**MVA'94** IAPR Workshop on Machine Vision Applications Dec. 13-15, 1994, Kawasaki

# OFF-LINE HANDWRITING RECOGNITION BY STATISTICAL CORRELATION

GEORGE SAON, ABDEL BELAÏD and YIFAN GONG

CRIN-CNRS /INRIA-LORRAINE
Email: saonandy,abelaid,gong@loria.fr

ABSTRACT

In this paper, we present a new approach to unconstrained, off-line handwriting recognition (HWR). It is based on the global plausibility estimation of a word knowing the local probability of each individual character. Having a set of character models issued from a training step and an input word pattern, we maximize for each word of the lexicon the sum of plausibilities of its component characters. This maximisation is made in terms of observation quality and extent of a symbol within the pattern. The proposed method operates in a top-down manner by giving segmentation hypotheses which induce local symbol extents for a given word of the lexicon against which the pattern is matched. The word which obtains the highest average plausibility per character is labeled as the recognized one. This method was applied with success on continuous speech recognition by [5].

## 1 INTRODUCTION

Automatic handwriting recognition (HWR) knew intensive studies this last decade. Many investigations are made on isolated characters and satisfactory solutions are obtained for specific applications. However, automatic recognition of natural writing with totally unconstrained words remains an even more challenging problem due to the large variety of handwriting styles where characters are often linked, broken and irregularly aligned. In such case, classical segmentation techniques may fail. Thus, to bypass the difficult problem of the segmentation, the tendancy was to consider globally the word based on particular features and, thence, know what the individual letters should be. This seems to correspond to the human perception of writing where recognition is made primarily using a global word shape rather than segmenting and recognizing each letter individually. On the other hand, researches were also carried out in *analytical* way by segmenting words into letters and recognizing them separately [7, 13, 3].

Hence, the most common way to recognize words is to use the context to limit possible intra- and extra-word relationships. For example, in applications such as check amount identification, context is given by reduced vocabulary size and number grammars [10],[8]. For mail sorting, where the vocabulary is large, address grammar limits the search.

Since there is a large similarity between HWR and speech recognition, techniques for recognition of unconstraint handwritten words can be borrowed from speech domain which has been very active during the last decade [12]. This is already the case with hidden Markov models where their application benefit from the large experience in speech recognition [1, 11, 2, 4]. In this paper, we present a letter based word recognition scheme for HWR on check amounts. This scheme falls in the framework of a stochastic method based on plausibility calculus used in speech recognition by Y. Gong [5]. This method gives for each input word the plausibility of each component letter with respect to letter models learned in a previous training process.

After describing the feature extraction phase in section 2, we give the method formulation in section 3 and we present in section 4 some promising results obtained on a 30-scriptor data base.

## 2 FEATURE EXTRACTION

As features, we use normalized distances $\in [0,1]$ along vertical axes between upper and lower word contours and typographical bands. These bands are obtained by histogram analysis of the vertical projection of the number of run-length horizontal segments. The figure 1 shows an example of continous character features. First, the system locates the bands and afterwards, by giving cut points between characters, surrounding frames for each character are determined. Distances are adjusted with respect to these rectangles in order to code more accurately the character shapes.

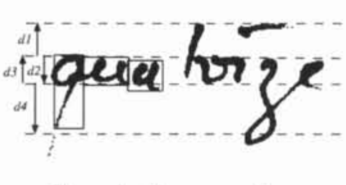


Figure 1: Feature coding.

## 3 METHOD DESCRIPTION

In the following formalism, we assume that a pattern can be described as an observation sequence of $\mathbb{R}^P$-parameterized features along one axis (i.e. contours on horizontal axis for handwriting).

Let $\mathbf{X} = x_0 x_1 \ldots x_N$ be an observation sequence of a signal where $x_i \in \mathbb{R}^P$. We call duration the length of the sequence in terms of the number of its vectors. Given a duration $d$ and a centred sequence $X_i$ at instant $i$, we define a $Q$-state observation matrix, $M_{X_i}(d) \in \mathbb{R}^{Q \times P}$,

$$M_{X_i}(d) = col(x^t_{i+\frac{d}{Q}j})_{-\frac{Q}{2} \leq j < \frac{Q}{2}} \quad (1)$$

