Fusing Multi-modal Features for Gesture Recognition

Jiaxiang Wu
National Laboratory of Pattern Recognition
Institute of Automation,
Chinese Academy of Sciences
Beijing, China
jiaxiang.wu@nlpr.ia.ac.cn

Jian Cheng*
National Laboratory of Pattern Recognition
Institute of Automation,
Chinese Academy of Sciences
Beijing, China
jcheng@nlpr.ia.ac.cn

Chaoyang Zhao
National Laboratory of Pattern Recognition
Institute of Automation,
Chinese Academy of Sciences
Beijing, China
chaoyang.zhao@nlpr.ia.ac.cn

Hanqing Lu
National Laboratory of Pattern Recognition
Institute of Automation,
Chinese Academy of Sciences
Beijing, China
luhq@nlpr.ia.ac.cn

ABSTRACT

This paper proposes a novel multi-modal gesture recognition framework and introduces its application to continuous sign language recognition. A Hidden Markov Model is used to construct the audio feature classifier. A skeleton feature classifier is trained to provided complementary information based on the Dynamic Time Warping model. The confidence scores generated by two classifiers are firstly normalized and then combined to produce a weighted sum for the final recognition. Experimental results have shown that the precision and recall scores for 20 classes of our multi-modal recognition framework can achieve 0.8829 and 0.8890 respectively, which proves that our method is able to correctly reject false detection caused by single classifier. Our approach scored 0.12756 in mean Levenshtein distance and was ranked 1st in the Multi-modal Gesture Recognition Challenge in 2013.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Interaction styles*; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis

General Terms

Algorithm, Experimentation

Keywords

Gesture Recognition; Multi-modal Fusion; Hidden Markov Model; Dynamic Time Warping

1. INTRODUCTION

Gesture recognition refers to recognizing meaningful motions executed by human, involving body, head, arm and/or hand movements. Gesture recognition has been a popular research field in recent years due to its promising application prospects in human-computer interaction.

In the early days of gesture recognition research, most approaches were controller-based, in which users had to wear or hold certain hardware for motion data capturing. In the recent few years, controller-free, especially vision-based gesture recognition has become the mainstream of the research. In vision-based approaches, users' motion data is captured by cameras and numerous computer vision methods have been successfully adopted into this area for further data analyzing and understanding.

With the development of input devices, more modalities have become available, which directly leads to the rise of multi-modal gesture recognition. Multi-modal gesture recognition tries to capture discriminative information from each modality and fuses them with certain strategies to obtain the final recognition result.

Kinect, the motion sensing input device developed by Microsoft corporation, features an RGB camera, a depth sensor and a multi-array microphone and is able to provide multi-modal data, including RGB image, depth image, skeleton, and audio. With all these features, Kinect provides an ideal experimental platform for multi-modal gesture recognition system's design and validation.

In 2013, ChaLearn organized the Multi-modal Gesture Recognition Challenge, which focused on recognizing "multiple instances, user independent learning" of 20 gestures categories of Italian signs. The dataset of this competition is captured by Kinect, including RGB video, depth video, skeleton and audio data. Our team scored 0.12756 in the

* is the corresponding author.

ICMI'13, December 9–13, 2013, Sydney, Australia
Copyright 2013 ACM 978-1-4503-2129-7/13/12 ...$15.00.
http://dx.doi.org/10.1145/2522848.2532589.

final evaluation phase and won the 1st prize of this competition.

In this paper, we describe our approach for this competition in detail. We construct classifiers respectively based on audio and skeleton feature and then fuse their results to obtain the final recognition result. In Section 2, we briefly review previously published methods in the gesture recognition field. We introduce our algorithm in Section 3 and experimental results in Section 4. Finally, we present a few conclusions and suggestions for future work in Section 5.

2. RELATED WORK

Gesture recognition systems can be roughly classified into two categories, based on their data capturing methods. Controller-based recognition systems constitute the first category, in which users have to wear or hold certain hardware while performing gestures. Kuroda *et al.* [9] introduced their low-price data glove, StrinGlove, which was able to obtain full degrees of freedom of human hand and had achieved satisfying performance for sign language recognition. Schreiber *et al.* [13] evaluated the potential of a gesture-based human computer interaction system with Wii Remote.

The second category is controller-free recognition systems, in which users do not need to hold any device while performing. Many sensors can be used for data capturing in these systems, such as cameras, laser sensors and infrared sensors. Among all these systems, camera-based or vision-based systems are more common in recent research. Considering the type and amount of camera(s) used, these systems can be further divided into single camera based, stereo cameras based, depth-aware camera based and so on.

