2015 8th International Symposium on Computational Intelligence and Design

Sign Language Recognition System Based on Weighted Hidden Markov Model

Wenwen Yang, Jinxu Tao, Changfeng Xi, Zhongfu Ye
Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, Anhui, People's Republic of China
Email: {yww2013, xcfeng}@mail.ustc.edu.cn, {tjingx, yezf}@ustc.edu.cn

***Abstract*—Sign language recognition (SLR) plays an important role in communication between deaf and hearing society. However, the recognition result is still worse for signer independent recognition. The reason is that there exists large variation between the signs from different subjects. In this paper, weighted hidden markov model (HMM) is proposed to deal with the variation. Unlike traditional HMM, WHMM assigns each sign samples with different weights. For the sign sample with big variation, the sample weight is big accordingly. Furthermore, we utilize Kinect to produce robust sign features to improve recognition rate. Our system is evaluated on one Chinese sign language dataset of 156 isolated sign words. Experimental result shows our proposed method outperforms other methods with a high recognition rate of 94.74%.**

Keywords- sign language recognition; hidden markov model; weighted hidden markov model; kinect

I. INTRODUCTION

In recent years, more and more researchers focus on developing sign language recognition system to enhance communication between the deaf and hearing people. Generally, manual signs are composed of four components, namely, hand shape, position, orientation and movements, when these four components are well defined for each sign. For a certain sign, some of these four components are different from other signs, so these discriminative information can be used to build sign models. Based on these components of signs, many machine learning and statistical methods are utilized to train sign models. Among these methods, HMM [2] [3], DTW [9] [10] and CRF [4] [5] have achieved better performance.

Starner et al. utilized HMM training sign model with feature vector consisted of each hand's x and y position in American sign language (ASL) recognition system [2]. To handle hand fragmented regions, Yang and Sarkar proposed frag-HMM model to allow for fragmented observations, via an intermediate grouping process [3]. Sminchisescu et al. utilized CRF to recognize human motion, which outperformed HMM [4]. For incorporating hidden structures of gesture sequences, Wang et al. proposed hidden state conditional random field (HCRF), which outperformed CRF [5]. Zhang et al. utilized adaptive boosting (AdaBoost) strategy in HMM training procedure at the subwords classification level for SLR, which outperformed conventional HMM [6]. To deal with multi-model gesture recognition question, Wu and Chen proposed a novel Bayesian Co-Boosting framework [7]. Jiang et al. utilized a sparse coding method for Chinese sign recognition, which achieved high performance [8]. Antonio et al. combined probability-based Dynamic Time Warping (PDTW) and the Bag-of-Visual-and-Depth-Words (BoVDW) model for human gesture recognition in RGB-D, where PDTW was utilized to segment gesture and BoVDW was employed to fuse multiple features in gesture classification step [9].

Most existing approaches or systems train sign models with equal weights for each training sign samples. However, in practice, signs performed by different signers have significant variation, which aggravates the difficulty of recognition. Thus, it is important to design a robust and effective system to deal with this variation. In this paper, we aim at building a robust sign language recognition system. To deal with sign variation from different signers, weighted hidden markov model is proposed to train sign models, which is inspired by Zhang's method [6]. In the training procedure of weighted hidden markov model, sign instances are assigned with different weights. Furthermore, to depict signs well, the RGB-D sensor, Kinect, is employed as the input equipment of sign video, and a representative sign feature is extracted, composed of normalized 3D hand trajectory and hand shape descriptors.

The arrangement of this paper is as follow: Section 2 gives an overview of the system. Feature representation is described in section 3. Section 4 introduces weighted hidden markov model. Section 5 depicts the experimental results. And the last section makes the summary and discussion.

II. SYSTEM OVERVIEW

In our system, Kinect is employed as sign sequence acquiring equipment, where color stream, depth stream and skeleton stream can be obtained from Kinect SDK. For convenience, we assume there is only one signer standing in front of Kinect. Our recognition system contains two parts, offline training and online recognition.

Offline training learns each sign's WHMM parameters, where these WHMM parameters are used in online recognition stage. The detail learning process will be described in Section 4. In online recognition, forward algorithm is employed to compute the log likelihood of input sign under each sign model. The recognition result is the sign class, corresponding to maximal log likelihood. Fig. 1 shows overview of our system.