The idea of such a matrix relies on a linear sampling of a subsequence in order to reduce the observation to a constant number of states $Q$. In this manner, a local observation is well parameterized by such a matrix and its corresponding duration. For example, in figure 1 the 5-state observation matrix for characters $q$ is:

$$M^q = \begin{pmatrix} -0.0004 & 0.6360 & 0.4542 & -0.0004 \\ -0.0003 & 0.8482 & 0.8179 & 0.7573 \\ -0.0002 & 0.9998 & 0.9392 & 0.2422 \\ -0.0002 & 0.8483 & 0.9998 & -0.0002 \\ -0.0001 & 0.6969 & 0.9090 & -0.0001 \end{pmatrix}$$

(zero centred noise was added to avoid small variations)

Let $\mathbb{L} = \{s_0, s_1, \ldots s_l\}$ be the set of characters and a word $\omega \in \mathbb{L}^*$ of length $L(\omega)$, $\omega = s_0\ s_1\ \ldots\ s_{L(\omega)-1}$. The conditional probability of $\omega$ given observation $\mathbf{X}$ is:

$$P(\omega/\mathbf{X}) = \max_{t_i}\Big(\prod_{i=0}^{L(\omega)-1} P(s_i/X_{\frac{t_i+t_{i+1}}{2}}, t_{i+1} - t_i)\Big)^{\frac{1}{L(\omega)}} \quad (2)$$

where $\{t_i\}_{0 \leq i \leq L(\omega)}$ represents a set of letter segmentation hypotheses with: $t_0 = 0,\ \ t_i < t_{i+1},\ \ t_{L(\omega)} = N$.
$t_{i+1} - t_i$ would be therefore the local duration of character $s_i$. Formula (2) reflects the conditional probability of a word as being the geometrical average of symbol probabilities. The most plausible word is the one which satisfies:

$$\omega^* = \operatorname*{argmax}_{\omega \in \mathbb{L}^*} P(\omega/\mathbf{X}) = \operatorname*{argmax}_{\omega \in \mathbb{L}^*} \log P(\omega/\mathbf{X}) \quad (3)$$

$$\omega^* = \operatorname*{argmax}_{\omega \in \mathbb{L}^*} \frac{1}{L(\omega)} \max_{t_i} \sum_{i=0}^{L(\omega)-1} \log P(s_i/X_{\frac{t_i+t_{i+1}}{2}}, t_{i+1} - t_i) \quad (4)$$

The probability in (4) can be expressed as:

$$P(s/X_i, d) = \frac{P(s, X_i, d)}{P(X_i, d)} \quad (5)$$

with $P(X_i, d)$ being constant during recognition therefore discardable. In the context of multi-models the marginal probability density function (pdf) $P(s, X_i, d)$ can be obtained by summing up on the set $K$ of all character models, $\mathcal{M}_k$, as follows:

$$P(s, X_i, d) = \sum_{k \in K} P(X_i, \mathcal{M}_k, d, s) \quad (6)$$

where $P(X_i, \mathcal{M}_k, d, s)$ is the joint pdf of vector sequence $X_i$, character model $\mathcal{M}_k$, duration $d$ and symbol $s$. Following the chain decomposition rule of conditional probability densities, one may obtain:

$$P(X_i, \mathcal{M}_k, d, s) = P(X_i/\mathcal{M}_k, d, s) P(d/\mathcal{M}_k, s) P(\mathcal{M}_k, s) \quad (7)$$

$$P(\mathcal{M}_k, s) = P(\mathcal{M}_k/s) P(s) \quad (8)$$

Moreover, some terms do not depend on $s$:

$$P(X_i, \mathcal{M}_k, d, s) = P(X_i/\mathcal{M}_k, d) P(d/\mathcal{M}_k) P(\mathcal{M}_k/s) P(s) \quad (9)$$

*-Calculus of $P(X_i/\mathcal{M}_k, d)$.* Since features are continuous, one naturally assumes that their corresponding quantified vectors belonging to a cluster follow a multi-variate Gaussian pdf with mean observation matrix $\overline{M}_k$ and state covariance matrix $\Sigma_k$ related to model $\mathcal{M}_k$.

$$P(X_i/\mathcal{M}_k, d) = \frac{\exp^{-\frac{1}{2}tr((M_{X_i}(d) - \overline{M}_k)^t \Sigma_k^{-1} (M_{X_i}(d) - \overline{M}_k))}}{\sqrt{(2\pi)^Q |\Sigma_k|}} \quad (10)$$

