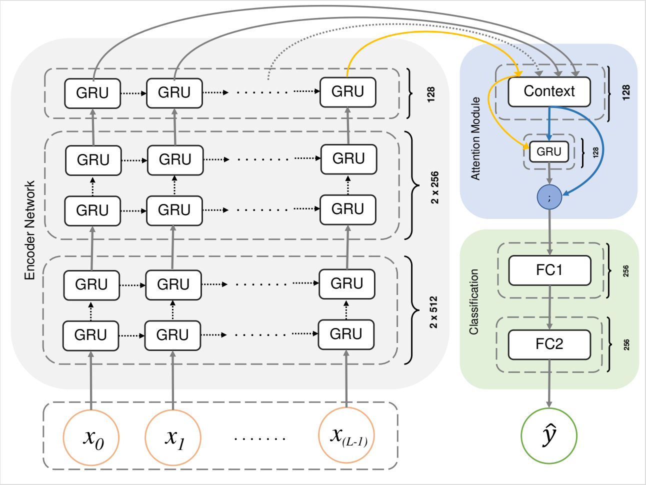
$$c' = \Gamma_{\text{attn}}\left(c, h_{(L-1)}\right)$$

$$o_{\text{attn}} = \left[c\ ;\ c'\right]$$

- Final output

$$\hat{y} = \text{softmax}\left(\text{FC}_2\left(\text{ReLU}\left(\text{FC}_1\left(o_{\text{attn}}\right)\right)\right)\right)$$

DeepGRU



Encoder Network
GRU
128
2 x 256
2 x 512
x_0
x_1
$x_{(L-1)}$
Attention Module
Context
128
Classification
FC1
FC2
256
$\hat{y}$

Experiments

- UT-Kinect
  - 10 gestures, 10 participants, 2 times (200 samples)
- NTU RGB+D
  - 60 action classes, 40 participants, multiple views/actors (56000 samples)
- SYSU-3D
  - 12 gestures, 40 participants (480 samples)
- DHG 14/28
  - 14/28 gestures, 28 participants (2800 samples)
- SBU Kinect Interactions
  - 8 two-person interactions, 7 participants (282 samples)

Results

UT-Kinect and SYSU-3D

Method	Accuracy	Method	Accuracy
Histogram of 3D Joints [35]	90.9	GCA-LSTM *(direct)* [17]	98.5
LARP + mfPCA [1]	94.8	CNN + Feature Maps [31]	98.9
ST LSTM + Trust Gates [18]	97.0	GCA-LSTM *(stepwise)* [17]	99.0
Lie Group [32]	97.1	CNN + LSTM [22]	99.0
ST-NBNN [33]	98.0	KRP FS [8]	99.0
DPRL + GCNN [29]	98.5	**DeepGRU**	**100.0**

Results on the UT-Kinect dataset

Method	Accuracy	Method	Accuracy
Dynamic Skeletons [12]	75.5	VA-LSTM [36]	77.5
ST LSTM + TG[18]	76.5	GCA-LSTM *(stepwise)* [17]	78.6
DPRL + GCNN [29]	76.9	**DeepGRU**	**80.3**

Results on the SYSU-3D dataset

Results

NTU RGB+D

Modality	Method	Accuracy		Modality	Method	Accuracy	
		CS	CV			CS	CV
Image	Multitask DL [21]	84.6	–	Pose	STA Model [28]	73.2	81.2
	Glimpse Clouds [4]	**86.6**	**93.2**		CNN + Kernel Feature Maps [31]	75.3	–
Pose+Image	DSSCA - SSLM [25]	74.9	–		SkeletonNet [13]	75.9	81.2
	STA Model (Hands) [3]	82.5	88.6		GCA-LSTM *(direct)* [17]	74.3	82.8
	Multitask DL [21]	**85.5**	–		GCA-LSTM *(stepwise)* [17]	76.1	84.0
Pose	Lie Group [32]	50.1	52.8		DPTC [34]	76.8	84.9
	HBRNN [11]	59.1	64.0		VA-LSTM [36]	79.4	87.6
	Dynamic Skeletons [12]	60.2	65.2		Clips+CNN+MTLN [14]	79.6	84.8
	Deep LSTM [26]	60.7	67.3		View-invariant [19]	80.0	87.2
	Part-aware LSTM [26]	62.9	70.3		DPRL + GCNN [29]	83.5	89.8
	ST LSTM + TG [18]	69.2	77.7		**DeepGRU**	**84.9**	**92.3**

Results on the NTU RGB+D dataset

Experiments

DHG 14/28 and SBU Kinect Interactions

Protocol	Method	Accuracy		Protocol	Method	Accuracy	
		C = 14	C = 28			C = 14	C = 28
Leave–one–out	Chen *et al*. [7]	84.6	80.3	SHREC'17 [10]	HOG^2 [23][10]	78.5	74.0
	De Smedt *et al*. [9]	82.5	68.1		HIF3D [5]	90.4	80.4
	CNN+LSTM [22]	85.6	81.1		De Smedt *et al*. [27][10]	88.2	81.9
	DPTC [34]	85.8	80.2		**DLSTM** [2]	**97.6**	**91.4**
	DeepGRU	**92.0**	**87.8**		**DeepGRU**	94.5	**91.4**

