SIGNER-INDEPENDENT SIGN LANGUAGE RECOGNITION BASED ON HMMs AND DEPTH INFORMATION

[1] *Ching-Tang Hsieh* (*謝景棠*), [1] *Peng-Hsiang Lin* (*林芃翔*) [1] *Li-Ming Chen (陳力銘)*
[1] *Hui-Chun Wang(王蕙君)* [1] *Chieh-Fu Hsieh (謝杰甫)*

[1] Department of Electrical Engineering, Tamkang University, Tanshui, Taipei, 25137, Taiwan
E-mail: hsieh@ee.tku.edu.tw

ABSTRACT

In this paper, we use the depth information to effectively locate the 3D position of hands in sign language recognition system. But the information will be changed by different signers and we can't do recognition well. Here, we use the incremental changes of the three-dimensional coordinates on a unit time as feature set to fix the above problem. And we use hidden Markov models(HMMs) as time-varying classifier to recognize the moving change of sign language on time domain. We also include HMMs with scaling factor to solve the underflow effect of HMMs. Experiments verify that the proposed method is superior then traditional one.

Keywords *Hidden Markov Models 、Sign Language Recognition 、Depth information.*

1. INTRODUCTION

Gesture is a nonverbal communication way. It can generally be classified to several types as follows: joining some auxiliary gesture in a dialogue, manipulation gesture and communication gesture. Here we try to manage the communication gesture. In communication gesture, the sign language has the integral structure and the most variation in time domain [1]. Stokoe et al. [2] define four parts of sign language in combination: the shape of hands, position, the moving direction and trajectory. The four parts is classified to two categories: hand gestures and space gestures. Hand gestures include the shape of hands and the moving direction; Space gestures include the position and trajectory. Most researches obtain these information by tracking the position of hands with skin-color detection, or wearing the specific color gloves[3][4]. For example, Koki Ariga et al. [3] use the HMMs as recognizer and skin-color as feature to detect face and hands. They use the centroid of face with K-means algorithm as the reference point to obtain 2D coordinates of hands' centroid. They also include the 1st-order and 2nd-order differential coefficients of the hands' coordinates as used in speech recognition system, which are called dynamic features, in their system. But the recognition is not well. M.Mohandes et al. [4] apply Gaussian Skin Color Model and the Region-growing Technique to track the face position of signers. Moreover, the signer's hands wear yellow and orange glove respectively to obtain higher recognition rate. However, background noise will affect the accuracy of feature extraction. Since, using colors as features, will be influenced easily in background and the light changing in environment.

In 2011, Microsoft released Kinect. This equipment has a RGB camera, a depth sensor and a multi-array microphone. Thus, it can track the action of the players, who can interact with the game station by the motion of the body and the voice. Sign language is also according to the difference of facial expressions with the variety of body gesture to expresses the meaning of the vocabulary and grammar. Simon Lang et al. [5] use Kinect to obtain body skeleton and depth information and combine HMMs to recognize the Deutsche Gebärdensprache Sign Language (DGS). They try to raise the recognition rate by including the depth information, but the features they chosed is not satisfied. Thus they can't achieve higher recognition rate in Signer-Independent experiments.

Also, W.Gao et al. [6][7][8] use a CyberGlove and three-dimensional position tracker released by Pohelmus to obtain the hands' features. They integrate the Simple Recurrent Network (SRN) and Hidden Markov Models (HMMs) to establish the recognition system for Chinese Sign Language (CSL). To raise the recognition rate, W.Gao et al. [8] apply the Self-Organizing Feature Maps (SOFMs) and HMMs, with Self-adjusting Recognition Algorithm as recognizer. The three-dimensional position can be correctly detected with the equipment, but the price is too high and it is hard to set up. So, it doesn't suit for the general users.

In sign language recognition, feature extraction and recognition are the most important. In this paper, we use Kinect to capture the features to avoid lighting effect of environment. We can easily turn the feature information from dimensional coordinates to three-dimensional coordinates with infrared depth sensor, and calculate the incremental change of these 3D data for each frame as feature set. Owing to the underflow effect of HMMs, we integrate HMMs with Scaling factor [9][10] as recognizer.

