Human Interaction Recognition

ShenShen

Shs2016f@bu.edu

1. Task

My project of the EC720 course is to try to explore the dataset from the "High-level Human Interaction Recognition Challenge". This challenge provides continuous videos captured by static camera on two scenario. Each video contains several human-human interactions (e.g. hand shaking) occurring sequentially and/or concurrently.[1]

The whole challenge requires contesters to correctly annotate the occurring activity of each frame and to locate where the activity is happening. This project will first focus on the annotation part of this challenge, which can be treated as a classification problem of each frame.

By exploring this challenge, I can learn about processing on video sequences to extract a respectively high level recognition feature. It requires me to combine the knowledge of video and image processing, DSP and machine learning techniques to develop a methodology of detection of specific behavioral states in humans. Many techniques from class such as motion detection and tracking may be applied on this project. Also, I will make practice on one of the most popular models on image and video proceeding problem, CNN.

The result of this challenge can be applied on many real world problems, such as behavior classification of surveillance camera and robot interaction.

2. Related Work

The behavior recognition problem has drawn attention recently because of its spread application on solving real world problems. However, to solve this problem is challenging due to the complexity of human behavior and the spatio-temporal features on video data.[2]

Although it has one higher temporal dimension of feature, we can refer to methods on image recognize. If we treat the task as a machine learning classification problem, the Deep Convolution Neural Network model could be applied. Convolutional neural network is a class of deep, feed-forward artificial neural networks[3]. CNN is proved to be a good model at doing image processing. Krizhevsky's[4] thesis provides us a basic reference of mainstream methodology to improve the behavior of CNN model.

One simple approach of solving this problem is to treat video frames as still images. Then we can apply CNNs to recognize actions at each individual frame. CNNs can also be applied on video sequence processing. However, if we make use of the connection between frames, we would acquire a more confident prediction result. One approach is to treat space and time as equivalent dimensions of the input and perform convolutions in both time and space[5]. Grey scale video frame could provide enough information of behavior classification. If the input video frames is in grey scale with time shift, a 3D convolution could be just like the convolution and pooling of a RBG image, which also has 3 dimension. In additional to use only adjunctive frames as one input of convolution, we can use different combination of the video frame in time scale as the input of the CNN model[6].

The course work of EC700 provide us knowledge of motion detection and because human action can be recognized only on the motion part of the video, motion detection will be a useful way to preprocessing the video frame.

3. Dataset

The dataset contains 6 classes of human-human interactions: shake-hands, point, hug, push, kick and punch. The ground truth labels for each interactions and its start and end time in each video is also provided. There are total 20 video sequences whose

lengths are around 1 minute. Each video contains at least one execution per interaction, providing us 8 executions of human activities per video on average.

The videos are taken with the resolution of 720*480, 30fps.

The dataset can be separated into two based on different back ground. The set 1 is composed of 10 video sequences taken on a parking lot. The set 2 (i.e. the other 10 sequences) are taken on a lawn in a windy day.[1] Both two scene involve slight camera movement. Also, the background like tree and grass would moving slightly.

4. Approach

I choose to use python as main coding language. Tensorflow and opencv are the main library on this project. Tensorflow is the deep learning framework to build the CNN model. Opencv is needed to do the demo and to demonstrate the video and the label.

In order to accelerate the model training, I may need to use the graphic processor on SCCC.

Challenging part of this project is:

a). Design and tune the CNN network to receive a converge and effective result: In order to receive a high performance CNN network, both the parameter, the structure, the optimization and regulization of the model should be carefully selected. In order to make a compare different CNN model, 2 or 3 CNN model will be apply on this project.

b). How to do the pre processing of the video using motion detection method: I will try both continuous frames subtraction, background subtraction, the adaptive threshold model, and no preprocessing video as the input of the CNN network to find if the motion detection method could provide a better result.

5. Detailed Timeline

10/25/2017 – 11/4/2017: Set up the deep learning frame work on my own computer, review the instruction of SCCC and try a simple model training, design the display of the processing result.

11/5/2017 – 11/25/2017: Design the CNN models(2 or 3), turn each model carefully, try data with different preprocessing and evaluate the result.

11/25/2017 – 12/5/2017: Make evaluation between models and result, write the report.

6. Code repository

https://github.com/shenmbsw/behavior_recognization

References

1. http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html#Data
2. Evaluation of Local Spatio-Temporal Features for Action Recognition, Proc. British Machine Vision Conf., p. 127, 2009.
3. https://en.wikipedia.org/wiki/Convolutional_neural_network
4. Krizhevsky, Alex. "ImageNet Classification with Deep Convolutional Neural Networks. Retrieved 17 November 2013.
5. Ji, Shuiwang; Xu, Wei; Yang, Ming; Yu, Kai (2013-01-01). 3D Convolutional Neural Networks for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 35 (1): 221–231.
6. Andrej Karpathy Large-scale Video Classification with Convolutional Neural Networks