Learning both Dynamic and Static Action Context for Gesture Recognition

July 7, 2017

1 Team details

- Team name
  SYSU_ISEE
- Team leader name
  Wanhua Li
- Team leader address, phone number and email
  address: 132 East Waihuan Road, Guangzhou Higher Education Mega Center, Guangzhou 510006, P.R. China
  phone number: (+86)15975554708
  email: lwh370@163.com
- Rest of the team members
  Jian-Fang Hu, Benchao Li and Wei-Shi Zheng
- Team website URL (if any)
- Affiliation
  Intelligence SciencE and systEm Lab (iSEE), School of Data and Computer Science, Sun Yat-sen University

2 Contribution details

- Title of the contribution
  Learning both Dynamic and Static Action Context for Gesture Recognition
- Final score
  Phase 1: **59.6992%** (Validation Accuracy)
  Phase 2: **67.0228%** (Test Accuracy)

- General method description
  In this implementation, we considered modeling both the dynamic and non-dynamic (static) action cues for the recognition of hand gestures. For the dynamic cues, we learned discriminative motion features from RGB videos, depth videos, optical flow sequences, and skeletons. For the non-dynamic action cues, we employed the rank pooling method to represent all the optical flow frames and depth frames in a sample by a static super-frame. All of them (except skeletons) are fed into the VGG-16 network separately to obtain a reliable network. For the skeletons, we fed them into a deep LSTM network to learn the dependencies among the observed skeletons. Indeed, each kind of the above features can provide us a recognition score indicating how likely does the given sample include a certain hand gesture. For achieving a robust recognition result, we select to fuse all the scores together and then make decision based on the fused recognition score.

- References
  [1] J. Wan, S. Z. Li, Y. Zhao, S. Zhou, I. Guyon, and S. Escalera, Chalearn looking at people RGB-D isolated and continuous datasets for gesture recognition, IEEE CVPR Workshops. IEEE, 2016, pp. 1–9.
  [2] Haoshu Fang, Shuqin Xie, Yuwing Tai and Cewu Lu. RMPE: Regional Multi-person Pose Estimation. arXiv preprint arXiv:1612.00137, 2016.
  [3] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi and Stephen Gould. Dynamic Image Networks for Action Recognition,CVPR, 2016.

- Representative image / diagram of the method
  Data preparation see Figure 1. Our model framework see Figure 2.

- Describe data preprocessing techniques applied (if any)
  Rank pooling[3] is used to produce a set of static super-frames from dynamic videos and sequences.
  RMPE[2] is used to generate skeleton data from RGB videos.

3 Visual Analysis

3.1 Gesture Recognition (or/and Spotting) Stage

3.1.1 Features / Data representation

Describe features used or data representation model FOR GESTURE RECOGNITION (OR/AND SPOTTING) STAGE (if any)

