An Off-Line Large Vocabulary Hand-Written Chinese Character Recognizer

Pak-Kwong WONG

Department of Computer Science
The University of Hong Kong
Pokfulam Road, Hong Kong
pkwong@cs.hku.hk

Chorkin CHAN

Department of Computer Science
The University of Hong Kong
Pokfulam Road, Hong Kong
cchan@cs.hku.hk

Abstract

An off-line hand-written Chinese character recognizer based on Contextual Vector Quantization (CVQ) supporting a vocabulary of 4,616 Chinese characters, alphanumerics and punctuation symbols has been reported. Trained with a sample for each character from each of 100 writers and tested on texts of 160,000 characters written by another 200 writers, the average recognition rate is 77.2%. Two statistical language models have been investigated in this study. Their performances in terms of their capabilities in upgrading the recognition rate by 8.8% and 12.0% respectively when used as post-processors of the recognizer will be reported in this paper.

1 Introduction

Chinese characters are complex patterns of strokes. The bit-map of a character image can be segmented into a number of regions each of which consists of either purely white or purely black pixels. An unknown character image is recognized by identifying its regions to that of the templates. The structural information of an image in terms of the inter-relationship between its regions is represented statistically. The location and size of a region are stochastic. Even if a pixel is known to belong to a particular region, the cellular features [1] considered as a feature vector, observed at the pixel are still stochastic and different feature vectors can be observed at different pixels of the same region. A region is not characterized by just the distribution of feature vectors observed at its pixels, but by the stochastic relationship between it and its neighbor regions also as well as its location and size. The totality of such stochastic properties of a region defines a codeword. Hence, a codeword and a region are synonymous. A character as a collection of regions corresponds therefore to a codebook.

2 CVQ Character Recognizer

A character image is abstracted into a matrix of cellular feature vectors $\mathbf{O} = [\mathbf{o}_{i,j}]$ with $\mathbf{o}_{i,j}$ observed at pixel (i, j). Each $\mathbf{o}_{i,j}$ is modeled as a realization of a random vector observable in $z_{i,j}$ which is the region where pixel (i, j) is located. $z_{i,j}$ takes one of the K qualitative values $\{G_1, G_2, \cdots, G_K\}$ each of which is a region of the character. Each region is characterized by three sets of attributes: $\{Pr(z_{i,j} = G_k),\ Pr(\mathbf{o}_{i,j} \mid G_k),\ Pr(z_{m,n} = G_l \mid z_{i,j} = G_k)\}$. m and n are integers equal to $i-1$ to $i+1$ and $j-1$ to $j+1$ respectively indexing the regions of the immediate neighbors of pixel (i, j). If these attributes of a region are fitted into the framework of a codeword, then a character is modeled by a codebook. Matching an unknown to a character becomes quantizing $\mathbf{O}$ with the codebook of the character. $Pr(z_{m,n} = G_l \mid z_{i,j} = G_k)$ supplies the contextual information and that leads to the name Contextual Vector Quantization.

$\mathbf{O}$ is quantized by quantizing each of its pixels individually. Pixel (i, j) is quantized to $z_{i,j}$ in order to maximize the posterior probability $\Pr(z_{i,j}|\mathbf{O})$. In order to reduce the complexity of the problem, $z_{i,j}$ is chosen to maximize $\Pr(z_{i,j}|\mathbf{o}_{i,j}, \mathbf{o}_{\eta_{i,j}})$, where $\eta_{i,j}$ is the immediate neighborhood of pixel (i, j). Under the assumption that feature vectors in the same neighborhood are related to each other through the regions they belong to only, one then has this posterior probability proportional to:

$$\sum_{z_{\eta_{i,j}}} \left\{ \Pr(z_{i,j}, z_{\eta_{i,j}}) \cdot \prod_{(m,n)\in\eta^{+}_{i,j}} Pr(\mathbf{o}_{m,n}|z_{m,n}) \right\} \quad (1)$$

where the summation is over all admissible values of

0-8186-8183-7/97 $10.00 © 1997 IEEE

$z_{\eta_{i,j}}$ defining the region membership of the pixels in the prescribed neighborhood $\eta_{i,j}$ of pixel (i,j). $\eta^+_{i,j}$ is the union of $\eta_{i,j}$ and (i,j). Even with this simplification, analytical progress is barred in general, because $\Pr(z_{i,j}, z_{\eta_{i,j}})$ is unavailable in closed form. For further simplification, it is assumed that $z_{m,n}$'s, where $(m,n) \in \eta_{i,j}$, are mutually independent given $z_{i,j}$. So,

$$\Pr(z_{i,j}, z_{\eta_{i,j}}) = \Pr(z_{i,j}) \cdot \prod_{(m,n)\in\eta_{i,j}} \Pr(z_{m,n}|z_{i,j}) \quad (2)$$

