# EECS 349 Project Final Report

Zhiyuan Wang, Haodong Wang and Xi Zheng

March 24, 2014

## 1 Method

We use tracklet feature for human action detection. The feature is 256-dimensional spatial-time interest points(STIP) based on HOG and HOF. HOG is short for histograms of oriented gradient which is an discrimitive and robust image descriptor and HOF is histograms of optical flow which is a motion descriptor. Tracklet

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 (Here "dictionary" can also be called as codebook or bag of features. It has 891 dimension or vocabulary).

Finally, we compute a video's feature vector by the following two steps: first assigning each feature point to its nearest vocabulary in the learned dictionary; then calculate the histogram of these visual words. 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). Each video is divided into 4 roughly equal segments, in our experiments we treat each video as an instance, so the dimension of one instance is $4 \times 891 = 3564$. We trained SVM with different kernels and variables. The variables cost and gamma are seleted by 10-fold validation. As Table 1-4 shows linear SVM achieved the best accuracy 86.1%. Second is RBF kernel SVM, which has 78% accuracy. Then is Sigmoid kernel with 75% accuracy. Polynomial kernel only get 72% accuracy.

We also test c-SVC and nu-SVC (regularized support vector classification using one vs one scheme for prediction). The linear one-vs-one multiclass SVM classifier achieved 88.9% accuracy on test set (see Table 5), but the other 3 kernels performed not well (under 40% accuracy, some of them even classified

all the instances into one class). That shows one-vs-all is more robust than one-vs-one in multiple class classification.

# 3 Conclusion and Discussion

The result, which linear SVM outperforms other more complicated classifier, shows that good feature is very important in classification. Discrimitive feature and simple classifier sometimes can attain suprisingly good performance. Intuitively our relatively small dataset, the way actions in video performed is similar, and we use the whole video rather than a smaller segment as instance, may also be a reason. From Table 1-4 we can also observe that handclapping and handwaving is get the highest recall, and handclapping has most false positive detected.

# 4 Future work

1. Though KTH dataset is a very popular action dataset, it has some drawbacks, which make it far enough to approximate natural video. We implement experiments in a more challenging dataset UCF101, which contains natural videos in 101 classes of actions, and make comparison with other models like GMM and Bayesian Network.

2. Sometimes we are not only interested in what event happens in a video, but want to know when and where does the event occur. We can further locate the event with spatial-temporal bounding box.

3. Detect complex event as a set of low-level events.

Table 1: Confusion matrix of SVM with RBF kernel (78% accuracy)

| | boxing | handclapping | handwaving | jogging | running | walking |
|---|---|---|---|---|---|---|
| boxing | 7 | 4 | 0 | 0 | 0 | 1 |
| handclapping | 1 | 11 | 0 | 0 | 0 | 0 |
| handwaving | 0 | 1 | 11 | 0 | 0 | 0 |
| jogging | 0 | 1 | 0 | 7 | 3 | 1 |
| running | 0 | 1 | 0 | 2 | 9 | 0 |
| walking | 0 | 1 | 0 | 0 | 0 | 11 |

Table 2: Confusion matrix of linear SVM (86.1% accuracy)

| | boxing | handclapping | handwaving | jogging | running | walking |
|---|---|---|---|---|---|---|
| boxing | 10 | 1 | 0 | 0 | 0 | 1 |
| handclapping | 0 | 12 | 0 | 0 | 0 | 0 |
| handwaving | 1 | 0 | 11 | 0 | 0 | 0 |
| jogging | 1 | 0 | 0 | 9 | 2 | 0 |
| running | 1 | 0 | 0 | 2 | 9 | 0 |
| walking | 1 | 0 | 0 | 0 | 0 | 11 |

Table 3: Confusion matrix of SVM with polynomial kernel(72.2% accuracy)

| | boxing | handclapping | handwaving | jogging | running | walking |
|---|---|---|---|---|---|---|
| boxing | 4 | 8 | 0 | 0 | 0 | 0 |
| handclapping | 0 | 12 | 0 | 0 | 0 | 0 |
| handwaving | 0 | 1 | 11 | 0 | 0 | 0 |
| jogging | 0 | 1 | 0 | 6 | 5 | 0 |
| running | 0 | 1 | 0 | 2 | 9 | 0 |
| walking | 0 | 2 | 0 | 0 | 0 | 10 |

Table 4: Confusion matrix of SVM with sigmoid kernel(75% accuracy)

| | boxing | handclapping | handwaving | jogging | running | walking |
|---|---|---|---|---|---|---|
| boxing | 5 | 6 | 0 | 0 | 0 | 1 |
| handclapping | 1 | 11 | 0 | 0 | 0 | 0 |
| handwaving | 0 | 1 | 11 | 0 | 0 | 0 |
| jogging | 0 | 1 | 0 | 7 | 4 | 0 |
| running | 0 | 1 | 0 | 2 | 9 | 0 |
| walking | 0 | 1 | 0 | 0 | 0 | 11 |

Table 5: Confusion matrix of C-SVC (88.9% accuracy)

| | boxing | handclapping | handwaving | jogging | running | walking |
|---|---|---|---|---|---|---|
| boxing | 10 | 2 | 0 | 0 | 0 | 0 |
| handclapping | 0 | 12 | 0 | 0 | 0 | 0 |
| handwaving | 0 | 1 | 11 | 0 | 0 | 0 |
| jogging | 1 | 0 | 0 | 11 | 0 | 0 |
| running | 1 | 0 | 0 | 1 | 10 | 0 |
| walking | 1 | 0 | 0 | 1 | 0 | 10 |