See Figure 1

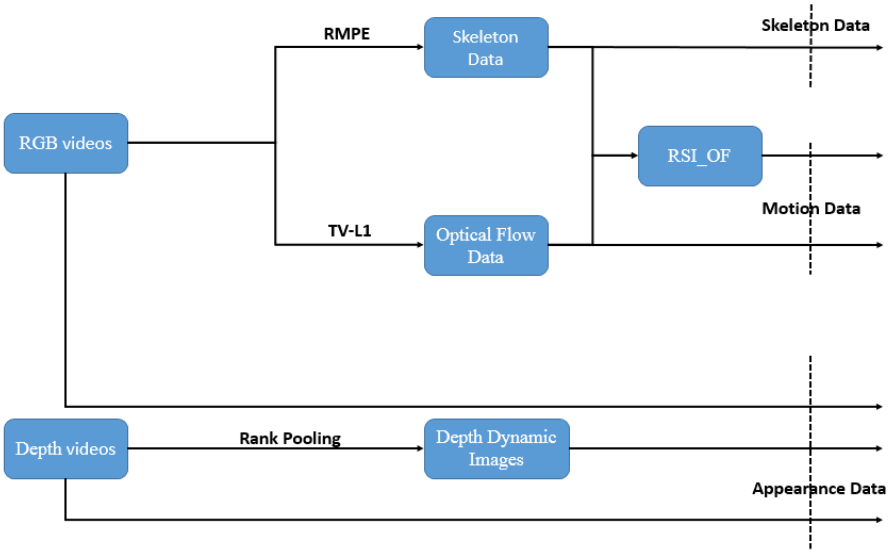

Figure 1: Data Preparation

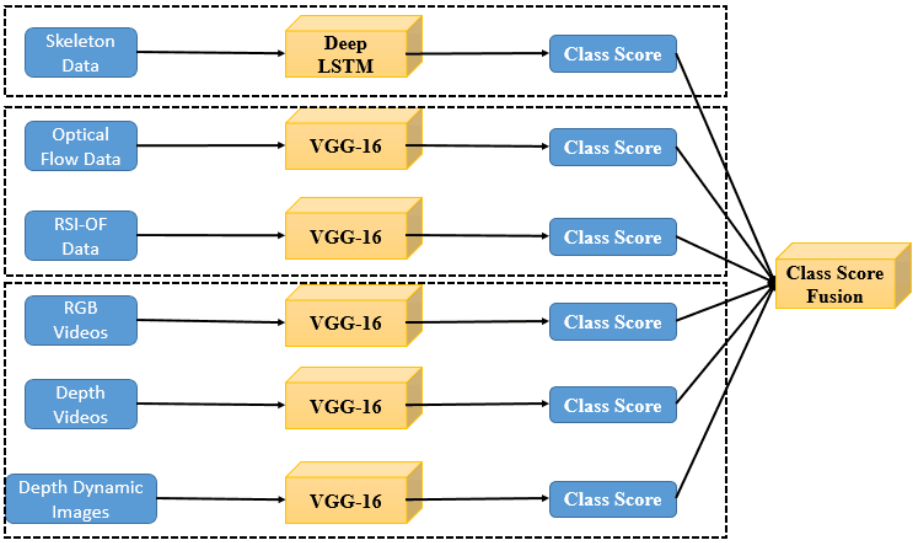

Figure 2: Model Framework

3.1.2 Dimensionality reduction

Dimensionality reduction technique applied FOR GESTURE RECOGNITION (OR/AND SPOTTING) STAGE (if any)

None

3.1.3 Compositional model

Compositional model used, i.e. pictorial structure FOR GESTURE RECOGNITION (OR/AND SPOTTING) STAGE (if any)

See Figure 2

3.1.4 Learning strategy

Learning strategy applied FOR GESTURE RECOGNITION (OR/AND SPOTTING) STAGE (if any)

VGG-16 and deep LSTM are used in our model. The parameters of our model are learned via stochastic gradient descent method.

3.1.5 Other techniques

Other technique/strategy used not included in previous items FOR GESTURE RECOGNITION (OR/AND SPOTTING) STAGE (if any)

Rank pooling[3] is used to generate depth dynamic images from depth videos. RMPE[2] is used to generate skeleton data from RGB videos.

3.1.6 Method complexity

Method complexity FOR GESTURE RECOGNITION (OR/AND SPOTTING) STAGE

The basic model we used is VGG-16. The count of parameter is about 135 millions.

3.2 Data Fusion Strategies

List data fusion strategies (how different feature descriptions are combined) for learning the model / network: Single frame, early, slow, late. (if any)

Each of the six features is used to train a discriminative recognition model separately. The final prediction result is achieved by fusing the recognition results of all the models.

3.3 Global Method Description

- Which pre-trained or external methods have been used (for any stage, if any)

Temporal stream ConvNet model in Two stream ConvNet pre-trained on UCF-101.

- Which additional data has been used in addition to the provided ChaLearn training and validation data (at any stage, if any)

  None. All the models are trained based on training sequences provided by the organizer. And the hyper parameters are determined based on the recognition on the validation set.

- Qualitative advantages of the proposed solution

  Our model makes use of both the dynamic and non-dynamic (static) action information to represent each gesture. We found that combining both dynamic and non-dynamic (static) action cues together can significantly improve the recognition performance.

- Results of the comparison to other approaches (if any)

  See Tabel 1

Table 1: Results

Rank	Team name	Accuracy
1	ASU	0.677085
2	**SYSU_ISEE**	**0.670228**
3	lostoy	0.656993
4	ICT_NHCI	0.642162
5	XDETVP	0.593207

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

  Have not been published.

4 Other details

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

  1). Language: C++, Matlab and python
  2). Platform: TensorFlow(mostly),caffe
  3). Memory: GTX TITAN X 12GB of memory

- Human effort required for implementation, training and validation?

  All can be executed automatically. Hunman effort is barely required.

- Training/testing expended time?

  The time for model preprocessing and model training is almost 10 days. It will take about 1-2 days to recognize all the test data.

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

  It's very exciting to see the progress we have made in this challenge. We believe that our recognition result will be better If a more powerful fusion method is employed. We hope the organizers can release the test labels so that the readers who are interested in this can use the set for research.