HMM BASED PYRAMID MATCH KERNEL FOR CLASSIFICATION OF SEQUENTIAL PATTERNS OF SPEECH USING SUPPORT VECTOR MACHINES

A. D. Dileep and C. Chandra Sekhar

Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai, India

ABSTRACT

Classification of varying length sequences using support vector machine (SVM) requires a suitable kernel that measures the similarity between a pair of sequences. In this paper we propose a novel approach to design a pyramid match kernel (PMK) using hidden Markov model. We study the performance of the SVM-based classifiers using the proposed PMK for recognition of isolated utterances of E-set in English alphabet and recognition of consonant-vowel segments of speech in Hindi and compare with that of the SVM-based classifiers using score-space kernels and alignments kernels.

Index Terms— varying length sequences, pyramid match kernel, support vector machine, speech recognition.

1. INTRODUCTION

Classification of varying length sequences of feature vectors using support vector machines (SVMs) requires design of a suitable kernel as a measure of similarity between a pair of sequences. The score-space kernels [1, 2] and alignment kernels [3–6] are the commonly used kernels. In this paper, we propose a novel approach to design pyramid match kernel (PMK) for sequences of feature vectors. In [7, 8], the PMK-based SVM is used for classification of varying length patterns represented as sets of feature vectors. Computation of PMK involves mapping each set of feature vectors onto a multi-resolution histogram pyramid. The Gaussian mixture model (GMM) based PMK in [8] is computed using a pyramid that consists of J+1 levels. The pyramid is constructed using a b^j-component class-independent GMM (CIGMM) at level j, with j=0, 1, 2, . . . , J. The CIGMM at level j is used to obtain a b^j-dimensional histogram vector representation of the input set of feature vectors. An element of the histogram vector corresponds to the effective number of feature vectors belonging to a component of the CIGMM. An histogram intersection function is used to compute the number of matches between a pair of histogram vectors corresponding to a pair of sets of feature vectors at each level. The PMK between a pair of examples is computed as a weighted sum of number of new matches at different levels of pyramid. The GMM-based PMK considered in [8] is not suitable for sequences of feature vectors because it does not use the sequence information while matching patterns. In this paper, we propose to design hidden Markov model (HMM) based PMK for sequences of feature vectors using continuous density HMMs (CDHMMs). In the HMM-based PMK, the pyramid is constructed using an N-state, left-to-right class-independent HMM (CIHMM) at each level. A b^j-component GMM is used for each state in the CIHMM at level j. A sequence of feature vectors is represented at level j by N histogram vectors with b^j elements in each vector. An element of a histogram vector corresponds to the effective number of feature vectors belonging to a component of the GMM of a state. The HMM-based PMK is constructed by matching the N histogram vector representations corresponding to a pair of sequences of feature vectors at consecutive levels of the pyramid. Our studies demonstrate the potential of the HMM-based PMK for classification of sequences of feature vectors using SVMs.

In Section 2, a review of kernels for sequences of feature vectors is presented. The proposed HMM-based PMK for sequences of feature vectors is described in Section 3. In Section 4, we present our studies. The conclusion is presented in Section 5.

2. KERNELS FOR SEQUENTIAL PATTERNS

In this section, we review the approaches to design kernels for varying length sequences. Score space kernels such as Fisher kernel (FK) [1] and likelihood ratio kernel (LRK) [2] for sequential patterns use a HMM for mapping a sequence onto a Fisher score-space. In FK, the Fisher score-space for a class is obtained using the first order derivatives of the log likelihood output of HMM for that class with respect to the HMM parameters. In LRK, the Fisher score-space corresponds to likelihood ratio score-space and is obtained by the first order derivatives of the ratio of the log likelihood outputs of HMMs for a pair of classes with respect to the HMM parameters.

Probability product kernel [9] for a pair of varying length sequences is computed by matching the distributions derived from the sequences. Each sequence is represented by a HMM and the probability product kernel is obtained by kernelizing the Kullback-Leibler (KL) divergence between the two HMMs.