*-Calculus of $P(d/\mathcal{M}_k)$.* Experience showed that duration is normally distributed within model $\mathcal{M}_k$ with mean value $\overline{d}_k$ and standard deviation $\sigma_k$.

$$P(d/\mathcal{M}_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp^{-\frac{(d-\overline{d}_k)^2}{2\sigma_k^2}} \quad (11)$$

*-Calculus of $P(\mathcal{M}_k/s)$ and $P(s)$.* They represent the ***a priori*** probability of the model $\mathcal{M}_k$ knowing $s$ and of $s$.

$$P(\mathcal{M}_k/s) = \frac{|N_k^s|}{|N^s|} \quad (12)$$

$$P(s) = \frac{|N^s|}{|N|} \quad (13)$$

$N_k^s$, $N^s$, $N$ representing respectively, the set of samples of $s$ belonging to $\mathcal{M}_k$, the set of samples of $s$ and the set of all character samples.

### 3.1 TRAINING PROCESS

It consists of defining the set of models $\mathcal{M}_k$, $k \in K$, from a training corpus $N$. Each character sample $i \in N$ is given by an observation matrix $M_i$ and duration $d_i$. Formally, a model $\mathcal{M}_k$, $k \in K$ is defined as: $\mathcal{M}_k = (\overline{M}_k, \Sigma_k, \overline{d}_k, \sigma_k)$.
-Mean observation matrix:

$$\overline{M}_k = \frac{1}{|N_k^s|} \sum_{i \in N_k^s} M_i \quad (14)$$

-State covariance matrix: $\Sigma_k = (a_{ij}^k)_{1 \leq i,j \leq Q}$

$$a_{ij}^k = \frac{1}{|N_k^s|} \sum_{l \in N_k^s} (M_l^i - \overline{M}_k^i)^t (M_l^j - \overline{M}_k^j) \quad (15)$$

where $N_k^s$ is the set of samples of $s$ related to model $\mathcal{M}_k$.
-Mean duration:

$$\overline{d}_k = \frac{1}{|N_k^s|} \sum_{i \in N_k^s} d_i \quad (16)$$

-Standard deviation for duration:

$$\sigma_k = \sqrt{\frac{1}{|N_k^s|} \sum_{i \in N_k^s} (d_i - \overline{d}_k)^2} \quad (17)$$

Linear sampling of each character subsequence leads us to an observation matrix. These matrixes are grouped into classes using a variation of the LBG clustering algorithm [14]. We used variable distances at each iteration based on a ***Mahalanobis*** measure:

$$d(X,Y) = \sqrt{tr((X-Y)^t \Sigma^{-1} (X-Y))} \quad (18)$$

with $X, Y \in \mathbb{R}^{Q \times P}$ and covariance matrix $\Sigma$ positively defined.

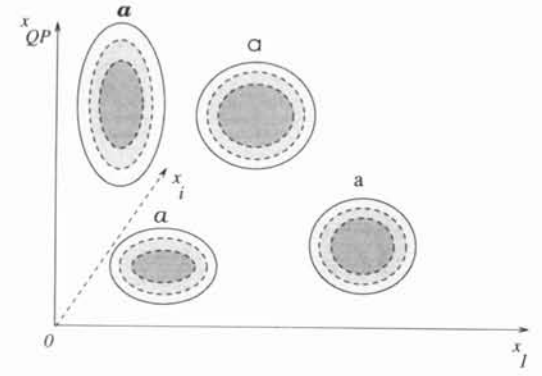


Figure 2: Clustering process regrouping allographs in classes

### 3.2 RECOGNITION

We need to maximize the sum of local duration probabilities. Applying *log* to equation (11) results, modulo a constant term, in :

$$f(t_0, t_1, \ldots, t_{L(\omega)}) = \sum_{i=0}^{L(\omega)-1} -\frac{(t_{i+1} - t_i - \overline{d}_i)^2}{2\sigma_i^2} \quad (19)$$

with the constraints $t_0 = 0$, $t_{L(\omega)} = N$, $t_i \leq t_{i+1}$. Following *Lagrange multiplicators theory* we deduce:

$$\begin{cases} t_0 = 0 \\ t_{i+1} = t_i + \overline{d}_i - \sigma_i^2 \dfrac{\left(\sum_{j=0}^{L(\omega)-1} \overline{d}_j - N\right)}{\sum_{j=0}^{L(\omega)-1} \sigma_j^2}, \quad 0 \leq i < L(\omega) \end{cases} \quad (20)$$

