A PROBABILISTIC MODEL FOR CURSIVE HANDWRITING RECOGNITION USING SPATIAL CONTEXT

Jigang Wang, Predrag Neskovic, Leon N Cooper

Department of Physics
Institute for Brain and Neural Systems
Brown University, Providence, RI 02912
jigang@brown.edu, pedja@brown.edu, Leon_Cooper@brown.edu

ABSTRACT

In this work we introduce a probabilistic model that utilizes spatial contextual information to aid recognition when dealing with ambiguous segmentations of handwritten patterns. The recognition problem is formulated as an optimization problem in a Bayesian framework by explicitly conditioning on the spatial configuration of the letters. As a consequence, and in contrast to HMMs, the proposed model can handle duration modeling without an increase in computational complexity. We test the model on real-world handwriting dataset and discuss several factors that affect the recognition performance.

1. INTRODUCTION

In many pattern recognition problems it is often very difficult to reliably identify local regions of a pattern due to ambiguous or insufficient information they contain. In such situations contextual information can greatly aid recognition.

In the statistical pattern recognition approach, the contextual information is usually encoded in some sort of joint distribution function of a local part and its surroundings. The joint distribution function is either parametrized based on the *a priori* knowledge of the problem domain or approximated by a consistent estimator. During the past decade, Hidden Markov models (HMMs) [1] have become the dominant model for speech recognition [2]. The power of HMMs lies in the fact that they provide a nearly universal parametrization of stationary processes, that the maximum-likelihood estimator (MLE) is known to be consistent and that there exists computationally efficient procedure to estimate the parameters [3]. Due to similarities between handwriting, especially cursive script, and speech, HMMs have been successfully applied to document analysis [4]-[6]. However, HMMs have several limitations: weak duration modeling, the assumption of conditional independence of observations given the state sequence, and the restrictions on feature extraction imposed by frame-based observation [7]. Although the semi-Markov models overcome the limitations by introducing explicit duration distributions and segment observation models, the computational complexity of such extension is high and the parameter estimation problem is much more difficult compared to standard HMMs.

In this paper, we present a novel probabilistic model for cursive handwriting recognition using spatial cues. The recognition problem is formulated as an optimization problem in a Bayesian framework by explicitly conditioning on the spatial configuration of the letters. As a consequence, and in contrast to segmental HMMs, the proposed model can handle duration modeling without an increase in computational complexity.

The rest of the paper is organized as follows. In section 2, we give a brief overview of the handwriting recognition system and introduce the problem we are going to address. In section 3, we present the probabilistic model for word representation and formulate the recognition problem in the maximum *a posteriori* (*MAP*) probability framework. In section 4, we describe how the system is implemented on a cursive handwriting dataset and discuss the experimental results. Concluding remarks are given in section 5.

2. SYSTEM OVERVIEW

An overview of the system is illustrated in Fig. 1. The input to the system is a raw data file which represents a handwritten word and contains the x and y positions of the pen recorded every 10 milliseconds. This input signal is first transformed by a preprocessor into strokes, which are defined as lines between points with zero velocity in the y direction. Each stroke is characterized by a set of features. The extracted features for each of the strokes of the input word are then fed into a segmentation network. The segmentation network we use is a Time Delay neural network

This work was supported in part by ARO under grants DAAD19-01-1-0754 and W911NF-04-1-0357.

0-7803-8874-7/05/$20.00 ©2005 IEEE

(TDNN) based on the weight sharing technique proposed by Rumelhart [8][9].

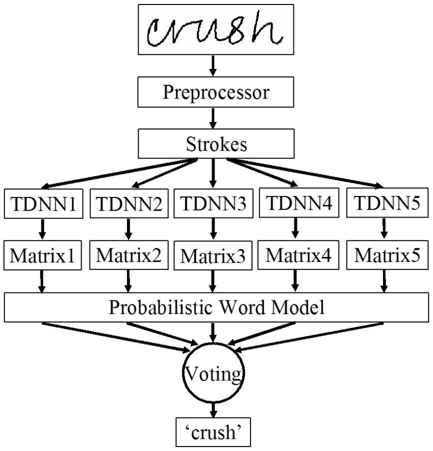

Fig. 1. System Overview.

