CONNECTED AND DEGRADED TEXT RECOGNITION USING PLANAR HIDDEN MARKOV MODELS

Oscar E. Agazzi[1] *Shyh-shiaw Kuo*[1] *Esther Levin*[2] *Roberto Pieraccini*[2]

[1]Signal Processing Research Department
[2]Speech Research Department
AT&T Bell Laboratories, Murray Hill, NJ 07974, U.S.A.

ABSTRACT

We present an algorithm for connected text recognition using enhanced Planar Hidden Markov Models (PHMMs). The algorithm we propose automatically segments text into characters (even if they are highly blurred and touching) as an integral part of the recognition process, thus jointly optimizing segmentation and recognition. Performance is enhanced by the use of state length models, transition probabilities among characters (bigrams), and grammars. Experiments are presented using: 1) A simulated database of over 24,000 highly degraded images of city names; 2) A database of 6,000 images rejected by a high performance commercial OCR machine with 99.5% accuracy. Measured performance on the first database is 99.65% for the most degraded images when a grammar is used, and 98.76% in the second database. Traditional OCR algorithms would fail drastically on these images.

1. INTRODUCTION

This paper describes a set of experiments for assessing the performance of the Planar Hidden Markov Model (PHMM) paradigm [1] in the recognition of connected printed text that has been severely degraded by noise, blur, and other stochastic distortions, as often happens with documents that have been repeatedly reproduced and/or transmitted by fax. Results show recognition accuracies in excess of 98% for images that have been so severely degraded that even a human observer has difficulty in recognizing them. The experiments were carried out using two large databases:

1] A simulated database of 24,600 images representing city names (Database I)

2] A database of 6126 images rejected by a high performance commercial optical character recognition (OCR) machine with 99.5% accuracy (Database II).

Database I has the advantage of allowing more control over parameters such as noise, blur, or character overlap, and was used for studying the variation of the recognition accuracy as a function of these parameters. Database II offers the advantage of testing our algorithm in a challenging "real world" application of significant commercial importance.

2. PLANAR HIDDEN MARKOV MODELS

Fig.1 shows the PHMM structure we use and the way it represents a character.

Each PHMM consists of a rectangular array of $N_X \times N_Y$ states, organized as N_X "superstates", each one of them consisting of N_Y states. In the vertical direction, transitions are allowed only among states of a given superstate. In the horizontal direction, transitions among arbitrary superstates are possible. These restrictions are necessary to ensure that the recognition algorithm runs in polynomial time [1]. In the application described in this paper, we further restrict the PHMMs, such that:

1] Superstates are strictly "top-to-bottom", i.e., from a certain state we allow transitions only to itself or to the next state.

2] The models are strictly "left-to-right", i.e., only transitions to itself or to the next superstate are allowed from a given superstate.

In an isolated character recognition application, the image is scanned from top to bottom and from left to right, and a score is computed for each PHMM representing a character. In the vertical scan, each superstate is scored using the Viterbi algorithm. These scores are then used in a second pass of the Viterbi algorithm while the image is scanned in the horizontal direction, yielding a score for the PHMM under test. All PHMMs are scored in a similar way, and the best one is chosen. The PHMMs are treated as nested one-dimensional models, rather than truly two-dimensional.

In this paper we deal with the recognition of strings, rather than isolated characters. This problem is similar to the problem of recognizing connected words in speech recognition, where the problem is solved by searching for the best path (the one with the highest likelihood) in a network that allows for all the possible combinations of phonetic HMMs [2]. Analogously, for our OCR problem, we built a network that allows for all the possible combinations of characters, each character being represented by a PHMM. The structure of this network is illustrated in Fig.2.