$\overline{d}_i$, $\sigma_i$ representing, respectively, global mean duration and standard deviation for symbol $s_i$. In practice, we may deplace the $t_i$'s around these values within a centred interval of variation assuring always $|t_{i+1} - t_i - \overline{d}_i| < 2\sigma_i$ in order to maximize also the observation quality. This method has the advantage of a global view on the initial segmentation hypotheses $\{t_i\}_i$ and afterwards it reinforces also the maximum calculus in (4).

## 4 EXPERIMENTS AND RESULTS

The corpus contains 627 words written by 30 scriptors without constraints but with a limited vocabulary (28, corresponding to the French bank check wording words). Characters in each word were manually segmented (about 2000 characters from the corpus, varying from 15 to 400 samples per character). We used 400 word images for the training phase and 227 for test distinct from the previous ones . For character recognition, the average rate is 90.63% top 3 choices. For word recognition, the average rate lies between 84.74% (top 1 choice) and 95.90% (top 3 choices). Figure 3 shows some recognition results for three different writing styles. Realignement techniques were used which consist to reiterate well recognized words in the training step together with their character segmentation hypotheses found by the system.

| *Choice* | *Character recognition rate %* | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | a | c | d | e | f | g | h | i | l | m | n |
| 1 | 94 | 77 | 81 | 70 | 74 | 80 | 86 | 69 | 93 | 93 | 77 |
| 2 | 98 | 86 | 88 | 85 | 89 | 80 | 95 | 81 | 94 | 98 | 86 |
| 3 | 100 | 91 | 93 | 90 | 89 | 93 | 95 | 85 | 97 | 98 | 90 |
| | o | p | q | r | s | t | u | v | x | z | |
| 1 | 85 | 53 | 87 | 74 | 84 | 75 | 73 | 53 | 81 | 61 | |
| 2 | 92 | 82 | 94 | 88 | 92 | 88 | 80 | 53 | 94 | 74 | |
| 3 | 96 | 94 | 95 | 93 | 95 | 92 | 82 | 53 | 97 | 83 | |

| | | | |
|---|---|---|---|
| un: | 84.00 | 100.00 | 100.00 |
| deux: | 92.31 | 92.31 | 100.00 |
| trois: | 73.08 | 84.62 | 84.62 |
| quatre: | 80.00 | 85.00 | 90.00 |
| cinq: | 96.00 | 100.00 | 100.00 |
| six: | 92.59 | 96.30 | 100.00 |
| sept: | 75.00 | 85.00 | 85.00 |
| huit: | 95.83 | 100.00 | 100.00 |
| neuf: | 90.00 | 95.00 | 95.00 |
| dix: | 72.00 | 96.00 | 100.00 |
| vingt: | 75.00 | 95.00 | 100.00 |
| quarante: | 94.74 | 94.74 | 100.00 |
| cinquante: | 85.71 | 90.48 | 95.24 |
| soixante: | 77.27 | 90.91 | 90.91 |
| cent: | 92.00 | 96.00 | 100.00 |
| mille: | 90.91 | 100.00 | 100.00 |
| million: | 68.18 | 81.82 | 86.36 |
| et: | 95.45 | 100.00 | 100.00 |
| centimes: | 80.00 | 85.00 | 95.00 |
| Average: | 84.74 | 93.06 | 95.90 |

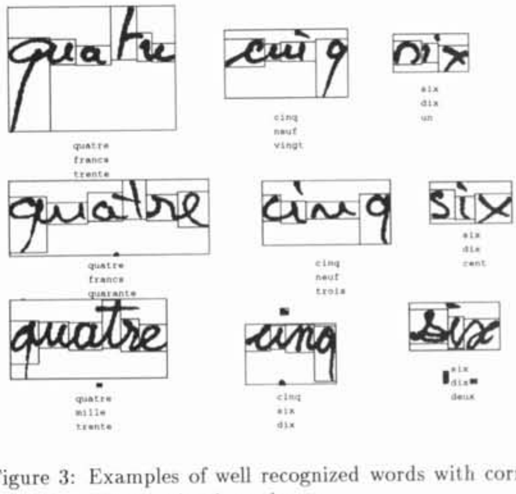


Figure 3: Examples of well recognized words with corresponding segmentation hypotheses

