SIGN LANGUAGE RECOGNITION BASED ON ADAPTIVE HMMS WITH DATA AUGMENTATION

Dan Guo[1]*, Wengang Zhou*[2]*, Meng Wang*[1]*, and Houqiang Li*[2]

[1]School of Computer and Information, Hefei University of Technology, Hefei, 230009, China
[2]EEIS Department, University of Science and Technology of China, Hefei, 230027, China
guodan@hfut.edu.cn, zhwg@ustc.edu.cn, eric.mengwang@gmail.com, lihq@ustc.edu.cn

ABSTRACT

Vision based sign language recognition (SLR) is a challenging task due to the complexity of signs and limited data collection. To improve the recognition precision, this paper proposes an adaptive GMM-based (Gaussian mixture model) HMMs (Hidden Markov Models) framework. We discover that inherent latent states in HMMs are not only related to the number of key gestures and body poses, but also related to the kinds of their translation relationships. We propose adaptive HMMs and obtain the hidden state number for each sign with affinity propagation clustering. Furthermore, to enrich the training dataset, we propose a data augmentation strategy by adding Gaussian random disturbances. Experiments on a vocabulary of 370 signs demonstrate the effectiveness of our proposed method over the comparison algorithms.

Index Terms— Sign Language Recognition, Hidden Markov Models, Gaussian Mixture Model, Data Augmentation

1. INTRODUCTION

Sign Language recognition (SLR) has attracted increasing research interests in computer vision and pattern recognition [1, 2, 3]. It is still a challenging task due to the complexity of signs and limited data collection [4, 5, 6, 7].

By limited data collection, many works were conducted on small datasets. For example, Kurakin *et al.* [8] proposed a realtime SLR system on a dataset with 12 American Sign Language (ASL) gestures. Dong *et al.* [9] obtained 90% recognition precision on 24 static ASL sign words. Sun *et al.* [10] achieved 85.5% accuracy in recognizing 73 ASL signs. Then two state-of-the-art works in large vocabulary SLR are released. One proposed by Ong *et al.* [11] achieved 74.1% accuracy on 982 signs and another proposed by Wang *et al.*[12] achieved 94% accuracy on 370 signs. However, most of these experiments are signer-dependent tests. The prior knowledge of signer that involved in recognition process leads to a high accuracy. In signer-independent experiments, Wang *et al.* [13] reported a application on 1,113 signs. Its accuracy of correct sign in the top 10 is 78%. But the dataset in [13] were marked whether the sign is one-handed or two-handed, and which the dominant hand is. It has used the prior knowledge of signs. In practice, it is not foreseeable that who the signer is, and whether the signer mark the label of hand gesture.

In addition, to get a powerful modeling ability, researchers adopt different models such as Dynamic Time Warping method (DTW) [14, 15], Curve Matching method [16], Hidden Markov Models (HMMs) [12, 17] and neural networks [18, 19, 20] have demonstrated promising results in automatic speech recognition. Salvador *et al.* [14] introduced a DTW (Dynamic Time Warping method) with a linear time and space complexity that can be used on gesture recognition. Lin *et al.* [16] proposed a curve matching method based on manifold analysis with trajectories of sign. Wang *et al.* [12] proposed Light-HMMs to select the key frames through low rank approximating and determine the number of hidden states by a Residual Sum of Squares (RSS) threshold. There are also more and more recurrent neural networks (RNNs) that are used to solve the SLR problem [19].

Generally speaking, in real application environment, it is difficult to get the prior knowledge of signs or signer. Besides,since expertise of sign language is lack and data collection with depth cue is rare, there is still lack of sufficient training samples [21, 22]. How to improve the recognition precision of SLR is still a challenging problem.

