EECS 349 Project Final Report

Zhiyuan Wang, Haodong Wang and Xi Zheng

March 21, 2014

1 Method

We use trackelet feature[2]. The features are 256-dimensional spatial-time interest points(STIP) based on HOG and HOF(see report for detail). Since video's temporal and spatial properties vary between each other, the number of STIP extracted from each video also varies. The first thing is to project all the feature vector of different videos into same space. We clustering these spatial-temporal interest points to learn a dictionary (finally get 891 vocabularies or bases). Then calculate histogram of projected points in a video as its feature vector. We use these feature vectors to train SVM classifier for each action. Prediction is made by one-vs-all strategy.

2 Experiments

The experiments are done on KTH action dataset, which contains 6 classes of actions with almost 100 videos for each class. We totally use 72 videos for training and 72 videos for test, the number of videos for different class is similar (11 to 12). We trained SVM with different kernels and variables, and linear model achieved the best performance, 86.1% accuracy on test set. Below is the confusion matrix. That shows good features' importance in classification, even simple classifier like linear SVM can work well with them.

3 Conclusion and Discussion