MULTILAYER METHOD BASED ON MULTI-RESOLUTION FEATURE EXTRACTING AND MVC DIMENSION REDUCING METHOD FOR SIGN LANGUAGE RECOGNITION

CHEN-XI ZHANG, HONG-XUN YAO, FENG JIANG, DE-BIN ZHAO, XIAO-TING SUN

Department of Computer Science & Engineering, Harbin Institute of Technology, Harbin, China
E-MAIL: {cxzhang, yhx, fjiang, dbzhao}@vilab.hit.edu.cn, sunxiaoting@hit.edu.cn

Abstract:

Hidden Markov Model (HMM) has been successfully used in the Sign Language Recognition (SLR). However, due to large vocabulary of the sign language, traditional one-layer HMM method is becoming limited with the increasing number of training samples. It is tardy when recognizing which cannot meet the real time requirement. In this paper, we present a multi-resolution feature extracting method and a reducing dimension method of Maximum Variance Criterion (MVC), which has better performance in Sign Language Recognition system than traditional reducing dimension methods of PCA or ICA. Our multilayer sign language recognition system increases the recognition accuracy by 3.42%, as well as reduces the recognition time by 0.992 second in average, compared with traditional HMM based system.

Keywords:

Sign language recognition; HMM; multi-resolution analysis; multilayer architecture

1. Introduction

Sign language is a static expressing system that is composed of signs by using hand motion aided by face expressions. It is a kind of language communicated through gestures and visions. Sign language is mainly employed by deaf-mutes to communicate with each other. However, communication with normal people is a big handicap for them since the normal people do not understand sign languages. Thus, the research of sign language becomes a necessity. With the development of computer techniques, Artificial Intelligence has been developed in a broad way. Through natural human-computer interface, data gloves as the input equipment for human-computer communication has made the computer possible to recognize and synthesize sign language. Also with the help of voice recognition and synthesis technique, deaf-mutes are able to communicate with the normal smoothly. Recently, there have been strong efforts in developing Multi-functional Perception and natural interfaces between users and systems that are based on gesture recognition [1-5].

The main methods of Chinese Sign Language Recognition (CSLR) are based on HMM. The earliest research was the combination of artificial nerve networks and HMM that leads to ANN/HMM [6]. Later, DGMM/HMM [7] and SOFM/HMM [5, 8] were presented. These two kinds of approaches both have good recognition accuracy by estimating the result from observing the probability at each HMM state of the gesture words' sequences. However, the above CSLR research all focuses on a large vocabulary so they have to search in very broad space and different classes. Potential words' resemblance and recognition difficulties increase with the increasing word number. The more the words, the more the cost of the searching operation as well as the cost of the memory that finally leads to the reduction of the recognition speed and recognition accuracy.

Aiming at alleviating the above problems, we adopt an idea of a multilayer classifier, which is based on multi-resolution analysis. Multi-resolution recognition means recognizing through different dimensions. First reduce the dimensions in order to reduce the amount of data being processed and thus increase the recognition speed. In specific, we recognize the gesture sequence by three layers step by step according to different dimensions: lower dimensions, middle dimensions and full dimensions. The gesture sequence should be trained before the lower and middle dimension recognition. Then make statistics of the lower and middle dimensional sign languages' Easily Confused Vocabulary (ECV) set. The concrete implementation of ECV set is discussed in Section 3. We choose PCA, ICA and MVC approaches respectively, to reduce the dimensions of the gesture sequences and find out the optimal MVC approach for our multilayer CSLR system.

The remainder of this paper is organized as following. Section 2 reduces the dimension using PCA, ICA and MVC approach respectively and compares their contributions to

0-7803-9091-1/05/$20.00 ©2005 IEEE

CSLR recognition accuracy. Section 3 presents the algorithm of building ECV set after training the lower and middle gesture sequence in detail. The designment of our multilayer CSLR system is described in Section 4. Section 5 is the experimental results of the recognition accuracy and recognition time. The conclusion is given in the last section.

