Classifier Fusion for Gesture Recognition using a Kinect Sensor

Ye Gu, Qi Cheng and Weihua Sheng
School of Electrical and Computer Engineering
Oklahoma State University
Stillwater, OK, 74078, U.S.A
{ye.gu, qi.cheng, weihua.sheng}@okstate.edu

Abstract

Gesture recognition is becoming a more and more popular research topic since it can be applied to lots of areas, such as vision-based interface, communication and interaction. In this paper, experiments are implemented to verify the potential to improve vision based gesture recognition performance using multiple classifiers. The proposed approach implements gesture recognition which combines decisions from a Dynamic Time Warping (DTW) based classifier and a Hidden Markov Model (HMM) based classifier. Both of these two classifiers share the same features extracted from the human skeleton model which is generated from a Kinect sensor. Then fusion rules are designed to make a global decision. The experiment results indicate that with the proposed fusion methods, the performance can be improved compared with either classifier.

1 Introduction

A gesture is a motion of the body that contains certain information. Edward T. Hall, a social anthropologist claims that 60% of all our communications are nonverbal [1]. Gestures are widely used from expressing emotions to conveying information. Therefore, more and more efforts are devoted to gesture recognition. It becomes a popular topic in computer science and language technology with the goal of interpreting human gestures via mathematical algorithms [2].

Gesture recognition has a wide application including Human Machine Interaction (HMI), Human Robot Interaction (HRI) and Social Assistive Robotics (SAR). This technology has the potential to change the way users interact with computers or robots by eliminating input devices such as joysticks, mouse and keyboards for HMI and robot controllers for HRI.

1.1 Related Works

Traditional gesture recognition is obtained through vision information. Depending on the type of the input data, the approach for interpreting a gesture could be done in different ways. Two most widely used approaches are skeletal-based and appearance-based algorithms. The former method makes use of 3D information of key body parts in order to obtain several important parameters, like palm position or joint angles. On the other hand, appearance-based systems take 2-D images or videos for direct interpretation [3]. Some vision-based related work are listed in [4–6].

Besides vision based approaches, wearable sensor based gesture recognition has been gaining attention. Due to the advancement in MEMS and VLSI technologies, several researchers use multiple sensors worn on a human body to record data of human movements. [7–9] is some related work.

In order to improve the recognition performance of the detection, the information fusion concept has been explored. Most of the fusion methods are decision-level fusion. Some fuse the decisions from different sensors. Zhang et al. [10] presented a framework for hand gesture recognition based on the information fusion of a three-axis accelerometer (ACC) and multichannel electromyography (EMG) sensors signals. A decision tree and multistream HMMs are utilized for decision-level fusion to generate a final decision. Chen et al. [11] presented a robust visual system that allows effective recognition of multiple-angle hand gestures in finger guessing games. Three Support Vector Machine (SVM) classifiers were trained for the construction of the hand gesture recognition system. The classified outputs were fused by proposed plans to improve system performance. The system presented can effectively recognize hand gestures, at over 93%, for different angles, sizes, and different skin colors. Feature-level fusion has also been explored. In He’s paper [12], a new feature fusion method for gesture recognition based on a single tri-axis accelerometer has been proposed. Both time-domain and frequency-domain features are extracted. Recognition of the gestures is performed using SVMs. The average accuracy results using the proposed fusion methods is 89.89% which is improved compared with the approach using only one of the features.

It is obvious that if multiple sensors are used together, the performance can be improved. Unlike the previous work, the focus of our paper is to seek improve-
ment on recognition results with mono data source. We try to find complementary information among different recognition algorithms, which may push the performance to maximum when limited sensors are available. Specifically in this work, we attempt to verify the possibility of performance improvement using a single sensor. We perform temporal human gesture recognition using a Kinect sensor. Despite the limitations of the device, it seems to have caused on to a large (and growing) extent in the marketplace which has brought gesture-recognition through 3-D sensors into the mainstream. Through this camera, the non-color based features of the human gestures can be extracted. The features are not sensitive to the changes of lighting condition and common image noise; also it is very convenient to integrate this platform with robots, such as mobile robots for human robot interaction purpose.

