DISCRIMINATIVE FEATURE SELECTION FOR HIDDEN MARKOV MODELS USING SEGMENTAL BOOSTING

Pei Yin, Irfan Essa, Thad Starner, James M. Rehg

School of Interactive Computing, College of Computing
Georgia Institute of Technology, Atlanta, USA

ABSTRACT

We address the feature selection problem for hidden Markov models (HMMs) in sequence classification. Temporal correlation in sequences often causes difficulty in applying feature selection techniques. Inspired by segmental k-means segmentation (SKS) [1], we propose Segmentally Boosted HMMs (SBHMMs), where the state-optimized features are constructed in a segmental and discriminative manner. The contributions are twofold. First, we introduce a novel feature selection algorithm, where the temporal dynamics are decoupled from the static learning procedure by assuming that the sequential data are piecewise independent and identically distributed. Second, we show that the SBHMM consistently improves traditional HMM recognition in various domains. The reduction of error compared to traditional HMMs ranges from 17% to 70% in American Sign Language recognition, human gait identification, lip reading, and speech recognition.

***Index Terms*—** Time-series, Pattern Recognition, Feature Extraction, Hidden Markov models

1. INTRODUCTION AND RELATED WORK

The ability of hidden Markov models (HMMs) to compensate for the variance in length of temporal sequences leads to good performance in speech processing, gesture recognition, DNA analysis, and other applications. An HMM is normally estimated by Maximum Likelihood Estimation (MLE). In practice, discriminative methods, which simply learn a decision boundary, are usually superior for classification than MLE. Previous attempts in introducing discriminative methods to HMMs can be classified into two categories: discriminative training of the model parameters and discriminative feature selection. Discriminative variants of HMM parameter training, *e.g.,* Minimum Classification Error (MCE), Maximum Mutual Information (MMI) [2] and Conditional Maximum Likelihood (CML) criteria directly adjust the model parameters for classification. A detailed review and comparison can be found in Sha and Saul [3]. Previous studies [4, 5] indicate that discriminative features will be able to improve discriminatively trained models: higher accuracy and efficiency can be achieved by emphasizing the informative features and filtering out the irrelevant ones. Therefore, selecting/extracting discriminative features for HMMs has been a focus of attention [6, 7, 8, 9]. Our Segmentally Boosted HMM (SBHMM) technique belongs to this category — our focus is to extract discriminative features for the classification with explicit consideration of temporal correlation.

In machine learning, automatic feature selection is usually cast as the optimization of a supervised classification problem. For x being the evidence and y being the labels, the pair (x, y) defines a classification problem $y = f(x)$, and the discriminative features are computed in order to minimize the classification loss. However, there are two major difficulties in applying such feature selection methods for time sequences.

ASLR (vision)	ASLR (accelerometer)	Gait Recognition	Lip Reading	Speech Recognition
36.4%+	17.1%+	70.1%	32.2%	39.2%

Table 1. The reduction of error by the SBHMM compared to the HMM baseline in the 5 experiments conducted in this paper. ASLR=American Sign Language recognition.

First, sequential data do not observe the basic assumption of supervised learning, *i.e.,* that the samples are independent and identically distributed (i.i.d.). Time sequences contain significant amount of temporal correlations (not independent sampling); some sequences may also contain several "phases" (states), where the discriminative features for one phase may be quite uninformative for another (not identically distributed). For instance, the sign "fish" in American Sign Language is expressed by moving the two hands asynchronously. If the state can be clearly identified, for example in word tagging [10] and video segmentation [11], conditional models such as Conditional Random Fields (CRF) can be successfully applied to perform *sequential* classification [12]. However, the meaning and the labeling of the states are mostly unavailable in *sequence* classification [12]; for example, to recognize the sign "brother", how can the human labeler precisely supervise the training for the first state when he does not even know the state's meaning or how many states comprise the sign? In such situations, SBHMMs provide a solution to obtain similar discriminative classifiers in a unsupervised manner. As we mention later, SBHMMs can be further modified to facilitate iterative refinement of selected features using Baum-Welch re-estimation.

