Gesture Recognition with Pyramidal 3D Convolutional Networks

August 17, 2016

1 Team details

- Team name
  XDETVP-TRIMPS
- Team leader name
  Liang Zhang
- Team leader address, phone number and email
  School of Software, Xidian University, No.2 South Taibai Road, Xi'an 710071, P.R.China
  +86-13700227912
  liangzhang@xidian.edu.cn
- Rest of the team members
  Guangming Zhu (Xidian University)
  Lin Mei (The Third Research Institute of Ministry of Public Security)
  Jie Shao (The Third Research Institute of Ministry of Public Security)
  Juan Song (Xidian University)
  Peiyi Shen (Xidian University)
- Team website URL (if any)
- Affiliation
  Schoole of Software, Xidian University
  The Third Research Institute of Ministry of Public Security

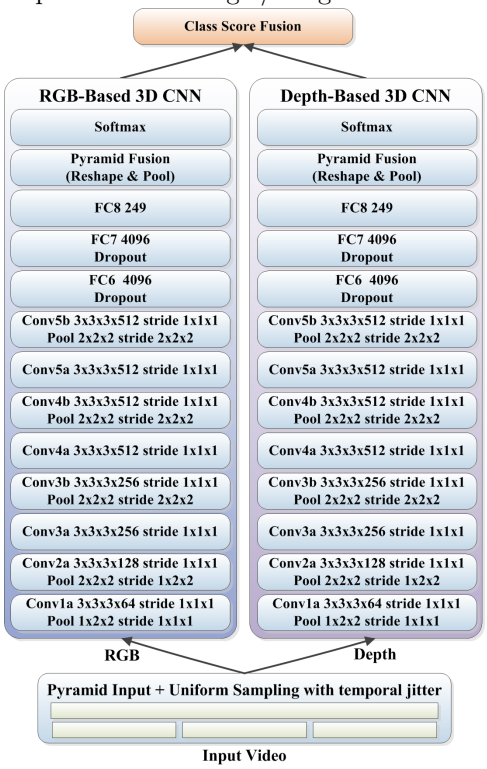
2 Contribution details

- Title of the contribution
  Gesture Recognition with Pyramidal 3D Convolutional Networks
- Final score
  Phase 1: 0.198997
  Phase 2: Unknown

- General method description
  1). 3D CNN model (C3D) is utilized.
  2). Pyramid input is employed. Specifically, each video file is sampled pyramidally: (a) Firstly, each file is segmented into three parts which may be overlapped in some degree according to the frame count of the video file;(b) Secondly, sixteen frames are sampled from each part and the whole video file respectively by use of uniform sampling with temporal jitter.(c) Lastly, four sixteen-frame batches are inputed into the 3D CNN model as the pyramid input for each video file.
  3). The outputs of the last fully-connected layer are fused among the pyramid input of each video file, before the final softmax layer or the softmax loss layer.
  4). The outputs of the two-stream, i.e. RGB and Depth, are fused later.

- References
  [1] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]. 2015 IEEE International Conference on Computer Vision (ICCV), 2015: 4489-4497.

- Representative image / diagram of the method



- Describe data preprocessing techniques applied (if any)
  1). Convert video files into image files for speeding up the training.

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)

No extra features except convolutional features are used in the method.

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)

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 total count of the parameters of the 3D CNN model is 105.8 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)

The proposed method is trained on the RGB and Depth modalities separately, and the final prediction result is obtained by fusing the pretiction result of the two modalities.

3.3 Global Method Description

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

  The pre-trained model "c3d_ucf101_finetune_whole_iter_20000" is used.

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

  None

- Qualitative advantages of the proposed solution
  1). The 3D CNN model learns spatial-temopral features.
  2). Pyramid input ensures that the most informative frames can be sampled when the length of each gesture video is greatly different.
  3). Uniform sampling with temporal jitter augments the original data.

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

  None

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

  The whole solution has not been published.

4 Other details

- Language and implementation details (including platform, memory, parallelization requirements)
  1). Language: C++ and Python
  2). Platform: Caffe
  3). Memory: TITAN X 12GB of memory

- Human effort required for implementation, training and validation?
  1). Convert video files into image files
  2). All can be executed automatically by the scripts if the environment has been configured correctly.

- Training/testing expended time?
  1). Training time: 45 hours for 45000 iterations
  2). Testing time: 10 minutes

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

  More than one chance to submit the test results may motivate challengers be more innovative.