Progression Modeling for Online and Early Gesture Detection

Vikram Gupta[1], Sai Kumar Dwivedi[1], Rishabh Dabral[2], Arjun Jain[2,3]
[1]Mercedes-Benz R&D India, [2]Indian Institute of Technology Bombay, [3]Axogyan AI

Overview

Goal: Online and Early detection and classification of hand gestures

Applications: Touchless gesture control, Immersive gaming experience, AR/VR

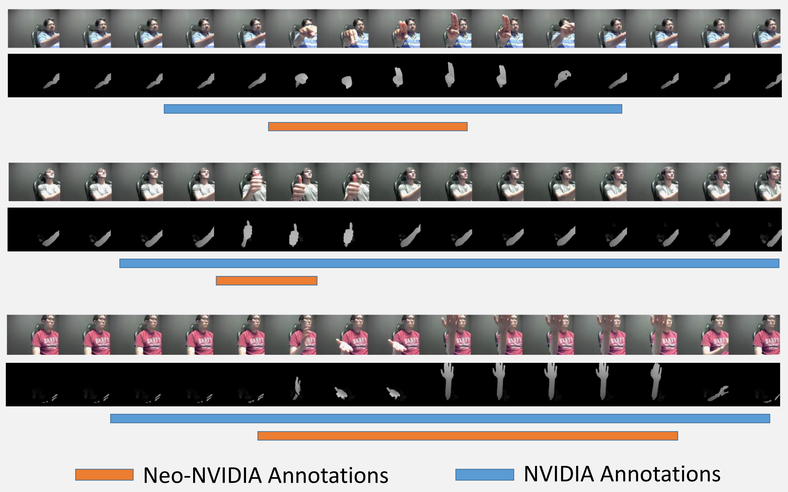
Abstract: We perform early gesture detection by modeling the progression level of the gesture along with frame level gesture recognition.

Background

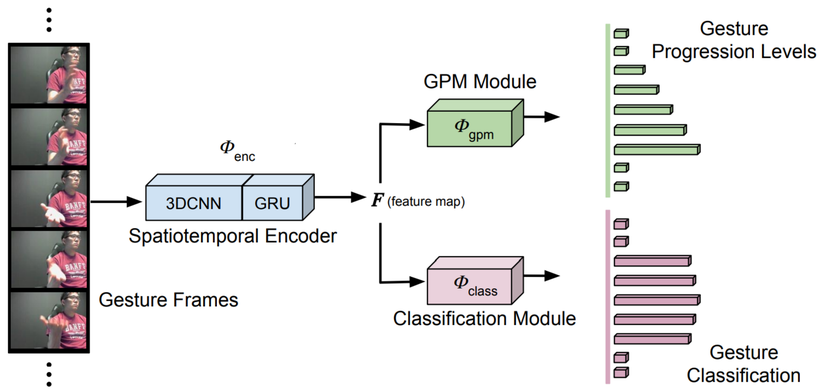
- *Molchanov et al.* used Connectionist Temporal Classification (CTC) for gesture detection where gesture trigger point can not be configured.
- Frame level gesture localization approaches (*Pigou et al.)* do not model gesture progression. Heuristic based thresholding is not effective.

Neo-NVIDIA Annotations

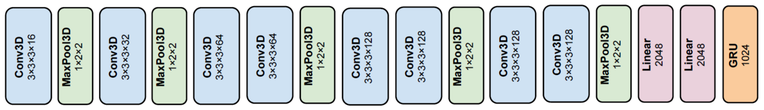
Tighter annotations for the NVIDIA gesture dataset.



Model Architecture



- We use 3D convolutions with 3x3x3 kernel to capture the local spatiotemporal features.
- GRU is used to model the long term temporal context.
- The encoded features are passed to Gesture Progression Module (GPM) and classification module.



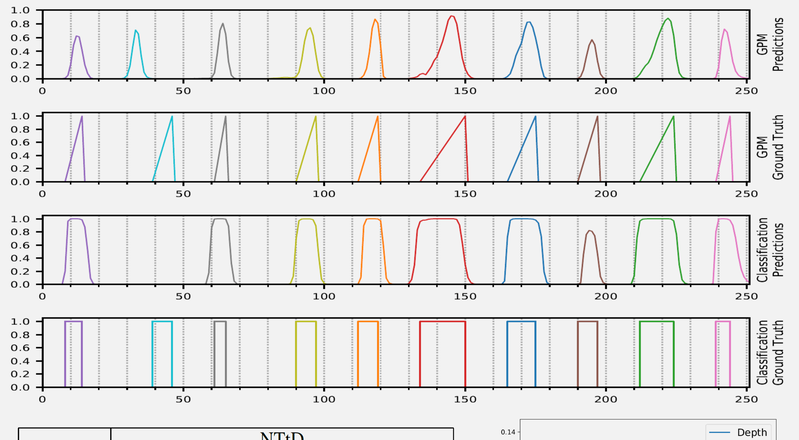
Gesture Progression Module (GPM)

$$\Phi_{gpm_t} = \begin{cases} \frac{t-t_s}{t_e-t_s}, & \text{if } t_s \leq t \leq t_e \\ 0, & \text{otherwise} \end{cases}$$

- Models the progression of the gesture
- *Offline setting*: Trigger the gesture prediction when the highest progression level is observed.
- *Online setting*: Trigger the gesture prediction when the gesture progression crosses the pre-configured threshold.

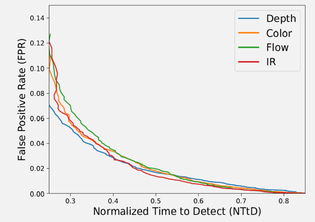
Results

Model Predictions vs Time



Modality	NTtD 0.25 FPR	NTtD 0.25 TPR	NTtD 0.50 FPR	NTtD 0.50 TPR	NTtD 0.75 FPR	NTtD 0.75 TPR
Depth	6.9	89.5	1.7	49.5	0.3	12.5
Color	11.3	91.2	1.8	43.6	0.3	7.9
Flow	11.1	92.5	1.9	45.6	0.2	7.1
IR	11.6	84.3	1.4	33.7	0.2	6.1

True Positive Rate (TPR) and False Positive Rate (FPR) across different Normalized Time to Detect (NTtD) values on the NVIDIA dataset.



NTtD vs False Positive Rate (FPR) on the NVIDIA dataset

* NTtD is the percentage of total gesture time taken to detect a gesture.

Modality	Ours	Molchanov *et al.*
IR	68.7	63.5
Color	75.9	74.1
Flow	78.2	77.8
Depth	85.5	80.3
IR Disparity (ID)	-	57.8
Flow + Color	80.3	79.3
Depth + Flow	85.5	82.4
Depth + Color	86.1	-
Depth + Color + Flow	86.3	81.5
Depth + Color + Flow + IR	**87.8**	83.4
Depth + Color + Flow + IR + ID	-	**83.8**
Human Accuracy		**88.4**

Comparison of classification accuracy (%) in *offline setting* on the NVIDIA gesture dataset.

Architecture	2DCNN-GRU	3DCNN-Linear	3DCNN-GRU
Acc (%)	77.4	81.5	85.5

Offline Classification accuracy(%) under different architecture settings of the Spatiotemporal Encoder on depth modality.

Modality	Jaccard Index
Depth	0.60
Flow	0.54
Color	0.53
IR	0.47
Depth + Color + Flow + IR	0.61

Gesture localization results on the Neo-NVIDIA annotations.