Second, sequences are variable in length, though the learning functions $f(\cdot)$ usually expect inputs x of fixed cardinality. The Fisher kernel [13] and its variants use a generative model to preprocess the sequences and construct a discriminative kernel according to the fisher score (local gradient) of that generative model. The "feature weighting" is encoded in the kernel matrix. The Fisher kernel has been successfully applied to domains such as Bioinformatics. However its reliance on a potentially imperfect generative model can cause problems.

Recently, Feature-space Minimum Phone Error (fMPE) [7] and Stereo-based Piecewise Linear Compensation for Environments (SPLICE) [6] produce good improvements in large vocabulary recognition. They compensate input features with posterior-based "correction vectors" [5]. However they are relatively expensive to compute in practice.

In this paper, we propose SBHMMs, which leverage both the dynamic nature of the sequential data and the static nature of the large-margin feature selection methods by assuming "piecewise i.i.d." (which

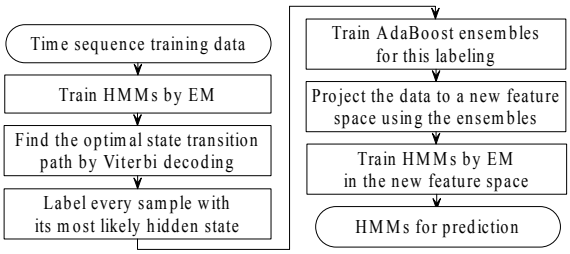

Fig. 1. The key steps to SBHMMs.

does not introduce additional approximations, because it is already assumed by HMMs). The key steps to our SBHMM technique are illustrated in Fig. 1.

Our experiments show that SBHMMs reduce the sequence recognition error by 17%-70% compared to HMMs in the application of American Sign Language recognition, human gait identification, lip reading and speech recognition (see Table 1, with details in Section 4). SBHMMs construct new features by comparing the feature value with a set of discriminatively chosen thresholds, which can be efficiently computed.

2. SEGMENTAL BOOSTING

The idea of segmental training [1] was first introduced in the 1990s. The motivation was to create a better initial estimation for the observation models. In this work, we extend the concept of segmental training in order to perform discriminative feature selection. We derive this strategy in the context of the first order HMM.

HMMs have been very successful in interpreting temporal data. An HMM builds a causal model for observation sequence $\mathbf{O} = (o_1 o_2 \cdots o_T)$ by introducing corresponding "hidden states" $\mathbf{q} = (q_1 q_2 \cdots q_T)$. Let $P(q_1) = P(q_1|q_0)$. The transition model is $P(q_t|q_{t-1})$ and the observation model is $P(o_t|q_t)$. Assuming that there are C types of sequences, recognition selects the one with the highest likelihood $c^* = \text{argmax}_{1\le c\le C} P(\mathbf{O}|\lambda_c)$, and $\Lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_C\}$ are the parameters of the HMMs.

For a type c sequence $\mathbf{O}^c$ with length T_c, we define the model distance (dissimilarity) [2] as

$$D(\lambda_c, \Lambda) = \frac{1}{T_c}[\log P(\mathbf{O}^c|\lambda_c) - \frac{1}{C-1}\sum\nolimits_{v\neq c} \log P(\mathbf{O}^c|\lambda_v)].$$

We intend to choose a subset of features that maximize $D(\lambda_c, \Lambda)$. Assuming an uninformative prior, it is equivalent to maximizing the "sequence margin"

$$M(\lambda_c, \Lambda) = \frac{1}{T_c}[\log P(\lambda_c|\mathbf{O}^c) - \frac{1}{C-1}\sum\nolimits_{v\neq c} \log P(\lambda_v|\mathbf{O}^c)].$$

Discriminative classifiers with logistic output, such as boosting $H(x) = \sum_j \alpha_j h_j(x) = \log P(y = y^*|x) - \log P(y \neq y^*|x)$, are capable of maximizing such a margin for the classification problems (x, y).

There are three natural choices for (x, y) representing different granularity: $(x = \mathbf{O}, y = c)$, $(x = o_t, y = c)$ and $(x = o_t, y = q_t)$. The first one is intractable since the length of the observation $\mathbf{O}$ varies. The second corresponds to the sliding window methods [12] with fixed [8] or empirically determined [14] size. Although improved results are reported, the oversimplified assumption limits its application to more complicated tasks where "the static features tend to cluster...without dynamic information" [9]. Therefore, the sequential dependency between the sliding windows, which conveys important information for recognition, can not be neglected.