978-1-4673-9587-8/15 $31.00 © 2015 IEEE
DOI 10.1109/ISCID.2015.254

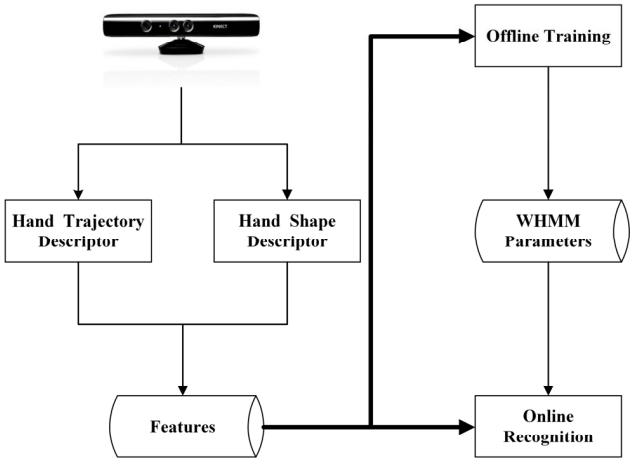

Fig. 1. System overview

III. FEATURE REPRESENTATION

As mentioned above, signs are composed of hand shape, position, orientation and movement. In this paper, hand shape and hand movement are utilized to depict signs, where hand shape describes sign local information when hand movement recodes global information.

A. Hand Trajectory

Hand movement, also called hand trajectory, recodes signs' trajectory, where most signs have different trajectories. Thus, hand trajectory is an effective feature to distinguish different signs. We can obtain 3D body skeleton information from Kinect SDK. In this paper, motion trajectory descriptor **T** is consisted of Right Hand, Right Wrist, Right Elbow, Left Hand, Left Wrist and Left Elbow.

$$\mathbf{T}=[\mathbf{P}_{RH},\mathbf{P}_{RW},\mathbf{P}_{RE},\mathbf{P}_{LH},\mathbf{P}_{LW},\mathbf{P}_{LE}] \tag{1}$$

where $\mathbf{P}_{(\cdot)}=[x,y,z]$ represents Cartesian coordinate of skeleton point and the subscript of **P** represents different skeletons.

As there has significant variation in the primitive hand trajectory due to several systemic and subjective factors, the primitive hand trajectory cannot be directly used in further steps. To build more robust hand trajectory, the scale and position normalizations are employed using (2).

$$\bar{\mathbf{P}}_{(\cdot)}=\frac{\mathbf{P}_{(\cdot)}-\mathbf{Pt}_{\text{Spine}}}{\left\|\mathbf{Pt}_{\text{Head}}-\mathbf{Pt}_{\text{Spine}}\right\|} \tag{2}$$

Then normalized hand trajectory $\bar{\mathbf{T}}$ can be obtained and sent to train sign models or do sign recognition.

$$\bar{\mathbf{T}}=[\bar{\mathbf{P}}_{RH},\bar{\mathbf{P}}_{RW},\bar{\mathbf{P}}_{RE},\bar{\mathbf{P}}_{LH},\bar{\mathbf{P}}_{LW},\bar{\mathbf{P}}_{LE}] \tag{3}$$

B. Hand Shape

As mentioned above, hand trajectory depicts the global information of signs. However, sometimes we cannot distinguish two signs only utilizing global information. The reason is that there may have similar trajectory for two different signs. Since hand shape describes the regional and edge information of hand image, we can utilize local information, hand shape, to enhance the ability to distinguish two signs of similar trajectory.

To depict shape and appearance of an object, several feature descriptors have been proposed. Considering the online performance of system, Histogram of Oriented Gradients (HOG) [11] is employed to depict hand shape in this paper. To obtain consistent HOG descriptors, the cropped hand image is resized to fixed size of 32×32. For each hand, the dimension of HOG is 252. Thus, the total hand shape descriptor **HS** is a 504D vector, concatenating HOG of two hands. Obviously, primitive **HS** is a super vector, with some dimension is not effective. In order to obtain more effective hand shape feature, PCA [12] is employed to reduce **HS** dimension and final hand shape descriptor $\bar{\mathbf{H}}\mathbf{S}$ has 40 dimensions.

In this paper, final sign feature **F** is obtained by concatenating hand trajectory feature and hand shape feature.

$$\mathbf{F}=[\bar{\mathbf{T}},\bar{\mathbf{H}}\mathbf{S}] \tag{4}$$

IV. WEIGHTED HIDDEN MARKOV MODEL

In speech recognition, HMM has achieved success. Since sign recognition task is similar to speech recognition, more and more researchers have employed HMM to train sign models, which also achieve better performance than other approaches. However, due to sign variation between different signers, basic HMM cannot model signs well. Thus, we propose weighted HMM.