The network is searched with the generalized Viterbi algorithm described in [1], [3]. This network interconnects all the character-level PHMMs, allowing arbitrary transitions from character to character. The transition probabilities among PHMMs correspond to the transition probabilities among characters, or probabilities of bigrams in the English language (or any other language according to the applica-

0-7803-0946-4/93 $3.00

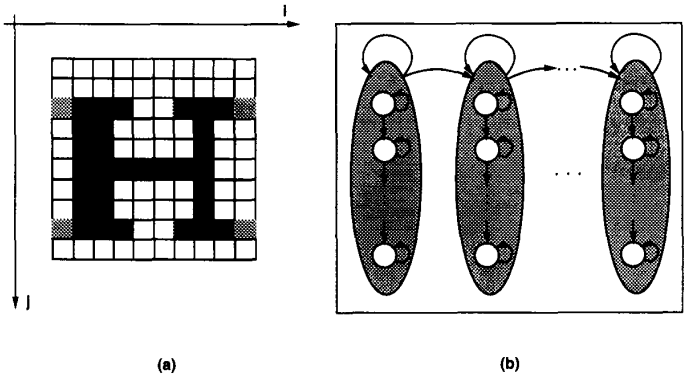

Figure 1: Structure of PHMMs. (a) The grid square located at ordinate j and abscissa i represents state j of superstate i. The grey levels represent the continuous range $[0, 1]$ of probabilities of black pixels. (b) State and superstate transition diagram corresponding to the array of (a).

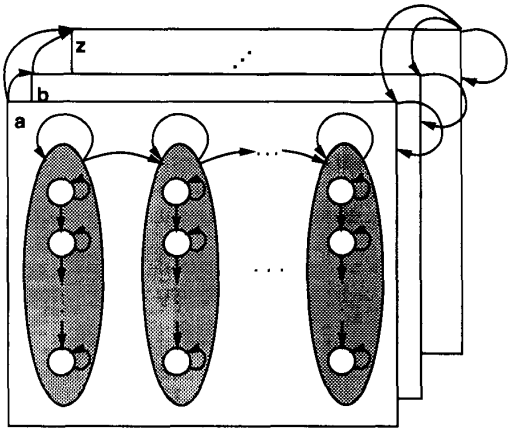

Figure 2: PHMM network used for connected character recognition

tion).

3. STATE AND SUPERSTATE LENGTH MODELING

The "dynamic planar warping" property of PHMMs [1] provides an excellent way to accommodate character shape variations, such as those encountered in multifont character recognition, or geometric distortions introduced by processes like fax transmission. However in their simplest form these models can also accommodate some unrealistic forms of distortion, like the transformation of a "p" into a "b", with very little penalty. The reason for this is that, in principle, when superstates are applied to different columns in the image, they are not forced to undergo the same or similar warping. This excessive flexibility would allow the descender of "p" to be transformed into an ascender, changing the identity of the character into a "b". Many other undesirable forms of distortion are also possible as a result of the same effect. This effect can be avoided completely while keeping all the flexibility to accommodate more realistic forms of distortion by modeling the lengths of the pixel sequences generated by states and superstates with a Gaussian distribution, instead of the exponential distribution inherent in the simple form of the models. The Gaussian distribution imposes a heavier penalty for extreme forms of distortion, thus preventing their occurrence.

The computational cost of introducing Gaussian state and superstate length densities can be made negligible by using the postprocessing approach described in [4].

4. SIZE NORMALIZATION

Traditionally, size normalization is applied to images *before* recognition. If text is not connected and therefore characters can be reliably segmented before recognition, it is usually not a problem to estimate the scaling factor and normalize the image. However, in the case of extremely connected and degraded text of unknown font, there is not enough information to estimate the scaling factor before recognition. As seen in section 2, the PHMM approach combines segmentation and recognition, thus avoiding one of the common causes of failure of traditional OCR systems when presented with connected text [5], namely, the inability of the recognition engine to recover from errors introduced by the segmentation engine. Another distinct advantage of PHMMs is that they also allow size normalization to be combined with recognition, therefore avoiding another possible failure mode of non-PHMM systems, i.e., the inability of the recognition engine to recover from inaccurate size normalization.

5. FEATURES

The feature vector can be simply the binary value of the pixels. However performance can be improved by the use of a more elaborate feature vector, where the pixel value is augmented by four more binary components representing the presence or absence of straight line segments at 0°, 45°, 90°, and 135° in the neighborhood of the pixel. These segments are extracted by a Hough transform applied in a small square window around the pixel.