To respect the temporal dependency while staying tractable, we argue that the samples inside every classification problem should be least correlated while preserving the temporal relationship between the problems (piecewise i.i.d.), *i.e.,* assign $(x = o_t, y = q_t)$. The idea is to *decouple* discriminative feature selection from the temporal dependencies, instead of *discarding* this information. The derivation follows the HMM's assumption on Markov property and conditional independence without introducing additional independence assumptions:

$$P(\mathbf{O}|\lambda_c) = \sum\nolimits_{\mathbf{q}} P(\mathbf{O}|\mathbf{q}, \lambda_c) P(\mathbf{q}|\lambda_c)$$
$$= \sum\nolimits_{\mathbf{q}} \prod_{t=1}^{T} P(o_t|q_t) P(q_t|q_{t-1})$$

So $M(\lambda_c, \Lambda)$ can be increased with some discriminative $P(o_t|q_t)$. The intuition is that HMMs decompose the evolving temporal trajectory into two types of behavior (1) loop within the same state, or (2) transition from one state to another. Thus, we can perform feature selection only in the segments of the same state, and those "static" segments are connected by the temporal transition $P(q_t|q_{t-1})$. Note that the concept of "hidden state" is still necessary to smooth the results of the observation model.

Therefore, we first train a set of HMMs with the original features, and label every observations o_t by its maximum *a posteriori* (MAP) state s_t computed by *Viterbi* decoding. Then we train a set of AdaBoost ensembles $\{H^{(s)}\}$ for such labeling. We ignore the superscript s when there is no ambiguity.

3. CONSTRUCTION OF THE NEW FEATURES

3.1. Data Aggregation by AdaBoost

AdaBoost linearly combines the weak learners $h_j(x_i) \in [-1, 1]$ to obtain a strong classifier (ensemble) $H(x) = \sum_j \alpha_j h_j(x)$ for each class s (state). The weak learners h used in SBHMMs are the decision stumps like "is the value of feature No.5 greater than 0.45?" The binary answers are then weighted according to their empirical discriminative power in separating class s from the other classes.

The margin of the ensemble with l weak learners at x_i is defined as $m_l(x_i) = \frac{y_i H_l(x_i)}{w_l}$, while $w_l = \sum_{j=1}^{l} \alpha_j$ is the sum of the learner weight, served as a normalization factor. During the AdaBoost training, the minimum margin $\min_i \{m_l(x_i)\}$ tends to increase [15], which leads to good generalization ability.

As the training proceeds, *the average margin tends to decrease*. The average margin of AdaBoost at round l is defined as

$$\overline{m_l} = \frac{\sum_{i=1}^{n} y_i H_l(x_i)}{n \cdot w_l} = \frac{\sum_{i=1}^{n} y_i \left[\sum_{j=1}^{l} \alpha_j h_j(x_i)\right]}{n \cdot \sum_{j=1}^{l} \alpha_j} = \frac{\sum_{j=1}^{l}\sum_{i=1}^{n} \alpha_j h_j(x_i) y_i}{\sum_{j=1}^{l}\sum_{i=1}^{n} \alpha_j y_i^2}$$

Denote

$$A = \sum_{j=1}^{l}\sum_{i=1}^{n} \alpha_j h_j(x_i) y_i, \quad B = \sum_{j=1}^{l}\sum_{i=1}^{n} \alpha_j y_i^2$$
$$C = \sum_{i=1}^{n} \alpha_{l+1} h_{l+1}(x_i) y_i, \quad D = \sum_{i=1}^{n} \alpha_{l+1} y_i^2$$

We have the average margin of round l and round $l+1$

$$\overline{m_l} = \frac{A}{B}, \quad \overline{m_{l+1}} = \frac{A+C}{B+D}$$

where

$$\overline{m_{l+1}} \gtreqless \overline{m_l} \Leftrightarrow \frac{A+C}{B+D} \gtreqless \frac{A}{B} \Leftrightarrow \frac{C}{D} \gtreqless \frac{A}{B}$$

In practice, "$<$" happens much more frequently than the other two cases, considering that (1) A/B is the classification loss of the ensemble composed by l weak learners, while C/D is the loss of one

weak learner at $l+1$ (2) AdaBoost gradually focus on the "harder" examples [15]. Therefore $\overline{m_l}$ will decrease as training proceeds. This effect can be observed from the margin distribution graph in Schapire *et al.* [15].