2. Dimension Reducing

Currently there are two major approaches to collect gesture data: the visual approach, and the instrumental approach. We choose the latter. American Virtual Technologies Company's CyberGlove with 18 sensors and three Polhemus FASTRAK 3-D position trackers are utilized as input devices. Position trackers collect the hand's movement trajectory, positions, the distance between each hand to the coordinate origin and the distance between two hands considering the neck's three-dimension orientation as the reference system. Finally, the computed fifty-one-dimension vectors got from the input equipment in every moment function as the input data. The range of each component is different and should be normalized to 0-1.

Here we discuss two traditional approaches and the new MVC approach respectively to reduce the 51-dimension data.

2.1. PCA approach

PCA is a classical unsupervised dimension reducing method searching "suitable" features in the data. The purpose of PCA is to denote higher-dimension data in lower-dimension sub space, to make lower-dimension representation best describe the original data in the sense of sum-of-squared-error criterion. In this Paper, because the data glove collects data with 51 sensors, each frame has a correlated 51-dimension vector. Every gesture sequence is composed of many frames. We regard frames of all the gesture sequences as the processing unit, whose dimension is going to be reduced. The implementation method is as following:

Definition: The form of a gesture sequence is $O = \{A_1, A_2, \cdots, A_i, \cdots, A_T\}$, where T represents the frames of a gesture sequence (T is different in different gesture sequences), A_i is the 51-dimension vector in the ith frame: $A_i = < a_1, a_2, a_3, \cdots, a_{51} >^T$. $O(j)$ denotes the jth gesture sequence, $O_j(A_i)$ denotes the ith frame's vector of the jth gesture sequence.

1). Merge all the frames of all the gesture sequences into a set, which becomes the whole sample space. Compute the data's expectation of each dimension $E[k]$, $k = 1, 2, \cdots, 51$, supposing that the probability is independently distributed among the frames of every gesture sequence and among the 51-dimension data.

$N =$ Overall number of the training gesture sequences

$M =$ Overall frames of the gesture sequences

$T(j) =$ Number of the jth frame, $j = 1, 2, \cdots, N$

$$M = \sum_{j=1}^{N} T(j), j = 1, 2, \cdots, N$$

$$E[k] = \frac{1}{sum} \sum_{j=1}^{N} \sum_{i=1}^{T(j)} O_j(A_i), i = 1, 2, \cdots, 51$$

2). Compute the covariance matrix $\Sigma_{51\times51}$ in the training space, in which Frame is the basic processing unit, $i = 1, 2, \cdots, M$.

$$A_i = A_i - E,\ \Sigma_{51\times51} = \frac{1}{M} \sum_{i=1}^{M} A_i \times A_i^T$$

3). According to the known $\Sigma_{51\times51}$, compute the eigenvalue λ_i and its correlated eigenvector e_i, $i = 1, 2, \cdots, M$. We have to get the transformed matrix $\mathbf{R}_{k\times51}$ if we want to reduce the gesture sequences' dimension. $\mathbf{R}_{k\times51}$ is the matrix that is composed of the first largest k eigenvalues' correlated eigenvectors.

4). Process the PCA dimension reducing method to all the gesture sequences.

$$A'_i = \mathbf{R}^T (A_i - E), i = 1, 2, \cdots, N$$

Finally, A'_i is the reduced k-dimension vector.

2.2. ICA approach

The principle of ICA is to search for the directions in which the data are independent in the eigenspace. ICA is actually trying to denote a group of random variables into a linear combination of statistically independent variables. ICA approach has minimized the statistical dependency of the analyzed signal's components, giving prominence of the original signal's essential structures. ICA approach we adopt in this paper uses the fast fixed-point algorithm[9].

ICA's prototype formula is $v_i = \mathbf{B} s_i$, where s_i is what we want. The computing formula is as following:

$$s_i = \mathbf{B}^T v_i$$

Here v_i denotes the whitened ith data vector. **B** is unknown. After independent statistic assumption of the original signals and preprocessing of v_i, we have the following relationship:

$$\mathrm{E}\{v_i v_i^T\} = \mathbf{B}\mathrm{E}\{s_i s_i^T\}\mathbf{B}^T = \mathbf{B}\mathbf{B}^T = \mathbf{I}$$

If **B** 's ith rank element is denoted as $b(i)$, then the ith independent component can be derived from $s_k(i) = w^T v_i$, where all of the **B** 's rank elements $w = b(i)$ make up of ICA's basis vector. w is computed through fast fixed-point algorithm in this paper. The iterative process is as following:

1). Initialization: set itineration variable $k = 1$, randomly initialize weight vector $w(k) = rand()$.