Gestures to be recognized can be either static (a stable body posture) or dynamic (a sequence of body movements). Static gesture recognition is also known as posture recognition. Just *et al.* [8] introduced an approach to hand posture classification and recognition tasks. They also proposed an illumination-invariant feature based on the Modified Census Transform and achieved encouraging results on a benchmark database in this field. Bretzner *et al.* [3] described their multi-scale color feature and its application in a prototype system for hand tracking and posture recognition. Hand states are simultaneously detected and tracked with particle filtering. Fang *et al.* [4] proposed a robust real-time hand gesture recognition method. Hand is detected with Adaboost and then tracked by adaptive hand segmentation with motion and color cues. Finally, hand posture type is determined by palm-finger configuration with scale-space feature.

For dynamic gesture recognition systems, most frequently used approaches include Hidden Markov Model, Finite State Machine, Particle Filtering and Time-Delay Neural Network, as concluded by Mitra *et al.* [10].

Hidden Markov Model was first applied to gesture recognition by Yamato *et al.* [15], in which a discrete HMM was used to recognize six classes of tennis strokes. Glomb *et al.* [5] proposed an unsupervised parameter selection approach for gesture recognition system, with Hidden Markov Model and Vector Quantization applied.

By modeling gestures as state sequences in Finite State Machine, Yeasin *et al.* [17] proposed a vision-based system for dynamic hand gestures automatic interpretation. The temporal signature is analyzed to automatically interpret the performed gesture. Hong *et al.* [6] presented a state-based approach for gesture recognition. The spatial information is learnt from training data and then grouped into segments, which is further integrated to FSM recognizer.

Based on Particle Filtering, the condensation algorithm was developed and further extended by Black *et al.* [1] for incremental recognition of human motions, which are modeled as temporal trajectories of some estimated parameters over time.

Yang *et al.* [16] applied the Time-Delay Neural Network (TDNN) to recognize 40 hand gestures of American Sign Language (ASL). Pixel-level motion trajectory is obtained across the image sequence by multi-scale motion segmentation and affine transformation. Then, the motion trajectory is matched to a given gesture model with TDNN.

With more data captured by different devices becomes available, it is a natural thought to combine these modalities together to enhance the recognition performance. Bolt *et al.* [2] proposed a framework in which both hand gestures and speech signals are used to augment the user's ability to communicate with computers. In their prototype, two-handed gestures, both static and dynamic, were designed to input concepts, manipulate items and specify actions to be taken. Tue Vo *et al.* [14] described the text editor they developed, which allowed users to manipulate text using a combination of speech and pen-based gestures. Jaimes *et al.* [7] gave an overview of major approaches to multi-modal human-computer interaction from a computer vision perspective and discussed several crucial issues such as user and task modeling, multi-modal fusion and emerging applications.

3. THE PROPOSED APPROACH

We proposed a multi-modal gesture recognition framework based on our solution for ChaLearn Multi-modal Gesture Recognition Challenge. We construct classifiers based on audio and skeleton feature separately and then combine them together to generate the final recognition result. In this section, we introduce the competition and its corresponding datasets, and then present our approach in detail.

3.1 Competition Introduction

ChaLearn organized the Multi-modal Gesture Recognition Challenge in 2013. This competition focused on spotting gestures drawn from a certain gesture vocabulary, based on multiple gesture instances performed by different people. This competition started on June 21st and ended on August 25th. In the final evaluation phase, our team scored 0.12756 (mean Levenshtein distance) and won the 1st prize of this competition. Table 1 gives more detailed information on the final scores of top-ranked teams.

There are 20 categories of gestures in the pre-defined gesture vocabulary. Each gesture is corresponding to a specific word or phrase in Italian, such as "ok" or "perfetto". While performing a gesture with his or her body movement, the performer also speaks out the corresponding Italian word or phrase. Thus, the audio data also contains useful information for gesture recognition.

Multi-modal Gesture Recognition Challenge provided 3 datasets: *Development*, *Validation* and *Final Evaluation*, for algorithm development and evaluation. Each dataset is consist of hundreds of zip files, and each file contains approximately one-minute-long multi-modal gesture data, including

Table 1: Final Scores of Top-ranked Teams

Team Name	Final Score
iva.mm	0.12756
wweight	0.15387
E.T.	0.17105
MmM	0.17215
pptk	0.17325
lrs	0.17727
MMDL	0.24452
telepoints	0.25841

audio, video and skeleton information. The detailed information of these datasets can be found in Table 2.