2. PROPOSED METHOD

The organization of the proposed method is shown in Fig.1. We separate the system into two parts. First, we extract the features by Kinect. We use the middle software NITE developed by PrimeSense to obtain the human skeleton. Then we normalize the coordinates to get the observation sequence. Second, we apply the well trained Hidden Markov Models as the recognizer to identify the sign language.

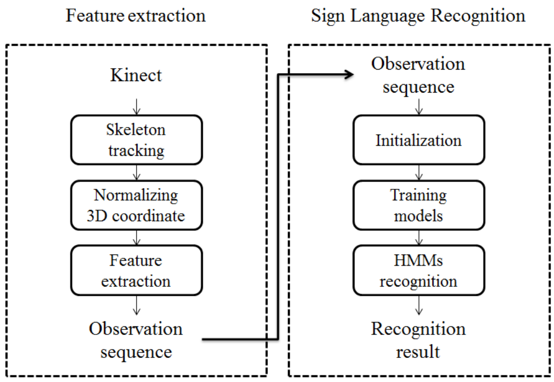


Fig. 1 The system flow chart.

2.1. Feature Extraction

In this section, we first apply Kinect to track the signer’s skeleton, then normalize the 3D depth coordinates of skeleton. We use the Homogeneous coordinates to replace the skeleton coordinate system. In the practical application, the user could be at different position, will become unstable feature. In this system, we do the geometric conversion to the Cartesian coordinate system for setting up the torso center as the coordinate origin. Figure 2 shows the geometry conversion in 3D Cartesian coordinates; Figure 3 shows the torso as the origin of the coordinate system [11]; Figure 4 shows the normalization result of Kinect skeleton coordinates.

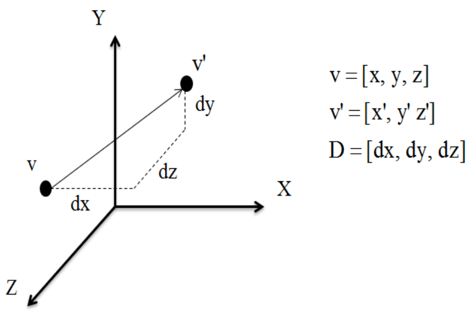


Fig. 2 Geometry conversion in the 3D Cartesian coordinates.

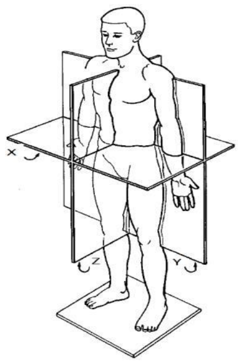
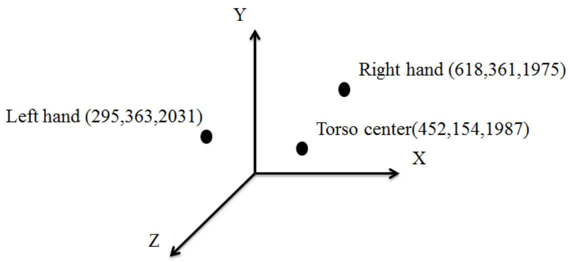
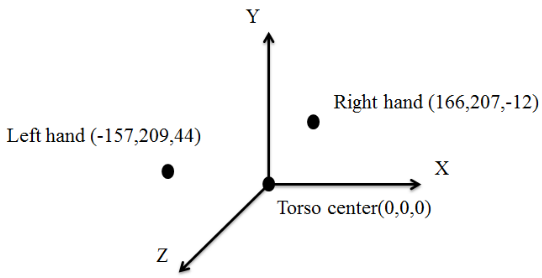


Fig. 3 The torso as the origin of the coordinate system [11].



(a)



(b)

Fig. 4 (a) Original coordinates of the Kinect skeleton.

(b) Normalization coordinates of the Kinect skeleton.

After, we build the three-dimensional space to capture the three-dimensional coordinates of the human joint points. Two feature sets are used in this paper. One is space gesture feature by choosing the absolute distance between the three-dimensional coordinates of the joint points. Become hands length and height for each signer could be different, the other feature set is calculating the variation of the hands movement to solve the location problem of hands caused by the different signers.

