Based on HMM and SVM Multilayer Architecture Classifier for Chinese Sign Language Recognition with Large Vocabulary

Jianjun Ye, Hongxun Yao, Feng Jiang
Department of Computer Science, Harbin Institude of Technology, Harbin, China
jjye@vilab.hit.edu.cn

Abstract

This paper has put forward a new architecture classifier method for Chinese sign language recognition (CSLR) to improve the performance of recognition. It is a signer-independent method, to recognize Chinese Sign Language with large vocabulary using multilayer architecture classifier and making use of the advantages both of HMM(Hidden Markov Model) and SVM(Support Vector Machines). Because HMM is good at dealing with sequential inputs, while SVM shows superior performance in classifying with good generalization properties especially for limited samples. Therefore, they can be combined to yield a better and effective multilayer architecture classifier. We apply SVMs to resolve the uncertainty of the remaining which are in confusable sets after the first-stage HMM-based recognizer. And the confusable sets would be updated dynamically according to the results of a recognition performance to optimize the discernment performance next time. Experimental results proved that it is an effective method for CSLR with large vocabulary

Keywords Sign language recognition, HMM, SVM, multilayer architecture classifier

1. Introduction

Sign language is composed of a number of basic gestures. The deaf people use kinds of combinations of hand gestures and hand motion for their speechless communication. The aim of sign language recognition is to offer an accurate and effective mechanism to convert sign language into text or speech.

Recently, attention has been paid to the research on sign language recognition with the development of multi-modal perception, human-machine interface and virtual reality. Some universities, companies and laboratories are embarking on research on algorithms and system implementation for sign language recognition [1-3]. But most researchers focus on small or medium vocabulary SLR in the signer-dependent field. People of our laboratory has been working on large vocabulary signer-independent CSLR and got considerable effects [4][5].

In this paper, we use HMM and SVM classification paradigms and develop a hybrid system. We get confusable sets from dynamic results of confusable words, and HMM and SVM are combined for two-stage recognizing on confusable sets. HMM are used to recognize signs in the usual way, and these signs are post-processed to identify regions of sign confusability in which the first-stage HMMs were unable to distinguish between competing word hypotheses. The goal of this work is to apply SVMs to solve the uncertainty remaining after the first-stage HMM-based recognizer.

This structure of this paper is organized as followings. First, we propose the application of SVM in CSLR. Second, the conception of confusable set is presented. And then we introduces the Multilayer architecture. The experimental results and their comparisons are shown then and at last the conclusion is given.

2. SVM classifier in CSLR

2.1 Support Vector Machine(SVM)

Support Vector Machine (SVM) is one of the latest and most successful statistical pattern classifiers that utilizes a kernel technique [6]. the basic form of SVM classifier which classifies an input vector $x \in R^n$ is expressed as

$$f(x) = \sum_{i=1}^{l} a_i y_i \phi(x_i) \cdot \phi(x) + b$$
$$= \sum_{i=1}^{l} a_i y_i K(x, x_i) + b \qquad (1)$$

where ϕ is a nonlinear mapping function $\phi(x) : R^n \to R^{n'}$, ($n \ll n'$), the operator " · " denotes the inner product operator, x_i ; y_i and a_i are the i-th training sample, its class label, and its Lagrange multiplier, respectively, K is a kernel function, and b b is a bias.

To employ SVM in CSLR, their simple formulation as binary classifiers with fixed-dimension vectors will have to be overcome . For in CSLR, we would like to classify a variable length sequence of fixed dimension patterns. These raw observation sequences can be expected neither to have fixed dimension nor to belong to one of only two classes. To apply the SVM which is basically formulated as a two-class classifier to themulti-class problem, we use a popular algorithm called Max Wins algorithm: each 1-v-1 classifier casts one vote for its preferred class, and the final result is the class with the most votes. And to deal with the variable length sequenes directly. we also proposed a approach that is a direct extension of the original SVM for variable length sequences in following section.

2.2 Dynamic Time-Warping Kernel (DTWK)

This section describes how the time aliment operation can be embedded into the SVM's kernel. The technique combines dynamic time warping (DTW) and support vector machines (SVMs) by establishing a new SVM kernel. We call this kernel Dynamic Time-Warping Kernel (DTWK).

This kernel approach has a main advantage over common HMM techniques. It does not assume a model for the generative class conditional densities. Instead, it directly addresses the problem of discrimination by creating class boundaries and thus is less sensitive to modeling assumptions. By incorporating DTW in the kernel function, general classification problems with variable-sized sequential data can be handled. In this respect the proposed method can be straightforwardly applied to all classification problems, where DTW gives a reasonable distance measure on sign recognition processing.