Considering the excellent capacity of the HMM model to capture sequential information, we choose the HMM based on Gaussian mixture model (GMM) as a baseline framework and denote it as GMM-HMM. Our work dedicates to two aspects: (1) A data augmentation strategy is proposed. We perturb the original samples by random noise variables subjected to Gaussian distributions. (2) In the SLR problem, the number of states in each sign's HMM model indicates key gesture and action changes. We propose a HMM-state adaptation strategy using affinity propagation clustering to adaptively determine appropriate state number for each sign.

The rest of this paper is organized as follows: we describe our method in Section 2. Section 3 gives the experimental results and analysis. In Section 4, we make the conclusion and discuss the future work.

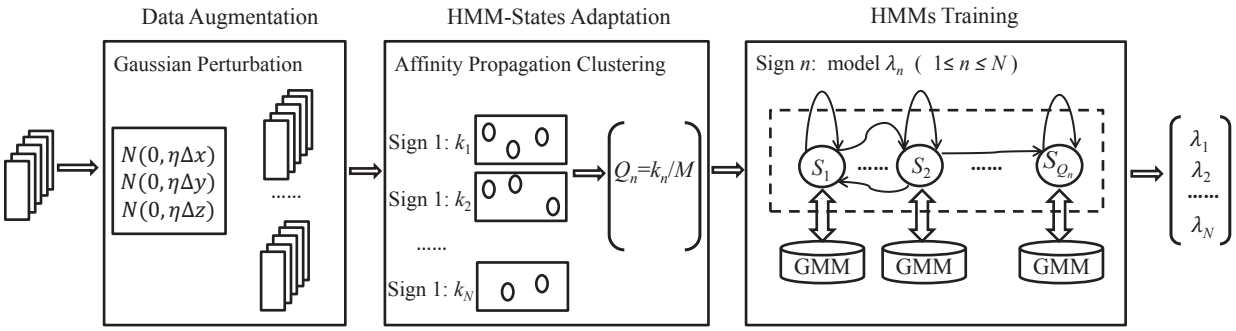


Fig. 1. The flowchart of our adaptive GMM-HMMs model.

2. OUR METHOD

This section gives a detailed description of our method. At first, we give the baseline framework of GMM-HMMs Model. Then we introduce the data augmentation technique into the SLR problem, as the detail is given in subsection 2.1. To further optimize the HMMs model, we describe the adaptive HMMs model in subsection 2.2.

The baseline framework of GMM-HMMs Model is as follow: given N signs' HMM models as $\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$, the most likely sign class λ^* of observation sequence O is given by Eq.(1). The recognition process is implemented by the famous Viterbi algorithm.

$$\lambda^* = \underset{\{\lambda_1,\lambda_2,\ldots,\lambda_N\}}{\operatorname{argmax}} P(\lambda|O). \tag{1}$$

2.1. Data augmentation

In our data preprocessing process, a data augmentation strategy is proposed. The data augmentation, (also known as data jittering or virtual sampling), aims at increasing information gain by generating additional samples [23]. In this paper, we use it to introduce appropriate noises and prevent overfitting trained by rare samples. We just explore the strategy on only 5 skeleton points (head, left elbow, right elbow, left hand, and right hand) to extract gesture position features. Here, we take (x, y, z) coordinates of a skeleton point as example, and the augmentation strategy is as follow:

Under all data samples (videos) of a sign, we look up the maximum x_{max} and the minimum x_{min} in all frames of video sequences and get the maximum range $\Delta x = x_{max} - x_{min}$. We suppose that a random variable X is subject to a Gaussian distribution $N(\mu_x, \sigma_x^2)$, where μ_x is expected value and σ_x^2 is variance. In our model, we have that $\mu_x = 0$ and $\sigma_x = \eta\Delta x$, where η is disturbance parameter. $X \sim N(0, (\eta\Delta x)^2)$ denotes the posed noise on x coordinate. Similarly, we get the $Y \sim N(0, (\eta\Delta y)^2)$ and $Z \sim N(0, (\eta\Delta z)^2)$ on y and z coordinates, respectinvely. An additional (x, y, z) coordinates of a skeleton point is generated as follows:

$$\begin{cases} x' = & x + N(0, (\eta\Delta x)^2) \\ y' = & y + N(0, (\eta\Delta y)^2) \\ z' = & z + N(0, (\eta\Delta z)^2) \end{cases} \tag{2}$$

With 5 skeleton points of a frame, we can generate an additional frame sample and then generate an additional video sample using Eq. 2 on each frame in a video sample. If we conduct Eq. 2 on the original dataset (videos) N times, we can augment the datasets by N times.

