CONTEXTUAL MODELING OF HAND WRITTEN CHINESE CHARACTER FOR RECOGNITION (II) - DISCRIMINATIVE TRAINING

Yan Xiong, Qiang Huo and Chorkin Chan

Department of Computer Science,
The University of Hong Kong, Hong Kong.

ABSTRACT

This is an extension of a companion paper entitled "Contextual Modeling of Hand Written Chinese Character for Recognition (I) - A Comparative Study" which is also submitted to this conference for presentation. In this investigation, contextual models are discriminatively trained using a gradient projection technique. Both open test and close test recognition rates are substantially upgraded when compared with the results of the decision directed training algorithm reported in the other paper.

1. INTRODUCTION

As an extension of the Contextual Vector Quantization (CVQ) modeling for speech recognition [1, 2], modeling complex and variant patterns like Chinese characters by contextual models has been proposed and demonstrated to be highly effective in [3]. In a companion paper submitted to this conference [4], a comparative study of several training methods and discriminant functions for contextual modeling-based character recognition is conducted, and their viability and usefulness are confirmed on a recognition task of 10 highly similar hand printed Chinese characters. In this paper, the recognition performance on the same task is further upgraded by discriminative training of the model parameters with an optimization technique originally developed for a speech recognition problem [5, 6].

2. DISCRIMINATIVE TRAINING OF A CONTEXTUAL MODEL

The training strategy of a contextual model used in [3, 4] is an approximate maximum likelihood estimation (AMLE) and the training algorithm to this end is a decision-directed (DD) one. It can be shown (e.g., [7]) that, if certain assumptions are met, one can argue intuitively that using the MLE and the MAP (maximum *a posteriori*) decision rule can lead to a recognition system that is asymptotically optimal. Nevertheless, inaccuracies in modeling the character pattern may lead to MLE that do not maximize the recognition accuracy, which is often observed in speech recognition (e.g., [8]). Recently, alternatives to ML training such as "Maximum Mutual Information (MMI) training" (e.g., [8]), "Minimum Discrimination Information (MDI) training" [9] and other methods (e.g., [10]) with the objective of lower recognition error rate have been proposed for speech recognition systems. Generally speaking, the purpose of contextual model training is to yield a recognizer of the lowest possible error rate. This objective is achieved by maximizing an objective function $R(\lambda)$. There are thus two important and difficult problems to consider. The first is to determine a meaningful objective function such that, if $R(\bar{\lambda}) > R(\lambda)$, then $\bar{\lambda}$ produces a better recognizer than that by λ. Once a function $R(\lambda)$ is chosen, the second problem (the estimation problem) is to find the parameter set $\bar{\lambda}$ which maximizes it.

The parameters of a contextual model $\lambda = (\beta, \mathbf{A}, \mathbf{B})$ (see the detailed explanations in [3, 4]), where

$$\beta = \{\beta_k = Pr(G_k)\}, \quad k = 1, 2, \cdots, K \tag{1}$$

$$\mathbf{A} = \{a_{k,l}^{m,n} = Pr^{m,n}(G_k \mid G_l)\}, \quad k, l = 1, 2, \cdots, K; \tag{2}$$

$$\mathbf{B} = \{b_{k,t} = b_k(v_t) = Pr(v_t \mid G_k)\}, \quad k = 1, 2, \cdots, K; \;\; t = 1, 2, \cdots, T \tag{3}$$

must satisfy the following constraints:

$$\sum_{k=1}^{K} \beta_k = 1 \;\; and \;\; \beta_k \geq 0, \quad k = 1, 2, \cdots, K \tag{4}$$

$$\sum_{t=1}^{T} b_{k,t} = 1 \;\; and \;\; b_{k,t} \geq \epsilon, \quad k = 1, 2, \cdots, K; \;\; t = 1, 2, \cdots, T \tag{5}$$

0-7803-4137-6/97/$10.00©1977 IEEE

$$\sum_{l=1}^{K} a_{k,l}^{m,n} = 1 \quad and \quad a_{k,l}^{m,n} \geq 0,$$
$$k, l = 1, 2, \cdots, K; \tag{6}$$

where ϵ is a small positive value. If one looks at the training problem of a contextual model as a problem of classical constrained optimization, then the standard optimization techniques can certainly be used to solve for the "optimal" model parameters. Classical optimization techniques are not only a viable alternative but may even be preferable in some cases. In particular they are virtually unrestricted by the forms of either the objective function or the constraints. So, it is clear that in the general case, there may be advantages in using classical optimization methods. Even under the situation of AMLE, such procedures have been shown to yield solutions comparable to that of a DD algorithm [4].

