PRINTED CHARACTER RECOGNITION USING MARKOV MODELS

K. KORDI C. XYDEAS M. HOLT

University of Technology, Loughborough, Leicestershire, U.K.

RESUME

La traduction des documents imprimés et dactylographiés en forme électronique a de l'importance pour l'automatisation du travail de bureau. Elle nécessite l'usage d'un système de reconnaissance optique de caractères efficace (R.O.C.). Cet exposé décrit un façon d'aborder la question de la reconnaissance de caractères au moyen d'une modélisation stochastique, où la séquence des directions, obtenue par tracer le contour d'une caractère, est modèlée comme fonction probabiliste d'une châine de Markov de premier ordre. Une simulation d'ordinateur a montré que la méthode proposée s'annonce bien sur le terrain de la R.O.C. de multipolice quand elle est combinée à un simple algorithme de préclassification déterministe.

ABSTRACT

The conversion of typewritten and printed documents into an electronic form is of importance in office automation and requires the use of an efficient optical character recognition (OCR) system

This paper describes a stochastic modelling approach for character recognition where the sequence of directions obtained from tracing the contour of a character is modelled as a probabalistic function of a first-order Markov chain. In a computer simulation, when combined with a simple deterministic pre-classification algorithm, the proposed scheme has shown a promising performance in multifont OCR.

1. INTRODUCTION

Increasingly decentralized computing power and digital communication capabilities are facilitating the introduction of electronic documents in the office. With the predominant component of existing office documents being typewritten and printed text, the conversion of these documents into an electronic form requires the use of an efficient optical character recognition (OCR) system. That is, an ideally low cost system able to cope with both the "extraneous" variability in character representation (grain, quality, ageing, and colour background of the document, quality of the ribbon, type of ink and adjustment of the typewriter, threshold drifts and errors in the scanning process, quantization noise) and with the variability which resides in the characters themselves (different fonts and sizes). As a result nunerous deterministic, statistical, structural and geometrical character recognition methods (1) have been proposed in an attempt to meet the above requirements.

This paper describes a stochastic modelling approach to character recognition where the sequence of direction codes, obtained from tracing the contour of a character, is modelled as a probabalistic function of a first-order hidden Markov chain. The contour of a character is therefore considered as being generated by an unconstrained eight-states Markov chain having all the possible transition paths between states. Each state of the model is associated with a random process whose output, when observed at a given instant of time, can be any of eight possible direction codes. The Markov process associated with each of R different characters expected to be found in the input documents, is specified in terms of i) an initial state distribution vector, $\bar{\pi}$, ii) a state transition matrix A that models the probability of change of direction along the contour of a character, and iii) a stochastic matrix B which represents the "properties" of the random processes associated with observing the states of the model.

Following this approach, the classification process of input patterns of direction codes employs R hidden Markov models whose $\bar{\pi}$, A and B elements are pre-determined from a training process operating on a large number of characters of different fonts and sizes. Thus given the sequence of direction codes of an unknown input character, the probability that the sequence was generated from each of the R models is estimated, and the character is recognized as that whose model provides the largest probability value.

A simple "deterministic" pre-classification algorithm was also employed in our experiments, and assigns the unknown input character into a subset of possible characters. This results in a simplification of the recognition process, since the number of Markov models to be tested is reduced.

2. FIRST-ORDER HIDDEN MARKOV MODELS AND CLASSIFICATION OF CHARACTERS

The sequence of direction codes obtained from following the contour of a character is considered to be the output of a stochastic first-order Markov process. The process consists of two interrelated mechanisms, an underlying Markov chain of eight states, q_1, q_2, ..., q_8 and a set of eight random functions one of which is associated to each state. At a given instant of time the Markov process is in a unique state and an observation is generated by the random function associated with this state, causing the process to change state. The states cannot be observed directly, only the outputs of the random functions. The eight states of the process are the direction codes D_0, D_1..., D_7 of figure 1 while the set V = (v_1, v_2, ..., v_m) from which the observations are drawn, also consists of the eight direction codes i.e. V = (D_0, D_1, ..., D_7).