A. Introduction of HMM

In traditional HMM, we often use $\lambda=(\mathbf{A},\mathbf{B},\boldsymbol{\pi})$ [1] to indicate the probabilistic parameter set. $\mathbf{A}=\{a_{ij}\}$ is a probability matrix of state transition, where a_{ij} means the probability from state i to j. $\mathbf{B}=\{b_j(O_t)\}$ is an observation probability matrix, with $b_j(O_t)=\sum_{m=1}^{M}c_{jm}G(O_t,\mu_{jm},\Sigma_{jm})$ denoting the probability that observation vector O_t is generated by state j. $\boldsymbol{\pi}=\{\pi_i\}$ is a vector of initial state probability, where π_i indicates the probability of considering state i as the first hidden state.

To train sign models, Baum-Welch algorithm [1] is employed to train HMM for each sign class. To recognition one test sign sample $O^{(n)}$, we often utilize (8).

$$\hat{y}_n=\arg\max_i P(O^{(n)}\mid\lambda_i) \tag{8}$$

where $O^{(n)}=(O_1^{(n)},O_2^{(n)},\cdots,O_T^{(n)})$ represents one unlabeled sign instance. $P(O^{(n)}\mid\lambda_i)$ measures the probability that $O^{(n)}$ is generated by model λ_i and forward algorithm [1] is employed to compute this probability usually.

B. Weighted HMM

Generally, conventional HMM consider each sign sample equally with the same weight. In practice, sign samples in the same class have a certain variation due to several systemic

and subjective factors. To deal with the variation, weighted HMM is proposed.

As conventional HMM, we also use $\lambda=(\mathbf{A},\mathbf{B},\boldsymbol{\pi})$ to indicate the probabilistic parameter set. The difference between conventional HMM and weighted HMM is that whether the sign instance weights are equal. The sign sample weights are based on the training error in weighted HMM, while it is constant in conventional HMM.

In weighted HMM, there exists three questions: a) parameter estimation, b) weight update, c) iteration condition.

Parameter Estimation. Given the weighted training sample set $\{O^{(n)},w^{(n)}\}$ with size N for one class, the log likelihood of the compete training samples can be defined as

$$\begin{aligned} l(\lambda) &= \sum_{n=1}^{N} w^{(n)} \log P\left(O^{(n)} \mid \lambda\right) \\ &= \sum_{n=1}^{N} w^{(n)} \log \sum_{Q^{(n)}} P\left(O^{(n)},\ Q^{(n)} \mid \lambda\right). \end{aligned} \tag{9}$$

where $O^{(n)}=\left(O_1^{(n)},O_2^{(n)},\cdots,O_{T_n}^{(n)}\right)$ represents training instance, $Q^{(n)}=\left(Q_1^{(n)},Q_2^{(n)},\cdots,Q_{T_n}^{(n)}\right)$ indicates corresponding hidden state, $w^{(n)}$ is the sign sample weight.

Our task is to find the optimal parameter λ to maximize $l(\lambda)$. Obviously, it is very difficult to optimize. To simplify this problem, the optimal state decoding probability is utilized to replace the sum item and the total log likelihood can be redefined as

$$l(\lambda)=\sum_{n=1}^{N} w^{(n)} \log P\left(O^{(n)},\ Q_*^{(n)} \mid \lambda\right). \tag{10}$$

where $Q_*^{(n)}$ represents the optimal decoding state sequence.

Before parameter estimation, we define forward and backward variables:

$$\alpha_t^{(n)}(i)=P\left(O_1^{(i)}O_2^{(i)}\cdots O_t^{(i)},q_t=s_i \mid \lambda\right) \tag{11}$$

$$\beta_t^{(n)}(i)=P\left(O_{t+1}^{(i)}O_{t+2}^{(i)}\cdots O_T^{(i)} \mid q_t=s_i,\lambda\right) \tag{12}$$