Consider a collection of P contextual models, $\Lambda = (\lambda_1, \lambda_2, \cdots, \lambda_P)$, where λ_p denotes the set of parameters of the p-th model. Let $\mathbf{x}^{(p,q)}$ denote the qth training observation sample associated with model p, and each model has W_p such observation samples. The objective function for discriminative training adopted in this paper is derived according to the minimum recognition error formulation recently proposed by Juang and Katagiri [10] which is a three-step procedure. The three-step definition emulates the classification/recognition operation as well as the performance evaluation, particularly in terms of classification errors.

The first step of the formulation is to prescribe an appropriate discriminant function $f_i(\mathbf{x}; \Lambda)$ which is used by the classifier to make its decision for each input $\mathbf{x}$ by choosing the largest of the discriminants evaluated on $\mathbf{x}$. This is often generically stated as

$$C(\mathbf{x}) = C_i, \quad \text{for} \quad f_i(\mathbf{x}; \Lambda) = \max_j f_j(\mathbf{x}; \Lambda) \tag{7}$$

where $C(\cdot)$ denotes a classification operation. The ith discriminant function $f_i(\mathbf{x}; \Lambda)$ is defined as:

$$f_i(\mathbf{x}; \Lambda) = ln(g_2(\mathbf{x}; \lambda_i)) \tag{8}$$

where $g_2(\mathbf{x}; \lambda_i)$ is defined in the companion paper submitted to this conference [4]. A misclassification measure is then introduced in the second step to embed the decision process in a function form. While there are many alternatives, one misclassification measure for each class i can be defined as:

$$d_i(\mathbf{x}; \Lambda) = -f_i(\mathbf{x}; \Lambda) + \ln \left[\frac{1}{M-1} \sum_{j \neq i} e^{f_j(\mathbf{x}; \Lambda)\zeta} \right]^{\frac{1}{\zeta}}, \tag{9}$$

where ζ is a positive value. This misclassification measure is a quantity that indicates whether an input token $\mathbf{x}$ of the ith class will be misclassified according to the decision rule of (7), implemented by the classifier parameter set Λ. $d_i(\mathbf{x}; \Lambda)$ measures the certainty of misclassifying $\mathbf{x}$. By varying the value of ζ, one can, to a degree, take all the competing classes into consideration in the process of optimizing the classifier parameter Λ.

The third step is to define the loss function $l_i(\mathbf{x}; \Lambda)$ for misclassifying a character of class i. One possibility is to choose

$$l_i(\mathbf{x}; \Lambda) = l_i(d_i(\mathbf{x}; \Lambda)) = \frac{1}{1 + e^{-\xi d_i(\mathbf{x}; \Lambda)}}, \tag{10}$$

where ξ is a positive value. Thus, for any unknown $\mathbf{x}$, the classifier performance is measured by:

$$l(\mathbf{x}; \Lambda) = \sum_{i=1}^{P} l_i(\mathbf{x}; \Lambda) 1(\mathbf{x} \in C_i), \tag{11}$$

where $1(\cdot)$ is an indicator function:

$$1(\hbar) = \begin{cases} 1 & \text{if } \hbar \text{ is true} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

and C_i is used to denote both the class and its data set.

At this point, the objective function of discriminative training is defined as the following *empirical average cost* for the entire training data set:

$$L(\Lambda) = \frac{1}{W} \sum_{p=1}^{P} \sum_{q=1}^{W_p} l_p(\mathbf{x}^{(p,q)}; \Lambda) \tag{13}$$

where $W = \sum_{p=1}^{P} W_p$ is the total number of training samples. By controlling parameters ζ and ξ and minimizing this *empirical average cost*, one can have an accurate approximation to the minimization of the classification error probability on the training set. The actual objective function adopted is

$$F(\Lambda) = -L(\Lambda) \tag{14}$$

3. OPTIMIZATION WITH THE GRADIENT PROJECTION METHOD

The training problem of a contextual model is just a general optimization problem with linear constraints. There are many general purposed procedures for linearly constrained optimization (e.g., [11]), that can be used to solve the training problem. However, in the case of contextual model training, there are typically a few thousand parameters to adjust and the evaluation of the objective function is always very

time-consuming. An important consideration that training a contextual model differs from the general problem that standard optimization techniques are designed for, is that one cannot afford to take nearly as many optimizing steps along each search direction as one would normally take. For example, a quasi-Newton method normally requires a number of steps roughly equal to the dimension of the parameter space to get a good estimate of the Hessian, which is usually out of the question in the case of contextual model training. As a result of this peculiarity, the simple gradient projection method (GPM) can be a reasonable and competitive choice.