2.2. HMMs Model Training

In this section, we will introduce the training method of HMMs model. In this paper, we use the continuous HMMs model. Therefore, we first initialize the parameters including mean vector, covariance and weight coefficient. After that, we can start to train the HMMs models. We will apply Baum-Welch method for training to adjust the parameters to get the maximum. We set up 5 signers and 20 words for training. Each signer can produce 10 sets of observation sequence to 20 words. And 20 words can train 20 HMMs models. The training data of every model respectively is the 6 sets of observation sequence. The others are for testing. In Singer-Independent experiment, we follow the above mentioned way to train the 5 signers in turn. Otherwise, the initialized data we use is the training data. After all, we can obtain a set of the initialized parameter from HMMs model. We can input this parameters and the training data, then we through the Baum-Welch method to produce a new parameter for recognizing. The Figure 6 is the HMMs model training flow chart.

2.3. Sign Language Recognition

The most difficult thing in sign language recognition is to recognize the same movement. Even we ask the same signers to do the same sign language; they may not do exactly the same. So, for solving this problem to get the higher recognition rate is how to calculate the statistical variations. In this system, we change the mean to the mean vector and the covariance to the covariance matrix. This solution is Continuous Hidden Markov Model. We have to initialize the mean vector, the covariance matrix and the weight coefficient. Also we need to re-estimate. After that, we use Forward-backward Algorithm to calculate the probability of the testing sequence in the HMMs models. Then we apply Viterbi Algorithm to search the state sequence corresponded to the testing sequence in the model. Therefore, the model and the state sequence are the recognition results. The Figure 7 is the recognition flow chart.

3. EXPERIMENTAL RESULTS

The input depth image size is 320x240 pixels, and the output color image size is 640x480 pixels. The hardware we used is the Microsoft Kinect and computer CPU of Intel ® Core (TM) i5-2400M 3.1GHz, RAM 3.49GB. The software we used is the Microsoft Visual Studio 2010, openCV2.3, OpenNI 1.5.4.0. In the experiment, we set up the Kinect at 140 cm high and 150 cm away from the signer. We have 5 signer, and each signer performs 20 isolated signs 10 times. In dependent test, we select 5 times as training the rest as testing for each signer and take average. In independent test, we select 1 from 5 signer as training, the rest four are testing, and then average these results. Table 1 shows the comparison of these two feature sets in Signer-dependent experiment. Where, A feature set represent the distance from two hands and elbows to head, and distance from right hand to left shoulders and left hand to right shoulders. B feature set represent the dynamic variation of 3D coordinates of both hands.

For comparison. We neglect the depth information from 3D feature set to obtain the 2D feature set. Table 2 shows that depth information is very important to get the high recognition rate in signer-independent experiment.

Table 3.1 and table 3.2 shows average of 5 signers' confusion matrix for singer-independent. We can find out that the similar actions and the same moving but different locations will be misjudgment, like "thanks" and "photograph" or "don't know" and "free".

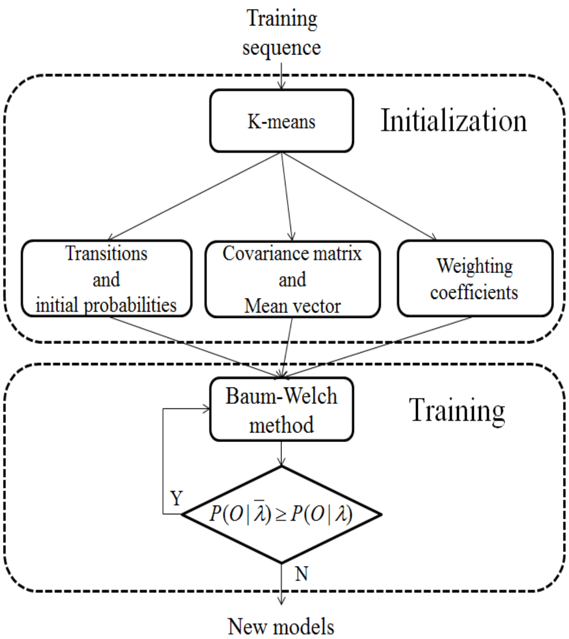


Fig. 6 The flow chart of HMMs model training part.