A CVQ method can be derived as follows. Given a character image with observed feature vectors $[\mathbf{o}_{i,j}]$, assign each $\mathbf{o}_{i,j}$ to region G_k if

$$G_k = argmax_{z_{i,j}} Pr(z_{i,j}) \cdot Pr(\mathbf{o}_{i,j}|z_{i,j}) \cdot$$
$$\prod_{(m,n)\in\eta_{i,j}} \sum_{z_{m,n}} \Pr(z_{m,n}|z_{i,j}) \cdot Pr(\mathbf{o}_{m,n}|z_{m,n}) \quad (3)$$

where the term on the second line of Eq.(3) represents the contribution of contextual information. The argument of the $argmax_{z_{i,j}}$ function:

$$Pr(z_{i,j}) \cdot Pr(\mathbf{o}_{i,j}|z_{i,j}) \cdot$$
$$\prod_{(m,n)\in\eta_{i,j}} \sum_{z_{m,n}} \Pr(z_{m,n}|z_{i,j}) \cdot Pr(\mathbf{o}_{m,n}|z_{m,n}) \quad (4)$$

is a pseudo-likelihood measurement of quantizing $\mathbf{o}_{i,j}$ to region G_k.

Upon matching an unknown image to a character template ω for identification, regions of the unknown image are matched to regions of ω. That in turn, is accomplished by identifying a region of ω for each pixel of the unknown image to be quantized to. That avoids segmenting an unknown image into regions explicitly and then matching them as a random graph as in [2]. This process of region identification for each pixel considers not just the pixel in question, but its neighboring pixels and the most suitable regions they belong to as well. Thus, recognizing a character becomes identifying the codebook that yields the minimum quantization error (measured in terms of the inverse of a pseudo-likelihood function) to the unknown image.

This algorithm has been implemented in an offline writer independent hand-written character recognizer supporting a vocabulary of 4,616 Chinese characters, alphanumerics and punctuation symbols [3]. The codebook for each character is trained with 100 samples written by 100 writers. When tested on 160,000 characters written by another 200 writers, the recognition rate is 77.2%

3 Post-Processing Language Models

If the input is a syntactically and semantically sound sequence of characters, its linguistic information can provide a useful basis for improving the recognition rate [4]. The second phase of the character recognizer is thus a language model which endows the recognizer with linguistic (just statistical at present) knowledge of Chinese. For each character image, the language model chooses the most suitable one out of the n-best candidates proposed by the image recognizer in order to arrive at a sequence of characters which is linguistically sound according to some criteria. There are two statistical language models experimented in this study as a post-processor of the image recognizer. They select a candidate according to its capability to form words with its neighboring images.

3.1 Lexical Analysis of a Lattice of n-best Candidates

The lexical analytic statistical language model bases on the usage frequency of each word in a large lexicon. This lexicon must cover most, if not all, of the Chinese words actively used in modern texts such as journals, newspapers, and literature. In order to determine the statistics of word-pairs, to enrich the lexicon of its vocabulary and to improve the estimates of word usage frequencies, a large Chinese text corpus of over 63 million characters has been acquired. The first step towards gathering such statistics is to segment text lines into words because different from texts in English, there is no explicit word marker in Chinese texts.

Maximum matching [5] is one of the most popular structural segmentation algorithms for Chinese texts. This method favors long words and is a greedy algorithm in nature, hence, sub-optimal. Segmentation may start from either end of the line without any difference in segmentation results. In this study, the forward direction is adopted. The major advantage of maximum matching is its efficiency while its segmentation accuracy can be expected to lie around 95%.

Most Chinese linguists accept the definition of a word as the minimum unit that is semantically complete and can be put together as building blocks to form a sentence. However, in Chinese, words can be united to form compound words, and they in turn, can combine further to form yet higher ordered compound words. As a matter of fact, compound words are extremely common and they exist in large numbers. It is impossible to include all compound words into the lexicon but just to keep those which are frequently used and have closely united word components. A lexicon, WORDDATA, was acquired from the Institute of Information Science, Academia Sinica in Taiwan. There are 78,410 word entries in this lexicon, each associated with a usage frequency. Due to cultural differences of the two societies, there are many words encountered in the text corpus but not in the lexicon. The latter must therefore be enriched before it can be applied to perform any lexical analysis. The first step towards this end is to merge a lexicon constructed in China into this one made in Taiwan, increasing the number of word entries to 85,855. This extended lexicon is then applied to segment the text corpus into words. In this process, when a word of a single character is encountered, word usage frequencies will be considered to decide if the single character should not be combined with it neighboring characters to form other words on the expense of the length of neighboring words. In this word segmentation process, words used in the text corpus but not found in the lexicon will be considered to be added to the latter which is eventually enriched to encompass 87,326 words.

