CONTEXTUAL MODELING OF HAND WRITTEN CHINESE CHARACTER FOR RECOGNITION (I) - A COMPARATIVE STUDY

Yan Xiong and Chorkin Chan

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

ABSTRACT

A hand-written Chinese character off-line recognizer based on contextual modeling of 2D images has been constructed. Each character is modeled by a collection of regions and their contextual relations. A relaxation training algorithm of the model is now investigated. In addition, a mixture-region training algorithm considering each pixel as belonging to a mixture of regions instead of any particular one is studied.

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 consisting of either purely white or purely black pixels. A black region is a stroke or just a segment of it. An unknown character image is recognized by identifying its regions to that of a model. The structural information of an image in terms of the contextual relationship between its regions is represented statistically. Even if a pixel is known to belong to a particular region, the cellular features [1], 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 its size and relationship with its neighbor regions also, both of which are stochastic. Upon matching an unknown image to a character model λ, regions of the unknown image are matched to regions of λ. That in turn, is accomplished by identifying a region of λ for each pixel of the unknown image to belong to. That avoids segmenting an unknown image into regions explicitly and then matching them as a random graph [2]. This process of region identification for each pixel considers not just the pixel in question, but its neighboring pixels and the regions they belong to as well, hence the name contextual modeling. Thus, recognizing a character becomes identifying the model with the largest discriminant function value on the unknown image.

2. CONTEXTUAL MODELING OF CHARACTERS

A character image is abstracted into a matrix of cellular feature vectors (the pixel color and the number of strokes encountered along the 4 horizontal and vertical directions from the pixel to the image boundary) $\mathbf{O} = [\mathbf{o}_{i,j}]$ 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 pixel (i,j) which belongs to region $z_{i,j}$. $V = \{v_1, ..., v_T\}$ is the complete set of observable feature vectors so that any $\mathbf{o}_{i,j}$ assumes the value of one of its members. $z_{i,j}$ 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 parameters:

$$Pr(G_k),\ Pr(v_t \mid G_k),\ \text{and}\ Pr^{m,n}(G_l \mid G_k)$$

$Pr(G_k)$ measures the relative size of a region that assumes the value G_k. $Pr(v_t|G_k)$ determines what will be observed at a pixel in such a region. $Pr^{m,n}(G_l|G_k)$ supplies the contextual information between regions $z_{i',j'}$ and $z_{i,j}$ when they assume the value of G_l and G_k respectively. Here, $i' = i + m, j' = j + n$ and $(i',j') \in \eta_{i,j}$ where $\eta_{i,j}$ is the immediate 8-neighborhood of pixel (i,j).

0-7803-4137-6/97/$10.00©1977 IEEE

Regions of $\mathbf{O}$ are identified by identifying each of its pixels independently. Pixel (i,j) should be identified 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 $Pr(z_{i,j}|\mathbf{o}_{i,j}, \mathbf{o}_{\eta_{i,j}})$. Under the assumption that feature vectors in the same neighborhood are related to each other through the regions they belong to **only**, one then approximates this posterior probability with:

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

The summation is over all admissible values of $z_{\eta_{i,j}}$ which defines 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 the neighbors of $z_{i,j}$, viz., $z_{i',j'}$'s, are mutually independent given $z_{i,j}$. So,

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

Given a character image with observed feature vectors $[\mathbf{o}_{i,j}]$, assign each $\mathbf{o}_{i,j}$ to region G_k if

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

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

$$g((i,j);\mathbf{O};\lambda) = Pr(z_{i,j}) \cdot Pr(\mathbf{o}_{i,j}|z_{i,j}) \cdot \prod_{(i',j')\in\eta_{i,j}} \sum_{z_{i',j'}} Pr(z_{i',j'}|z_{i,j}) \cdot Pr(\mathbf{o}_{i',j'}|z_{i',j'}) \quad (4)$$

is a measurement of the appropriateness of identifying $\mathbf{o}_{i,j}$ to region G_k. The overall measurement of the similarity between the unknown image and a character model is the discriminant function:

$$g(\mathbf{O};\lambda) = \prod_{(i,j)} g((i,j);\mathbf{O};\lambda) \quad (5)$$

2.1. Training of A Contexual Model

A decision-directed (DD) method is adopted to estimate the contextual model parameters. In order to start the training process for a character, the initial model parameters can be specified according to the initial pixel identifications of an arbitrarily chosen training sample of the character which are obtained as follows:

Each row of the sample image as a bit map is decomposed into alternate white and black segments. Each segment in the first row is assigned a unique region identity. For each segment in the next and subsequent rows, if there is a segment in the previous row having the same color and approximately the same starting and ending columns, say, differing by no more than one pixel position, the same region identity will be inherited from the segment of the previous row, otherwise, a new region identity will be created for the segment of the new row. Pixels of the same stroke may therefore belong to more than one region and blank spaces between strokes will be divided into regions also. Thus, a region map is created for the image. The training algorithm is as follows:

Step 1. Given N feature vectors from a training sample, based on its initial region map created by the region segmentation algorithm described, one computes the initial parameter estimates of the contextual model by going to **Step 3** using just this chosen training sample.