2). $w(k) = rand()$, standardization: $w = w / \| w \|$, $w_0 = 0$.

3). Iterative computation: converge (two directions of w) while $\|w(k) - w_0\| > \varepsilon \wedge \|w(k) + w_0\| > \varepsilon$, where ε is a constant smaller than 1. $w_0 = w(k)$.

$$w(k+1) = E\{v_k (w(k)^T w_k)^3\} - 3w(k)$$

4). Restandardization: $w, w = w / \| w \|$.

5). Circulate till the end.

Now we get k independent elements from the above algorithm and put them in $s_i = \mathbf{B}^T v_i$ to get dimensional reduced data.

2.3. MVC approach

When begin lower resolution recognition in the first round, directly select k-dimension data from 51-dimension data can reach the best performance. If we test every possible situation, the answer space would be C_{51}^{k}. If the time to train and test 5000 gesture sequences were 4 hours, it would cost 20 billion years to solve this problem! Here we present a new solution that would have better performance basing on data feature selection, Maximum Variance Criterion (MVC) approach. The difference between MVC and traditional PCA/ICA is that MVC is based on probability statistics rather than Matrix projection operations. MVC directly selects some dimensions that contribute most in gesture sequences recognizing, and it doesn't change the original data value which differs from PCA/ICA approach.

The principle of MVC is to regard all the frames from all the gesture sequences as the sample space, make statistics of the most fiercely fluctuating situation in every dimension, then choose the dimensions that frustrate most fiercely, where variance is employed to measure the frustrating extent. The feature selection process is as follows, where O, A_i, $O(j)$, $O_j(A_i)$ have the same meanings as the above definition in Section 2.1.

1). Compute the expectation $E[k]$ in every dimension, where $k = 1, 2, \ldots, 51$ supposing that the probability is independently distributed among every frames of the gesture sequences, and among the 51-dimension data.

$N =$ Number of the trained gesture sequences

$T(j) =$ The number of frames of the jth gesture sequences. $j = 1, 2, ..., N$

$$sum = \sum_{j=1}^{N} T(j), j = 1, 2, ..., N$$

$$E[k] = \frac{1}{sum} \sum_{j=1}^{N} \sum_{i=1}^{T(j)} O_j(A_i), i = 1, 2, ..., 51$$

2). Compute the variance in every dimension: $cov[k]$.

$$cov[k] = \frac{1}{sum} \sum_{j=1}^{N} \sum_{i=1}^{T(j)} (O_j(A_i) - E[k])^2, k = 1, 2, ..., 51$$

3). Return k dimensions from the biggest $cov[k]$ as the feature selecting result.

Here we exhibit the result of the gesture sequence after the dimension reducing process

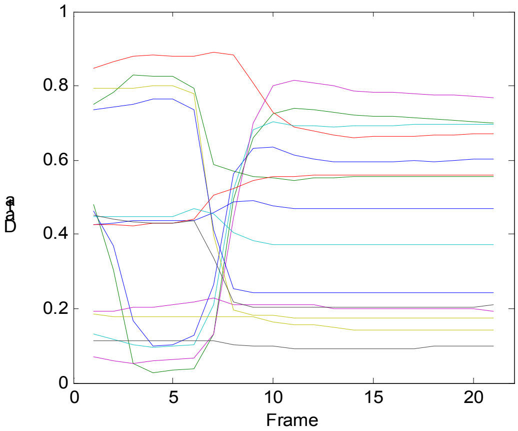


Figure 1. 15-dimension gesture sequence 'library' acquired from the sensors.

The horizontal axis of Figure 1 denotes the time sequence; the vertical axis denotes the normal data acquired from the sensors. It is obviously shown in Figure 1 that

MVC approach reserves the intensively fluctuated curves and filters lots of curves that do not fluctuated intensively.

Section 2.4 shows that these intensively fluctuated curves are more important to recognition accuracy.