Using EM algorithm, the parameters of weighted HMM can be obtained in the following:

$$\bar{\pi}_i=\frac{\sum_{n=1}^{N} w^{(n)}\sum_{m=1}^{M}\gamma_1^{(n)}(j,m)}{\sum_{n=1}^{N} w^{(n)}\sum_{m=1}^{M}\sum_{j}\gamma_1^{(n)}(j,m)} \tag{13}$$

$$\bar{a}_{ij}=\frac{\sum_{n=1}^{N} w^{(n)}\sum_{t=1}^{T_n-1}\xi_t^{(n)}(i,j)}{\sum_{n=1}^{N} w^{(n)}\sum_{t=1}^{T_n-1}\sum_{j}\xi_t^{(n)}(i,j)} \tag{14}$$

$$\bar{c}_{jm}=\frac{\sum_{n=1}^{N} w^{(n)}\sum_{t=1}^{T_n}\gamma_t^{(n)}(j,m)}{\sum_{n=1}^{N} w^{(n)}\sum_{t=1}^{T_n}\sum_{m=1}^{M}\gamma_t^{(n)}(j,m)} \tag{15}$$

$$\bar{\mu}_{jm}=\frac{\sum_{n=1}^{N} w^{(n)}\sum_{t=1}^{T_n}\gamma_t^{(n)}(j,m)O_t^{(n)}}{\sum_{n=1}^{N} w^{(n)}\sum_{t=1}^{T_n}\gamma_t^{(n)}(j,m)} \tag{16}$$

$$\bar{\Sigma}_{jm}=\frac{\sum_{n=1}^{N} w^{(n)}\sum_{t=1}^{T_n}\gamma_t^{(n)}(j,m)\left(O_t^{(n)}-\mu_{jm}\right)\left(O_t^{(n)}-\mu_{jm}\right)^{\mathrm{T}}}{\sum_{n=1}^{N} w^{(n)}\sum_{t=1}^{T_n}\gamma_t^{(n)}(j,m)} \tag{17}$$

where,

$$\begin{aligned} \xi_t^{(n)}(i,j) &= P\left(q_t=s_i,q_{t+1}=s_j \mid O^{(n)},\lambda\right) \\ &= \frac{\alpha_t^{(n)}(i)a_{ij}b_j(O_{t+1}^{(n)})\beta_{t+1}^{(n)}(j)}{P(O^{(n)}\mid\lambda)} \end{aligned} \tag{18}$$

$$\gamma_t^{(n)}(j,m)=\frac{\alpha_t^{(n)}(j)\beta_{t+1}^{(n)}(i)}{\sum_{i}^{N_s}\alpha_t^{(n)}(j)\beta_{t+1}^{(n)}(i)}\cdot\frac{c_{jm}G\left(O_t^{(n)},\mu_{jm},\Sigma_{jm}\right)}{\sum_{m=1}^{M}c_{jm}G\left(O_t^{(n)},\mu_{jm},\Sigma_{jm}\right)} \tag{19}$$

Weight Update. In order to deal with the sign variation from different subjects, the sign sample weights are updated by training error. In each iteration, when sign models are obtained, training samples are tested using these sign models and we update training instance weights using (20).

$$\bar{w}^{(n)}=\begin{cases} w^{(n)} & \text{, if the sample is labeled correctly} \\ \dfrac{P\left(O^{(n)}\mid\lambda_c\right)}{P\left(O^{(n)}\mid\lambda_{true}\right)} & \text{, if the sample is labeled incorrectly} \end{cases} \tag{20}$$

If the training sample is labeled correctly, the sample weight is not changed. Otherwise, the weight is set as the ratio of probability under test label and probability under true label. Then updated sign sample weights and train samples will be sent to train weighted HMM in next iteration.

Iteration Condition. In order to stop the iteration of training weighted HMM, termination condition must be set. In this paper, the training error ε and iteration times t are employed as termination conditions. The iteration can be terminated when ε is smaller than threshold T_ε, or t is beyond the maximal iteration times T_i.

The weighted HMM training process is list in algorithm 1.

Algorithm 1. Train weighted HMM

Input: training set $\{O^{(n)},w^{(n)}\}$, training error threshold T_ε, maximal iteration times T_i, sign class number N_s

Initialize: $w^{(n)}=1$, $t=0$

step 1. train weighted HMM $\{\lambda_k\}$ using (13)-(17), $k=1,2,\cdots,N_s$.

step 2. label training samples using (8).

step 3. update training sample weights using (20).

step 4. calculate training error ε and update iterations $t=t+1$.

step 5. if $\varepsilon<T_\varepsilon$ or $t<T_i$, go to next step, otherwise, go back to step 1.

output: the sign model set $\{\lambda_k\}$.

V. EXPERIMENTAL RESULTS

In this paper, we test on a Chinese sign language dataset recorded by depth camera, Microsoft Kinect. The dataset contains 8892 sign samples over 156 isolated signs from 8 signers. Fig. 2 shows the screenshot of three sign samples: welcome, everyone and happy.