A hidden Markov model is defined in terms of three parameters $\bar{\pi}$, A and B. $\bar{\pi}$ is the initial distribution vector $\bar{\pi}$ = (π_1,π_2 ...,π_8) where π_1 = prob (q_1 at t=0). A is an 8x8 state transition matrix A = $[a_{ij}]$ $1 \leq i,j \leq 8$ where a_{ij} = prob (q_j at t+1$|$ q_i at t) and models the probability of change of direction along the contour of a character. B is a stochastic 8x8 matrix B = $[b_{j,k}]$, $1 \leq j,k \leq 8$, whose element b_{jk} provides the probability of observing the kth direction D_k at the current state q_j i.e. b_{jk} = prob(D_k at t$|$ q_j at t).

Now, in order to recognize R different characters, the parameters $(\bar{\pi}, A, B)$ of R different Markov models are first estimated using a training process that operates on many observation sequences $\bar{O}_d^1, \bar{O}_d^2, \cdots \bar{O}_d^n$ of contour direction codes, obtained for each of the R characters, $1 \leq d \leq R$. Thus the kth observation sequence of the dth character will be $\bar{O}_d^k = [O_1, O_2, \ldots, O_t, \ldots, O_{T_k}]$ where each $O_t \in V$.

Once the R models M_d $1 \leq d \leq R$ are defined, an input sequence of direction codes $\bar{O}_{in}$ of length T is classified to one of the R characters by first computing $P_d = \text{prob}(\bar{O}_{in} | M_d)$ for $1 \leq d \leq R$ and then assigning the input pattern to the character for which P_d is maximum. P_d is calculated as

$$P_d = \sum_{i=1}^{8} \alpha(i) \qquad (1)$$

using the forward probabilities $\alpha_t(i)$ for $1 \leq i \leq 8$ and $i \leq t \leq T$, where $\alpha_t(i) = \text{prob}(O_1, O_2 \ldots, O_t \text{ and } q_i \text{ at } t | M_d)$. $\alpha_T(i)$ is computed recursively from

$$\alpha_1(i) = \pi_i b_i(O_1) \qquad (2a)$$

$$\alpha_{t+1}(j) = \sum_{i=1}^{8} \alpha_t(i) a_{ij} \; b_j(O_{t+1}) \qquad 1 \leq j \leq 8, \; 1 \leq t \leq T-1 \qquad (2b)$$

where $b_j(O_t) = b_{jk}$ iff $O_t = D_k$.

3. MODEL ESTIMATION

Given a set of sequences $\bar{O}_d^1, \bar{O}_d^2, \ldots, \bar{O}_d^n$ of different "versions" of the dth character, the parameters $(\bar{\pi}^d, A^d, B^d)$ of the Markov model are estimated recursively, so that the probability $P = \prod_{i=1}^{n} P_i$ is maximised, where $P_i = \text{prob}(O_d^i | M_d)$. The training process for the estimation of th dth model can be described as follows.

Step 1: The optimization process commences with an initial set $(\bar{\pi}_{in}^d, A_{in}^d, B_{in}^d)$. $\bar{\pi}_{in}^d$ is set to $(1,0,\ldots 0)$ since the contour of a character is always traced clockwise starting from the upper left-hand corner of a window whose sides are tangent to the character. The initial estimate for A_{in} is given in Table 1a and was defined heuristically giving the highest values a_{ji} for $j=i$ and the lowest values for $|i-j| = 4$ (a reversal of direction is extremely unlikely). B_{in} is initially assumed to be equal to A_{in}.

Step 2: The forward probabilities $\alpha_t^k(i)$ $1 \leq i \leq 8$, $1 \leq t \leq T_k$ are estimated using Eq. 2a, 2b for each $\bar{O}_d^k$ sequence, $1 \leq k \leq n$. The backward probabilities $\beta_t^k(i)$ $1 \leq i \leq 8$ where $\beta_t^k(i) = \text{prob}(O_{t+1}^k, O_{t+2}^k, \ldots, O_{T_k}^k | q_i \text{ at } t, \text{ and } M_d)$ are also estimated recursively, for each $\bar{O}_d^k$, according to:

$$\beta_T^k(i) = 1 \qquad 1 \leq i \leq 8 \qquad (3a)$$

$$\beta_t^k(i) = \sum_{j=1}^{8} a_{ij} b_j(O_{t+1}^k) \beta_{t+1}^k(j), \quad 1 \leq i \leq 8, \; T_k - 1 \geq t \geq 1 \qquad (3b)$$