2.4. Comparisons of the experimental results

We compared the above dimension reducing approaches' experimental results. The data are collected from 7 signers with each signer performing 4942 isolated words twice. The vocabularies are chosen from Chinese sign language dictionary. We select 4 from 7 signers as the registered signers. The rest are referred to as the unregistered signers. Each of the registered signers contributes two groups of data as training samples; the samples from the unregistered signers are referred to as the unregistered test set. CSLR system is a recognition system that is based on DGMM/HMM [7]. In this experiment, we reduce the gesture sequence to 15 dimensions and 35 dimensions respectively. The results are as follows:

Table 1. Comparison of recognition accuracy of 15 dimensions

Signer	Recognition accuracy of three methods in %		
	PCA	ICA	MVC
A	38.89	21.83	52.25
B	31.00	22.76	44.35
C	41.54	20.52	47.71
Average	37.14	21.70	48.10

Table 2. Comparison of recognition accuracy of 35 dimensions

Signer	Recognition accuracy of three methods in %		
	PCA	ICA	MVC
A	66.45	59.45	76.51
B	59.67	46.99	77.78
C	65.84	49.96	79.34
Average	63.99	52.13	77.88

Table 1 and Table 2 present the recognition results being dimension reduced to 15 dimensions and 35 dimensions from full dimensions by adopting these three dimension reducing methods, where A,B,C represent 3 unregistered signers respectively.

It is obviously shown in Table 1 and Table 2 that the recognition accuracy of Maximum Variance Criterion approach is much better than PCA and ICA. Moreover, MVC apparently surpasses the other two in the running time, because it selects from the input gesture sequence directly rather than does matrix multiply operations.

3. Building Easily Confused Vocabulary (ECV) Set

While doing multilayer recognition, we first recognize with lower resolution in lower dimensional data after reducing dimensions. Features become less after reducing the dimensions, which leads to lower recognition accuracy instead. Suppose the recognizing result of a certain word A is O_i, it does not mean that the word is sure to be O_i, it can be any other easily confused words. We build every word's easily confused word sets after testing lower dimensional data at first, then build every word's ECV set accordingly. Therefore, we need go on to recognize O_i's easily confused words in its ECV set with higher resolution in higher dimensional data.

The ECV set building algorithm is described as following:

1). $N=$ Number of the gesture sequences.

2). Build an ECV $Table(i)$ for each gesture sequence O_i, initialize $Table(i)$ as Null, where $i = 1, 2, \dots, N$.

3). If the gesture sequence O_i is mistaken as gesture sequence O_j, put O_i in O_j's ECV set.

4). If there isn't any gesture sequence mistaken as gesture sequence O_j, O_j's ECV set is null; if O_j is not null, put O_j itself in the ECV set.

5). Return ECV $Table(i), i = 1, 2, \dots, N$.

4. Multilayer architecture in sign language recognition

The multilayer architecture with three-stage hierarchy is presented as follows. Different from the usual way of recognizing the gesture sequence with full dimensions directly, we recognize the gesture sequence in three steps. The first step locates the observation sequence with the reduced 15 dimensions in the overall searching space. The second step, which is based on the 15-dimension ECV set that has been discussed in detail in the above section, recognizes the gesture sequences utilizing the reduced 35 dimensions in a smaller searching space. The third step, which is similar to the second step, recognizes the gesture sequence utilizing the full 51 dimensions in the 35-dimension ECV set in the even smaller searching space. Because all the search work is restricted in a single confusion set, less time is needed.

How to select the number of the lower and middle dimensions? In this paper, we reduce the collected

51-dimension data from 45 to 10 dimensions, with every 5 dimensions as interval. As shown in Figure 2, the horizontal axis represents the change of dimensions; the vertical axis represents their recognition accuracy (%). We can see that, from the curve between dimension 10 and dimension 25, the change of dimension 15 is the biggest, thus 15 is chosen as low dimensions. The curve is becoming smooth from dimension 25 to 45, thus the mean value 35 is chosen as middle dimensions.