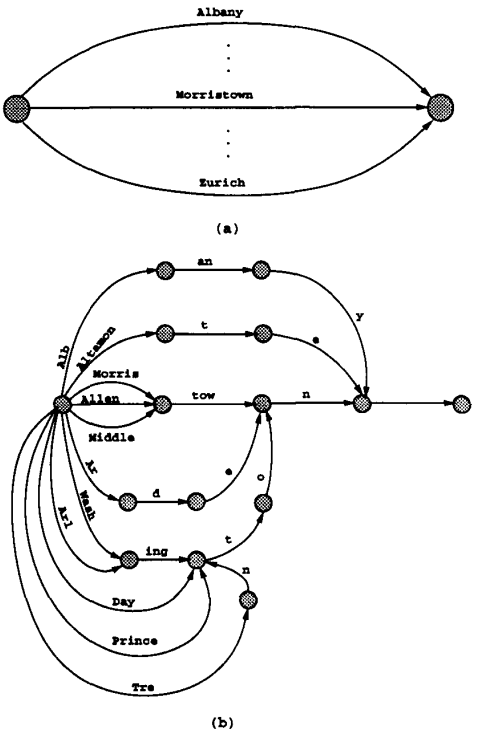

Figure 3: Grammars used in the city names experiment. For clarity, only a few nodes are shown. (a) Simple grammar. (b) Grammar obtained by minimization of (a).

6. USE OF GRAMMARS

In character recognition systems it is common to correct recognition errors by searching in a dictionary for the strings that best match the ones provided by the recognizer. Dictionary search is in these cases a postprocessing operation. When using PHMMs, it is possible to constrain the recognized strings to be members of a certain lexicon by using a grammar. Unlike the traditional postprocessing approach, this method prevents the recognizer from generating erroneous strings, instead of correcting them a posteriori, resulting in enhanced performance. Fig.3 shows a simplified diagram of a grammar used in the experiments with Database I described in the next section.

7. EXPERIMENTS

Experiments to evaluate the performance of the algorithm have been conducted on a database of 24,600 images representing blurred and noisy instances of 205 city names printed in Times-Roman size 10 characters (Database I). The original binary images represented by 250×40 pixel arrays, were corrupted by pixel-flip noise with a probability P_N, then they were blurred using a Gaussian filter with standard deviation σ_B, and finally thresholded at a level T. Values of σ_B and T were kept fixed at 1.5 pixels and 160 (in a scale where 0 represents black and 255 represents white), whereas P_N was varied between 0. and 0.25 in intervals of 0.05. These parameter values were chosen to produce realistic forms of distortion, closely resembling the effect of multiple reproductions and/or fax transmissions. The largest values of P_N generate severe levels of distortion, making some of the images difficult to recognize even for a human observer (Fig.4). For each value of P_N, a set of 4100 images was generated (20 instances of each city name). Half of them were used for training and the rest for testing.

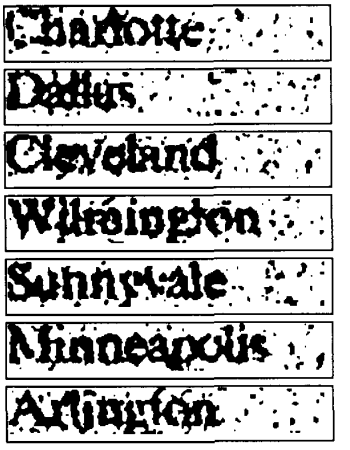

Figure 4: Images of city names Charlotte, Dallas, Cleveland, Wilmington, Sunnyvale, Minneapolis, and Arlington in the simulated database. In these images the probability of noise is $P_n = 0.25$.

P_n	Char. Accuracy	String Accuracy
≤ 0.20	100.00%	100.00%
0.25	99.65%	99.62%

Table 1: Recognition accuracy on Database I. In these experiments a grammar was used to improve performance.

The results of our experiments are given in Table 1.

The accuracy is extremely high even for highly degraded images such as the ones shown in Fig.4, and it is 100% for $P_N \leq 0.20$. Traditional OCR algorithms would fail drastically on these images. A major factor contributing to this excellent performance is the use of a grammar (Fig.3) that constrains the recognized string to be one of the 205 city names.