Results on the DHG 14/28 dataset

Method	Accuracy	Method	Accuracy
HBRNN [11]	80.4	Clips + CNN + MTLN [14]	93.5
Deep LSTM [26]	86.0	GCA-LSTM *(direct)* [17]	94.1
Co-occurance Deep LSTM [37]	90.4	CNN + Kernel Feature Maps [31]	94.3
STA Model [28]	91.5	GCA-LSTM *(stepwise)* [17]	94.9
ST LSTM + Trust Gates [18]	93.3	**VA-LSTM** [36]	**97.2**
SkeletonNet [13]	93.5	DeepGRU	95.7

Results on the SBU Kinect Interactions dataset

Experiments

Small Training Sets and Runtime

- Training with a very limited number of examples (at most 4 per-class)
  - Inspired by the $-Family of recognizers
  - Useful for gesture customization
- Datasets
  - **Acoustic:** Over-the-air hand gestures via Doppler shifted soundwaves
  - **Wii Remote**: Wii controller gestures
- Runtime experiments:
  - How long to converge?
  - Is training possible without powerful hardware?

Experiments

Small Training Sets and Runtime (cont'd)

Dataset	Method	Accuracy		Dataset	Method	Accuracy	
		$\mathcal{T}=2$	$\mathcal{T}=4$			$\mathcal{T}=2$	$\mathcal{T}=4$
Acoustic [24]	**Jackknife** [30]	**91.0**	94.0	Wii Remote [6]	Protractor3D [16]	73.0	79.6
	DeepGRU	89.0	**97.4**		$3 [15]	79.0	86.1
					Jackknife [30]	**96.0**	98.0
					DeepGRU	92.4	**98.3**

Small training sets evaluation

Device	Configuration	Dataset	Time	Device	Configuration	Dataset	Time
CPU	12 threads	Acoustic [24]	1.7	GPU	2× GTX 1080	SHREC 2017 [10]	5.5
		Wii Remote [6]	6.9			NTU RGB+D [26]	129.6
					1× GTX 1080	SHREC 2017 [10]	6.2
						SYSU-3D [12]	9.0
						NTU RGB+D [26]	198.5

Training times ($\mathcal{T}=4$ where applicable)

Experiments

Ablation Study

- Study the effects of various components
- Clearly shows the advantage of GRUs

Attn.	Rec. Unit	# Stck	# FC	Time	Acc.	Attn.	Rec. Unit	# Stck	# FC	Time	Acc.
-	LSTM	3	1	162.2	91.7	✓	LSTM	3	1	188.2	92.7
-	LSTM	3	2	164.0	91.0	✓	LSTM	3	2	192.1	92.0
-	LSTM	5	1	246.4	91.9	✓	LSTM	5	1	277.3	92.3
-	LSTM	5	2	251.6	89.5	✓	LSTM	5	2	283.3	92.2
-	GRU	3	1	143.8	93.4	✓	GRU	3	1	170.4	94.1
-	GRU	3	2	148.0	93.3	✓	GRU	3	2	174.0	93.8
-	GRU	5	1	210.8	93.6	✓	GRU	5	1	243.1	93.9
-	GRU	5	2	212.9	93.8	✓	**GRU**	**5**	**2**	**248.6**	**94.5**

Ablation study on DHG 14/28 dataset. Time is in seconds.

Future Outlook

- Requires segmented input
  - Unsegmented training is straightforward
  - Achieved the highest accuracy in SHREC'19 Online Gesture Recognition challenge
- Study the different aspects of the network
  - Sensitive to input dimensionality
    - Works better with high-dimensional inputs
  - Effects of regularization
- Reduce the need for paramter tuning

References (1)

Anirudh, R., Turaga, P., Su, J., Srivastava, A.: Elastic functional coding of human actions: From vector-fields to latent variables. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3147–3155 (2015)

Avola, D., Bernardi, M., Cinque, L., Foresti, G.L., Massaroni, C.: Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures. IEEE Transactions on Multimedia pp. 1–1 (2018)

Baradel, F., Wolf, C., Mille, J.: Human action recognition: Pose-based attention draws focus to hands. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). pp. 604–613 (2017)

Baradel, F., Wolf, C., Mille, J., Taylor, G.W.: Glimpse clouds: Human activity recognition from unstructured feature points. In: The IEEE Conference on Computer Vision and Pattern Recognition (2018)

Boulahia, S.Y., Anquetil, E., Multon, F., Kulpa, R.: Dynamic hand gesture recognition based on 3d pattern assembled trajectories. In: 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA). pp. 1–6 (2017)

Cheema, S., Hoffman, M., LaViola, J.J.: 3d gesture classification with linear acceleration and angular velocity sensing devices for video games. Entertainment Computing **4**(1), 11 – 24 (2013)