## 5 Conclusion

We developed a theoretical framework and a system prototype for HWR by stochastic modeling with a close relation with similar researches in continuous speech recognition. Problems encountered are analogous but HWR has particularities due to its bidimensional aspect. Despite modest recognition scores, the method remains promising and is open to improvements concerning feature analysis and context modeling. Further developments should cope with more discriminant continuous features taking into account geometrical and topological properties inherent to writing. Convolution masks for finding basic components such as strokes, cavities, peaks and curves are actually in study. Future extensions dealing with grammar constraints for search space limitation (context at syntactic level) should be foreseen.

## References


[1] O. E. Agazzi and S. Kuo. Hidden Markov Model Based Optical Character Recognition in the Presence of Deterministic Transformation. *Pattern Recognition*, 26(12):1813–1826, February 1993.

[2] J. C. Anigbogu and A. Belaïd. Recognition of Multifont Text Using Markov Models. In *7$^{th}$ Scandinavian Conference on Image Analysis*, volume I, pages 469–476, August 1991.

[3] J. R. Bellegarda, D. Nahamoo, K. S. Nathan, and E. J. Bellegarda. Supervised hidden Markov modeling for online handwriting recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume V, pages 149–152, 1994.

[4] R.M. Bozinovic and S.N. Srihari. Off-Line Cursif Script Word Recognition. In *IEEE Transactions on PAMI*, volume 11, 1 1989.

[5] M. Y. Chen and A. Kundu. An Alternative to Variable Duration HMM in Handwritten Word Recognition. In *In Proc. of the International Workshop on Frontiers in Handwriting Recognition IWFHR-3*, pages 82–91, Paris, 1993.

[6] M. Y. Chen, A. Kundu, J. Zhou, and S. N. Srihari. Off-line handwritten word recognition using hidden Markov model. In *USPS'92*, 1992.

[7] E. Cohen, J. J. Hull, and S. N. Srihari. Interpretation of handwritten addresses. Technical report, Center for Document Analysis and Recognition, Buffalo, New York, 1993.

[8] A. Gillies. Cursive word recognition using hidden Markov models. Technical report, Environmental Research Institute of Michigan, Michigan, 1991.

[9] A. M. Gillies. Cursive Word Recognition Using Hidden Markov Models. In *In USPS'92*, pages 557–562, 1992.

[10] M. Gilloux, J-M. Bertille, and M. Leroux. Recognition of handwritten words in a limited dynamic vocabulary. In *In Proc. of the International Workshop on Frontiers in Handwriting Recognition IWFHR-3*, pages 417–422, Paris, 1993.

[11] Y. Gong and J. P. Haton. Stochastic trajectory modeling for speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 57–60, Adelaide, Australia, 1994.

[12] T. K. Ho. Recognition of handwritten digits by combining independent learning vector quantizations. In *International Conference on Document Analysis and Recognition (ICDAR), Tsukuba, City Science Japan*, pages 818–821, 1993.

[13] F. Kimura, M. Shridhar, and Z. Chen. Improvements of a lexicon directed algorithm for recognition of unconstrained handwritten words. In *International Conference on Document Analysis and Recognition (ICDAR), Tsukuba, City Science Japan*, pages 18–22, 1993.

[14] Y. Linde, A. Buzo, and R. M. Gray. An Algorithm for Vector Quantized Design. *IEEE Transactions on Computer*, 28(1):84–95, 1980.

[15] L.R. Rabiner. Mathematical Foundations of Hidden Markov Models. *NATO ASI Series. Vol. F46, Recent Advances in Speech Understanding and Dialog Systems*, F46:183–205, 1988.

[16] T. Starner, J. Makhoul, R. Schwartz, and G. Chou. On-line cursive handwriting recognition using speech recognition methods. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume V, pages 125–128, 1994.

[17] S. Tsuruoka, F. Kimura, M. Shridhar, and Z. Chen. Segmentation and recognition for handwritten 2-letter state names. In *International Conference on Document Analysis and Recognition (ICDAR), Tsukuba, City Science Japan*, pages 814–817, 1993.