The rest of the paper is organized as follows. In Section II, the recognition and fusion approaches are introduced. Section III presents the implementations of the experiments. In Section IV, the experiment results are analyzed. Finally in Section V, the conclusion is drawn and some potential research directions are discussed.

2 Methodology

The problem arises naturally as a need for improvement of classification rates obtained from individual classifiers. Fusion of data/information can be carried out on three levels of abstraction closely connected with the flow of the classification process: data level fusion, feature level fusion, and classifier fusion [13]. We apply the same features for different classifiers. Therefore, it is reasonable to adopt classifier fusion approaches for performance improvement.

We have already developed a gesture recognition system in which HMMs are chosen to model and recognize the dynamics of the gestures [14]. In addition to the HMM-based classifier, a DTW classifier is developed which uses the same features i.e., time sequences of the four left arm related joint angles. Currently we consider 4 joint angles. However, even we have more joint angles involved, after data-preprocessing, they will always be mapped into 1-D symbols. So the feature space does no change much. The flowchart is shown in Fig. 1. In the data preprocessing procedure of the DTW classifier flow, besides segmentation and symbolization, GMM (Gaussian Mixture Model)/GMR (Gaussian Mixture Regression) are applied to the multiple training data sets to generate the template for each gesture.

2.1 Dynamic Time Warping (DTW)

DTW is an algorithm for measuring similarity between two sequences which may differ in time or speed. DTW has been applied to video, audio, and graphics. Indeed, any data which can be turned into a linear representation can be analyzed with DTW [15]. A well known application has been automatic speech recognition, to cope with different speaking speeds. Due to the similarity between the gesture and speech signals, it can be applied to gesture recognition too. The first step is to find the DTW templates for each gesture given the training data. Here the GMM/GMR is used to find the generalized trajectory. In the recognition phase, the new testing data is compared with each template, and the total distance is calculated to measure the similarity.

2.2 Fusion Method

Two mainstream fusion methods suitable for our case are selected. The first one attempts to determine a single classifier, which is the most likely one to produce the correct classification label for an input sample. While the other one takes into the information from each classifier into consideration to obtain a combined result.
2.2.1 Dynamic Classifier Selection (DCS)

DCS methods reflect the tendency to extract a single best classifier instead of mixing many different classifiers [16,17]. As a result, only the output of the selected classifier is taken as a final decision. In our case, the two classifiers used take different measure for classification. The HMM based classifier decides on the model with the maximum likelihood, while the DTW classifier selects the template with the minimum total distance. Therefore, a measure which can be used to compare the performance of different classifiers should be defined for the fusion purpose.

Here we define a parameter called decision confidence. The purpose is to compare the classification confidence between these two classifiers. If one of the classifiers has higher confidence than the other one, then the contribution to the classification from the other classifier will be ignored. It means that the final global decision is made based on the best local decision. The definition of the parameter is shown in the following equations.

For the HMM based classifier, the decision confidence $\alpha_H$ is defined as follows,

$$\alpha_H = \frac{1}{\sum_{i=1}^{5} \frac{1}{\log P(O|\lambda_i)}}$$

(1)

$O$ is the observation sequence. $\lambda_i$ is the $i$th HMM. The model $\lambda^*$ is the HMM with the highest likelihood for the given observation sequence $O$.

Similarly for the DTW based classifier, the decision confidence $\alpha_D$ is defined as follows:

$$\alpha_D = \frac{1}{\sum_{i=1}^{5} \frac{1}{Dist_i}}$$

(2)

$Dist_i$ is the normalized total distance when compared to each template $i$. $Dist^*$ is the minimum total distance among all when compared with each template. The parameters defined for the local classifiers can reflect the degree of the ambiguity of the local decisions. Each parameter is within [0,1] interval. The higher the parameter is, the more confidence the classifier shows. The output of the classifier with higher decision confidence will be chosen as the global decision.