In order to achieve best performance, configurations are set as follow. The training error threshold T_ε is set as 0.02% and maximum iterations T_i is set 30. For each weighted HMM, the state number is set as 4 and gauss number is set as 1 for each state.

We first test feature effectiveness of trajectory, HOG and combined features via 3 fold cross validation. As shown in Table I, the combined feature, trajectory+HOG, achieves best performance, which demonstrates the effectiveness of our feature. And trajectory performs worse than HOG, since there are several signs with similar trajectory in the dataset.

In order to demonstrate the effectiveness of weighted HMM, we compare our method with three approaches: HMM, DTW, LC-KSVD. As shown in Table II, our proposed method outperforms others with a high accuracy of 94.74%.

Welcome Everyone Happy

Fig. 2. Screenshots of three sign samples

TABLE I. EFFECTIVENESS FOR DIFFERENT FEATURES

Feature	Only Trajectory	Only HOG	Trajectory + HOG
Accuracy	81.04%	89.10%	94.74%

TABLE II. COMPARISON OF ACCURACY FOR DIFFERENT METHODS

Method	DTW	HMM	LC-KSVD	Our Method
Accuracy	73.26%	92.56%	86.97%	94.74%

VI. CONCLUSION

In this paper, we utilize Kinect to produce one robust feature consisted of hand shape and hand trajectory. Then, weighted HMM is proposed to deal with the variation between signs from different signers. Test on Chinese sign language dataset, experimental result shows our method outperforms others.

REFERENCES

[1] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of IEEE, vol. 77, issue 2, Feb. 1989, pp. 257-289, doi:10.1109/5.18626.

[2] T. Starner, J. Weaver, and A. Pentlan, "Real-time American sign language recognition using desk and wearable computer based video," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol 20, issue 12, Dec. 1998, pp. 1371-1375, doi:10.1109/34.735811.

[3] R. Yang and S. Sarkar, "Gesture Recognition using Hidden Markov Models from Fragmented Observations," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 06), IEEE Press, Jun. 2006, pp. 766-773, doi:10.1109/CVPR.2006.126.

[4] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Conditional models for contextual human motion recognition," Computer Vision and Image Understanding, vol. 104, issue 2, Nov. 2006, pp. 1808-1815, doi: doi:10.1016/j.cviu.2006.07.014.

[5] S. B. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell, "Hidden Conditional Random Fields for Gesture Recognition," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006), IEEE Press, 2006, pp. 1521-1527, doi:10.1109/CVPR.2006. 132.

[6] L. Zhang, X. Chen, C. Wang, Y. Chen, and W. Gao, "Recognition of sign language subwords based on boosted hidden Markov models," Proc. 7th International Conference on Multimodal Interfaces (ICMI 05), ACM Press, 2005, pp. 282-287, doi: 10.1145/1088463.1088511.

[7] J. Wu and J. Cheng, "Bayesian co-boosting for multi-modal gesture recognition," The Journal of Machine Learning Research, vol. 15, issue 1, Jan. 2014, pp. 3013-3036.

[8] Y. Jiang, J. Tao, W. Ye, W. Wang, and Z. Ye, "An Isolated Sign Language Recognition System Using RGB-D Sensor with Sparse Coding," Proc. 17th IEEE International Conference on Computational Science and Engineering (CSE 2014), IEEE Press, Dec. 2014, pp. 21-26, doi:10.1109/CSE.2014.38.

[9] H.V. Antonio, A. B. Miguel, P. S. Xavier, P. L. Victor, E. Sergio, B. Xavier, P. Oriol, and A. Cecilio, "Probability-based Dynamic Time Warping and Bag-of-Visual-and-Depth-Words for Human Gesture Recognition in RGB-D," Pattern Recognition Letters, vol. 50, Dec. 2014, pp.112-121, doi:10.1016/j.patrec.2013.09.009.

[10] J. F. Lichtenauer, E. A. Hendriks, and M. J. T. Reinders, "Sign Language Recognition by Combining Statistical DTW and Independent Classification," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 30, issue 11, May 2008, pp. 2040-2046, doi:10.1109/TPAMI.2008.123.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005), IEEE Press, Jun. 2005, pp. 886-893, doi:10.1109/CVPR.2005.177.

[12] J. Yang, D. Zhang, A. F. Frangi, and J. Y. Yang, "Two-dimensional PCA: a new approach to appearance-based face representation and recognition," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 26, issue 1, Jan. 2004, pp. 131-137, doi:10.1109/TPAMI.2004.1261097.