Table 2: Detailed Information of All Datasets

Dataset Name	Gesture Amount	Label Availability	
		Phase 1	Phase 2
Development	7,754	Yes	Yes
Validation	3,362	No	Yes
Final Evaluation	2,742	No	No

Figure 1 displays a group of sample files in the dataset. From left to right are respectively the image selected from the RGB video, depth video and user-index video.

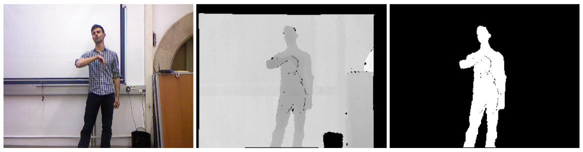


Figure 1: Samples from Training Dataset

The competition consists of two phases. In the first phase, competitors are required to train their models with *Development* dataset and predict labels in *Validation* dataset, and the prediction score is calculated instantly. In the second phase, both *Development* and *Validation* dataset are used as training data and all teams are required to submit their final prediction on *Final Evaluation* dataset. The final ranking is based on the score of their final prediction, which will not be published until the end of the competition.

3.2 Audio Feature Classifier

Since every gesture in video data is corresponding to a word or phrase spoken in audio data, and the performance of speech recognition has reached a satisfying level in recent years, it is natural for us to consider audio-based approaches at first.

First of all, we perform end-point detection in order to remove non-speech intervals. For end-point detection, there are mainly three categories of approaches [19]: time domain approach, frequency domain approach and time-frequency domain approach. Time domain approach is simple and easy to implement, but often performs badly under noisy background environment. Frequency domain approach is more robust to noise, but the computational complexity is much higher than time domain approach. Time-frequency domain approach tries to combine these two approaches together and improve robustness to noise while keeping the computational expense at a relatively low level.

Considering we can filter out false detections of speech intervals in later processing, we choose time domain approach for end-point detection for higher processing speed. We apply a 25-millisecond-long slide window for short-time energy calculation, and smooth the short-time energy sequence with Gaussian function. After Gaussian smoothing, we calculate the average short-time energy E_{ave} and set high threshold $T_H = 0.8E_{ave}$ and low threshold $T_L = 0.6E_{ave}$. After that, we scan the short-time energy sequence: any element larger than T_H indicates the beginning of a candidate speech interval and this interval ends when any of the following short-time energy values drops below T_L. Figure 2 presents one of the end-point detection result.

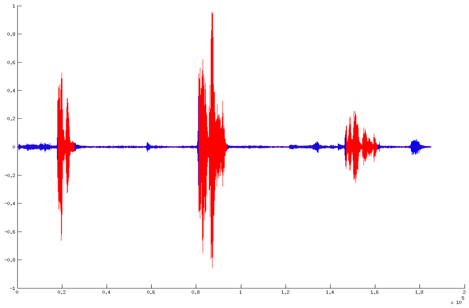


Figure 2: End-Point Detection Result

After end-point detection, we can obtain numbers of candidate speech intervals. For simplicity, we suppose all candidate speech intervals are corresponding to specific words drawn from the vocabulary. Therefore, the recognition problem is converted into a classification problem and we only need to classify each interval into one category.

Each word contains several phonemes. Therefore, we use different states to represent different phonemes, and the pronunciation process of any word can be modeled as a series of state transmissions in Hidden Markov Model framework [12][18], as shown in Figure 3.

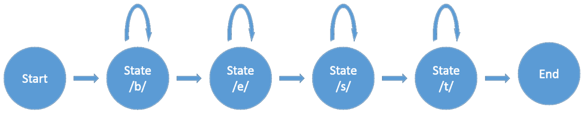


Figure 3: Hidden Markov Model for Word "Best"

We choose MFCC (Mel Frequency Cepstral Coefficients) as our audio feature used in Hidden Markov Model. The MFCC feature we use is consist of 39 dimensions: 12 for cepstral coefficients, 12 for delta cepstral coefficients, 12 for delta-delta cepstral coefficients, 1 for log energy, 1 for delta log energy and 1 for delta-delta log energy.