Historically, the gradient projection method was proposed and extensively analyzed by Rosen [12]. Its main idea is to search along the projection of the gradient on the constraint space for a local maximum. The method has been tailored for estimation of hidden Markov model parameters in [5, 6]. It is also suitable for contextual model parameter estimation because of their linear constraint properties. In this paper, this optimization technique is adopted for contextual model training.

Given the above objective function, one now can apply the GPM to discriminatively adjust the model parameters Λ to equivalently minimize the cost function. Apart from the evaluation of $F(\Lambda)$, the computation of its derivatives is also needed in the GPM. To compute the gradient $\nabla F(\Lambda)$, let θ_k denote a particular parameter of model k, then one has

$$\frac{\partial F(\Lambda)}{\partial \theta_k} = -\frac{1}{W}\sum_{p=1}^{P}\sum_{q=1}^{W_p}\frac{\partial l_p(\mathbf{x}^{(p,q)};\Lambda)}{\partial \theta_k}. \tag{15}$$

After some algebraic manipulation, one gets

$$\begin{aligned}\frac{\partial F(\Lambda)}{\partial \theta_k} = {} & \frac{\xi}{W}\sum_{q=1}^{W_k}\{l_k(\mathbf{x}^{(k,q)};\Lambda)\cdot \\ & [1-l_k(\mathbf{x}^{(k,q)};\Lambda)]\cdot\frac{\partial f(\mathbf{x}^{(k,q)};\lambda_k)}{\partial \theta_k}\} \\ -\frac{\xi}{W}\sum_{p\neq k}^{M}\sum_{q=1}^{W_p}\{ & l_p(\mathbf{x}^{(p,q)};\Lambda)[1-l_p(\mathbf{x}^{(p,q)};\Lambda)]\cdot \\ & \frac{e^{g_k(\mathbf{x}^{(p,q)};\lambda_k)\zeta}}{\sum_{j\neq p}^{P} e^{g_j(\mathbf{x}^{(p,q)};\lambda_j)\zeta}}\cdot\frac{\partial f(\mathbf{x}^{(p,q)};\lambda_k)}{\partial \theta_k}\}\end{aligned} \tag{16}$$

By substituting the relevant derivatives of $\frac{\partial f(\mathbf{X};\lambda_k)}{\partial \theta_k}$ into the above equation, the final derivatives used in the gradient projection method will be obtained. The explicit expressions for the derivatives are:

$$\frac{\partial f}{\partial \beta_k} = \sum_{i,j}\frac{b_k(\mathbf{o}_{i,j})Y_{i,j,k}}{X_{i,j}} \tag{17}$$

$$\frac{\partial f}{\partial a_{k,l}^{m,n}} = \sum_{i,j}\frac{\beta_k b_k(\mathbf{o}_{i,j})b_l(\mathbf{o}_{i',j'})Y_{i,j,k}^{m,n}}{X_{i,j}} \tag{18}$$

$$\begin{aligned}\frac{\partial f}{\partial b_{k,t}} = {} & \sum_{i,j}\sum_{k'}^{K}[1(k'\equiv k)1(\mathbf{o}_{i,j}\equiv v_t)\cdot \\ & \beta_k Y_{i,j,k}+\beta_{k'}b_{k'}(\mathbf{o}_{i,j})\cdot \\ & \sum_{(i',j')\in\eta_{i,j}}1(\mathbf{o}_{i',j'}\equiv v_t)a_{k',k}^{m,n}Y_{i,j,k'}^{m,n}]/X_{i,j}\end{aligned} \tag{19}$$

where $\eta_{i,j}$ is the neighborhood of pixel (i,j).

$$Y_{i,j,k} = \sum_{(i',j')\in\eta_{i,j}}\sum_{l=1}^{K}a_{k,l}^{m,n}b_l(\mathbf{o}_{i',j'}) \tag{20}$$

$$\begin{aligned}Y_{i,j,k}^{m,n} = {} & \sum_{(i'',j'')\in\eta_{i,j}}\sum_{l=1}^{K}1(m',n'\neq m,n)\cdot \\ & a_{k,l}^{m',n'}b_l(\mathbf{o}_{i'',j''})\end{aligned} \tag{21}$$

$$X_{i,j} = \sum_k \beta_k b_k(\mathbf{o}_{i,j})Y_{i,j,k} \tag{22}$$

$i' = i+m;\ j' = j+n;\ i'' = i+m';\ j'' = j+n'$

and $1(\cdot)$ is an indicator function as in (12).

