Learning Adaptive Hidden Layers for Mobile Gesture Recognition

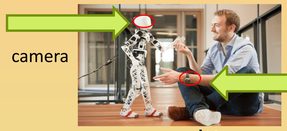
Ting-Kuei Hu, Yen-Yu Lin, and Pi-Cheng Hsiu
Research Center for IT Innovation, Academia Sinica, Taiwan

Data & Codes

Goal

- Gesture recognition provides an intuitive and convenient means for human-machine interaction in various environments.
- Three major challenges:
  - Large intra-class variations
  - Subtle inter-class variations
  - Ubiquitous environments
- Our goal in this project is to
  - Leverage multi-modal signals captured by diverse mobile sensors
  - Develop a new network layer for adaptive multi-modal learning.

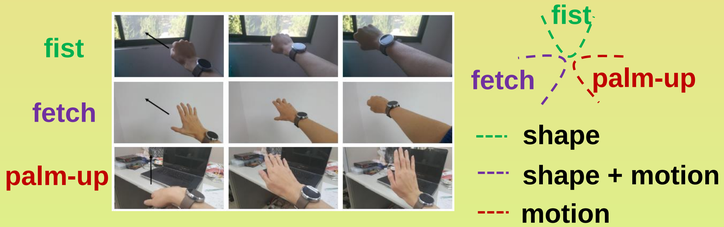
source: Chalearn Gesture challenge



source: sculpteo

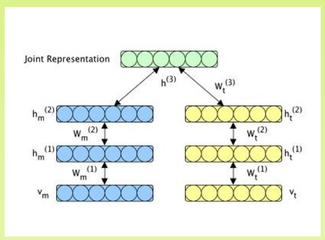
Motivation

- Large data variations from diverse user behaviors and ubiquitous environments. → **multi-modal signals**
- Optimal features for recognition often varies from gesture to gesture. → **adaptive learning**

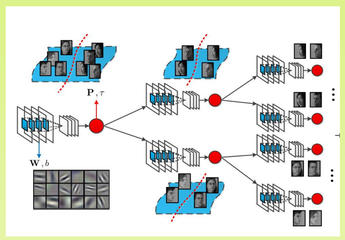


Related Work

- Most multi-modal approaches seek an immutable combination of multi-modal information.
- Adaptive tree-structured CNNs models have exponentially many sub-networks.



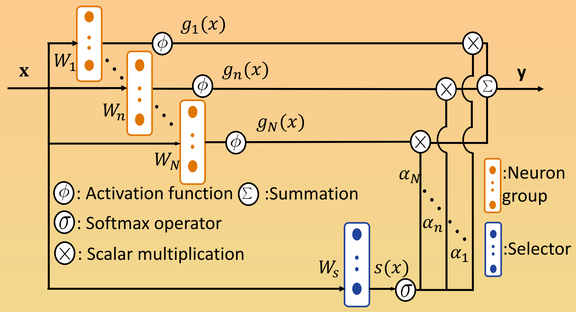
[Srivastava et al., NIPS 2012]



[Xiang et al., ICCV 2015]

Our Approach: Adaptive Hidden Layers

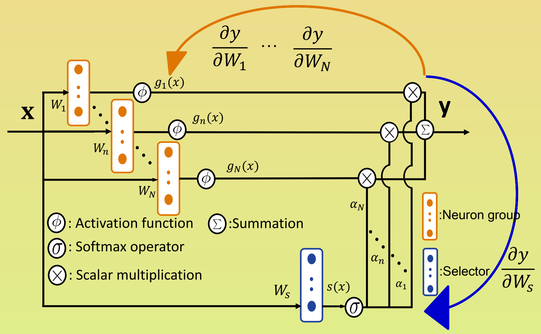
- We propose a new network layer, called the adaptive hidden layer (AHL), which is composed of **multiple neuron groups** and **an extra selector**.
  - Neuron groups: generate different activation maps.
  - Selector: adaptively picks a plausible group for each input.



Backward Propagation

$$\frac{\partial y}{\partial W_n} = \frac{\partial \sum_{i=1}^{N} \alpha_i g_i(x)}{\partial W_n} = \alpha_n \frac{\partial g_n(x)}{\partial W_n}$$

$$\frac{\partial y}{\partial W_s} = \frac{\partial \sum_{i=1}^{N} \alpha_i g_i(x)}{\partial W_s} = \sum_{i=1}^{N} \frac{\partial \alpha_i}{\partial W_s} g_i(x)$$



- Differential module → end-to-end trainable
- When stacking multiple AHLs
  - Linearly increased #parameters → computationally feasible
  - Exponentially many forward paths → high flexibility

Training Issues

- Data balance issue: the selector might assign most data to a or few subsets of neuron groups due to bad initialization of weights.
- Two training tips to resolve this issue
  - *k*-means clustering: partition the training data at the first epoch
  - SBR: an entropy-based function, called selection balancing regularizer (SBR), to encourage even distribution of data over neuron groups

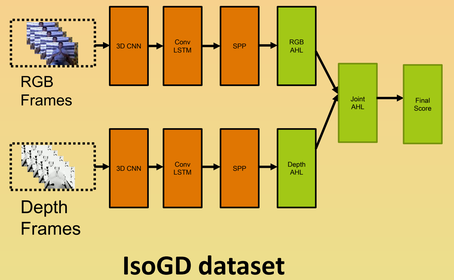
$$-\beta \sum_{n=1}^{N} (P_n + \varepsilon)\, log(P_n + \varepsilon)$$

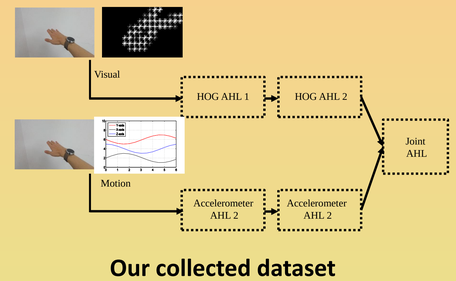
Two Datasets for Performance Evaluation

Modalities: RGB and depth videos
Data: 47933 gestures of 249 classes

Modalities: RGB videos & motion signal
Data: 4704 gestures of 14 classes

Network Architecture





IsoGD dataset **Our collected dataset**

Experimental Results

approach	accuracy
C3D + ConVLSTM(RGB)	43.88%
Ours (RGB)	**44.88%**
C3D + ConVLSTM(Depth)	44.66%
Ours (Depth)	**48.96%**
C3D+ConvLSTM(RGB + Depth)	51.02%
Our (RGB+Depth)	**54.14%**

approach	accuracy
DAE + HOG	81.52%
DAE + ACCE	76.24%
DAE + conc. feat.	82.34%
multi-modal DAE	86.48%
double-sized multi-modal DAE	86.02%
Ours	**90.57%**

IsoGD dataset **Our collected dataset**

More Results