We model the MFCC feature sequence of the candidate speech interval as observation sequence $O = \{o_1, o_2, \ldots, o_T\}$, then the classification problem can be formulated as

$$c^{A*} = \arg\max_{1\leq c\leq 20} P\left(O|\lambda_c^A\right) \quad (1)$$

where λ_c^A represents the trained HMM model for the c-th category of gesture vocabulary and T is the length of MFCC feature vector sequence of the candidate speech interval.

The calculation of $P\left(O|\lambda_c^A\right)$ can be solved with both Forward Algorithm and Backward Algorithm [18]. One thing to note here is that the observation probability $b_j\left(o_t\right)$, which stands for the probability of the j-th state generating observation o_t, is defined as

$$b_j\left(o_t\right)=\frac{1}{\left(2\pi\right)^{D/2}\left|\Sigma_j\right|^{1/2}}\exp\left[-\frac{1}{2}\left(o_t-\mu_j\right)^T\Sigma_j^{-1}\left(o_t-\mu_j\right)\right] \tag{2}$$

where μ_j is the mean MFCC feature vector and Σ_j is the covariance matrix. Another thing to note is that we restrict the covariance matrix Σ_j to be a diagonal matrix to reduce the model complexity, in order to overcome the shortage of training data.

In training period, we first run end-point detection on the continuous audio data and obtain several candidate speech intervals. Then, we compare the segmentation result with the label information, remove false detections of speech intervals and assign the remaining intervals to the corresponding categories. After that, we train Hidden Markov Model classifiers based on the MFCC feature vector sequences of each category, using Baum-Welch Method [18].

3.3 Skeleton Feature Classifier

Since the topic of this competition is to develop a multimodal gesture recognition system, it is natural that we attempt to dig more information from other modalities, rather than only audio data. Among all the video features available, skeleton feature is easiest to obtain, and also provides a compact and informative representation of human body posture in each video frame. Hence, we choose skeleton feature to construct another gesture classifier.

To be specific, we only extract 4 points from all 20 skeleton points available, respectively are left elbow, left wrist, right elbow and right wrist. According to our observation, other skeleton points are either less informative or more unstable than these 4 points. The 3D positions of these 4 points are used to make up a feature vector of 12 dimensions.

Firstly, we need to divide the continuous skeleton data sequence into meaningful intervals. Due to the synchronization of audio and skeleton data, every candidate speech interval in audio data is corresponding to a candidate gesture interval in skeleton data, only with tiny variation in time. Thus the result of end-point detection can also be used for the segmentation of skeleton data sequence.

Secondly, we also assume that every candidate gesture interval contains a meaningful gesture, similar to the assumption of audio feature classifier, and attempt to classify each interval into 1 of the 20 gesture categories. For classification task, we need to train a model for each gesture category, compare the candidate gesture interval with all models and choose the most similar category as the classification result, as shown below

$$c^{S*}=\arg\max_{1\leq c\leq 20}Sim\left(S,\lambda_c^S\right) \tag{3}$$

where $S=\{s_1,s_2,\ldots,s_T\}$ stands for the skeleton data sequence of candidate interval and λ_c^S stands for the trained model of skeleton feature for the c-th gesture category.

We define the similarity measuring function $Sim\left(S,\lambda_c^S\right)$ in a neighborhood-based approach

$$Sim\left(S,\lambda_c^S\right)=\frac{1}{K}\sum_{i=1}^{K}Sim\left(S,S_{c,i}\right) \tag{4}$$

where $S_{c,i}$ stands for the i-th nearest neighbor of all training data in the c-th gesture category, and K is the size of neighborhood. $Sim\left(S,S_{c,i}\right)$ is the similarity between two skeleton feature vector sequences, and can be calculated with Dynamic Time Warping [11].

3.4 Classifier Combination Framework

Up to now, we have proposed two gesture classifiers, separately based on audio and skeleton features. Both classifiers rely on an important assumption, that all candidate intervals are corresponding to a gesture category drawn from the vocabulary. However, this assumption does not hold for all situations. First, the end-point detection may produce false speech intervals due to noisy background. Second, the performer may speak out-of-vocabulary words. The assumption fails under these situations and causes false recognition.

In order to overcome this problem, we propose a classifier combination framework, which combines the audio feature classifier and skeleton feature classifier, aiming to filter out meaningless candidate intervals. If the candidate interval does contain a gesture from the vocabulary, then two classifiers should produce same classification result, while this phenomenon should not appear when the candidate interval contains only background noise or out-of-vocabulary word.