Skeleton Pair (SP) feature Extraction: After the above processing, we get a bigger dataset. Then for a frame sample, we extract mutual distances of 5 skeleton points as a 10-dimension SP feature [12, 24]. The feature has a good invariance to rotation, scaling, and translation [13]. To scale SP by different signers' sizes, each SP feature is normalized by its maximum element. Then we get each sign sample's SP observation sequence O for succeeding GMM-HMMs training.

2.2. Adaptive HMMs

Each sign has its own GMM-HMM model. Before GMM-HMMs training, we try to find appropriate latent states Q_n in each GMM-HMM model ($\lambda_n, 1 \le n \le N$). A good GMM-HMM model with discriminative states is learned by dividing data samples into reasonable clusters. To adaptively determine the state number $\{Q_n\}$ $(1 \le n \le N)$ for signs 1,2,..., N, we design a HMM-states adaptation strategy using affinity propagation (AP) clustering [25]. Under the data samples of sign n $(1 \le n \le N)$, we first construct a "net similarity" as deformed distance function between frame pairs. The net similarity computing is viewed to obtain the responsibility and availability log-probability ratios between frame f_i and candidate cluster "exemplar" f_k. Then we maximize the similarity function to find the best exemplars $\{f_k\}$ and automatically determine the number of different exemplars as the number of clusters k_n.

As shown in Fig.2, the gesture of sign "people" is more complicated than "I"'s. But its fitness of net similarity con-

verges faster and has not much change. The cluster convergence of sign "I" has quite large jumps in early iterations. Although with different distributions, they are assigned approximate state numbers. It indicates that inherent latent states in a HMM model are not only related to the number of key gestures and body poses, but also related to the kinds of their translation relationships. Fixed M-component GMM, the number of variant state Q_n is proportional to the number of clusters k_n.

Our HMM-states adaptation strategy has good stability and consistency. In our experiments, the number of clusters barely changed with the same data samples. In the 10 times test on 370×5 group training datasets, only 0.11% of the clusters changed and their difference was very small. Thus training data samples can be preprocessed to determine $\{Q_1, Q_2, \ldots, Q_N\}$. Our adaptive GMM-HMM model is given in Algorithm 1.

Algorithm 1 Adaptive GMM−HMMs (HMMs+AP)

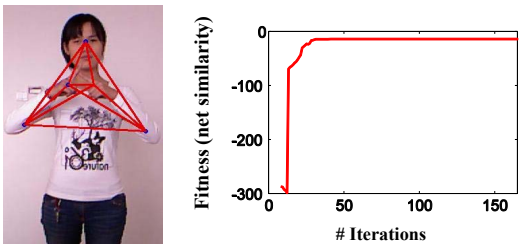
Input: N signs' training samples $\{Set\_O_n\}$ $(1 \leq n \leq N)$;
Output: N signs' HMM models $\{\lambda_1, \lambda_2, \cdot, \lambda_N\}$

1: **for** each sign n **do**
2: **for** $\forall$ observation sequence (video) $O \in Set\_O_n$ **do**
3: generate an additional O' by Eq.2;
4: **end for**// get an additional training samples $Set\_O'_n$;
5: extract SP feature set SP_n of $<Set\_O_n, Set\_O'_n>$;
6: compute the number of clusters k_n on SP_n by AP clustering;
7: $Q_n = k_n/M$;
8: learn the GMM-HMM model $\lambda_n = (A_n, B_n, \pi_n)$ [1] with training samples $\{<Set\_O_n, Set\_O'_n>\}$ and Q_n;
9: **end for**