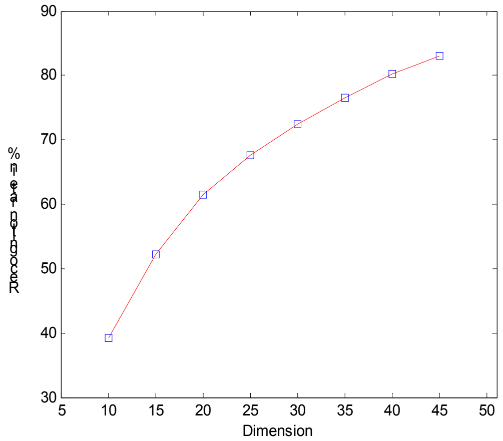


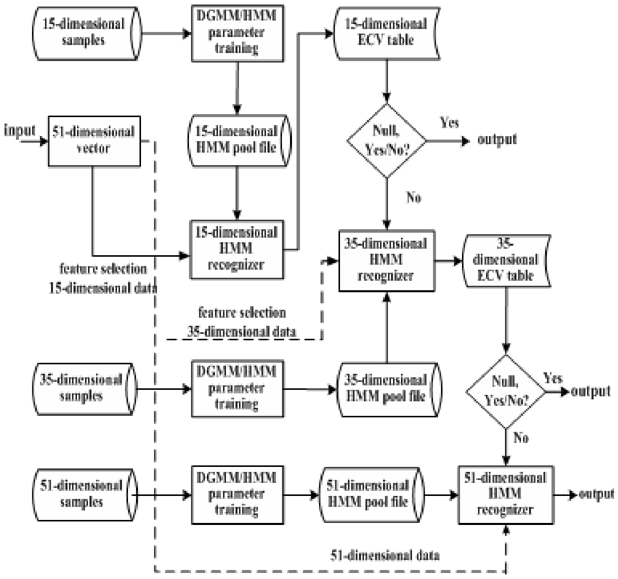
Figure 2. Recognition accuracy changing with dimensions

The gesture sequences' multiplayer recognition procedures are divided into three steps: training, building ECV set and multiplayer recognizing.

Training: Build a vocabulary set as the index for each gesture sequence. Train a HMM and its correlated DGMM in every state, regarding the single gesture sequence as the processing unit. DGMM is used for estimating each state's emitted probabilities for each frame of each gesture sequence under its HMM. The training formula for DGMM refers to reference[7]. Build three HMMs and DGMMs for every gesture sequence of the lower, middle and full dimensional words respectively.

Building ECV set: Test the lower and middle dimensional gesture sequences under a large number of testing data. Build the ECV set using the algorithm that has been discussed in section 3 after retrieving the test results.

Recognizing: The above two procedures are the preparation of CSLR. The system continues to recognize after those two procedures. The recognizing procedure is also the process to test and evaluate the system, whose detailed working flow is shown in Figure 3.



15-dimensional samples
DGMM/HMM parameter training
15-dimensional ECV table
input
51-dimensional vector
15-dimensional HMM pool file
Null, Yes/No?
Yes
output
No
15-dimensional HMM recognizer
35-dimensional HMM recognizer
35-dimensional ECV table
feature selection 15-dimensional data
feature selection 35-dimensional data
35-dimensional samples
DGMM/HMM parameter training
35-dimensional HMM pool file
Null, Yes/No?
Yes
output
No
51-dimensional samples
DGMM/HMM parameter training
51-dimensional HMM pool file
51-dimensional HMM recognizer
output
51-dimensional data

Figure 3. Multilayer-architecture classifier

The working flow shown in Figure 3 is divided into 7 steps:

1). Get a 15-dimension vector after feature selecting from the gesture sequence's 51-dimension sequence.

2). The vector being feature selected is put in the 15-dimension recognizer. Find the biggest $P(O \mid \lambda)$ in all the 15-dimension HMM pools

3). Find the gesture sequence's correlated 15-dimension data's ECV set of the biggest $P(O \mid \lambda)$ in step 2. Check in this word's ECV set whether it has its easily confused word. If not, end this step and this gesture sequence is the final output. Or else, go to step 4.

4). Get a 35-dimension vector after feature selecting from the gesture sequence's 51-dimension sequence.

5). The vector being feature selected is put in the 35-dimension recognizer. Find the biggest $P(O \mid \lambda)$ in all the 35-dimension HMM pools.