For any candidate interval, both classifiers are able to calculate a confidence score for each gesture category, indicating how confident the classifier is about its classification result. However, the confidence scores need to be normalized for later fusion operation, since the data range of these scores may differ a lot.

We assume that for each classifier, 20 confidence scores (representing 20 categories) follow a Gaussian distribution. The mean and variance of this Gaussian distribution can be estimated with

$$\hat{\mu}=\frac{1}{20}\sum_{c=1}^{20}S_c \tag{5}$$

$$\hat{\sigma}^2=\frac{1}{19}\sum_{c=1}^{20}\left[S_c-\hat{\mu}\right]^2 \tag{6}$$

where S_c is the confidence score for the candidate interval and the c-th category.

Then we can use this Gaussian distribution to normalize these confidence scores. The normalized scores are defined as

$$S_c^*=\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{S_c-\hat{\mu}}{\sqrt{2\hat{\sigma}^2}}\right)\right] \tag{7}$$

After normalization, all scores are located in interval $(0,1)$ and higher confidence score leads to higher normalized score.

Therefore, we can obtain the normalized score S_c^{A*} and S_c^{S*}, which is achieved by the audio feature classifier and skeleton feature classifier, respectively. The final score is a linear combination of these two scores

$$S_c^*=\alpha S_c^{A*}+\left(1-\alpha\right)S_c^{S*} \tag{8}$$

where coefficient α controls the weight of two classifiers and is determined with a 4-fold cross validation on *Development* dataset.

After obtaining the final score for the candidate interval, we use a threshold θ to determine whether this interval does contain a valid gesture. Any interval that obtains a higher score than θ is classified into the most similar category; otherwise, this interval is canceled. This threshold θ is also determined with cross validation.

4. EXPERIMENTAL RESULTS

In this section, we present the experimental results to evaluate the performance of each individual classifier and their combination. Because *Final Evaluation* dataset lacks label information, we use *Development* as training set and *Validation* as testing set in the following experiments.

4.1 Single-modal Gesture Recognizer

We begin with the performance evaluation of audio feature classifier and skeleton feature classifier.

Firstly, we apply end-point detection to continuous audio data to obtain candidate intervals. Since the label data of *Validation* dataset is available, we can label each candidate interval with its corresponding gesture ID [1] or "NULL" if it contains no meaningful gesture.

Secondly, we give each candidate interval a predicted label with these two classifiers separately. All intervals are treated as an occurrence of a valid gesture, so the predicted label will never be "NULL". We compare the predicted labels with the actual labels and obtain one confusion matrix for each classifier. Figure 4-5 is the visualization of confusion matrixes of each classifier.

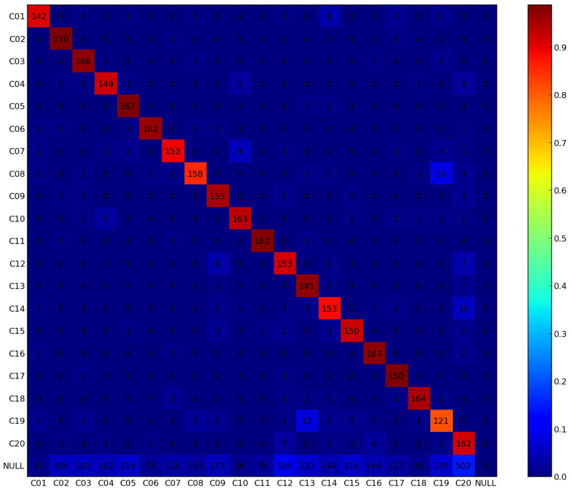


Figure 4: Single-modal Gesture Recognizer based on Audio Features

From Figure 4-5, we can see that the overall performance of audio feature classifier is superior to the performance of skeleton feature classifier.

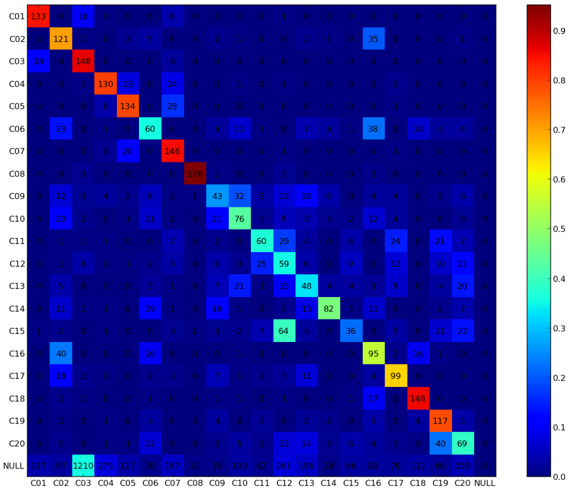


Figure 5: Single-modal Gesture Recognizer based on Skeleton Features

For detailed analysis of these two confusion matrixes, we calculate the precision and recall scores for each gesture category, as shown in Table 3 and 4.[2]