Step 2. Based on the current estimate of model parameters, a region map for each training sample of the character is generated with pixels identified according to Eq.(3). All training samples will be utilized in Step 3 from now on.

Step 3. The model is updated as follows:

$$\hat{Pr}(G_k) = N_k/N \quad (6)$$

$$\hat{Pr}^{m,n}(G_l|G_k) = \frac{|\{(i,j) \mid z_{i',j'} = G_l, z_{i,j} = G_k\}|}{N_k} \quad (7)$$

$$\hat{Pr}(v_t|G_k) = \frac{|\{(i,j)|\mathbf{o}_{i,j} \sim v_t, z_{i,j} = G_k\}|}{N_k} \quad (8)$$

where N_k denotes the total number of pixels from the training sample(s) assigned to region G_k, and pixels (i', j') lie within $\eta_{i,j}$. $|\{...\}|$ is the number of pixels in the set defined within the two bars.

$\hat{Pr}(v_t|G_k)$ for any v_t not observed in region G_k should be assigned a small constant ϵ followed by normalization instead of leaving it at zero because such a lack of observation may simply be due to the finite size of the training sample set.

Step 4. Repeat **Step 2** and **Step 3** until convergence (i.e the change in Eq.(4) over all pixels of all samples drops below a predefined threshold).

2.2. Recognition of an Unknown Image

By using the above training method, one can generate a contextual model for each character. Let there be a collection of C such models, $\Lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_C\}$. A discriminant function for class λ_c is defined as:

$$g(\mathbf{O}; \lambda_c) = \prod_{(i,j)} Pr(z_{i,j}^{(c)}) \cdot Pr(\mathbf{o}_{i,j}|z_{i,j}^{(c)}) \cdot \prod_{(i',j')\in\eta_{i,j}} \sum_{z_{i',j'}^{(c)}} Pr(z_{i',j'}^{(c)}|z_{i,j}^{(c)}) \cdot Pr(\mathbf{o}_{i',j'}|z_{i',j'}^{(c)}) \quad (9)$$

where $z_{i,j}^{(c)}$ is chosen to maximize Eq.(4). An unknown image $\mathbf{O}$ will be classified to λ_d if

$$g(\mathbf{O}; \lambda_d) > g(\mathbf{O}; \lambda_c) \quad \forall c \neq d \quad (10)$$

2.3. Relaxation Training

The training process above, and so is the recognition process, despite its success, has a flaw in the sense that Eq.(3) does not produce the optimal region map of an image required by the parameter reestimation process because of the summation process. An alternative algorithm of a relaxation nature successfully applied to speech recognition [5] can determine the optimal region map for an image by modifying **Step 2** of the above training algorithm as follows:

Step a. Get an initial region map for the image. This can be achieved by quantizing pixel i, j to codeword G_k according to Eq.(3).

Step b. Let (i, j) move from the top left corner of the image to the bottom right corner along a row-wise raster scan. Re-identified pixel (i, j) to G_k according to:

$$G_k = argmax_{z_{i,j}} Pr(z_{i,j}) \cdot Pr(\mathbf{o}_{i,j}|z_{i,j}) \cdot \prod_{(i',j')\in\eta_{i,j}} Pr^{m,n}(z_{i',j'}|z_{i,j}) \cdot Pr(\mathbf{o}_{i',j'}|z_{i',j'}) \quad (11)$$

and replace the old $z_{i,j}$ with G_k. Then repeat this process of pixel re-identification following a path of row-wise raster scan but in the opposite direction, from bottom to top. Repeat again along column-wise raster scans in two opposite directions. This equation differs from Eq.(4) only in the absence of the summation over the pixel neighborhood. Such a change is introduced to assure the derivation of an optimal region map.

Step c. Repeat **Step a** until convergence.

Recognition of an unknown image is done by determining the optimal region map for the image by a relaxation process as in training. The suitability of λ_c is measured by the following discriminant function:

$$g_1(\mathbf{O}; \lambda_c) = \prod_{(i,j)} Pr(z_{i,j}^{(c)}) \cdot Pr(\mathbf{o}_{i,j}|z_{i,j}^{(c)}) \cdot \prod_{(i',j')\in\eta_{i,j}} Pr^{m,n}(z_{i',j'}^{(c)}|z_{i,j}^{(c)}) \cdot Pr(\mathbf{o}_{i',j'}|z_{i',j'}^{(c)}) \quad (12)$$

where every $z_{i,j}^{(c)}$ is the region that pixel i, j is identified to in the optimal region map using λ_c.

