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 to solve 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 video signal has temporal dimension, we can refer to methods for image recognition. 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 get 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 are 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 addition 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.

**3.1 bounding box**

For the classification problem, the ground truth of when and where the interaction is happening and witch type of act is it will be provided to the training set. When and where the interaction is happening will be feed for the testing set and which type of act is it need to be predicted.

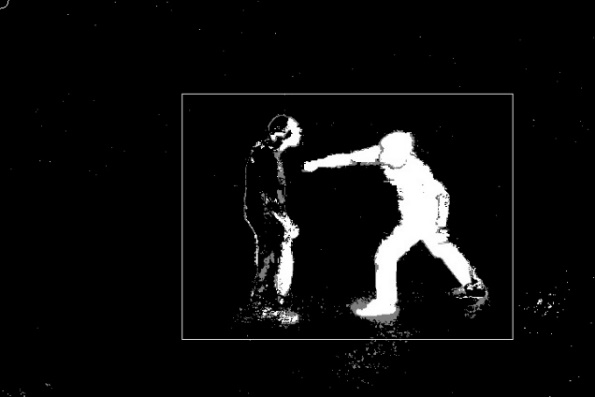


Fig1. The bounding box for a Punch action in seq1



Fig2. The bounding box for a handshake action in seq7

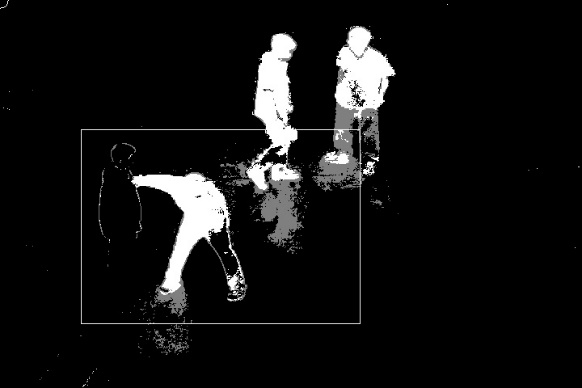


Fig3. The bounding box for a Punch action in seq8

From the data set we found that the area and the shape of the bounding box vary a lot between different video sequences. Another challenging issue for the data set is that in some video sequence, irrelevant pedestrians may appear and more than one pair pairs of interacting persons may execute the activities simultaneously.

**3.2 Action**

Among all 6 activities, only “Pointing” could be finished by only one person. The ground truth bounding box of Pointing has a very different height weight ratio with other activity because it only cover one person.

**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.

The first attempt for the project is to throw the whole image frame with the label of action class into the CNN. The network I design refer to is Alexnet. It is a 6 convolution layers + 3 fully connection layers convolution network.

This model guarantee the sufficient number of samples. If this model works, it would be helpful for detect time and location of activities.

The problem is that: 1.The whole image will first reshape into 224 \* 224. This may cause distortions and detail lost. 2. It may be confused by simultaneously activities and pedestrians issue. 3. Some frame of action (the very beginning or end of an action) may not provide information for the action type.

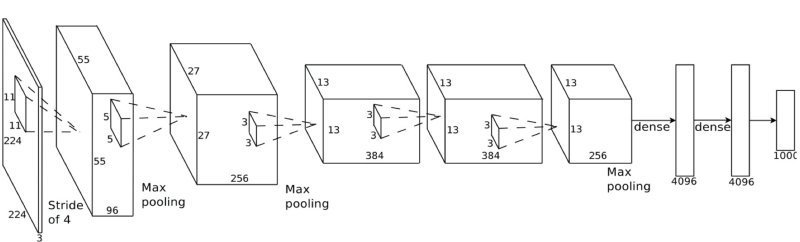
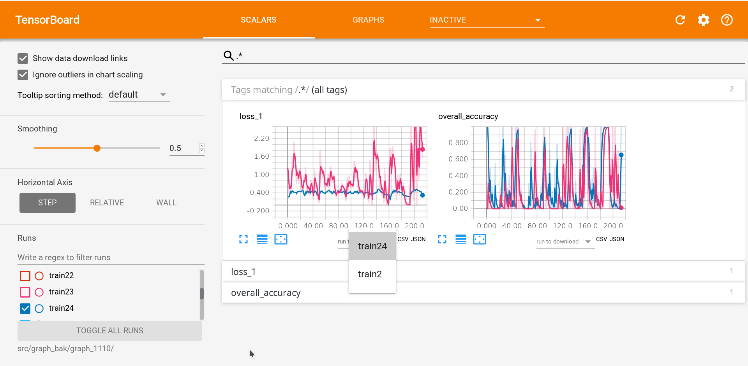


Fig4: the 2D CNN model



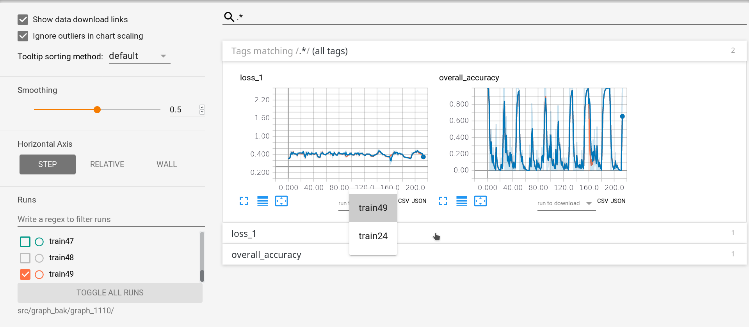


Fig5.Training loss and accuracy of frame to frame model

The graph above is the training accuracy for the frame to frame model. It shows that the loss of the CNN network do converge from the 2 epoch to 24 epoch, after 24 epoch, the loss does not decrease so much. The overall training accuracy is only 30%, which is just a little better than randomly guessing. It indicates that applying the frame to frame model could not provide an acceptable result.

**5. Plan for next stage**

Work for the next stage is to change the presentation of data to the CNN and change the CNN model.

The presentation of data could be:

1. Treat all data in each spatio-temporal bounding box as one sample.

2. Generating sample in spatio-temporal bounding box with sliding window. And make final decision by voting.

The problem of 1’ is that the sample number will drop rapidly. The problem of 2’ is that it might be hard for the model to learn action from the window.

The CNN model would be modified as:

1. 3D convolution
2. 2D convolution connected by FC layer
3. Recurrent structure

The measurement of the model will be not only the training loss, but also:

1. Confusion matrix
2. Cross validation by leave one out

**6. Code repository**

https://github.com/shenmbsw/behavior\_recognization

**7. 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