The alignment kernels such as dynamic time warping kernel (DTWK) [3], dynamic time alignment kernel (DTAK) [4], global alignment kernel (GAK) [5] and triangular global alignment kernel (TGAK) [6] compute a kernel between a pair of sequences in a parametric vector space using the dynamic time warping (DTW) method. The DTWK is obtained by exponentiating the DTW distance between the sequences. The DTAK is obtained by using the DTW distance in the Gaussian kernel feature space. The DTWK and DTAK are shown to be positive definite kernels only under some favorable conditions [5, 6]. The GAK considers the sum of all possible alignment scores in the Gaussian kernel feature space. The TGAK is an extension to the GAK to perform faster computation of kernel. Both GAK and TGAK are shown to be positive definite kernels [5, 6]. In our studies, we compare the performance of the proposed HMM-based PMK with the performance of kernels reviewed in this section.

3. HMM-BASED PYRAMID MATCH KERNEL

In designing the pyramid matching kernel (PMK) in [7], an example represented as a set of feature vectors is mapped onto a multi-resolution histogram pyramid. The histogram at a level is computed by binning the feature vectors of an example into discrete regions [7]. In [8], the class-independent GMMs (CIGMMs) built with increasingly larger number of components are used to construct the histograms at different levels. At level j, a CIGMM of b^j components is built using the feature vectors in the training examples of all the classes. The assignment of a feature vector of an example to components follows the soft assignment technique. For a set of feature vectors, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, a b^j-dimensional histogram vector at the jth level is formed using the b^j components of CIGMM at that level. An element in the histogram vector corresponds to the effective number of feature vectors from $\mathbf{X}$ assigned to a component. An histogram intersection function [10] is then used to compute the number of matches between a pair of histogram vectors corresponding to a pair of examples $\mathbf{X}_m$ and $\mathbf{X}_n$ at each level. The matching is a hierarchical process from the bottom of the pyramid to the top of the pyramid. The number of new matches at a level is calculated by computing the difference between the number of matches at that level and the number of matches at its immediately higher level. The number of new matches at each level is weighted according to the number of components of CIGMM at that level. The GMM-based PMK between a pair of examples is computed as a weighted sum of the number of new matches at different levels of pyramid.

As the GMM-based PMK does not use the sequence information in matching the examples, it is not suitable for sequential patterns. In the proposed approach to computation of HMM-based PMK for sequential patterns, we use class-independent HMMs with increasingly larger number of components in the state-specific GMMs to construct the histograms at different levels. In acoustic modeling of subword units of speech such as phonemes, triphones and syllables using HMMs, a state is associated with an acoustic event. The number of events in different classes for a type of subword unit is the same. We propose to build a class-independent HMM (CIHMM) with the same number of states at each level of pyramid used to construct the HMM-based PMK. At different levels of pyramid, the number of components in state-specific GMMs is different. A CIHMM, λ, is an N-state, left-to-right, continuous density HMM with state-specific GMMs. The CIHMM at level j, λ_j, is built using the sequences of feature vectors corresponding to the training examples of all the classes. The CIHMMs are used only for matching the sequences to compute the kernel.

The effective number of feature vectors from a sequence $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_T)$ assigned to the qth component in the GMM of ith state of λ_j is obtained using the responsibility term and it is given by

$$R_{iq}(\mathbf{x}_t|\mathbf{X}, \lambda_j) = \gamma_{jit}\ \gamma_{jiq}(\mathbf{x}_t) \tag{1}$$

Here the term γ_{jit} is the probability of being in state i at time t given $\mathbf{X}$ and λ_j. The term γ_{jit} can be computed using the parameters of λ_j [11]. The term $\gamma_{jiq}(\mathbf{x}_t)$ is the probability that a feature vector $\mathbf{x}_t$ is generated by the component q of GMM of state i in λ_j. For a sequence of feature vectors $\mathbf{X}$, the effective number of feature vectors, h_{jiq}, assigned to a component q of state i in λ_j is given by

$$h_{jiq}(\mathbf{X}) \quad = \quad \sum_{t=1}^{T} R_{iq}(\mathbf{x}_t|\mathbf{X}, \lambda_j) \tag{2}$$

Let $J+1$ be the number of levels in the multi-resolution histogram pyramid. At the top level (j=0), a CIHMM with a single component state-specific GMM is built using the training data of all the classes. At level j, a CIHMM with b^j number of components in each state-specific GMM is built using the training data of all the classes. The process of constructing histogram vectors at different levels of pyramid for a sequence of feature vectors, $\mathbf{X}$, is illustrated in Figure 1. Let $\mathbf{h}_{ji}(\mathbf{X}_m)$ and $\mathbf{h}_{ji}(\mathbf{X}_n)$ be the b^j-dimensional histogram vectors corresponding to a pair of sequences $\mathbf{X}_m$ and $\mathbf{X}_n$ at the jth level for the ith state formed using b^j components of GMM Ψ_{ji}. The qth elements in the histogram vectors, $h_{jiq}(\mathbf{X}_m)$ and $h_{jiq}(\mathbf{X}_n)$, correspond to the effective number of feature vectors from $\mathbf{X}_m$ and $\mathbf{X}_n$ respectively assigned to the qth component in ith state at jth level, and are computed using (2). The effective number of matches in the qth component of Ψ_{ji} is given by the histogram intersection function [10], defined as follows:

$$s_{jiq} = \min\left(h_{jiq}(\mathbf{X}_m), h_{jiq}(\mathbf{X}_n)\right) \tag{3}$$

Total number of matches in state i at level j is obtained as

$$\mathcal{S}_{ji} = \sum_{q=1}^{b^j} s_{jiq} \tag{4}$$

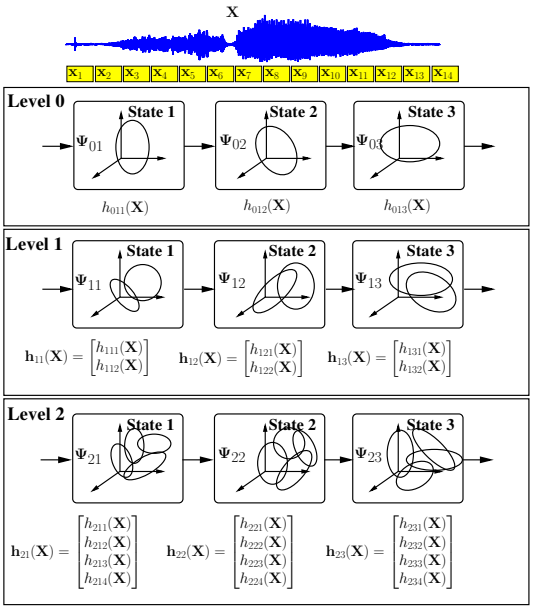

Fig. 1. Illustration of construction of histogram vectors at different levels of pyramid for a sequence of feature vectors $\mathbf{X}$. The values of J and b are 2 and 2 respectively. A 3-state left-to-right CIHMM with 2^j-component GMM for each state is used at each level j.

The effective number of new matches in state i at level j is calculated as

$$G_{ji} = \mathcal{S}_{ji} - \mathcal{S}_{(j+1)i} \tag{5}$$

At the bottom most level (j=J), $G_{Ji} = \mathcal{S}_{Ji}$. The total number of new matches in all the states at level j is computed as

$$M_j = \sum_{i=1}^{N} G_{ji} \tag{6}$$

The HMM-based PMK for $\mathbf{X}_m$ and $\mathbf{X}_n$ is now computed as:

$$K_{\text{PMK}}(\mathbf{X}_m, \mathbf{X}_n) = \sum_{j=0}^{J} w_j M_j \tag{7}$$

The weight at level j is considered as $w_j = \frac{1}{2^{J-j}}$. The proof for the HMM-based PMK as positive definite kernel is excluded due to the limitation of pages.

4. EXPERIMENTAL STUDIES ON SPEECH RECOGNITION

The proposed HMM-based PMK (HMMPMK) is used in building the SVM-based classifiers for recognition of isolated utterances of the E-set of English alphabet and recognition of consonant-vowel (CV) segments of continuous speech in Hindi. Mel frequency cepstral coefficients (MFCC) are used as features. A frame size of 20 ms and a shift of 10 ms are used for feature extraction from the speech signal of an utterance. Every frame is represented using a 39-dimensional feature vector. Here, the first 12 features are Mel frequency cepstral coefficients and the 13th feature is log energy. The remaining 26 features are the delta and acceleration coefficients. The Oregon Graduate Institute (OGI) spoken letter database is used in the study on recognition of E-set [12]. The E-set includes the 9 letters: B, C, D, E, G, P, T, V, and Z. The training and test sets include 240 and 60 utterances for each letter respectively. The E-set recognition accuracy is presented as the classification accuracy obtained for 540 test examples. The continuous speech corpus of broadcast news in Hindi [13] is used for the study on recognition of CV segments. We considered 103 CV classes [13] and the data set consists of total 24,324 CV segments. The CV segment recognition accuracy presented is the average classification accuracy along with 95% confidence interval obtained for 5-fold stratified cross-validation.