From Table 3, we can see that the recall scores of audio feature classifier are satisfying; however, the precision scores are rather low. This is mainly because of the false detections caused by meaningless candidate intervals, since "Precision-2" scores are much higher than "Precision-1" scores. We can observe similar phenomenon in Table 4.

The precision scores confirm that our concern about the previous assumption (that all candidate intervals contain valid gestures) is necessary. There are indeed many situations that this assumption will fail and cause the deterioration of recognition result.

4.2 Multi-modal Gesture Recognizer

Now we examine the recognition performance of our proposed multi-modal gesture recognizer.

The experiment procedures are similar to the experiment above. We also obtain a confusion matrix using our multi-modal gesture recognizer and its visualization is shown in Figure 6. However, since our gesture recognizer rejects candidate intervals with low scores, the predicted label may sometimes be "NULL". This maybe be either correct detection of meaningless intervals or incorrect rejection of meaningful intervals.

In order to compare with previous single-modal gesture recognizers, we also calculate the precision and recall scores for each gesture category, as shown in Table 5.

It is obvious that the "Precision-1" scores have improved significantly, comparing with single-modal gesture recognizers. In addition, the recall scores only drop slightly, and still maintain an acceptable level. This indicates that the multi-modal gesture recognizer is able to reject most of meaningless candidate intervals without much loss in recall rate. Therefore, to some degree, the problem caused by the previ-

[1] Gesture ID is determined according to alphabetical order. For instance, "basta" is the 1st category and "vieniqui" is the 20th category.

[2] In Table 3-5, row "Precision-1" contains the precision scores with the false detections caused by meaningless candidate intervals included, while row "Precision-2" contains the precision scores with these false detections excluded.

Table 3: Precision and Recall of Audio Feature Classifier

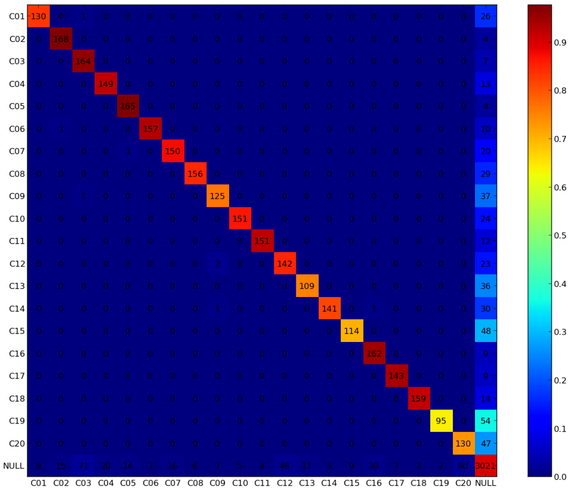
Gesture ID	1	2	3	4	5	6	7	8	9	10	
Precision-1	0.7245	0.4658	0.4598	0.4671	0.4260	0.6807	0.5464	0.5918	0.4519	0.6996	
Precision-2	0.9530	0.9714	0.9540	0.9490	0.9653	0.9878	0.9563	0.9634	0.9118	0.9209	
Recall	0.9045	0.9884	0.9708	0.9198	0.9882	0.9643	0.8947	0.8541	0.9509	0.9314	
Gesture ID	11	12	13	14	15	16	17	18	19	20	Mean
Precision-1	0.6667	0.3097	0.3588	0.4873	0.4043	0.5252	0.5396	0.7885	0.3135	0.2314	0.5069
Precision-2	0.9938	0.9107	0.8812	0.9000	0.9677	0.9598	0.9317	0.9762	0.8231	0.8182	0.9348
Recall	0.9816	0.9162	0.9724	0.8844	0.9259	0.9766	0.9868	0.9480	0.8121	0.9153	0.9343