2.2.2 Classifier Structuring and Grouping (CSG)

Meanwhile, another fusion method belonging to CSG is also designed. Instead of only taking the best decision, outputs of the classifiers are combined into one decision. The idea of CSG fusion method is to organize different classifiers in parallel, simultaneously and separately get their outputs as an input for a combination function or alternatively sequentially apply several combination functions [16]. Here a combination function is designed as follows,

$$\arg\max_k \left( \frac{1}{\sum_{i=1}^{5} \frac{1}{\log P(O|\lambda_i)}} + \frac{1}{\sum_{i=1}^{5} \frac{1}{Dist_i}} \right)$$

(3)

In this case, the global decision is collaboratively made by these two classifiers. Given an observation sequence, the decision confidence for each class is calculated for both of the two classifiers. Then the “total confidence” is obtained by adding up the decision confidences from each classifier. Finally, the decision with the highest “total confidence” is the global decision.

3 Implementation

Five gestures are defined for the experiments, i.e., come, go, wave, rise up and sit down. One demonstration for gesture “wave” is shown in Fig. 2. Each gesture is modeled in an HMM. Meanwhile, a template for each gesture is created for the DTW matching purpose. All the gestures are made by left arm. Features are extracted from four joint angles: left elbow yaw and roll, left shoulder yaw and pitch.

Figure 2: Snapshots for the gesture ”wave”.

3.1 Training phase

The flowchart of the HMM training phase is shown in Fig. 3.

Figure 3: Flowchart of the HMMs training phase.
Starting pose is defined. Each training data set consists of thirty data point (around 1.5s) after the starting point. Then K-means clustering is used to convert the vectors into the observable symbols for HMMs. The centroids are saved for clustering further testing data. To balance the computation complexity, efficiency and the accuracy, we set up parameters for HMM as follows: the number of states in the model is thirty; the number of distinct observation symbols is six.

3.2 Recognition phase

For real-time processing, the Robot Operation System (ROS) framework is adopted. The difference compared with our previous work [14] is the real-time classification node. Both HMMs based classifier and DTW based classifier are applied here followed by the implementation of two different fusion methods.

To get rid of noise and reduced the false alarm rate, for HMM-based decisions, first we use the variance of the input to judge if it is a gesture or not. Secondly we set a threshold for each HMM. If the likelihood is smaller than the threshold, it is treated as noise. For DTW-based decisions, a similar threshold is also set. The flowchart of the recognition phase is shown in Fig. 4.

4 Experiment Results

4.1 Training results

The training results of the HMMs are shown in Fig. 5. As the training iteration increases, the likelihood for each model converges. The templates created for the DTW algorithm are shown in Fig. 6. The GMM/GMR algorithm outputs the generalized templates for each gesture. Then the centroids obtained from the training data are used to cluster the 4-D templates into 1-D templates.

4.2 Recognition results

• Offline results

For the offline experiment, the data is saved to a file for post processing. Since it is offline-processing, we ignore the computation cost and use a sliding window of 30 data points with the step of one data point. Two offline testing results are shown in Fig. 7 and Fig. 8. Each gesture is made with one stroke; that is the subject will stay still for a couple of seconds between any two consecutive gestures. Most of the gestures are recognized by both classifiers. There are some gestures that have been recognized by only one classifier. Type I error occurs to the HMM based classifier. And both Type I and Type II errors occur to the DTW based classifier. One of the gestures shown in Fig. 7 and Fig. 8 is recognized by neither of the classifiers. With these two fusion method used, most errors are eliminated. Therefore, the performance of the fused decision is better than that of either single classifier. As the offline results indicate, the performance of these two fusion methods has no obvious difference. In the testing phase, we have two subjects, the one who participated in the training data collection (trainer) and the one who did not (tester). However, it is found that once the tester is familiar with the predefined gestures, the performance of the trainer and tester
has no big difference.

Figure 7: Offline recognition results 1.

Figure 8: Offline recognition results 2.