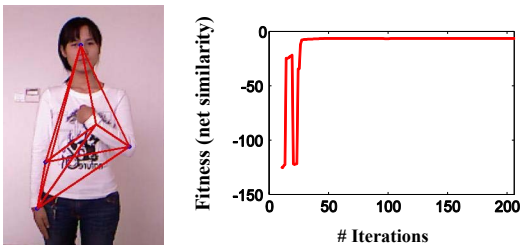
3. EXPERIMENTS

We conduct experiments on a Kinect dataset with large vocabulary, which contains 370 signs played by 5 signers with 5 repetitions [12]. The 5 signers contain both female and male. Their heights and gesture habits are completely different. We adopt leave-one-out cross-validation (LOO) to test DTW [14], traditional GMM-HMMs (HMMs), Light-HMMs [12], and our method (HMMs+AP(Q)). The size of dataset in each LOO experiment is shown in Table 1. We also use the HMMs+AP framework to get adaptive M with fixed Q, and denote it as HMMs+AP(M). A good solution for traditional HMMs is $Q = M = 3$ [12]. In the following experiments, we set HMMs+AP(Q) with $M = 3$ and set HMMs+AP(M) with $Q = 3$.

[1] A public HMM matlab package: `http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html`. Parameter Q_n and M are discussed in the paper, other general parameters A_n,B_n and π_n can be handle by this code package.



(a) Sign "people"



(b) Sign "I"

Fig. 2. Cluster convergence on SP features with net similarity computing. (a). Sign "people". (b) Sign "I".

Dataset	Signs	Signers	Repetitions	Samples
Training	370	4	5	7400
Testing	370	1	5	1850

Table 1. The dataset sizes in each LOO validation experiment.

3.1. Experiment with HMM-states adaptation

Experiment on test data is signer-independent test. Tables 2 and 3 show that the performance is less influenced by GMM's parameter M than HMM's parameter Q. Q indicates the latent changes of each sign and M simulates the data distribution model. Due to the chaos of real sign samples and the characteristic of Gaussian simulation, the influence of M is not very obvious in our problem. Thus we do not extend discussion on classical methods that determine the number of Gaussian components, such as Akaike information criterion (AIC)[26] and Bayesian information criterion (BIC) [27]. We

Methods	Test Validation	Training Validation	Time (ms/sign)
DTW	0.3158	/	277
HMMs	0.2874	0.5402	16
Light_HMMs	0.2827	0.5210	19
HMMs+AP(M)	0.2912	0.5570	43
HMMs+AP(Q)	**0.3354**	**0.7228**	51

Table 2. Performance comparison among different methods.

Methods	Top 1	Top 5	Top 10	Time (ms/sign)
DTW [14]	0.3159	0.5284	0.6245	277
HMMs	0.2869	0.5498	0.6583	16
Light_HMMs [12]	0.2827	0.5256	0.6322	19
HMMs+AP(M)	0.3022	0.5761	0.6662	43
HMMs+AP(Q)	**0.3354**	**0.5979**	**0.6941**	51

Table 3. The comparison of recognition precision in the top 1,top 5, and top 10.

Parameters	Top 1	Top 5	Top 10	Time (ms/sign)
(Q,1,0.01)	**0.3538**	0.6069	0.7063	47
(Q,1,0.02)	0.3520	0.6131	0.7056	55
(Q,1,0.03)	0.3514	0.6126	**0.7109**	55
(Q,1,0.04)	0.3482	0.6129	0.7080	61
(Q,1,0.05)	0.3521	**0.6134**	0.7103	65

Table 4. The comparison of HMMs+AP's recognition precision on different disturbance parameters.

focus on the adaptive Q. HMMs+AP(Q) is the best method whether on the training data or testing data. DTW is the most time consuming which searches the whole dataset to find the nearest sample's sign class. All HMMs-based methods just need to search N HMM models, where N is the number of sign words. HMMs+AP consumes more time than HMMs and Light-HMMs due to its model complexity with increasing state numbers.