The increase of the minimum margin and the decrease of the average margin indicates that boosting generates a natural clustering of data in its output space according to their labels [11].

3.2. The Discriminative New Feature Space

SBHMMs use boosting ensembles to *construct* a new feature space $\mathcal{V}$. We define $\mathcal{V}$ as the output space spanned by the S ensembles $\mathcal{V} = (H^{(1)}, \cdots, H^{(S)})$, where S is the total number of the classes (states). For example, boosting constructs two ensembles $H^{(1)}$ and $H^{(2)}$ corresponding to a two-class problem. Suppose the outputs of the ensembles are $H^{(1)}(x) = 0.2$, $H^{(2)}(x) = 0.7$ for x, then x is projected to coordinate $(0.2, 0.7)$ in $\mathcal{V}$. Note that (1) this projection is *nonlinear* because the ensemble is a combination of the nonlinear thesholding functions on the feature values, and (2) the standard dimensionality reduction methods such as PCA can be applied to $\mathcal{V}$ to further improve efficiency.

The intuition of using $\mathcal{V}$ as the new feature space is twofold (1) the composition of the ensembles (weighting of weak learners) contains important information about discrimination (2) the aggregation of the samples in $\mathcal{V}$ refines the data distribution to fit the Gaussian observation model. Our tests show that the original data is hyperleptokurtic (high kurtosis), which requires a dense mixture of Gaussians. After the nonlinear projection by SBHMMs, the distribution become more Gaussian in the new feature space, and the data kurtosis is reduced by 70%. The training of the (mixture of) Gaussian observation models is in turn improved.

4. EXPERIMENTAL RESULTS

We evaluate the performance of SBHMMs across four domains and compare it with the results reported previously by other researchers. Note that the size of boosting ensembles are empirically determined for each application.

4.1. American Sign Language Recognition Results

In the application of American Sign Language recognition (ASLR), we compare SBHMMs with two baseline HMMs [1] [16, 17]. In the first continuous recognition experiment, 500 sentences of 40 different signs are performed by one subject in five-sign phrases. 16 features including the position, velocity and the size of the two hands are computed from a color-based hand tracker. We choose the same 400 sentences for training and 100 sentences for testing as the original researchers [16]. 4-state HMMs are used for recognition. In the second experiment, accelerometers mounted on a glove, elbow, and shoulder provide 17 features such as wrist rotation and hand movement [17]. This dataset contains 665 sentences with 141 different signs. We use 3-state HMMs on this dataset with 10-fold cross-validation.

The experimental results are listed in Table 2. Despite the high accuracy of the original HMM baselines, SBHMMs are able to reduce the error rate by about 20% on both datasets, with or without postprocessing by grammar. SBHMMs assign heavy weights on features like the rotation of the wrists, which is considered meaningful by sign language experts.

[1]Both datasets are available at `http://wiki.cc.gatech.edu/ccg/projects/asl/asl`.

	with grammar			without grammar		
error	HMM	SBHMM	reduction	HMM	SBHMM	reduction
vision	2.2%	**1.4%**	36.4%	3.2%	**2.0%**	37.5%
accel.	2.2%	**1.8%**	17.1%	4.9%	**3.8%**	22.4%

Table 2. Comparison of the test error on vision-based ASLR (top) and accelerometer-based ASLR (bottom)

Total Length	30m45s	Sampling Rate	120Hz
Training Data	24m42s	Testing Data	06m03s
Total Sentences	275	Total Phones	8468
Total Phonemes	39	Total Samples	>200,000