2.4. Mixture-region Training

Each pixel can be considered belonging to all regions stochastically instead of a particular one. The associated training method is to re-estimate the model parameters so that

$$g_2(\mathbf{O}; \lambda_c) = \prod_{(i,j)} \sum_{z_{i,j}^{(c)}} Pr(z_{i,j}^{(c)}) \cdot Pr(\mathbf{o}_{i,j}|z_{i,j}^{(c)}) \cdot \prod_{(i',j')\in\eta_{i,j}} \sum_{z_{i',j'}^{(c)}} Pr^{m,n}(z_{i',j'}^{(c)}|z_{i,j}^{(c)}) \cdot Pr(\mathbf{o}_{i',j'}|z_{i',j'}^{(c)}) \quad (13)$$

is maximized by first computing a region mixture map $Pr(z_{i,j} = G_k)$ of each training sample followed by re-estimating the model parameters as:

$$\hat{Pr}(G_k) = \frac{\sum_{as} \sum_{i,j} Pr(z_{i,j} = G_k)}{\sum_{k'} \sum_{as} \sum_{i,j} Pr(z_{i,j} = G_{k'})} \quad (14)$$

where

$$Pr(z_{i,j} = G_k) = Pr(G_k) \cdot Pr(\mathbf{o}_{i,j}|G_k) \cdot \prod_{(i',j') \in \eta_{i,j}} \sum_{z_{i',j'}} Pr^{m,n}(z_{i',j'}|G_k) \cdot Pr(\mathbf{o}_{i',j'}|z_{i',j'}) \quad (15)$$

$$\hat{Pr}^{m,n}(G_l|G_k) = \frac{\sum_{as} \sum_{i,j} Pr(z_{i',j'} = G_l) \cdot Pr(z_{i,j} = G_k)}{\sum_{as} \sum_{i,j} \sum_{l'} Pr(z_{i',j'} = G_{l'}) \cdot Pr(z_{i,j} = G_k)} \quad (16)$$

$$\hat{Pr}(v_t|G_k) = \frac{\sum_{as} \sum_{\{(i,j)|\mathbf{o}_{i,j} \sim v_t\}} Pr(z_{i,j} = G_k)}{\sum_{all\ v_t} \sum_{as} \sum_{\{(i,j)|\mathbf{o}_{i,j} \sim v_t\}} Pr(z_{i,j} = G_k)} \quad (17)$$

Here, $\sum_{as}$ stands for the summation over all samples. Recognition of an unknown image is done as above except that Eq(9) is replaced by Eq(13).

3. EXPERIMENTAL RESULTS

The contextual model with the algorithms in Section 2.1 and Section 2.2 has been implemented in a writer independent hand-written character off-line recognizer supporting a vocabulary of 4,616 Chinese characters, alphanumerics and punctuation symbols, where the average recognition rate is 78% [3] [4]. To examine the characteristics of the various discriminant functions and training methods, 5 pairs of highly similar characters as shown in Figure 1 are used as the vocabulary in the present study. Each character is written by 200 writers with 150 of them used for training and the rest for testing. In Table 1, the performances of these algorithms can be seen when they are tested on the training data themselves (close test) as well as the testing data (open test).

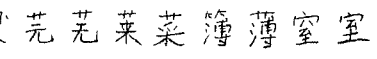

Figure 1: Vocabulary of highly similar characters

Table 1: Performances of the 3 algorithms

Recognition Rate	g	g_1	g_2
Close test (%)	96.33	97.00	96.33
Open test (%)	88.40	87.60	90.40

$\{Pr(G_k),\ Pr(v_t \mid G_k)$, and $Pr^{m,n}(G_l \mid G_k)\}$ for all G_k can be considered variables of function g_2 subject to linear constraints normalizing the probabilities. A general purpose linear constraint optimization algorithm has been employed to maximize g_2. The recognition results are almost the same as those listed under g_2 in Table 1 suggesting the validity of the mixture-region approach for recognizer training.

4. DISCUSSION

From Table 1, it can be observed from the close test results that relaxation training can produce a better fitting to the training data than the other two methods. However, from the open test results, it is obvious that the relaxation training method tends to over-train the model leading to a weaker generalization capability. In other words, if there is no shortage of training data, one can expect this method of contextual model training superior to the other two methods. The mixture-region approach seems to be a good compromise when there are insufficient training data which is usually the case in the real world. In conclusion, one can claim that a powerful contextual modeling method has been found for complex and variant pattern classes like hand-written Chinese characters.

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] P.K. Wong & C. Chan; "Off-line Hand-Written Chinese Character Recognition as a Compound Bayes Decision Problem"; submitted to *IEEE Trans. on PAMI*.

[5] Q.Huo and C. Chan; "A Study on the Use of Bi-directional Contextual Dependence in Markov Random Field-based Acoustic Modelling for Speech Recognition"; *Computer Speech and Language*, Vol. 10, pp. 95-105, 1996.