Step 3: The A^d, B^d and $\bar{\pi}^d$ elements of the model are re-estimated using

$$a_{ij} = \frac{\sum_{k=1}^{n} \sum_{t=1}^{T_k - 1} \alpha_t^k(i) a_{ij} \; b_j(O_{t+1}^k) \beta_{t+1}^k(j)}{\sum_{k=1}^{n} \sum_{t=1}^{T_k} \alpha_t^k(i) \; \beta_t^k(i)} \qquad (4)$$

$$b_{ij} = \frac{\sum_{k=1}^{n} \sum_{t \ni O_t^k = v_j} \alpha_t^k(i) \; \beta_t^k(i)}{\sum_{k=1}^{n} \sum_{t=1}^{T_k} \alpha_t^k(i) \; \beta_t^k(i)} \qquad (5)$$

and

$$\pi_i = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{P_j} \alpha_1^j(i) \; \beta_1^j(i) \qquad (6)$$

where

$$P_j = \text{prob}(\bar{O}_d^j | M_d) = \sum_{i=1}^{8} \alpha_{T_j}^j(i)$$

and the model estimation process returns to step 2 iff the ratio $\frac{P^{(k+1)}}{P^{(k)}}$ of probabilities obtained at the (k+1) and (k) iterations, is larger than a threshold ε whose value is very close to unity. $(P^{(k)} = \prod_{i=1}^{n} \text{prob}(\bar{O}_j^i | M_j))$.

The implementation of both the classification and the model estimation algorithms is likely to yield arithmetic underflow since the forward and backward probabilities are very small. Thus at each stage of the forward/backward probabilities estimation process i.e. for each t in Eq. 2b, 3b the values of $\alpha_t(i)$ and $\beta_t(i)$ are normalized[2] with $C_t = [\sum_{i=1}^{8} \alpha_t(i)]$.

When the R models $(\bar{\pi}^d, A^d, B^d)$, $1 \leq d \leq R$ are estimated, the B matrices are averaged to produce a B_c matrix that is "common" to all the R models. In addition, the B_c matrix forms the initial condition $B_{in}^d = B_c$ for an initial set $(\bar{\pi}_{in}^d, A_{in}^d, B_{in}^d)$ $1 \leq d \leq R$, and $\bar{\pi}^d$, A^d are re-estimated once again using steps 2 and 3 of the above procedure. B_c can be considered as a "quantization noise" modelling matrix.

4. RESULTS AND DISCUSSION

The proposed stochastic modelling approach for character recognition has been simulated on an ICL PERQ 1 system. The training data base, used in our preliminary experiments for the estimation of the models of the 26 lower-case characters, was formed from three different fonts per character and four samples per font giving a total number of 12 samples per character. Figures 2(a), 2(b) and 2(c) provide examples of the three fonts used in the data base while Table 1b gives the stochastic matrix B_c estimated using this particular data base.

In order to reduce the number of Markov models tested for maximum P_d (during the classification process of am input character), a simple deterministic pre-classifier was also employed that divides the 26 characters into six classes. The first five classes, specified as class 1 = (j,y), class 2 = (f,h,k,ℓ,t), class 3 = (a,e,o), class 4 = (p,q,g), class 5 = (b,d) are determined according to a character having a descender, an ascender, a loop, a loop and descender, a loop and ascender, respectively. Class 6 contains all the remaining characters.

When using as input a set of 250 characters (which are different from those employed in training) including characters of the font shown in figure 2d, the system performed reasonably well with a recognition accuracy up to 96%.

Some examples of cases where the proposed scheme failed to identify the correct input character are shown in figures 3a,3b and can be attributed to the small number of characters used in the training process. Figure 3c shows the interesting case where a "t" is classified as "f", since the contour sequence of direction codes for the two characters is similar, although the starting points when tracing the contour are different.

It is believed that an extended experiment, using an adequately large data base for the definition of the Markov models, will allow the introduction of a separate matrix B per model and will result in an improved performance.

5. REFERENCES

1. J. MANTAS "A survey of character recognition methodologies" Melecon 85 Volume,II: Digital Signal Processing, Eds: A. Luque, A.R. Figueiras Vidal, V. Cappellini. Elsevier Science Publishers B.V. (North Holland) IEEE 1985.