Table 4: Precision and Recall of Skeleton Feature Classifier

Gesture ID	1	2	3	4	5	6	7	8	9	10	
Precision-1	0.4586	0.3399	0.1061	0.3044	0.4589	0.3315	0.3510	0.8421	0.2739	0.2934	
Precision-2	0.8693	0.4672	0.8000	0.8553	0.7657	0.4196	0.6667	0.9888	0.3525	0.4872	
Recall	0.8471	0.7035	0.8655	0.8025	0.7929	0.3571	0.8538	0.9514	0.2638	0.4343	
Gesture ID	11	12	13	14	15	16	17	18	19	20	Mean
Precision-1	0.3509	0.1245	0.1433	0.6165	0.2222	0.3065	0.4108	0.5103	0.4209	0.2156	0.3541
Precision-2	0.5505	0.2770	0.3453	0.7810	0.5455	0.4130	0.6000	0.8268	0.5519	0.4182	0.5991
Recall	0.3681	0.3533	0.3310	0.4740	0.2222	0.5556	0.6513	0.8555	0.7852	0.3898	0.5929

Table 5: Precision and Recall of Multi-modal Gesture Recognizer

Gesture ID	1	2	3	4	5	6	7	8	9	10	
Precision-1	0.9320	0.8895	0.6735	0.8539	0.9071	0.9814	0.8902	0.9471	0.9306	0.9686	
Precision-2	1.0000	0.9941	0.9880	0.9935	0.9940	1.0000	1.0000	0.9938	0.9781	1.0000	
Recall	0.8726	0.9826	0.9649	0.9383	0.9822	0.9405	0.9006	0.8703	0.8221	0.8800	
Gesture ID	11	12	13	14	15	16	17	18	19	20	Mean
Precision-1	0.9682	0.6837	0.8511	0.9481	0.8844	0.8250	0.9363	0.9817	0.9717	0.6347	0.8829
Precision-2	1.0000	0.9932	1.0000	1.0000	1.0000	0.9940	1.0000	1.0000	1.0000	0.9858	0.9957
Recall	0.9325	0.8802	0.8276	0.8439	0.8025	0.9649	0.9671	0.9306	0.6913	0.7853	0.8890



	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	NULL
C01	130	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26
C02	0	168	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
C03	0	0	164	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7
C04	0	0	0	149	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13
C05	0	0	0	0	165	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
C06	0	1	0	0	0	157	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10
C07	0	0	0	0	1	0	150	0	0	0	0	0	0	0	0	0	0	0	0	0	20
C08	0	0	0	0	0	0	0	156	0	0	0	0	0	0	0	0	0	0	0	0	29
C09	0	0	1	0	0	0	0	0	125	0	0	0	0	0	0	0	0	0	0	0	37
C10	0	0	0	0	0	0	0	0	0	151	0	0	0	0	0	0	0	0	0	0	24
C11	0	0	0	0	0	0	0	0	0	0	151	0	0	0	0	0	0	0	0	0	12
C12	0	0	0	0	0	0	0	0	2	0	0	142	0	0	0	0	0	0	0	0	23
C13	0	0	0	0	0	0	0	0	0	0	0	0	109	0	0	0	0	0	0	0	36
C14	0	0	0	0	0	0	0	0	1	0	0	0	0	141	0	1	0	0	0	0	30
C15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	114	0	0	0	0	0	48
C16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	162	0	0	0	0	9
C17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	143	0	0	0	9
C18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	159	0	0	14
C19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95	0	54
C20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	130	47
NULL	8	15	71	20	14	3	16	6	7	5	4	48	17	5	9	30	7	3	2	60	3021

Figure 6: Multi-modal Gesture Recognizer

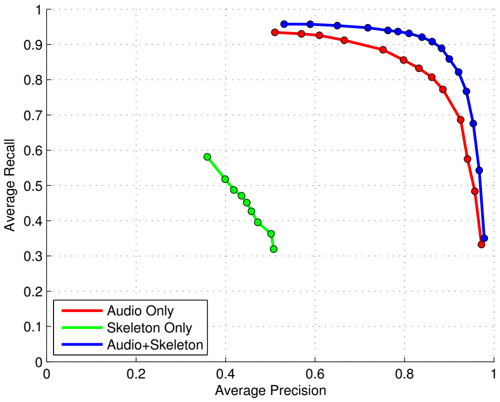


Figure 7: Precision-Recall Curves

ous assumption has been solved by our classifier combination framework.

Now we plot the precision-recall curves of each gesture recognizer for more intuitive comparison in Figure 7.