In DTW a distance D(X, Y) from two vector sequences $X = (X_1, X_2, \ldots, X_{Nx})$, $Y = (Y_1, Y_2, \ldots, Y_{N_y})$ is determined. Given a so-called warping path $\varphi = (\varphi(1), \ldots, \varphi(N))$ with $\varphi(n) = (\varphi_x(n), \varphi_y(n)) \in \{1, ..., N_x\} \times \{1, ..., N_y\}$ and a local distance measure d , $d(X_i, Y_j) = \|X_i - Y_j\|^2$, we define the alignment distance D_φ as the mean distance along a particular alignment path φ

$$D_\varphi(X, Y) = \frac{1}{N} \sum_{n=1}^{N} d(X_{\varphi_x(n)}, Y_{\varphi_y(n)}) \qquad (2)$$

The DTW distance (or Viterbi distance) $D(X, Y)$ is defined as the alignment distance along the Viterbi path φ^*

$$D(X, Y) = D_{\varphi^*}(X, Y) = \min_{\varphi} \{D_\varphi(X, Y)\} \qquad (3)$$

It is convenient to model φ as a sequence of local transitions. Usual dynamic programming and beam search strategies are applied to reduce the computational complexity when minimizing.the DTW technique itself in combination with a minimum distance classifier, as well as the incorporation of statistical knowledge to this concept. have been successfully applied to sign language recognition [7].

As indicated above, when dealing with sequential sign date we can not simply employ the basic SVM framework given above. And we find an important property that the sign vectors sequences X, Y appear only in form of kernel evaluations. Thus our objective, when adopting SVMs to sequential sign language data, can be to state a kernel definition suitable to the particular properties of the sequential data. An obvious modification is to replace the squared Euclidean distance $\|X - Y\|^2$ in Gaussian kernel with the equivalent when dealing with temporally distorted, sequential signals—the DTW distance $D(X, Y)$ described above. So we apply this modification and define the new Gaussian kernel for sequential data by

$$K(X, Y) = \exp\{-\lambda D(X, Y)\} \qquad (4)$$

We can use the Dynamic Time-Warping Kernel(DTWK) to the discriminant function of SVM for sequential pattern. And the SVM discriminant function for time sequence has the same for the form with the original SVM except for the difference in kernels. So same training algorithms for the original SVM can be used to solve the problem.

3. Confusable sets used for multilayer architecture

How to combine HMM and SVM effectively, take the most of advantage of the relative strengths of these two classification paradigms to improves the recognition performance of the system, is the main problem we should resolve. And the key is how to build confusable sets effectively.

When the number of signs becomes larger, the potential similarity between signs becomes more, and it is the more difficultly to distinguish them. If some signs are very similar in some common features, the whole of signs will also be similar and it will be difficultly to distinguish them. Especially in the HMMs training, each word model is estimated separately using the corresponding labeled training observation sequences without considering the confused data. So the discrimination becomes poorer. We call these words that are confusable set. And also signer independence increases gesture confusability by increasing the variability inherent in each gesture unit, because of bad generalization of HMM models and individual features of sign words, though not to the same degree [8].

As to a sign, because of the different confusion degree with others caused by large vocabulary and independent signer, the number of other signs difficulty to be distinguished from this sign is also different. So we can construct confusable sets with different size according to the different confusion degree of signs to be used for two-stage recognition, which avoid the serious increasing of recognition time as the result of two-stage recognition at a certain extent and also improves the recognition rate effectively.

For the results of recognition on sample data can show the confusion degree of a sign directly, we can focus on modeling the most frequent errors found in training. Doumpiotis's work [9] has verified that training set errors found in the same way are good predictors of errors that will be encountered in unseen data. We can test the recognition performances on large vocabulary using training data with HMM, and then according to the results we synthesize and classify the signs confused with others to construct the initial confusable sets for formal two-stage recognition with HMM and SVM. At the same time, the confusable sets can be updated dynamically according to the results after a recognition performance. To keep the recognition rate, we also can restrict the size of confusable sets. With the dynamitic updating by testing, the confusable sets perform better classification on confusable signs and also the recognition performance of the system can be optimized progressively.

.

4. Multilayer architecture in sign language recognition

As the most popular and effective method in pattern recognition, HMM and SVM are intrinsically related with each other,. HMM have the advantage of being able to handle dynamic data with certain assumptions about stationarity, but when the samples of a sign is not sufficient. and origin from different signer, the HMM doesn't ferformance very well. While SVM can generalize well and improving discrimination of these data effecitively. Taking advantage of the good qualities of these two classification paradigms we have developed a hybrid SVM/HMM system.

A Multilayer architecture with two-stage hierarchy is presented as follows. As the usual way of recognizing the gesture sequence directly, the first stage is accomplished by HMM. If the observation sequence is not located at a confusable set, we consider it as the final recognizing result. Else we goto the second stage, which is accomplished by SVM on the confusable set. In this stage we finally recognize the gesture sequence from the confusable signs in the confusable set. In the training process, we must build an HMM λ_v for each word v in the vocabulary as well as the confusable sets and corresponding SVM models before recognition.