Table 4. Georgia Tech Speech Reading Database

4.2. Human Gait Identification Results

We compare SBHMMs with human gait recognition results previously reported by Kim and Pavlovic [18]. In their paper, performance of several cutting-edge discriminative training methods for mixtures of Bayesian network classifiers such as CML are evaluated using the gait data [19]. This dataset [2] consists of 9 trials of 15 subjects walking in 4 different speeds. The data record the 3D position of 22 markers on the subject at 120Hz. Exactly following the authors' convention, we obtain 180 subsequences for the five subjects with various speeds. Each sequence contains 6 dimensional feature vector describing the joint angle of torso-femur, femur-tibia and tibia-foot. 100 sequences are randomly chosen for training, and the other 80 for testing. A 3-state HMM with single Gaussian is used to identify the subjects. The test error averaged over 10 random training/testing splits is reported in Table 3. It shows that the SBHMM outperforms all the other 6 algorithms, and the reduction of the test error ranges from 14% (MixCML [18]) to 70% (HMM).

4.3. Audio and Visual Speech Recognition Results

We also test our SBHMM algorithm on Georgia Tech Speech Reading Database [3] [8] with two tasks: lip reading [22] (visual feature only) and speech recognition (acoustic feature only). The visual features are 18 infrared trackers around the lip. Their 3D positions are recorded at 120Hz. The audio features are the first 13 orders of MFCCs [2] and their derivatives, computed at 120Hz from a 16kHz sound track. Since the goal is to illustrate automatic feature selection on the baseline HMM, we didn't perform elaborate preprocessing steps as most state-of-the-art speech recognizers do. The dataset is described by Table 4. Both visual and acoustic recognition system are naïvely implemented using 3-state HMMs with diagonal Gaussians mixtures. Note that the visual phoneme is defined the same as the acoustic phonemes. SBHMM reduces the test error by 30% compared to HMM in Table 5. Table 3 and Table 5 also illustrate that, by assuming piecewise i.i.d. instead of i.i.d. for the entire sequence, SBHMM has higher accuracy than Boosted HMM (BHMM) [8], which selects features using $(x = o_t, y = c)$.

In order to validate our approach, we compute the Generalized Rayleigh Quotient [23] (the ratio of the interclass variance and the intraclass variance) on Georgia Tech Speech Reading dataset. The higher generalized eigenvalue of the quotient reflects the better separability of the data. SBHMMs manage to obtain an order of magnitude higher generalized eigenvalue than traditional HMM, which

[2]The dataset is available at `ftp://ftp.cc.gatech.edu/pub/gvu/cpl/walkers/speed_control_data/`.

[3]The dataset is available at `http://www.cc.gatech.edu/cpl/projects/speechreading/index.html`.

1-NN DTW	HMM	BML [20]	MixCML [18]	BoostML [21]	BHMM [8]	SBHMM
8.38±3.68%	11.50±4.78%	10.13±3.61%	4.00±3.48%	11.87±5.11%	5.93±6.64%	**3.44±1.43%**

Table 3. Comparison of test error on Georgia-Tech Speed-Control Gait dataset. The first 5 columns are directly from [18]

	HMM	BHMM [8]	SBHMM
Visual	50.36%	42.56%	**34.16%**
Acoustic	32.30%	26.54%	**19.65%**

Table 5. Comparison of the test error on visual lip reading (top) and acoustic speech recognition (bottom).

illustrates the effectiveness of the discriminative feature space.

5. CONCLUSION AND FUTURE WORK

In this paper, we presented a segmental boosting algorithm to perform discriminative feature selection for time sequences. Assuming piecewise i.i.d. segments corresponding to HMM state boundaries, SBHMM constructs a new discriminative feature space. Experiments on various applications illustrate that SBHMM achieves up to 17%-70% reduction of error compared to HMM; SBHMM is also compared favorably with many other HMM-based discriminative methods such as BHMM.

If we view the discriminative learning as a logistic regression [24] of the posterior probability, the global convergence of SBHMM can be achieved in an EM manner. The proof generally follows that in Juang and Rabiner [1], and the details are omitted due to the space limit.

Although the derivation and experiments in this paper are performed using HMMs and boosting, they can be generalized to other Markovian models and other discriminative algorithms. In the future, we are interested in integrating segmental feature selection techniques with segmental discriminative training in more challenging gesture recognition tasks.

