Continuous Gesture Recognition Fact Sheet

July 6, 2017

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 62600555
  E-mail: chaixiujuan@ict.ac.cn
- Rest of the team members
  Zhipeng Liu, 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
  3D Convolutional Networks for Continuous Gesture Recognition with Efficient Segmentation.
- Final score
  0.609681(on testing)
- General method description
  First, the RGB and depth image frames are calibrated. The hands are detected through a two-streams Faster R-CNN method. Thus the continuous gesture sequence is segmented into several isolated gestures to realize the temporal segmentation. In order to represent each gesture by the hand posture and location information, the face region is located and thus the relative hand locations are encoded into the 3D convolution features. Specifically, the face region only considered in the RGB image.

While in the depth channel, we do not add face region because of the coarse calibration. Then the hand feature are extracted by C3D model, a 3D convolutional network model that learns spatiotemporal features. We also fuse the RGB and depth feature to boost the performance. The final classification is achieved by SVM classifier.

- References
  Chalearn continuous gesture dataset[5].
  Faster R-CNN[3].
  C3D[4].
  Caffe[2].
  Face Detection[6].
  SVM[1].

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

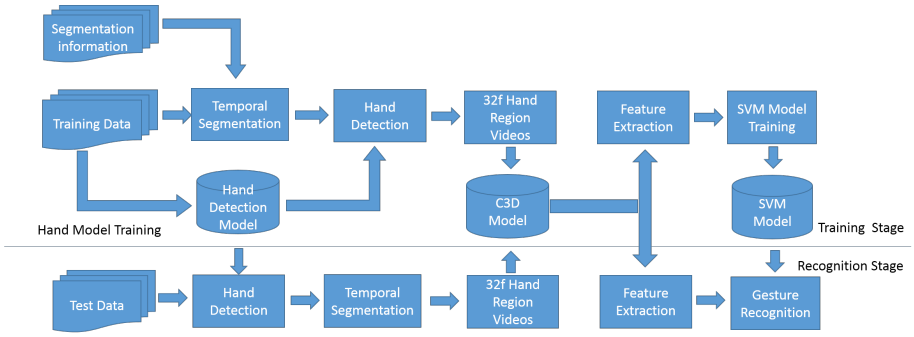

Figure 1: Diagram of the method.

- Describe data preprocessing techniques applied (if any)
  For each image frame, the hand is extracted and other region are blocked in RGB and depth. And we add face region to RGB. Then all video frames are converted into 32-frames.

3 Visual Analysis

3.1 Gesture Recognition (or/and Spotting) Stage

3.1.1 Features / Data representation

In gesture segmentation stage, the feature is hand position. We detect the hands in the video by using two-streams Faster R-CNN method. Figure 2 shows the hand detection pipeline of our two-streams Faster R-CNN.

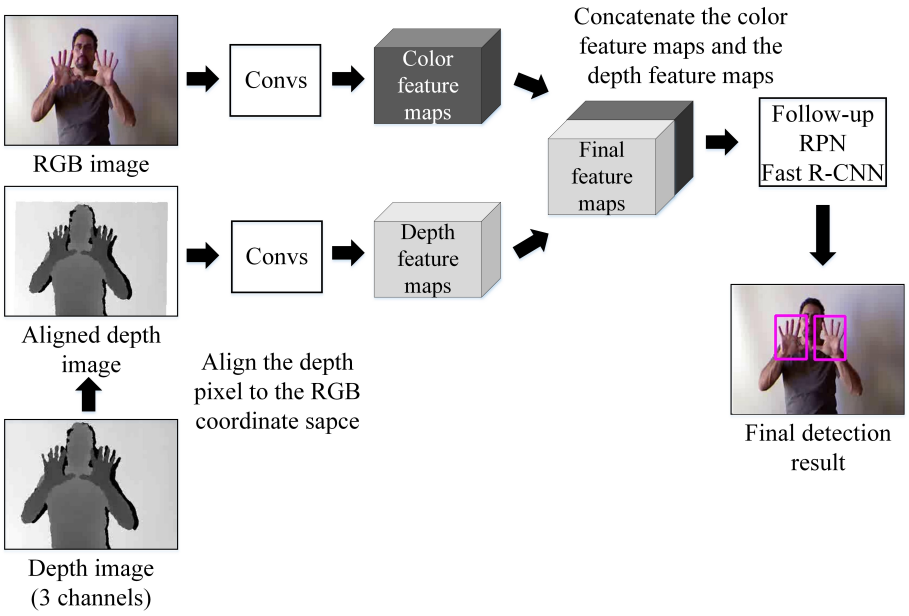

Figure 2: The detection pipeline of two-streams Faster R-CNN.