The output of a segmentation network is a matrix M, called detection matrix. Each detection matrix has 26 rows, with each row corresponding to a lowercase letter in the english alphabet. Its column number is determined by the number of strokes in the input pattern and the structure of the network used. If the detection matrix M has L columns, it means that the input pattern is divided into L segments. A letter detector that receives input from the i-th segment outputs the probability that this segment represents a particular letter positioned at location x_i.

The output of the segmentation network represents an over-segmentation of the input word pattern. Due to large variability of the shape of each letter, at any position x_i, when viewed in isolation, the identity of the letter is often ambiguous and misleading. The goal of the paper is to present a probabilistic model that exploits the spatial constraint of letters to reduce ambiguity.

3. THE PROBABILISTIC MODEL

3.1. The Objective Function

In the maximum *a posteriori* probability framework, the goal of word recognition is to find the word hypothesis $\vec{W}^* = (w_1, \ldots, w_n)$ with maximum a posteriori probability, i.e.,

$$\vec{W}^* = \arg\max_{\vec{W}} P(\vec{W}|O), \tag{1}$$

where O is the observed input pattern. Note that a word is completely determined by specifying its constituent letters $\vec{W} = (w_1, \ldots, w_n)$ and their locations $\vec{X} = (x_1, \ldots, x_n)$ within the detection matrix, where w_i denotes the i^{th} letter within the word and x_i its position. Therefore the word recognition problem reduces to the problem of finding $(\vec{W}, \vec{X})^*$ such that the *a posteriori* probability $P(\vec{W}, \vec{X}|O)$ is maximized, namely,

$$(\vec{W}, \vec{X})^* = \arg\max_{(\vec{W},\vec{X})} P(\vec{W}, \vec{X}|O). \tag{2}$$

The problem is intractable without further reduction. In the following, we use the Bayes rule and make a few necessary simplifications to decompose the *a posteriori* probability into a manageable form. We first give the decomposition:

$$\begin{aligned}\log P(\vec{W}, \vec{X}|O) &= \log P(\vec{X}|O) + \log P(\vec{W}|\vec{X}, O)\\ &= -\sum_{i=1}^{n} V(x_{i-1}, x_i) + \sum_{i=1}^{n} \log P(w_i|x_i, O) + nC. \end{aligned} \tag{3}$$

The first equality follows from the Bayes rule. To model $\log P(\vec{X}|O)$, we assume that the position of a letter in a word is conditionally independent of the positions of other letters given the positions of its immediate neighbors, i.e., $\vec{X}$ is modeled as a first-order Markov chain. In our model, we further assume that the width of successive characters can be modeled as a Gaussian random variable, i.e.,

$$V(x_{i-1}, x_i) = \frac{(x_i - x_{i-1} - d)^2}{\sigma^2}, \tag{4}$$

where d is interpreted as the average width between two neighboring letters, σ is the standard deviation and $x_0 \equiv 0$ is a dummy variable.

Given the specific spatial configuration $\vec{X}$ of the letters, we assume that the detection of a letter is conditionally independent of the detection of other letters, i.e.,

$$\log P(\vec{W}|\vec{X}, O) = \sum_{i=1}^{n} \log P(w_i|x_i, O).$$

This term can be directly obtained from the detection matrix M by identifying $P(w_i|x_i, O)$ with $M_{w_i x_i}$.

Note that there is a third term nC in Eq. 3. Part of this term is due to the normalization constant of the Gaussian random variable in Eq. 4. However, it is also used to compensate for the fact that, compared to short words, long words tend to have smaller probabilities because of the larger number of terms. It will become important when we compare the *a posteriori* probabilities of word hypotheses with different lengths.