Chen, X., Guo, H., Wang, G., Zhang, L.: Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 2881–2885 (2017)

References (2)

Cherian, A., Sra, S., Gould, S., Hartley, R.: Non-linear temporal subspace representations for activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2197–2206 (2018)

De Smedt, Q., Wannous, H., Vandeborre, J.P.: 3d hand gesture recognition by analysing set-of-joints trajectories. In: Understanding Human Activities Through 3D Sensors. pp. 86–97 (2018)

De Smedt, Q., Wannous, H., Vandeborre, J.P., Guerry, J., Le Saux, B., Filliat, D.: Shrec'17 track: 3d hand gesture recognition using a depth and skeletal dataset. In: 10th Eurographics Workshop on 3D Object Retrieval (2017)

Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1110–1118 (2015)

Hu, J., Zheng, W., Lai, J., Zhang, J.: Jointly learning heterogeneous features for rgb-d activity recognition. IEEE transactions on pattern analysis and machine intelligence **39**(11), 2186–2200 (2017)

Ke, Q., An, S., Bennamoun, M., Sohel, F., Boussaid, F.: Skeletonnet: Mining deep part features for 3-d action recognition. IEEE Signal Processing Letters **24**(6), 731–735 (2017)

Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3d action recognition. In: Computer Vision and Pattern Recognition, 2017 IEEE Conference on. pp. 4570–4579. IEEE (2017)

Kratz, S., Rohs, M.: The $3 recognizer: Simple 3d gesture recognition on mobile devices. In: Proceedings of the 15th International Conference on Intelligent User Interfaces (2010)

References (3)

- Kratz, S., Rohs, M.: Protractor3d: A closed-form solution to rotation-invariant 3d gestures. In: Proceedings of the 16th International Conference on Intelligent User Interfaces (2011)
- Liu, J., Wang, G., Duan, L., Abdiyeva, K., Kot, A.C.: Skeleton-based human action recognition with global context-aware attention lstm networks. IEEE Transactions on Image Processing **27**(4), 1586–1599 (2018)
- Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal lstm with trust gates for 3d human action recognition. In: Computer Vision – ECCV 2016. pp. 816–833 (2016)
- Liu, M., Liu, H., Chen, C.: Enhanced skeleton visualization for view invariant human action recognition. Pattern Recogn. **68**(C), 346–362 (2017)
- Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (2015)
- Luvizon, D.C., Picard, D., Tabia, H.: 2d/3d pose estimation and action recognition using multitask deep learning. In: The IEEE Conference on Computer Vision and Pattern Recognition. vol. 2 (2018)
- Núñez, J.C., Cabido, R., Pantrigo, J.J., Montemayor, A.S., Vélez, J.F.: Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recogn. **76**(C), 80–94 (2018)
- Ohn-Bar, E., Trivedi, M.M.: Joint angles similarities and hog2 for action recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (2013)

References (4)

Pittman, C.R., LaViola, Jr., J.J.: Multiwave: Complex hand gesture recognition using the doppler effect. In: Proceedings of the 43rd Graphics Interface Conference. pp. 97–106 (2017)

Shahroudy, A., Ng, T., Gong, Y., Wang, G.: Deep multimodal feature analysis for action recognition in rgb+d videos. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(5), 1045–1058 (2018)

Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+d: A large scale dataset for 3d human activity analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)

Smedt, Q.D., Wannous, H., Vandeborre, J.: Skeleton-based dynamic hand gesture recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1206–1214 (2016)

Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: AAAI. vol. 1, pp. 4263–4270 (2017)

Tang, Y., Tian, Y., Lu, J., Li, P., Zhou, J.: Deep progressive reinforcement learning for skeleton-based action recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (2018)

Taranta II, E.M., Samiei, A., Maghoumi, M., Khaloo, P., Pittman, C.R., LaViola Jr., J.J.: Jackknife: A reliable recognizer with few samples and many modalities. In: Proceedings of the 2017 Conference on Human Factors in Computing Systems. pp. 5850–5861 (2017)

Tas, Y., Koniusz, P.: Cnn-based action recognition and supervised domain adaptation on 3d body skeletons via kernel feature maps. In: BMVC (2018)

References (5)

Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 588–595 (2014)

Weng, J., Weng, C., Yuan, J.: Spatio-temporal naive-bayes nearest-neighbor (st-nbnn) for skeleton-based action recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. pp. 445–454 (2017)

Weng, J., Liu, M., Jiang, X., Yuan, J.: Deformable pose traversal convolution for 3d action and gesture recognition. In: Europen Conference on Computer Vision (ECCV) (2018)

Xia, L., Chen, C., Aggarwal, J.: View invariant human action recognition using histograms of 3d joints. In: Computer Vision and Pattern Recognition Workshops , 2012 IEEE Computer Society Conference on. pp. 20–27. IEEE (2012)

Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., Zheng, N.: View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2136–2145 (2017)

Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., Xie, X.: Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. pp. 3697–3703 (2016)

Questions?

https://github.com/Maghoumi/DeepGRU