In gesture recognition stage, both RGB and calibrated depth frames are highlighted hand region and blocked other regions. To learn relative motion between hands and face, we add face region to RGB hand-images.

Then both RGB and depth hand features are learned by C3D model. The architecture of C3D model is illustrated in Fig.3. After 8 convolution and 5 pooling, the input video is converted into a 1×4096 dimension feature vector, and that is exact what we used for classification.

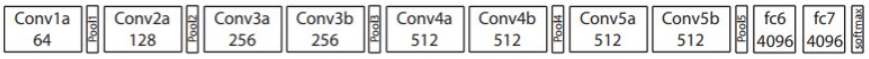

Figure 3: The architecture of C3D model[4]. It consists of 8 convolution layers, 5 pooling layers, 2 fully-connected layers and a softmax loss layer. The feature we extracted is from fc6 layer, i.e., the first fully-connected layer.

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

Use the features described above in training dataset to train the SVM.

3.1.5 Other techniques

Face detection.

3.1.6 Method complexity

Our method mainly consists of hand detection and fine-tuning c3d model. The part of hand detection takes about 3 days in a Titan X GPU. The part of fine-tuning c3d model needs about 50 hours in a Titan X GPU.

3.2 Data Fusion Strategies

The data we used for fusion is hand image from RGB and depth channels. Extracted hand images are generated based on accurate hand detection in order to alleviate the disturbing influence of non-motion area, such as, background, clothing and body. Since the RGB and depth hand are not matched well though depth frames are calibrated, we choose to fuse features in the later stage - after hand features of RGB, depth are extracted by C3D model.

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 in face detection step. The C3D model is pre-trained with the Sports-1M dataset. At last, the VGG network of Faster R-CNN is pre-trained in ImageNet.

- 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
  1) The continuous gesture recognition is transformed into the isolated gesture recognition problem with the help of the accurate hand detection.
  2) We extract hand region and block other regions in order to alleviate the disturbing influence of non-motion regions, such as, background, clothing and body.
  3) 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

  1) We propose two-stream Faster R-CNN to realize more accurate hand detection.
  2) The gesture segmentation is realized with the accurate hand detection.
  3) We extract only-hand-image to avoid the background disturbing influence of non-motion area. Then C3D model is used to extracted spatiotemporal hand feature.
  The work has not been published.

4 Other details

- Language and implementation details (including platform, memory, parallelization requirements)
  Hand detection and C3D are implemented in caffe.
  Face detection SDK and temporal segmentaion are programmed in Visual Studio 2012 with C++.
  Data preprocessing, extracting hand image and SVM are programmed with python.
  Segmenting the training data into isolated gestures, reading and fusing fc6 features are programmed with MATLAB.
- Human effort required for implementation, training and validation?

  The hand regions of 31750 images from training data are annotated by human and used for hand detection model training.
- Training/testing expended time?
  For training, it takes about 5 hours to segment the training data into isolated gestures in MATLAB, then it takes about 80 hours to train the RGB and depth hand detection model (40 hours per model) using two-streams Faster-RCNN. After getting the detection models, it takes about 60 hours to detect hands (one Titan X GPU) and 4 hours to detect face on these gestures. Preprocessing and Extracting hand region need about 3 hours. Fine-tuning C3D model and extracting fc6 feature needs about 50 hours and 1.5 hours respectively. At last, it just takes about 20 mins to train the final SVM model.

  For testing, it takes about 15 hours to detect hands (one Titan X GPU) and 0.5 hour to detect face on test data and about 0.5 hour to segment the test into isolated gestures using the detection result. Then it takes about 0.5 hours to extract features from test data. At last, it takes 5 mins 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?

References

[1] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[2] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[3] 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.

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[5] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Z. Li. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 56–64, 2016.

[6] S. Wu, M. Kan, Z. He, S. Shan, and X. Chen. Funnel-structured cascade for multi-view face detection with alignment-awareness. *Neurocomputing*, 221:138–145, 2017.