- **Online results**

For application purpose, the system should allow the user do certain gestures with any strokes, and most of the gestures should be recognized. Some statistical results are collected from the real-time experiment. Table 1 and Table 2 show the accuracy of the HMM and DTW classifiers respectively. The sum of each row $i$ is the total number of the gesture $i$ has been made. The column $i$ of row $i$ gives the number of gesture $i$ being detected. The column 6 gives the number of missed detections for each gesture. In the final column, the accuracy is given.

The results show that the HMM classifier has better performance than the DTW classifier. The statistical results for the two fusion approaches are shown in Table 3 and 4, respectively. It indicates that the fusion approaches can improve the recognition accuracy compared to either single classifier. For the DCS based fusion method, it improves the performance because it catches the complementary characteristics of these two classifiers. As shown in offline results, one of the classifier detect certain gesture which is missed by the other one. For the CSG based classifier, it improves the performance because it can detect the gesture which is missed by both of the classifiers. Since it takes into account the contribution of these two individual classifier, the total decision confidence can be high enough, even when neither of the classifier is confident with the detection result.

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Gesture Recognized</th>
<th>Test accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>37 0 0 0 0 4</td>
<td>.9024</td>
</tr>
<tr>
<td>2</td>
<td>0 41 2 0 0 2</td>
<td>.8696</td>
</tr>
<tr>
<td>3</td>
<td>2 0 46 0 0 2</td>
<td>.8790</td>
</tr>
<tr>
<td>4</td>
<td>0 0 0 45 0 5</td>
<td>.9000</td>
</tr>
<tr>
<td>5</td>
<td>3 0 0 0 44 2</td>
<td>.8800</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Gesture Recognized</th>
<th>Test accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35 0 0 0 0 6</td>
<td>.8537</td>
</tr>
<tr>
<td>2</td>
<td>0 38 4 0 0 4</td>
<td>.8261</td>
</tr>
<tr>
<td>3</td>
<td>0 0 36 0 0 5</td>
<td>.8780</td>
</tr>
<tr>
<td>4</td>
<td>0 3 0 42 0 8</td>
<td>.8400</td>
</tr>
<tr>
<td>5</td>
<td>0 0 0 30 0 2</td>
<td>.7800</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Gesture Recognized</th>
<th>Test accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>38 0 0 0 0 3</td>
<td>.9268</td>
</tr>
<tr>
<td>2</td>
<td>0 43 0 0 0 4</td>
<td>.9370</td>
</tr>
<tr>
<td>3</td>
<td>0 0 39 0 0 2</td>
<td>.9512</td>
</tr>
<tr>
<td>4</td>
<td>0 0 0 46 0 4</td>
<td>.9200</td>
</tr>
<tr>
<td>5</td>
<td>0 0 0 0 44 6</td>
<td>.8800</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Gesture Recognized</th>
<th>Test accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>38 0 0 0 0 3</td>
<td>.9268</td>
</tr>
<tr>
<td>2</td>
<td>0 42 0 0 0 4</td>
<td>.9370</td>
</tr>
<tr>
<td>3</td>
<td>0 0 38 0 0 3</td>
<td>.9268</td>
</tr>
<tr>
<td>4</td>
<td>0 0 0 45 0 5</td>
<td>.9000</td>
</tr>
<tr>
<td>5</td>
<td>0 0 0 0 44 6</td>
<td>.8400</td>
</tr>
</tbody>
</table>
5 CONCLUSIONS AND FUTURE WORK

In this work, two fusion methods are proposed for non-intrusive human gesture recognition through a Kinect sensor. Both HMMs and DTW algorithms are used for preliminary classification. HMMs are statistical models while DTW is a deterministic method. The results show that with appropriate fusion approaches designed, the performance of the classification can be improved.

The fusion method is very basic and straightforward. More efforts should be devoted to the theoretical explanation of the impact of the methods to the experiment results. Currently these two classifiers use the same information source. Different sensors such as motion sensors are encouraged to be used with the Kinect sensor to increase the diversity of the data we need to verify the possibility of further improvement, since fusion of data with large diversities may end up with more promising results.

References