For multi-modal gesture recognizer, we can adjust the value of threshold θ to obtain different precision-recall score pairs. However, for the default single-modal gesture recognizers, there is no corresponding precision-recall curve since the recognizer does not reject any candidate interval, regardless of how low the confidence score is. Therefore, we first normalize the confidence scores and then set a threshold for candidate interval rejection. In this way, we can obtain the precision-recall curves for the single-modal gesture recognizers, based on audio and skeleton feature respectively.

From Figure 7, we can see that our multi-modal gesture recognizer outperforms other single-modal gesture recognizers obviously, which also proves that our classifier combination framework has achieved the desired performance.

5. CONCLUSION

In this paper, we present our approach employed in the Multi-model Gesture Recognition Challenge in 2013. Our approach makes full exploration of both audio and skeleton data, and with a novel classifier combination framework, our multi-modal gesture recognizer achieves satisfying performance in the final evaluation phase in this competition.

However, due to the time limitation, many other modalities still remain unused. We are looking forward to mining more information from video data, so as to improve the overall gesture recognition performance to a new level. Also, the skeleton feature should be able to produce better recognition results and more experiments are needed in this aspect.

6. ACKNOWLEDGMENTS

This work was partially supported by the key project of Chinese Academy of Sciences (Grant No.KGZD-EW-103-5(8)) and "Strategic Priority Research Program" of Chinese Academy of Sciences (Grant No.XDA0601040).

7. REFERENCES

[1] M. J. Black and A. D. Jepson. A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions. In *Computer Vision-ECCV'98*, pages 909–924. Springer, 1998.

[2] R. A. Bolt and E. Herranz. Two-handed gesture in multi-modal natural dialog. In *Proceedings of the 5th annual ACM symposium on User interface software and technology*, pages 7–14. ACM, 1992.

[3] L. Bretzner, I. Laptev, and T. Lindeberg. Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 423–428. IEEE, 2002.

[4] Y. Fang, K. Wang, J. Cheng, and H. Lu. A real-time hand gesture recognition method. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 995–998. IEEE, 2007.

[5] P. Głomb, M. Romaszewski, A. Sochan, and S. Opozda. Unsupervised parameter selection for gesture recognition with vector quantization and hidden markov models. In *Human-Computer Interaction–INTERACT 2011*, pages 170–177. Springer, 2011.

[6] P. Hong, M. Turk, and T. S. Huang. Gesture modeling and recognition using finite state machines. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 410–415. IEEE, 2000.

[7] A. Jaimes and N. Sebe. Multimodal human–computer interaction: A survey. *Computer vision and image understanding*, 108(1):116–134, 2007.

[8] A. Just, Y. Rodriguez, and S. Marcel. Hand posture classification and recognition using the modified census transform. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 351–356. IEEE, 2006.

[9] T. Kuroda, Y. Tabata, A. Goto, H. Ikuta, and M. Murakami. Consumer price data-glove for sign language recognition. In *Proc. of 5th Intl Conf. Disability, Virtual Reality Assoc. Tech., Oxford, UK*, pages 253–258, 2004.

[10] S. Mitra and T. Acharya. Gesture recognition: A survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 37(3):311–324, 2007.

[11] M. Müller. Dynamic time warping. *Information Retrieval for Music and Motion*, pages 69–84, 2007.

[12] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[13] M. Schreiber, M. von Wilamowitz-Moellendorff, and R. Bruder. New interaction concepts by using the wii remote. In *Human-Computer Interaction. Novel Interaction Methods and Techniques*, pages 261–270. Springer, 2009.

[14] M. T. Vo and A. Waibel. Multi-modal hci: combination of gesture and speech recognition. In *INTERACT'93 and CHI'93 Conference Companion on Human Factors in Computing Systems*, pages 69–70. ACM, 1993.

[15] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*, pages 379–385. IEEE, 1992.

[16] M.-H. Yang and N. Ahuja. Recognizing hand gestures using motion trajectories. In *Face Detection and Gesture Recognition for Human-Computer Interaction*, pages 53–81. Springer, 2001.

[17] M. Yeasin and S. Chaudhuri. Visual understanding of dynamic hand gestures. *Pattern Recognition*, 33(11):1805–1817, 2000.

[18] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. The htk book. *Cambridge University Engineering Department*, 3:175, 2002.

[19] M.-z. Zhou and L.-x. Ji. Real-time endpoint detection algorithm combining time-frequency domain. In *Intelligent Systems and Applications (ISA), 2010 2nd International Workshop on*, pages 1–4. IEEE, 2010.