Video Gesture Recognition with RGB-D-S Data Based on 3D Convolutional Networks

August 16, 2016

1 Team details

- Team name
  FLiXT
- Team leader name
  Yunan Li
- Team leader address, phone number and email
  address: Xidian University, 2th South Taibai Road, Xi'an, Shaanxi, China
  phone number: 18710849937
  email: xdfzliyunan@163.com
- Rest of the team members
  Kuan Tian, Yingying Fan, Xin Xu, Rui Li.
- Team website URL (if any)
- Affiliation
  School of Computer Science and Technology, Xidian University

2 Contribution details

- Title of the contribution
  Video Gesture Recognition with RGB-D-S Data Based on 3D Convolutional Networks
- Final score
  48.62% (on validation)
- General method description

  Our method recognizes gestures by employing both RGB and depth videos and learning with the features extracted by the 3D CNN model. To learn

more about the detail of motions, we make a pre-processing on the inputs and convert them into 32-frames videos. Since the variations in background, clothing, skin color and other external factors may disturb the recognition, we employ saliency video to concentrate on the gestures. The features of the videos are learnt by C3D model, a 3D convolutional network model that learns spatiotemporal features. Then we also blend the RGB feature, depth feature and saliency features together to boost the performance. The final classification is implemented by SVM classifier.

- References
[1] Jun Wan, Yibing Zhao, Shuai Zhou, Isabelle Guyon, Sergio Escalera and Stan Z. Li, "ChaLearn Looking at People RGB-D Isolated and Continuous Datasets for Gesture Recognition", CVPR workshop, 2016.
[2] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV 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 2014.
[4] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014.
[5] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk,"Frequency-tuned salient region detection," CVPR 2009.
[6] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.

- Representative image / diagram of the method

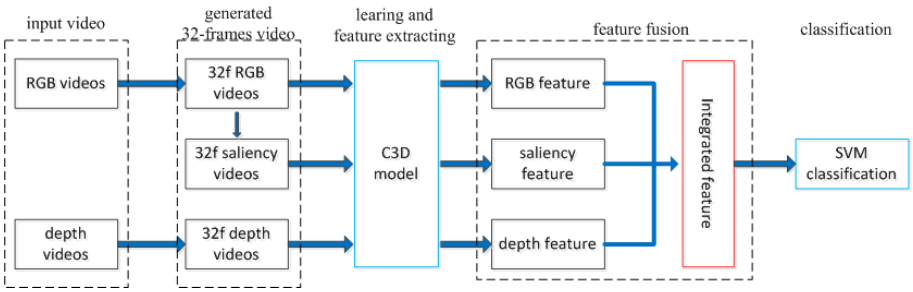

Figure 1: The diagram of our method

- Describe data preprocessing techniques applied (if any)
We make a pre-processing on the videos to convert them into 32-frames. As the statistics of all 35878 train videos show, most of them are with 29-39 frames and the peak is 33-frames video. For easier processing in

C3D learning, we choose 32 as a benchmark and sampling or extending input videos in terms of the length of them.

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)

The features we use for gesture recognition is extracted by C3D model. The architecture of C3D model is illustrated in Fig.2. After 8 convolution and 5 pooling, the input video is converted into a 1×4096 dimension feature vector, and that is exact what we use for classification.

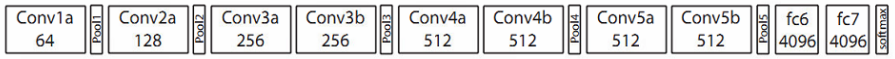

Figure 2: The architecture of C3D model (from Tran et al.'s paper). It consists of 8 convolution layers, 5 pooling layers, 2 fully-connected layers and a softmax loss layer. The feature we extract is from fc6 layer, i.e., the first fully-connected layer.

3.1.2 Dimensionality reduction

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

3.1.3 Compositional model

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

3.1.4 Learning strategy

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

We adopt a "learning-extracting-fusing" strategy. Since blending the three kinds of input videos may result unreasonable data (because the objects in RGB and depth videos are not presented with the same size), we first use three kinds of video to finetune the model and extract features respectively, then we blend these three features together and use the integrated feature to train SVM and obtain the final classification result.

3.1.5 Other techniques

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

3.1.6 Method complexity

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

The part of our method that most likely to have high complexity is the fine-tuning process of C3D model. It needs about 50.9 hours to finetune and update nearly parameters. It also takes about 8G graphic memory. The classification is implemented by a linear-SVM classifier so that the complexity of it is not very high.

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)

The data we use for fusion is RGB, depth and saliency data. The saliency data is generated from RGB data to alleviate the disturbing influence of background, clothing, skin color and etc. Since the RGB and depth videos are not matched well (objects in RGB video is a little bigger), we choose to fusion in the later stage - after features of RGB, depth and saliency data are extracted by C3D model, we blend these three 1×4096 feature vectors and calculate the average.

3.3 Global Method Description

- Which pre-trained or external methods have been used (for any stage, if any)
  The C3D model is pre-trained with the Sports-1M dataset.
- Which additional data has been used in addition to the provided ChaLearn training and validation data (at any stage, if any)
- Qualitative advantages of the proposed solution
  The advantages of our method are:
  1. The 32-frames strategy achieves better results than the original C3D model which requires 16-frames input.
  2. The fusion feature make a significant progress compared with any single feature in boosting accuracy.
- Results of the comparison to other approaches (if any)
- Novelty degree of the solution and if is has been previously published
  The 32-frame scheme and saliency video used for fusion are novel proposed and have never been published before.

4 Other details

- Language and implementation details (including platform, memory, parallelization requirements)
  Our experiments are processed on a PC with Intel Core i7-6700 CPU @ 3.40GHz × 8, 16 GB RAM and Nvidia Geforce GTX TITAN X GPU. The experiments of C3D model training and feature extracting are processed under caffe framework on Linux Ubuntu 14.04 LTS, others including 32-frames video generation, feature fusion and SVM classification are implemented by matlab R2012b on 64-bit Windows 7.

- Human effort required for implementation, training and validation?
  As our method is divided into four modules (data pre-processing module, feature extraction module, classification module and prediction generation module), the intermediary data needs transporting among those modules, especially for the feature extraction, since the network needs data and creates data in specific directory.

- Training/testing expended time?
  The training time of C3D model is about 50.9 hours, the feature extraction time is about 1 hour for 35878 training data. The classification time of SVM classifier is about 6-9 hours (depending on the amount of training data and the rate that CPU is occupied).

- General comments and impressions of the challenge? what do you expect from a new challenge in face and looking at people analysis?
  The challenge is really fantastic. Some of the gestures are really difficult to distinguish so we need to try our best to make the details of each gesture be learnt by the network. It might be better that the face and looking at people analysis challenge is different from existing face recognition scheme. A specific application like criminal discrimination may be interesting and realistic.