In our studies, the SVMs using the proposed HMMPMK are built using different values for N corresponding to the number of states in CIHMM, J corresponding to the number of levels in the histogram pyramid and the branching factor b. Parameters of CIHMMs are estimated using maximum likelihood (ML) method. Diagonal covariance matrices are considered for the GMM of each state. We consider LIBSVM [14] tool to build the SVM classifiers. The one-against-rest approach is considered for multi-class pattern classification task. The value of trade-off parameter C in SVM is chosen empirically as 10. The classification accuracies for the HMMPMK based SVMs for E-set recognition are given in Table 1. It is seen that the HMMPMK constructed using the CIHMMs for N=5, J=4 and b=3 gives the best performance of 94.81%. The classification accuracies with 95% confidence interval for CV segment recognition are given in Table 2. It is seen that the HMMPMK constructed using the CIHMMs for N=6, J=5 and b=3 gives the best performance of 56.91%.

Table 1. Classification accuracy (in %), of the SVM classifiers using HMMPMK for E-set recognition for different values of N, J and b. Q_{Ji} indicates the number of GMM components in each state at bottom most level.

N=5				N=6			
J	b	Q_{Ji}	Accuracy	J	b	Q_{Ji}	Accuracy
5	2	32	92.41	5	2	32	89.48
6	2	64	92.59	6	2	64	91.84
4	3	81	**94.81**	4	3	81	92.18
3	4	64	92.59	3	4	64	91.84

Tabel 3 compares the accuracies of E-set recognition obtained using the continuous density HMM (CDHMM) based system and SVM-based classifiers using Fisher kernel (FK), likelihood ratio kernel (LRK), discrete time warping kernel (DTWK), dynamic time alignment kernel (DTAK),

Table 2. Classification accuracy (in %), estimated at 95% confidence interval, of the SVM classifiers using HMMPMK for CV segment recognition for different values of N, J and b. Q_{Ji} indicates the number of GMM components in each state at bottom most level.

N=5				N=6			
J	b	Q_{Ji}	Accuracy	J	b	Q_{Ji}	Accuracy
7	2	128	53.91±0.86	7	2	128	54.89±0.85
8	2	256	55.33±0.82	8	2	256	56.23±0.84
5	3	243	55.12±0.85	5	3	243	**56.91±0.81**
4	4	256	55.33±0.82	4	4	256	56.23±0.84
3	5	125	53.53±0.91	3	5	125	54.36±0.92

global alignment kernel (GAK), triangular alignment kernel (TGAK) and the proposed HMMPMK. The parameters of CDHMM built for each class are estimated using ML method. For CDHMM-based classifier, we consider diagonal covariance matrices for the GMMs at each state. The accuracies of CDHMM based system are observed for different values of N corresponding to the number of states and Q corresponding to the number of GMM components for each state. The FK and LRK are computed using CDHMMs for each class with the different values of N and Q. The DTWK, DTAK and GAK are computed for different values of the parameter σ [6]. The TGAK is computed for different values of parameters σ and R [6] (notation R is similar to T in [6]). The probability product kernel is not considered in our study. The reason is that, in tasks such as E-set recognition and CV unit recognition, patterns are extracted from short duration (200 to 500 milliseconds) segments of speech and building a HMM for each example is difficult. We have considered the one-against-one approach for 9-class E-set recognition task using LRK based SVM, as LRK is computed for every pair of classes. For SVMs using other kernels, we considered the one-against-rest approach. The best accuracies and the corresponding values of parameters are given in Table 3.

Table 3. Comparison of classification accuracy (in %), of the CDHMM-based systems and SVM classifiers using score-spce kernels, alignment kernels and HMMPMK for E-set recognition task.

Classification model		(N,Q)/(σ,R)/(N,J,b)	Accuracy
CDHMM		(12,2)	90.30
SVM using	FK	(12,2)	91.57
	LRK	(12,2)	**95.00**
	DTWK	(1000,-)	85.37
	DTAK	(15,-)	87.11
	GAK	(20,-)	87.41
	TGAK	(45,0.5)	87.30
	HMMPMK	(5,5,3)	**94.81**

Tabel 4 compares the best observed accuracies of CV segment recognition obtained using the CDHMM-based system and SVM-based classifiers using FK, DTWK, DTAK, GAK, TGAK and the HMMPMK. We have not considered LRK based SVM for the CV segment recognition task. Because, for 103 CV classes, it leads to 10712 pairs of classes and generating such a large number of LRKs is computationally intensive.