A second set of experiments was conducted on another database consisting of 6,126 strings rejected by a commercial OCR machine (Database II). This machine has a 99.5% accuracy, however in the high volume application where it is used, the cost of manually processing the 0.5% rejects is high. A robust algorithm that can recognize a large fraction of these rejects would be desirable. Fig.5 shows some typical images of this database. Although characters are not very connected, the resolution and image quality are poor, and there is a mixture of different fonts and sizes. In this case no probabilities of bigrams, grammars, or any other form of contextual information was available to improve

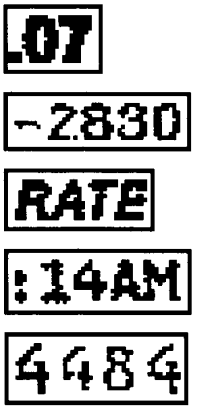

Figure 5: Images from Database II.

Character Accuracy	98.76%
String Accuracy	96.13%

Table 2: Recognition accuracy on database of rejects of commercial OCR machine.

the recognition results. Approximately half of the database was used for training and the rest for testing. The results of the experiment are summarized in Table 2.

If this level of accuracy could be maintained in the field, the compounded error rate of the combination of the commercial OCR machine with the HMM algorithm would be reduced to 0.0062%, or 1 error in 16,000 characters. This integration of the two algorithms, however, has not been done.

More than half of the residual errors reported in Table 2 are substitutions caused by highly confusable pairs, such as O and 0, which could be corrected by the use of contextual information, so that a reduction of the error rate of at least a factor two could be achieved if this information were incorporated.

Experiments were also conducted on Database II to assess the relative importance of some of the enhancements of the basic algorithm that were described in previous sections. Table 3 shows the impact on performance of removing these enhancements.

It can be seen that the use of Gaussian state duration

Enhancement Removed	Char. Acc.	Str. Acc.
State dur. post.	91.73%	75.80%
Size normaliz.	94.46%	80.96%
Vector features	91.40%	75.53%

Table 3: Effect on performance of eliminating algorithm enhancements.

Beam Width	Char. Acc.	Str. Acc.	Time/char
50	97.62%	93.87%	0.65s
100	98.76%	96.13%	0.81s
200	98.83%	96.41%	1.15s
500	98.94%	96.75%	2.17s

Table 4: Accuracy and computation time as functions of the beam width.

postprocessing and the use of vector features are the most important. The importance of size normalization depends on the size variability in the test set.

In these experiments, beam search was used to reduce computation time. Table 4 shows the computation time on an SGI Indigo workstation for several values of beam width, and the associated effect on performance (the results of Table 2 correspond to a beam width of 100).

8. CONCLUSION

We described an algorithm for the recognition of connected and degraded text based on PHMMs. A number of enhancements have been added to the basic algorithm of [1], including incorporation of Gaussian state duration densities, size normalization done jointly with recognition, and the use of a feature vector that incorporates information about the orientation of line segments. We also reported on the results of extensive experiments on two databases. In one of them, use of a PHMMs and a grammar results in recognition accuracy of 99.65% in the presence of text so degraded that it is almost unrecognizable even for a human observer. In the other, accuracy is 98.76%.

REFERENCES

[1] E.Levin and R.Pieraccini, "Dynamic planar warping for optical character recognition", *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, California, March 23-26, 1992.

[2] L.R.Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257-286, Feb. 1989.

[3] S.Kuo and O.E.Agazzi, "Machine vision for keyword spotting using pseudo 2D hidden Markov models", *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing* Minneapolis, April 27-30, 1993.

[4] L.R.Rabiner, B.H.Juang, S.E.Levinson, and M.M.Sondhi, "Recognition of isolated digits using hidden Markov models with continuous mixture densities", *AT&T Technical Journal*, Vol. 64, No.6, pp. 1211-1233, July-August 1985.

[5] M.Bokser, "Omnidocument technologies", *Proceedings of the IEEE*, vol.80, No.7, pp.1066-1078, July 1992.