The image recognizer supplies the n-best candidates for each character image scanned. A line of text as a sequence of m images delimited by a pair of punctuation symbols correspond to m by n candidates. Starting from the first image position, the longest word that can be formed with a candidate of the image as the first character of the word is accepted. This repeats starting from the next image position lying beyond the last image of the word just formed until the end of the line is reached.

As n increases, the number of coincidental word formations increases also, thus bringing down the recognition rate instead of upgrading it. On the other hand, for pages poorly recognized, n must be large enough to include the true candidate. A compromise on the optimal choice of n is reached by experimenting the effect of n on the recognition rates on another 100 pages earmarked for language model tuning. Consequently, n is chosen to be 6. The recognition rate over the test text of 160,000 characters is upgraded to 86% from 77.2%.

3.2 A Language Model of Word Class Bigram Statistics

The limitation of maximum matching word segmentation as a language model is its failure to capture the inter-dependence of words in a line of text. The use of bigram statistics in a language model is a step towards overcoming this shortcoming. Since there are over 80,000 words in the lexicon, the number of parameters in such a language model will be astronomical. A common practice is to employ the bigram statistics between word-classes instead. If a sequence of character images $\mathbf{o}_1, ..., \mathbf{o}_T$ is segmented into $\mathbf{o}_1^{w_1}, ..., \mathbf{o}_{k_1}^{w_1}, \mathbf{o}_1^{w_2}, ..., \mathbf{o}_{k_2}^{w_2}, ..., \mathbf{o}_1^{w_h}, ..., \mathbf{o}_{k_h}^{w_h}$ correpsonding to a word sequence of $w_1, ..., w_h$ which in turn, belonging to word-classes $s_1, ..., s_h$ respectively, the soundness of the segmentation is measured in terms of:

$$L = p(s_0 \mid s_h) \prod_{i=1}^{h} p(s_i \mid s_{i-1}) p(\mathbf{o}_1^{w_i}, ..., \mathbf{o}_{k_i}^{w_i} \mid s_i) \quad (5)$$

where s_0 is a word-class of punctuation symbols appearing before and after the sequence of character images. $p(s_i \mid s_{i-1})$ and $p(s_0 \mid s_h)$ can be collected from the segmented text corpus while the suitability of $\mathbf{o}_1^{w_i}, ..., \mathbf{o}_{k_i}^{w_i}$ forming a word w_i in s_i is defined as:

$$p(\mathbf{o}_1^{w_i}, ..., \mathbf{o}_{k_i}^{w_i} \mid s_i) = p(w_i \mid s_i) \prod_{j=1}^{k_i} p(\mathbf{o}_j^{w_i} \mid c_j^{w_i}) \quad (6)$$

Here, word w_i is a character sequence $c_1^{w_i}, ..., c_{k_i}^{w_i}$. $p(w_i \mid s_i)$ is computed from the segmented text corpus. $p(\mathbf{o}_j^{w_i} \mid c_j^{w_i})$ is a measure of similarity between the observed image $\mathbf{o}_j$ and the character $c_j^{w_i}$ of the word w_i supplied by the image recognizer. The principle of dynamic programming is employed to determine the optimal segmentation of the character images into words.

Originally, words in WORDDATA are grouped into 192 syntactic/semantic word-classes with each word belonging to mostly one but up to four word-classes. In this investigation, each word is assigned the membership of the most important class indicated in WORDDATA. A natural and objective criterion in measuring the soundness of any clustering is that all members within a cluster should have a similar pattern of associations with all clusters. From the text corpus, the probability of observing word w_j of s_i placed before any word of class s_q, can be computed for all q. Associated with word w_j of s_i, there is therefore a probability vector $\mathbf{p}_j^{s_i}$ of 192 components, viz., $p_{j_k}^{s_i}$ for $k = 1, 2, ..., 192$. $p_{j_k}^{s_i}$ is the probability of seeing w_j of class s_i before any word of class s_k in an average line of the corpus. Since each word belongs to one class only in this investigation, there is no ambiguity if the superscript s_i is dropped in $\mathbf{p}_j^{s_i}$ and its components. These vectors are normalized so that they lie on the surface of a unit hyper-sphere.