In summary, the word recognition problem reduces to the following optimization problem

$$\arg\max_{(\vec{W},\vec{X})} \sum_{i=1}^{n} [-\frac{(x_i - x_{i-1} - d)^2}{\sigma^2} + \log P(w_i|x_i, O) + C], \tag{5}$$

where d, σ and C are free parameters of the model that will be learned from training samples by cross-validation. Although our model appears to be similar to semi-Markov models, we want to stress their differences. As extensions to conventional HMMs, semi-Markov models are still based on the notion of transition probability. In semi-Markov models, there are as many terms as the number of observations and the coupling between transition and emission probability makes the parameter estimation difficult. However, in our model, by conditioning on the spatial configuration of letters, the *a posteriori* probability is decomposed into two terms: one term, modeled as a Markov chain $\vec{X}$ in the spatial domain, describes the spatial configuration of the word; the other term, conditioned on the realization of the spatial configuration, characterizes the local appearance of each component, which can be easily obtained by a letter classifier. This decomposition makes the parameter estimation problem much easier.

3.2. Search Strategy

Given the objective function, the search for the *MAP* configuration $(\vec{W}, \vec{X})^*$ can be solved in two steps. In the first step, for each valid word hypothesis $\vec{W}$, select the spatial configuration $\vec{X}$ of the letters that gives the highest score; the highest score will be taken as the optimal matching score for the particular word hypothesis. In the second step, from all the valid words in the dictionary choose the word hypothesis with the highest optimal matching score.

Under the Markov assumption, for a given M and word hypothesis $\vec{W} = (w_1, \ldots, w_n)$ the search for the optimal spatial configuration $\vec{X}$ reduces to

$$\arg \max_{\vec{X} \in \{0,\ldots,L-1\}^n} \sum_{i=1}^{n} [-V(x_{i-1}, x_i) + \log P(w_i|x_i, O)]. \tag{6}$$

Standard Dynamic Programming (DP) techniques, such as the Viterbi algorithm, can be easily applied to give the optimal solution. If L is the average column number of the detection matrices, n the average word length, K the number of words to be recognized and N the size of the dictionary, the time complexity of the DP algorithm is $O(KNnL^2)$.

3.3. Voting

With the same training handwriting dataset, we train several segmentation networks with different sizes of the receptive fields of the letter detectors and their overlaps. Therefore, for the same input pattern, fed into different segmentation networks, we have several different detection matrices (see Fig. 1). We apply the probabilistic model to each detection matrix to find the word hypothesis with the highest score. The word with the majority vote is chosen as the final output of the system with ties broken at random.

4. IMPLEMENTATION AND RESULTS

The experiments were carried out on a handwriting dataset originally obtained from David Rumelhart [8]. The dataset consists of words written by 100 different writers. The size of the dictionary is 1000 words. The segmentation networks were trained on 70 writers and an independent group of writers was used as a cross-validation set [10].

4.1. Baseline Model

To see how much the spatial information improves the recognition rate, we implemented a baseline model that does not take into account detailed spatial information as long as the left-right order of letters in a word is observed. To be specific, the optimization form of the baseline model is formulated as

$$\arg \max_{(\vec{W}, \vec{X})} [\sum_{i=1}^{n} \log P(w_i|x_i, M) + nC_b] \tag{7}$$

subject to $x_i \leq x_{i+1}, \forall i = 1, \ldots, n-1$, where C_b is determined by cross-validation.

4.2. Results and Discussion

We tested our model on 6 datasets, where each dataset consists of 905 words written by a different writer. The recognition rates of our model and the baseline model obtained with the Dynamic Programming algorithm on the 6 writers are reported in Table 1. The results show an average of 7.3 percent improvement on the recognition accuracy by incorporating the spatial information.

Table 1. Recognition rates obtained with DP algorithm

Writer	**aeb**	**ak**	**cdb**	**lml**	**ses**	**yuko**
Baseline	85.0	89.2	84.4	88.1	88.7	64.8
Our Model	92.8	93.6	90.9	95.0	92.3	79.3
Improvement	7.8	4.4	6.5	6.9	3.6	14.5

We also compared the recognition rates obtained with the majority voting scheme with the mean and highest recognition rates obtained on 5 individual segmentation networks with different sizes of the receptive fields and their overlaps. Detailed results are reported in Table 2. The voting method improves the mean recognition rate by 4.7 percent and improves the best single recognition rate by 3.1 percent.