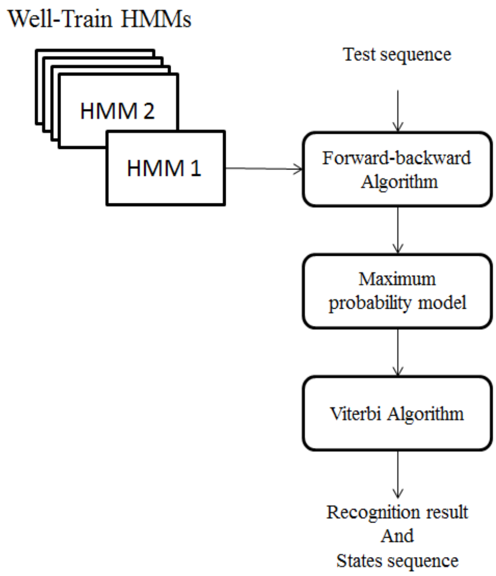


Fig. 7 The flow chart of recognition part.

Table 1 The average of the recognition rate for Singer-dependent.

Words	A	B
Work	96%	100%
No good	96%	100%
Don’t know	94%	98%
Center	96%	100%
Hello	98%	98%
Free	98%	100%
Patience	100%	98%
Photograph	100%	98%
Service	98%	100%
Know	100%	96%
Consult	98%	100%
Visit	96%	100%
Enter	96%	100%
Open	96%	98%
Sorry	94%	100%
Excuse me	96%	98%
Thanks	94%	98%
Welcome	96%	100%
Hear	96%	98%
Toilets	100%	100%
Mean	96.9%	99%

Table 2 The average of the recognition rate for Singer-independent.

Words	2D	3D
Work	90%	95.5%
No good	20%	94.5%
Don’t know	50%	98%
Center	65%	93%
Hello	15%	97%
Free	85%	98%
Patience	55%	98.5%
Photograph	40%	94%
Service	95%	98.5%
Know	75%	99.5%
Consult	45%	98%
Visit	75%	97%
Enter	45%	97.5%
Open	90%	98.5%
Sorry	90%	99.5%
Excuse me	70%	100%
Thanks	20%	94%
Welcome	95%	99%
Hear	20%	95%
Toilets	65%	97.5%
Mean	60.3%	97%

Table 3.1 The average of 5 signers’ confusion matrix for Singer-independent.

The average of 5 signers HMMs Confusion matrix		Actual output									
		Work	No good	Don’t know	Center	Hello	Free	Patience	Photograph	Service	Know
Expectation output	**Work**	191	0	0	0	0	0	0	0	0	0
	No good	0	189	1	0	0	0	0	0	0	6
	Don’t know	0	0	196	0	0	2	0	0	0	0
	Center	1	0	1	186	0	0	0	1	0	0
	Hello	0	5	0	0	194	0	0	0	0	0
	Free	0	0	2	0	0	196	0	0	0	0
	Patience	0	0	0	0	0	1	197	0	0	1
	Photograph	1	0	0	0	0	0	0	188	0	0
	Service	0	0	0	0	0	0	0	0	197	0
	Know	0	0	1	0	0	0	0	0	0	199
	Consult	0	0	0	0	0	0	0	0	0	0
	Visit	0	0	5	0	0	0	0	0	0	1
	Enter	0	0	0	3	0	0	0	0	0	0
	Open	0	0	0	0	0	0	0	0	3	0
	Sorry	0	0	0	0	0	0	0	1	0	0
	Excuse me	0	0	0	0	0	0	0	0	0	0
	Thanks	0	0	0	0	0	0	0	4	4	0
	Welcome	0	0	0	0	0	0	0	0	0	0
	Hear	0	0	1	0	0	2	0	0	0	5
	Toilets	0	0	0	0	0	0	0	0	0	0

Table 3.2 The average of 5 signers’ confusion matrix for Singer-independent.

