Isolated Gesture Recognition Fact Sheet

August 17, 2016

1 Team details

- Team name: ICT_NHCI
- Team leader name: Xiujuan Chai
- Team leader address, phone number and email
  Address: No.6 Kexueyuan South Road Zhongguancun,Haidian District Beijing,China
  Phone number: +86 10 62600553
  E-mail: chaixiujuan@ict.ac.cn
- Rest of the team members
  Zhipeng Liu, Fang Yin, Zhuang Liu and Xilin Chen
- Affiliation
  Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS

2 Contribution details

- Title of the contribution
  Two streams RNN for Isolated Gesture Recognition
- Final score
- General method description
  Taken the multi-modal sequential hand features as input, RNN is used for obtain the isolated gesture recognition results.
- References
  Faster Recurrent Convolutional Neural Network(Faster R-CNN)[4].
  Recurrent Neural Network (RNN)[2].
  Keras: Deep Learning library[1].
  Caffe[3].
  Face Detection[6].

- Representative image / diagram of the method
  Figure 1 is the diagram of our method.

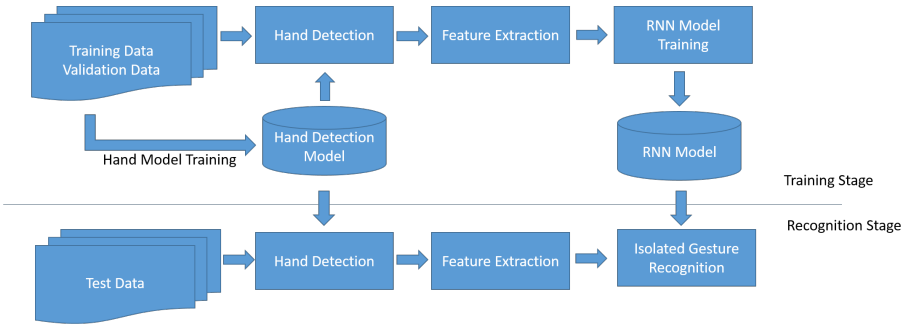

Figure 1: Diagram of the method.

- Describe data preprocessing techniques applied (if any)
  A hand detection model is trained with some manually labeled hand images from the provided training and validation data. Therefore, the hand regions of training, validation and test data are detected with this model.

3 Visual Analysis

3.1 Gesture Recognition Stage

3.1.1 Features / Data representation

In each frame, the features are represented by the hand shape and positions from two separated channels, i.e. RGB and depth.

For the hand shape representation, HOG is extracted from the detected hand regions.

For the hand position representation, skeleton pairwise feature[5] is used. The face and two hands are selected as the key points and the skeleton pairwise feature is constructed by the distances between each pair of three points.

3.1.2 Dimensionality reduction

PCA is used for HOG feature dimensionality reduction. The feature dimension for the final hand shape representation is reduced to 81 from 324 with 90% energy reserved.

3.1.3 Compositional model

Figure 2 illustrates the structure of two stream RNN, which is the main model in our method.

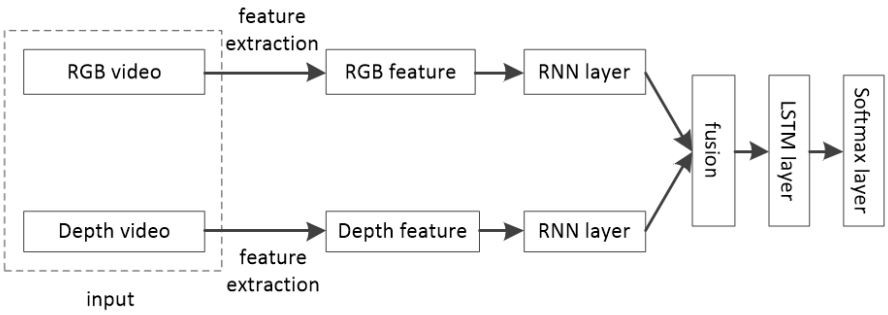

Figure 2: Pictorial structure for two streams RNN model.

3.1.4 Learning strategy

Use the two feature streams described above in train and validation data to train the RNN.

3.1.5 Other techniques

Face detection[6] technique is used for skeleton pair feature extraction as described in Section 3.1.1.
Faster R-CNN[4] is used for hand detection.

3.1.6 Method complexity

The architecture of our method has four layers. The first layer has two independent RNN channels with 330 neurons, which corresponding to RGB and Depth channel respectively. The second layer is the fusion layer. The third LSTM layer has 165 neurons and the last layer is softmax layer.

3.2 Data Fusion Strategies

The hand shape and position features are extracted for both RGB and depth videos. In each separated channel, the hand shape feature and position feature are fused by concatenating directly. While the features from different channels are fused by the RNN model. Concretely speaking, they are fed into two RNN layers respectively and fused by the fusion layer.

3.3 Global Method Description

- Which pre-trained or external methods have been used (for any stage, if any)
  The face detection model[6] is pre-trained.

- Qualitative advantages of the proposed solution
  Firstly, RNN can model the contextual information of gesture. Secondly, the fusion can make full use of RGB and Depth information.

- Novelty degree of the solution and if is has been previously published
  1)Two streams RNN fuses the RGB and Depth information effectively and it can model the contextual information of the temporal gesture sequences.
  2)The hand detection module gives the precise hand positions, which is very important for the correct recognition.
  3)Hand HOG and skeleton pair feature is integrated to describe the gesture well by avoiding the background noise.
  The work has not been published.

4 Other details

- Language and implementation details (including platform, memory, parallelization requirements)
  Hand detection is implemented in Caffe[3].
  Face detection SDK, HOG and skeleton pair feature extraction are programmed in Visual Studio 2012 with C++.
  RNN classifier training and testing are implemented in keras with cuDNN on a Titan X GPU.

- Human effort required for implementation, training and validation?
  The hand regions of roughly 50000 images from training and validation data are annotated manually and used for hand detection model training.

- Training/testing expended time?
  In the training stage, it takes about 16 hours to train the RGB and depth hand detection model (8 hours per model) using Faster R-CNN. It takes about 80 hours and 4 hours for hands and face detection respectively on train and validation data(one Titan X GPU). After getting the detection results, it takes about 9 hours to extract features from train and validation data. At last, it just takes about 20 minutes to train the final two streams RNN model.
  In the test stage, it takes about 12 hours to detect hands (one Titan X GPU) and 30 minutes to detect faces on test data. Then it takes about 1 hours to extract features from test data. At last, it just takes 5 minutes to get the recognition result on test data.

- General comments and impressions of the challenge? what do you expect from a new challenge in face and looking at people analysis?
  Given the complicated environments and the large variations between different subjects, the dataset is quite challenging.

References

[1] F. Chollet. Keras. https://github.com/fchollet/keras, 2015.

[2] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber. Lstm: A search space odyssey. *Computer Science*, 2015.

[3] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[4] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2016.

[5] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1290–1297. IEEE, 2012.

[6] S. Wu, M. Kan, Z. He, S. Shan, and X. Chen. Funnel-structured cascae for multi-view face detection with alignment awareness. *Neurocomputing(Under review)*.