In cursive handwriting recognition, one of the modeling difficulties lies in the fact that the word length is not available directly from the over-segmented detection matrix. In our model, comparison of word hypotheses with different lengths is modulated by the penalty term C (see Eq. 3).

Table 2. Effect of voting

Writer	aeb	ak	cdb	lml	ses	yuko
Voting	92.8	93.6	90.9	95.0	92.3	79.3
Mean	87.9	90.8	86.3	90.5	89.2	71.1
Max	90.5	91.9	87.3	91.6	91.6	72.6

Cross-validation results showed that there exists a unique value of C that achieves the highest performance and the value is stable across different writers. To evaluate the influence of word length, we compared the recognition rates of our model with the results obtained by *assuming* that the input word lengths are available (therefore, only word hypotheses with the correct lengths are considered and the penalty term does not come into play). The results are reported in the Table 3. The results show an improvement of 3.6 percent by restricting the search space into word hypotheses with the correct lengths, implying that there is still room for further improvement.

Table 3. Effect of knowing the word length

Writer	aeb	ak	cdb	lml	ses	yuko
n Unknown	92.8	93.6	90.9	95.0	92.3	79.3
n Known	95.7	96.1	94.4	96.9	95.0	87.2
Difference	2.9	2.5	3.5	1.9	2.7	7.9

5. CONCLUSION

In this paper, we introduced a probabilistic model that utilizes spatial contextual information to aid recognition of cursive handwriting. By explicitly conditioning on the spatial configuration of the letters, the *a posteriori* probability of a word is decomposed into two terms using Bayesian inference: one term encodes the spatial relationships between letters and the other is the conditional probability of a word given the specific spatial configuration. In contrast to segmental HMMs, the proposed model deals with duration modeling easily without increasing the computational complexity. We also showed how combining results obtained with different sizes of the segments can be used to improve the recognition rate.

Acknowledgments

Jigang Wang would like to thank Stuart Geman, Nathan Intrator, Nicola Neretti and other memebers of the instute for helpful comments.

6. REFERENCES

[1] Baum, L. E. & Petrie, T. (1966) Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, **37**, pp. 1554-1563.

[2] Rabiner, L. R. (1989) A tutorial on hidden Markov model and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2), pp. 257-286.

[3] Künsch, H. , Geman, S. & Kehagias, A. (1995) Hidden Markov random fields. *Annals of Applied Probability*, **5**(3), pp. 577-602.

[4] Gillies, A. M. (1992) Cursive word recognition using Hidden Markov models. *Proceedings of the 5th U.S. Postal Service Advanced Technology Conference*, pp. 557-562.

[5] Chen, M. Y., Kundu, A. & Zhou, J. (1994) Off-line handwritten word recognition using a Hidden Markov model type stochastic network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **16**(5), pp. 481-496.

[6] El-Yaccoubi, A. , Sabourin, R. , Gilloux, M. & Suen, C. Y. (1999) Off-Line handwritten word recognition using Hidden Markov Models. In *Knowledge-based Intelligent Techniques in Character Recognition*, Jain and Lazzerini (Eds.), CRC Press, pp. 193-229.

[7] Ostendorf, M. , Digalakis, V. , Kimbal, O. A. (1996) From HMMs to Segment Models: A unified view of stochastic modeling for speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 4, pp. 360–378.

[8] Rumelhart, D. E. (1993) Theory to practice: A case study - recognizing cursive handwriting. In E. B. Baum, editor, *Computational Learning and Cognition: Proceedings of the Third NEC Research Symposium.* SIAM, Philadelphia.

[9] Schenkel, M. , Guyon, I. & Henderson, D. (1995) Online cursive script recognition using time delay neural networks and hidden Markov models. *Machine Vision and Applications*, **8**, pp. 215-223.

[10] Neskovic, P. , Davis, P. C & Cooper, L. N (2001) Interactive parts model: an application to recognition of on-line cursive script. *NIPS 13*.