3.2. Experiment with data augmentation

We test HMMs+AP on data augmentation and denote it as HMMs+AP(Q, num, η), where num is the number of times of data augmentation. Note that with too many times of augmentation, data quality should pull down by its own duplication and redundancy. And large η will introduce too much noise. In this paper, we just test doubled data samples by augmenting once ($num = 1$) with disturbance parameter values $\eta \sim$ (0.01, 0.02, 0.03, 0.04, 0.05). As the LOO validation revealed in Table 4, the strategy keeps good stability and consistency with different η. The recognition precision improves by 1.84% in top 1, 1.55% in top 5 and 1.68% in top10.

3.3. Summary results

As shown in Figs. 3 and 4, HMM-based methods have a good learning ability. All HMM-based methods surpass DTW. Even if the traditional HMMs has lower recognition precision than DTW in the top 1, but it surpasses DTW since top 3. Our method further learns the variant state numbers. Working with data augmentation, HMMs+AP has a good recognition precision and stability.

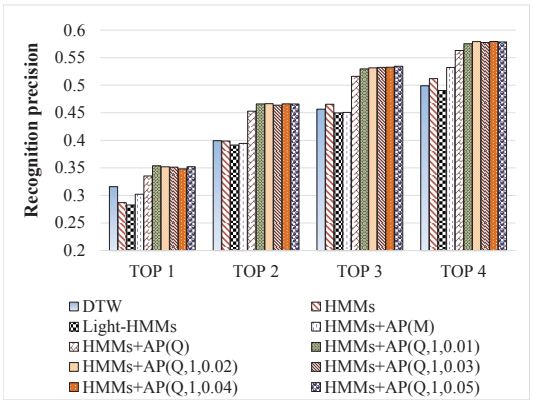


Fig. 3. Recognition precision comparison in the top-1∼4.

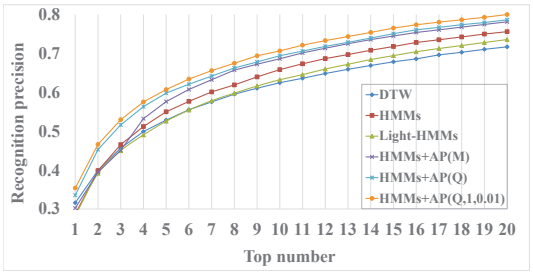


Fig. 4. Contrasting curves of recognition precision in the top-20.

4. CONCLUSIONS

An adaptive GMM-HMM model based on data augmentation with Gaussian disturbances is proposed. We verify it on the signer-independent test on the dataset with a large vocabulary but limited samples. Our method improves the recognition precision by 6.69% over traditional HMMs with only 10-dim skeleton pair feature. There maybe a big promotion potential on recognition precision by using multi-channel features, such as hand visual feature (HOG), trajectory, facial expression, and so on. We will explore more features and investigate features fusion in our further work.

5. ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (NSFC) under grant 61305062, Anhui Provincial Natural Science Fundation under contract No. 1508085MF109 and the Fundamental Research Funds for the Central Universities.

6. REFERENCES

[1] T. Pfister, J. Charles, and A. Zisserman, "Large-scale learning of sign language by watching tv," in *BMVC*, 2013, pp. 20.1–20.11.

[2] S. Sarkar, B. Loeding, R. Yang, S. Nayak, and A. Parashar, "Segmentation-robust representations, matching, and modeling for sign language," in *CVPR Workshops*, 2011, pp. 13–19.