The centroid $\mathbf{C_i}$ of class s_i is defined as a unit vector along the direction of the average probability vector of all the words (weighted by the prior probability of the word) of the class. With this concept in mind, the homogeneity of class s_i, a word-class of M_i words, can be defined as:

$$H_i = \sum_{j=1}^{M_i} P(w_j) \mathbf{C}_i \cdot \mathbf{p}_j^{s_i} \quad (7)$$

Various thresholds are chosen over a number of iterations so that any word-class with a homogeneity below it will be split into two as in ISODATA, except that the feature space is confined to a unit hyper-sphere surface. A newly formed word-class with a homogeneity still below the threshold will be further split repeatedly. At the end of an iteration corresponding to a particular homogeneity threshold, the effect of the word-class bigram statistics language model on the recognition of the 100 earmarked pages is measured. Finally, 470 word-classes are formed and bigram statistics between them are collected from the text corpus when the process converges.

The word-class splitting process discussed above is hierarchical. To mitigate the ill effect caused by any mis-classification of words in WORDDATA, after the number of classes has stabilized at 470, each word is re-assigned to a word-class whose centroid has the minimum inner product with the probability vector of the word. As soon as a word has been re-assigned, the centroids of the two word-classes affected are updated accordingly. With the newly defined word memberships, the probability vector of each word is re-computed by going over the text corpus again and so are the homogeneities of all word-classes consequently. This process repeats over several iterations. The average recognition rate is upgraded to 89.2% after the word-class reassignment process.

4 Man-Machine Contest

9 students from the Chinese Department of The University of Hong Kong have taken part in a contest against the language model as post-processors to upgrade the recognition rate. There are two sessions in the contest. In the first session, each student is given news lines of 1000 characters in three blocks of approximately equal sizes. These lines are selected at random from the text database of 160,000 characters. Instead of seeing the original character images of each line, a student sees a number (3, 6 or 9 for the three blocks respectively) of proposed candidates for each character image and the student selects one for each image in order to arrive at a line that makes sense. The students are told that 80% accuracy can be expected if the they always select the first candidate. They are also told that the true candidate can lie outside the proposed set of candidates but they are not allowed to select beyond the proposed ones.

In the second session, another 3 blocks of newlines of about the same amount are presented to each of the students in the same manner as above. This time, lines on the same page are related because they come from the same piece of news. Selecting the right candidate becomes easier for the students because they have a more global knowledge of the content.

The same materials are processed by the language model and the table below shows the performances of the students and the language model.

Type of Lines	Students Rec. Rate %			Lang. Model Rec. Rate %
	n=3	n=6	n=9	n=6
Unrelated	81.44	83.70	84.74	87.02
Related	82.31	85.26	91.37	86.73

Table 1: **Contest between Students and the Language Model**

5 Discussion

For the blocks of related news lines with 9 candidates for each image, the average recognition rate using the language model as a post processor is 87.13% compared to 91.37% achieved by the students. From this contest, one sees that human beings out-perform the software language model only when global information as well as abundant candidates for selection are available which cannot be utilized by a statistical language model effectively. Other than that, the language model is better than a human being in terms of helping the recognizer to select suitable character candidates.

References

[1] T.H. Hildebrandt & W.T. Liu, "Optical Recognition of Handwritten Chinese Characters: Advances since 1980", *Pattern Recognition*, Vol. 26, No. 2, pp. 205-225, 1993.

[2] A.K.C. Wong & M. You, "Entropy and Distance of Random Graphs with Application to Structural Pattern Recognition", *IEEE Trans. on PAMI*, Vol. 7, No. 5, pp. 599-609, 1985.

[3] S.L. Leung, P.C. Chee, Q. Huo & C. Chan; "Contextual Vector Quantization Modeling of Handprinted Chinese Character Recognition"; *Procs. of IEEE International Conference on Image Processing*, pp. 432-435, Washington, D.C., Oct. 1995.

[4] K.T. Lua, "From Character to Word – An Application of Information Theory", *Computer Processing of Chinese and Oriental Languages*, Vol. 4, No. 4, pp. 304-313, March 1990.

[5] Y. Liu, Q. Tan and K.X. Shen, "The Word Segmentation Rules and Automatic Word Segmentation Methods for Chinese Information Processing (in Chinese)", *Tsinghua University Press and Guangxi Science and Technology Press*, page 36, 1994.