4. EXPERIMENTAL RESULTS

In this study, the same experimental setup has been adopted as in [4] where the task is the recognition of 10 highly similar hand written Chinese characters. The parameters ζ and ξ used in equation (9) and (10) are respectively set to be ∞ and 0.1. When ζ approaches ∞, the misclassification measure for each class i becomes:

$$d_i(x) = -f_i(x;\lambda) + f_j(x;\lambda) \tag{23}$$

where C_j is the class with the largest discriminant value among those classes other than C_i.

The training process starts with initial models well trained by the DD algorithm [3, 4]. After 20 iterations, the close and open test recognition rates are 99.47% and 93.80% respectively. Figure 1 illustrates the rate of convergence of the discriminative training process in terms of the objective function and close and open test results. About 85% error rate reduction is achieved by the discriminative training for the close test and 35% for the open test.

The very high close test rate suggests the power of discriminative training in tuning the model parameters to the training data. This is not accomplished on the expense of model generalization to unseen samples because effectively, the model of each

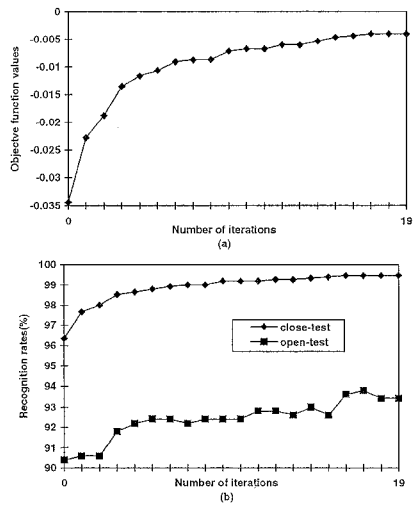

Figure 1: Learning curves of discriminative training based on GPM with DD-trained initial models: (a) objective function values, (b) close-test and open-test recognition rate (% correct).

character is now trained with not only its own samples but also those of the similar characters. This makes the training much more robust. Like any local optimization procedure, the final result of GPM-based training highly depends on the initial values of the contextual model parameters. This also suggests that the algorithm based on GPM is most attractive for final "tune-up" and will usually be bootstrapped from well-trained initial models trained with other methods such as DD algorithm.

5. CONCLUSION

The capability of contextual modeling of complex and variant patterns like hand written Chinese characters has been demonstrated. The performance of such a recognizer can be further upgraded by parameter fine-tuning through optimizing a minimum classification error oriented objective function by means of a gradient projection algorithm.

REFERENCES

[1] Q. Huo and C. Chan, "Contextual vector quantization for speech recognition with discrete hidden Markov model," *Pattern Recognition*, Vol. 28, pp.513-517, 1995.

[2] Q. Huo and C. Chan, "A study on the use of bi-directional contextual dependence in Markov random field-based acoustic modeling for speech recognition," *Computer Speech and Language*, Vol. 10, pp. 95-105, 1996.

[3] S.-L. Leung, P.-C. Chee, C. Chan and Q. Huo, "Contextual vector quantization modeling of hand-printed Chinese character recognition," *Proc. ICIP-95*, Washington, D.C., Oct. 1995, pp. 432-435.

[4] Y. Xiong and C. Chan, "Contextual modeling of hand written Chinese character for recognition (I) - a comparative study," submitted to *DSP'97*, Santorini, Greece, July 2-4, 1997.

[5] Q. Huo and C. Chan, "The gradient projection method for the training of hidden Markov models," *Speech Communication*, Vol. 13, pp.307-313, 1993.

[6] Q. Huo and C. Chan, "Discriminative training of HMM based speech recognizer with gradient projection method", *Proc. Eurospeech-95*, Madrid, Spain, September 1995, pp.101-104.

[7] A. Nadas, "A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-31, No. 4, pp.814-817, 1983.

[8] P. F. Brown, *The Acoustic Modeling Problem in Automatic Speech Recognition*, Ph.D. thesis, Department of Computer Science, Carnegie Mellon University, 1987.

[9] Y. Ephraim, A. Dembo and L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," *IEEE Trans. on Information Theory*, Vol. 35, No. 5, pp.1001-1013, 1989.

[10] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. on Signal Processing*, Vol. 40, No. 12, pp. 3043-3054, 1992.

[11] P. E. Gill, W. Murray and M. H. Wright, *Practical Optimization*, Academic Press, 1981.

[12] J. B. Rosen, "The gradient projection method for nonlinear programming–part i: linear constraints," *SIAM*, Vol. 8, No. 1, pp.181-217, 1960.