Table 4. Comparison of classification accuracy (in %), estimated at 95% confidence interval, of the CDHMM-based systems and SVM classifiers using score-space kernels, alignment kernels and HMMPMK for CV segment recognition.

Classification model		(N,Q)/(σ,R)/(N,J,b)	Accuracy
CDHMM		(5,3)	48.87±0.77
SVM using	FK	(5,3)	52.58±0.86
	DTWK	(1000,-)	52.51±0.79
	DTAK	(15,-)	54.10±0.89
	GAK	(20,-)	54.76±0.85
	TGAK	(25, 0.5)	54.77±0.87
	HMMPMK	(6,6,3)	**56.91±0.81**

It is seen that performance of the SVM classifiers is better than that of the CDHMM-based systems. This is mainly because the CDHMM-based classifier is trained using a non-discriminative learning based technique, where as the SVM classifier is built using a discriminative learning based technique. It is also seen that the performance of SVM classifiers using the proposed HMMPMK is significantly better than that of the SVM classifiers using FK and alignment kernels for both E-set recognition and CV segment recognition. For the E-set recognition, the SVM using the proposed HMMPMK performed close to SVM using the LRK.

5. CONCLUSIONS

In this paper, we proposed an approach to compute the pyramid match kernel for sequences of feature vectors. We proposed to use class-independent CDHMMs with increasingly large number of components in the state-specific GMMs to construct the pyramid. Results of studies on recognition of isolated utterances of letters in E-set of English alphabet and recognition of CV segments from continuous speech corpus of Hindi demonstrate the effectiveness of the proposed HMM-based pyramid match kernel in building the SVM classifiers for sequential pattern classification. These classifiers can be used to build the acoustic models for subword units of speech such as triphones and syllables in developing a continuous speech recognition system. For the speech classes where the number of acoustic events is different, the class-specific CDHMMs with increasingly large number of components in the state-specific GMMs can be used to compute the PMK for that class.

6. REFERENCES

[1] T. Jaakkola, M. Diekhans, and D. Haussler, "A discriminative framework for detecting remote protein homologies," *Journal of Computational Biology*, vol. 7, no. 1-2, pp. 95–114, 2000.

[2] N. Smith and M. Gales, "Speech recognition using SVMs," in *Advances in Neural Information Processing Systems*. 2002, pp. 1197–1204, MIT Press.

[3] B. Haasdonk and C. Bahlmann, "Learning with distance substitution kernels," in *Proceedings of 26th Annual Pattern Recognition Symposium, DAGM'04*, Tübingen, Germany, 2004, pp. 220–227.

[4] H. Shimodaira, K.-I. Noma, M. Nakai, and S. Sagayama, "Dynamic time-alignment kernel in support vector machine," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. 2002, pp. 921–928, MIT Press.

[5] M. Cuturi, J. P. Vert, O. Birkenes, and T. Matsui, "A kernel for time series based on global alignments," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, Honolulu, Hawaii, USA, April 2007, vol. 2, pp. 413–416.

[6] M. Cuturi, "Fast global alignment kernels," in *Proceedings of the 28th International Conference on Machine Learning (ICML-2011)*, Lise Getoor and Tobias Scheffer, Eds., New York, NY, USA, June 2011, pp. 929–936, ACM.

[7] K. Grauman and T. Darrell, "The pyramid match kernel: Efficient learning with sets of features," *The Journal of Machine Learning Research*, vol. 8, pp. 725–760, 2007.

[8] A. D. Dileep and C. C. Sekhar, "Speaker recognition using pyramid match kernel based support vector machines," *International Journal of Speech Technology*, vol. 15, no. 3, pp. 365–379, 2012.

[9] T. Jebara, R. Kondor, and A. Howard, "Probability product kernels," *Journal of Machine Learning Research*, vol. 5, pp. 819–844, December 2004.

[10] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, pp. 11–32, 1991.

[11] L. Rabiner and B-H. Juang, *Fundamentals of Speech Recognition*, Pearson Education, 2003.

[12] ISOLET Corpus, *Release 1.1*, Center for Spoken Language Understanding, Oregon Graduate Institute, July 2000.

[13] S. V. Gangashetty, "Neural network models for recognition of consonant-vowel units of speech in multiple languages," PhD thesis, Dept. of Computer Science and Engg., IIT Madras, Chennai, February 2005.

[14] C-C. Chang and C-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.