2. S. E. LEVINSON, L.R. RABINER, M.M. SONDHI "An introduction of the theory of probabilistic functions of a Markov process to automatic speech recogntiion", BSTJ, Vol.62, No. 4,April 1983

0.60000	0.17500	0.01000	0.01000	0.01000	0.01000	0.01000	0.17500
0.17500	0.60000	0.17500	0.01000	0.01000	0.01000	0.01000	0.01000
0.01000	0.17500	0.60000	0.17500	0.01000	0.01000	0.01000	0.01000
0.01000	0.01000	0.17500	0.60000	0.17500	0.01000	0.01000	0.01000
0.01000	0.01000	0.01000	0.17500	0.60000	0.17500	0.01000	0.01000
0.01000	0.01000	0.01000	0.01000	0.17500	0.60000	0.17500	0.01000
0.01000	0.01000	0.01000	0.01000	0.01000	0.01000	0.60000	0.17500
0.17500	0.01000	0.01000	0.01000	0.01000	0.01000	0.17500	0.60000

TABLE 1a

0.80518	0.13140	0.02600	0.00001	0.00001	0.00001	0.00120	0.03490
0.06700	0.69128	0.21649	0.00360	0.00001	0.00001	0.00001	0.02030
0.00001	0.01560	0.93496	0.04570	0.00240	0.00001	0.00001	0.00001
0.00001	0.00100	0.20079	0.69408	0.08210	0.02050	0.00001	0.00001
0.00001	0.00001	0.00060	0.08450	0.86377	0.04950	0.00030	0.00001
0.00001	0.00001	0.00220	0.00750	0.13020	0.74059	0.11060	0.00760
0.00330	0.00001	0.00001	0.00001	0.00280	0.03420	0.93327	0.02510
0.16250	0.00450	0.00001	0.00001	0.00001	0.00980	0.20999	0.61188

TABLE 1b

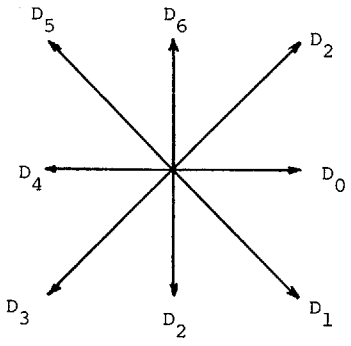


FIGURE 1

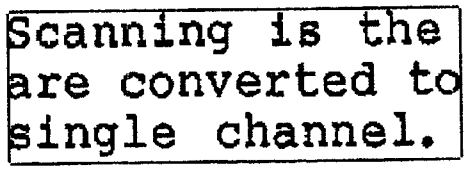


FIGURE 2a

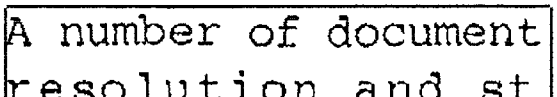


FIGURE 2b

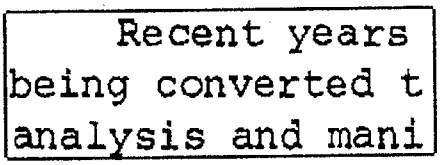


FIGURE 2c

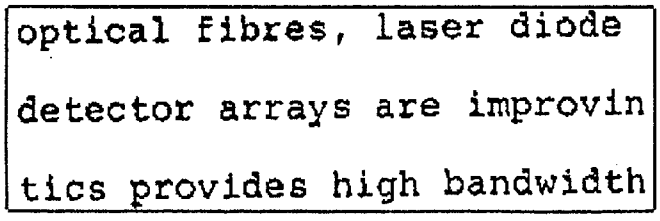
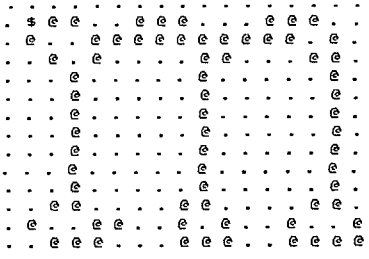
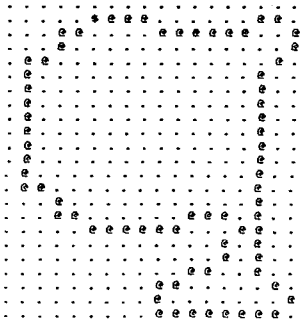


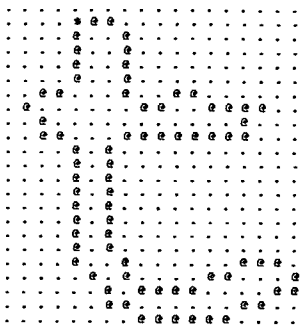
FIGURE 2d



(a) Classified as "n"



(b) Classified as "g"



(c) Classified as "f"

FIGURE 3