6). Find the gesture sequence's correlated 35-dimension data's ECV set of the biggest $P(O \mid \lambda)$ in step 5. Check in this word's ECV set whether it has its easily confused word. If not, end this step and this gesture sequence is the final output. Or else, go to step 7.

7). Put the input 51-dimension vector in the 51 HMM recognizer. Find the biggest $P(O \mid \lambda)$ of the easily confused words found in step 6. Output the biggest $P(O \mid \lambda)$'s correlated gesture sequence as the final recognition result.

5. Experimental result

The experiment is to test the recognition performances on large vocabulary signer-independent CSLR with HMM and multilayer architecture approach respectively. The trained data and tested data are the same as described in Section 2.4.

Table 3. Result comparison between traditional HMM and multilayer recognition method in recognition accuracy and speed

signer	Recognition accuracy		Recognition speed	
	HMM	Multilayer	HMM	Multilayer
A	89.48	90.46	2.369	1.413
B	86.26	92.37	2.366	1.302
C	90.45	93.61	2.358	1.402
Average	88.73	92.15	2.364	1.372

Table 3 reports the test results of HMMs and multilayer architecture approach, where HMMs have 3 states and 5 mixture components. 88.73% and 92.15% of the mean recognition accuracy are respectively observed. The multilayer recognition system increases the recognition accuracy by 3.42%, compared with traditional HMM based system. An exciting performance of the processing time can be seen in the table, the average recognition time has been reduced by 0.992 second in average.

Why does the multi-resolution recognition system both increase the recognition speed and recognition accuracy? Firstly, during the lower dimensional recognition, the dimensions that contribute little to the recognition (equal to noises) are filtrated in the process of feature selecting of the gesture sequence. Therefore, the recognition accuracy is increased. In addition, the data processed in this system approximately equals to half of the traditional HMM recognition method from the view of data volume, the recognition speed is increased accordingly.

6. Conclusions

This Paper presents a multilayer sign language recognition method based on multi-resolution consideration and offers an effective dimension reducing method: MVC approach. This approach gets better performances than traditional PCA/ICA in CSLR application through experiments. Multilayer architecture in CSLR increases the average recognition speed by 41.96% and the recognition accuracy is 3.42% higher than the HMM-based recognition method.

References

[1] S. S. Fels and G. Hinton, “Glove Talk: A neural network interface between a DataGlove and a speech synthesizaer”, IEEE Transactions on Neural Networks, 1993, Vol. 4, pp.2-8.

[2] M. W. Kadous, “Machine recognition of Auslan signs using PowerGlove: Towards large-lexicon recognition of sign language”, proceeding of workshop on the Integration of Gesture in Language and Speech, Wilmington, DE, 1996, pp.165-174.

[3] C. Vogler, D. Metaxas, “Toward Scalability in ASL Recognition: Breaking Down Signs into Phonemes”, In Proceedings of Gesture Workshop, Gif-sur-Yvette, France, 1999, pp. 400-404.

[4] R.H. Liang, M. Ouhyoung, “A Real-time Continuous Gesture Recognition System for Sign Language”, In Proceeding of the Third International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998, pp. 558-565.

[5] G.L. Fang, W. Gao, “A SOFM/HMM System for Person-Independent Isolated Sign Language Recognition”, INTERACT2001 Eight IFIP TC.13 Conference on Human-Computer Interaction, Tokyo, Japan, 2001, pp.731-732.

[6] Wu Jiangqin and Gao Wen. “Sign Language Recognition Method on ANN/HMM”. Computer science and application.No.9, pp 1-5.1999.

[7] Wu Jiang-Qin,Gao Wen. “A Hierarchical DGMM Recognizer for CSLR” . Journal of Software. Vol.11.No.11, pp 552-551.2000

[8] Gaolin Fang, Wen Gao, Jiyong Ma, “Signer-Independent Sign Language Recognition Based on SOFM/HMM”, IEEE ICCV Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems (RATFG-RTS 2001), Vancouver, Canada, 2001: 90-95

[9] A Hyvarinen. “A family of fixed-point algorithms for independent component analysis”. In Proc. IEEE Int .Conf. on Acoustics, Speech and Signal Processing (ICASSP’97), Munich, Germany, IEEE, 1997. 3917-3920