6. ACKNOWLEDGEMENT

The authors would like to thank Prof. Aaron Bobick, Jianxin Wu and Qiang Fu for the insightful discussions. This work is partly supported by NSF grants IIS#0205507 and IIS#0511900.

7. REFERENCES

[1] B. Juang and L. Rabiner, "The segmental k-means algorithm for estimating parameters of hidden markov models," *IEEE Trans. Acoust. Sp. Sign. Process*, vol. 38, pp. 1639–1641, 1990.

[2] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ, Printice Hall, 1993.

[3] F. Sha and L. Saul, "Comparison of large margin training to other discriminative methods for phonetic recognition by hidden Markov models," in *Proc.of ICASSP*, pp. IV313–316.

[4] C. Meyer, "Towards 'large margin' speech recognizers by boosting and discriminative learning," in *Proc.of ICML*, 2002, pp. 419–426.

[5] L. Deng, J. Wu, J. Droppo, and A. Acero, "Analysis and comparison of two speech feature extraction/compensation algorithms," *IEEE Signal Processing Letters*, vol. 12, no. 6, pp. 477–479, 2005.

[6] L. Deng, J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 569–580, 2003.

[7] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for continuous speech recognition," in *Proc. of ICASSP*, 2005.

[8] P. Yin, I. Essa, and J. Rehg, "Asymmetrically boosted HMM for speech reading," in *Proc.of CVPR*, 2004, pp. II:755–761.

[9] F. Lv and R. Nevatia, "Recognition and segmentation of 3-D human action using hmm and multi-class adaboost," in *Proc. of ECCV*, 2006, pp. IV: 359–372.

[10] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilitstic models for segmentaing and labeling sequence data," in *Proc.of 18th ICML*, 2001, pp. 282–289.

[11] P. Yin, A. Criminisi, J. Winn, and I. Essa, "Tree-based classifiers for bilayer video segmentation," in *Proc. of CVPR*, 2007.

[12] Thomas G. Dietterich, "Machine learning for sequential data: A review," *LNCS*, vol. 2396, pp. 15–30, 2002.

[13] T. S. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Proc. of NIPS*, 1999.

[14] P. Smith and M. Shah N. Lobo, "Temporalboost for event recognition," in *Proc. of ICCV*, 2005, pp. 733–740.

[15] R. Schapire, Y. Freund, P. Bartlett, and W. Lee, "Boosting the margin: a new explanation for the effectiveness of voting methods," in *Proc. 14th ICML*, 1997, pp. 322–330.

[16] T. Starner, J. Weaver, and A. Pentland, "Real-time American Sign Language recognition using desk and wearable computer based video," *IEEE Trans. on PAMI*, vol. 20, no. 12, pp. 1371–1375, 1998.

[17] T. Westeyn, H. Brashear, A. Atrash, and T. Starner, "Georgia Tech Gesture Toolkit: Supporting experiments in gesture recognition," in *Proc. of ICMI*, 2003, pp. 85–92.

[18] M. Kim and V. Pavlovic, "Discriminative learning of mixture of bayesian network classifiers for sequence classification," in *Proc.of CVPR*, 2006, pp. 268–275.

[19] R. Tanawongsuwan and A. Bobick, "Performance analysis of time-distance gait parameters under different speeds," *LNCS*, vol. 2688, pp. 1060–1068, 2003.

[20] Y. Jing, V. Pavlovic, and J. Rehg, "Efficient discriminative learning of bayesian network classifier via boosted augmented naive bayes," in *Proc.of ICML*, 2005, pp. 369–376.

[21] V. Pavlovic, "Model-based motion clustering using boosted mixture modeling," in *Proc.of CVPR*, 2004, pp. I:811–818.

[22] G. Potamianos, C. Neti, G. Gravier, and A. Garg, "Automatic recognition of audio-visual speech: Recent progress and challenges," *Proc.of the IEEE*, vol. 91, no. 9, 2003.

[23] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, Wiley-Interscience, second edition, 2000.

[24] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *Annals of statistics*, vol. 38, pp. 337–374, 2000.