[3] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff, "A unified framework for gesture recognition and spatiotemporal gesture segmentation," *TPAMI*, vol. 31, no. 9, pp. 1685–1699, 2009.

[4] H. Cheng, L. Yang, and Z. Liu, "A survey on 3d hand gesture recognition," *TCSVT*, 2015.

[5] J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using real-sense," in *ChinaSIP*, 2015, pp. 166–170.

[6] J. Zhang, W. Zhou, and H. Li, "A new system for chinese sign language recognition," in *ChinaSIP*, 2015, pp. 534–538.

[7] J. Zhang, W. Zhou, C. Xie, J. Pu, and H. Li, "Chinese sign language recognition with adaptive hmm," in *ICME*, 2016.

[8] A. Kurakin, Z. Zhang, and Z. Liu, "A real time system for dynamic hand gesture recognition with a depth sensor," in *EUSIPCO*, 2012, pp. 1975–1979.

[9] C. Dong, M. Leu, and Z. Yin, "American sign language alphabet recognition using microsoft kinect," in *CVPR Workshops*, 2015, pp. 44–52.

[10] C. Sun, T. Zhang, B. Bao, C. Xu, and T. Mei, "Discriminative exemplar coding for sign language recognition with kinect," *IEEE T Cybernetics*, vol. 43, no. 5, pp. 1418–1428, 2013.

[11] E. J. Ong, H. Cooper, N. Pugeault, and R. Bowden, "Sign language recognition using sequential pattern trees," in *CVPR*, 2012, pp. 2200–2207.

[12] H. Wang, X. Chai, Y. Zhou, and X. Chen, "Fast sign language recognition benefited from low rank approximation," in *FG*, 2015, pp. 1–6.

[13] H. Wang, A. Stefan, S. Moradi, V. Athitsos, C. Neidle, and F. Kamangar, "A system for large vocabulary sign search," in *ECCV*, 2010, pp. 342–353.

[14] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[15] S. Celebi, A.S. Aydin, T.T. Temiz, and T. Arici, "Gesture recognition using skeleton data with weighted dynamic time warping," in *VISAPP*, 2013, pp. 620–625.

[16] Y. Lin, X. Chai, Y. Zhou, and X. Chen, "Curve matching from the view of manifold for sign language recognition," in *Computer Vision-ACCV Workshops*, 2014, pp. 233–246.

[17] J. Pu, W. Zhou, J. Zhang, and H. Li, "Sign language recognition based on trajectory modeling with hmms," in *MMM*, 2016, pp. 686–697.

[18] K. Murakami and H. Taguchi, "Gesture recognition using recurrent neural networks," in *CHI*, 1991, pp. 237–242.

[19] M. Maraqa, F. Al-Zboun, M. Dhyabat, and R. A. Zitar, "Recognition of arabic sign language (arsl) using recurrent neural networks," *Journal of Intelligent Learning Systems and Applications*, vol. 4, pp. 41–52, 2012.

[20] J. Huang, W. Zhou, H. Li, and W. Li, "Sign language recognition using 3d convolutional neural networks," in *ICME*, 2015, pp. 1–6.

[21] Y. Sun, N. Kuwahara, and K. Morimoto, "Development of recognition system of japanese sign language using 3d image sensor," in *HCI*, 2013, pp. 286–290.

[22] A. C. Kak, "Purdue rvl-slll asl database for automatic recognition of american sign language," in *ICMI*, 2002, pp. 167–167.

[23] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[24] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *CVPR*, 2012, pp. 1290–1297.

[25] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[26] H. Bozdogan, "Akaike's information criterion and recent developments in information complexity," *Journal of mathematical psychology*, vol. 44, no. 1, pp. 62–91, 2000.

[27] K. P Burnham and D. R Anderson, "Multimodel inference understanding aic and bic in model selection," *Sociological methods & research*, vol. 33, no. 2, pp. 261–304, 2004.