# Simple Trick: Masked C3D for Isolated Sign Language Recognition

July 6, 2017

## 1 Team details


- lostoy
- Yingwei Li
- Department of Electrical and Computer Engineering, UCSD, La Jolla CA; 8584316308;lostoy.li@gmail.com
- UCSD


## 2 Contribution details

- Title of the contribution: Masked C3D
- Final score: 65.69
- Final score: Regions other than hands in the input RGB/depth images are masked to zero. The masked images are used to learn C3D model for classification.
- References: D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV 2015.
- Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." arXiv preprint arXiv:1611.08050 (2016).
- Describe data preprocessing techniques applied: Hand locations are detected by pose estimation and regions outside the hand bounding boxes are set to 0.

# 3 Visual Analysis

## 3.1 Gesture Recognition Stage

### 3.1.1 Features / Data representation

Describe features used or data representation model:
Finetuned C3D network with masked input images.

### 3.1.2 Dimensionality reduction

None

### 3.1.3 Compositional model

None

### 3.1.4 Learning strategy

1. RGB stream is finetuned from C3D reference model provided by the original author.
2. Depth stream is finetuned from C3D reference model provided by the original author.
3. RGB stream is finetuned from depth stream model in step 2.
4. Depth stream is finetuned from RGB stream model in step 1.

### 3.1.5 Other techniques

None

### 3.1.6 Method complexity

Each stage of the training takes around 6 hours on 4 Titan X(Maxwell) GPUs.

## 3.2 Data Fusion Strategies

Late fusion is used to fuse RGB and depth stream with equal weights for both streams.

## 3.3 Global Method Description

- Which pre-trained or external methods have been used : C3Dv1.1
- Which additional data has been used in addition to the provided ChaLearn training and validation data: None

- Qualitative advantages of the proposed solution:

  CNN based models can easily overfit to background, clothing etc. of the ISOGD dataset. The masked C3D trick is simple to implement and yet provide useful guidance for CNN. The whole pipeline of the final submission is simple to replicate and the fusion is simply two stream late fusion without ensemble of many different models.

- Results of the comparison to other approaches:
  Validation accuracies of different approaches:
  RGB C3D w/ masked input and finetuned on depth stream: 55.07
  RGB C3D w/ masked input: 50.10
  RGB C3D w/o mask and finetuned on depth stream: 51.83
  RGB C3D w/o mask: 45.83

  Depth C3D w/ masked input and finetuend on RGB stream: 56.71
  Depth C3D w/ masked input: 50.62
  Depth C3D w/o mask and finetuned on RGB stream: 51.04
  Depth C3D w/o mask: 46.99

- Novelty degree of the solution and if is has been previously published:

  Using simple mask operation for preprocessing is novel. Masking can enforce C3D to attend to the most important regions for gesture recognition and avoid overfitting to non-informative appearance. Also, the trajectories of non-zero regions in the masked image also provide important information about motion of the posture for C3D to capture.

# 4 Other details

- Language and implementation details (including platform, memory, parallelization requirements):

  The whole system is implemented with pytorch. Training is done on 4 x Titan X(Maxwell) GPUs with 6GB GPU memory footage for each GPU.

- Human effort required for implementation, training and validation?

  None

- Training/testing expended time?

  Each training stage costs 6 hours. And testing time is 1-2 minutes.

- General comments and impressions of the challenge? what do you expect from a new challenge in face and looking at people analysis?

  The gesture challenge is a good attempt to guide researchers to look closer at fine-grained level human behaviors. It also provides an opportunity to diagnose current CNN architecture for video understanding. I would expect a dataset with more variations: a lot of the videos here are almost

repetitions of the same gesture performed by the same person in the same background. A more proper dataset will facilitate models that can generalize to other applications of human gesture analysis.