The average of 5 signers HMMs Confusion matrix		Actual output										
		Consult	Visit	Enter	Open	Sorry	Excuse me	Thanks	Welcome	Hear	Toilets	Recognition rate
Expectation output	**Work**	1	1	0	0	2	0	2	3	0	0	95.5%
	No good	0	1	0	0	0	0	0	1	0	2	94.5%
	Don’t know	0	2	0	0	0	0	0	0	0	0	98%
	Center	0	0	7	0	1	0	0	3	0	0	93%
	Hello	1	0	0	0	0	0	0	0	0	0	97%
	Free	0	1	0	0	0	0	0	0	1	0	98%
	Patience	0	0	0	0	0	0	0	1	0	0	98.5%
	Photograph	5	0	1	0	0	0	5	0	0	0	94%
	Service	0	0	1	2	0	0	0	0	0	0	98.5%
	Know	0	0	0	0	0	0	0	0	0	0	99.5%
	Consult	196	0	0	0	0	0	0	0	0	4	98%
	Visit	0	194	0	0	0	0	0	0	0	0	97%
	Enter	0	0	195	0	2	0	0	0	0	0	97.5%
	Open	0	0	0	197	0	0	0	0	0	0	98.5%
	Sorry	0	0	0	0	199	0	0	0	0	0	99.5%
	Excuse me	0	0	0	0	0	200	0	0	0	0	100%
	Thanks	1	0	0	0	3	0	188	0	0	0	94%
	Welcome	0	0	1	0	0	0	1	198	0	0	99%
	Hear	0	1	0	0	1	0	0	0	190	0	95%
	Toilets	5	0	0	0	0	0	0	0	0	195	97.5%
Correct proportions		3885/4000										97.12%

4. CONCLUSIONS

In this paper, we use the depth information obtained from Kinect and combine the skeleton tracking system by OpenNI to capture the human skeleton coordinates. And we get the feature parameters through the simple algorithm for the system. We apply the HMMs to conduct the independent sign language recognition experiment. Since we use the feature parameters have low correlation with the space position to reduce the problem about the size of the signers. Also we solve the problem of the different postures. Therefore, we can increase the recognition rate.

We hope that we can join the feature information of the speed and the direction, etc. And we cannot be affected by the signer's habit or the speed of the sign language movement. Moreover, we hope that we can recognize the much fine movement to increase the number of words. In addition to using the HMMs as the recognizer, we can combine another algorithm. So that, we can make our system more complete and apply more widely.

REFERENCES

[1] Kelly D., McDonald J., Markham C. "Recognizing Spatiotemporal Gestures and Movement Epenthesis in Sign Language," *IMVIP '09, ISBN: 978-0-76953-769-2, IEEE Computer Society Washington, DC, USA, pp. 145-150,* 2009.

[2] William C. Stokoe, Jr., "Sign Language Structure:An Outline of the Visual Communication Systems of the American Deaf," *Journal of Deaf Studies and Deaf Education, v10 n1 p3-37Win 2005*.

[3] Koki Ariga, Shinji Sako, Tadashi Kitamura, "HMM-based Sign Recognition in Consideration of Motion Diversity,"*IUCS2010,pp. 258-261*, 2010.

[4] M. Mohandes, M. Deriche,U. Johar, S. Ilyas, "A signer-independent Arabic Sign Language recognition system using face detection, geometric features, and a Hidden Markov Model," *Computers and Electrical Engineering 38 (2012) 422–433*, 2012.

[5] Lang S., Block-Berlitz M., Rojas R., "Sign Language Recognition using Kinect," *Artificial Intelligence and Soft Computing (ICAISC), pp. 394-402*, 2012.

[6] Gaolin Fang, WenGao,:"A SRN/HMM System for Signer-independent Continuous Sign Language Recognition," *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR02), pp. 312-317*, 2002.

[7] Chunli Wang, Wen GAO, Shiguang Shan, "An Approach Based on Phonemes to Large Vocabulary Chinese Sign Language Recognition," *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR02), pp. 411-416*, 2002.

[8] Gaolin Fang, Wen Gao, Jiyong Ma, "Signer-Independent Sign Language Recognition Based on SOFM/HMM, "*Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 2001. Proceedings. IEEE ICCV Workshop on, pp. 90-90,* 2001.

[9] Rabiner L.R. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *IN: Proceedings of the IEEE, Vol. 77, No. 2, pp. 257-286,* 1989.

[10] Rahimi A. "An Erratum for 'A Tutorial on Hidden Markov Modelsand Selected Applications in Speech Recognition," *website of Ali Rahimiat MIT. MediaLaboratoryhttp://xenia.media.mit.edu/~rahimi/rabiner/rabinererrata/rabiner-errata.html,* 2000.

[11] Ching-Tang Hsieh, Ruei-Chi Chung, "Physical rehabilitation assistant system based on Kinect," *Proceedings of 2012 National Symposium on System Science and Engineering National Taiwan Ocean University, Keelung, pp. 336-339,* 2012.