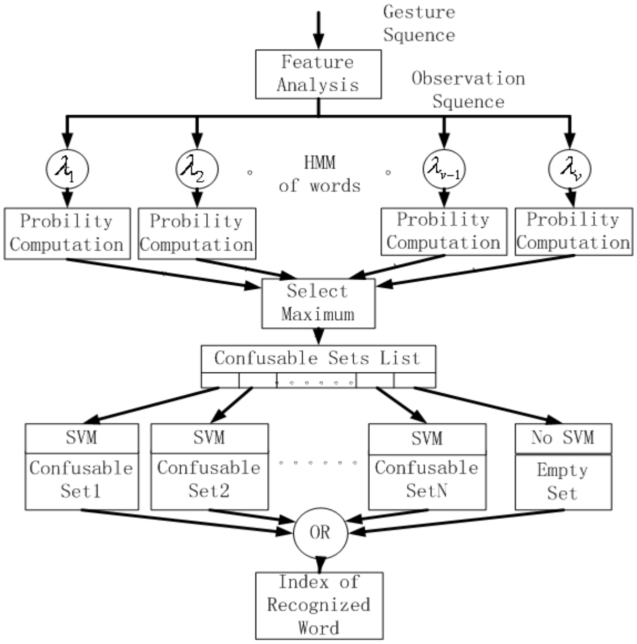

Figure １． Multilayer architecture classifier

5. Experimental result

We use two Cyber Gloves and three space-position trackers as input devices. Two trackers are positioned on the wrist of each hand and another is fixed at back. The Cyber Gloves collect the variation information of hand shapes with the 18-dimensional data at each hand, and the position trackers collect the variation information of orientation, position, movement trajectory.

The experiment was carried on a large vocabulary with 4942 signs. Experimental data consist of 59304 samples over 4942 signs from six signers with each performing signs twice. The words of vocabulary are chosen from sign language dictionary of China. One group data form the six signers are referred to as the test set and the other 11 group data are used as the training samples. The experiment is to test the recognition performances on large vocabulary signer-independent CSLR respectively with HMM and multilayer architecture**.**

Table 1**. The comparison of different results**

Signer	Recognition rate in %		Recognition speed in second	
	HMM	Multilayer architecture	HMM	Multilayer architecture
A	83.48	89.01	2.369	2.386
B	80.26	88.18	2.366	2.382
C	85.88	91.01	2.285	2.286
Average	83.21	89.40	2.340	2.351

Table 1 reports respectively test results of single HMMs and the multilayer architecture based on HMM and SVM, where HMMs have 3 states and 5 mixture components. 83.21% and 89.40% of mean recognition rates are respectively observed and the average recognition speed has not decreased distinctly under the architecture. From the experiments above, we know that HMM/SVM has better performance than HMM. For the multilayer architecture take advantage of the good qualities of these two classification paradigms and the effective using of confusable sets also optimize the discernment performance the system. Therefore, the multilayer architecture is more suitable for signer-independent CSLR, and it has better performance than single HMM on large vocabulary signer-independent CSLR.

6. Conclusion

This paper presents the multilayer architecture in Sign Language Recognition for the signer-independent CSLR, where classical HMM and SVM is combined within an initiative scheme. In the two-stage hierarchy, we define the confusable sets and build them in the vocabulary space to combined HMM and SVM effectively. The experiments show that the multilayer architecture in Sign Language Recognition increase the recognition accuracy 6.19% (from 83.21% to 89.40%) than single HMM-based recognition method. We are in the process of apply this multilayer architecture to continuous sign language recognition.

7. References

[1] 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.

[2] 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.

[3] 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.

[4] G.L. Fang, W. Gao, “Large Vocabulary Sign Language Recognition Based on Fuzzy Decision Trees”, IEEE Trans. Syst., Cybern. A, vol. 34, pp. 305-313, May. 2004

[5] G.L. Fang, W. Gao, “A SRN/HMM System for Signer-Independent Continuous Sign Language Recognition”, In Proceeding of the Fifth International Conference on Automatic Face and Gesture Recognition, Washinton DC, America, 2002,pp 312-317.

[6] V. N. Vapnik, Statistical learing Theory. Wiley, 1998

[7] Claus Bahlmann, Bernard Haasdonk and Hans Burkhardt, “On-line Handwriting Recognition with Support Vector Machines--a Kernel for Approach,” in IWFHR,2002..

[8] A. Ganapathiraju and J. Picone. “Hybrid SVM/HMM architectures for speech recognition,”in ICSLP2000,2000

[9] V. Doumpiotis, S. Tsakalidis, and W. Byrne, “Discriminative training for segmental minimum Bayes risk decoding”, in ICASSP, Hong Kong, 2003.