OCR of Degraded Documents using HMM-Based Techniques

**Issam Bazzi, Premkumar Natarajan, Richard Schwartz,
Andras Kornai, Zhidong Lu, John Makhoul**

BBN Technologies, GTE Corporation
70 Fawcett Street, Cambridge MA, 02138
ibazzi@bbn.com

Abstract

We present an OCR system for handling degraded documents, such as faxed text. The basic system utilizes the BBN BYBLOS OCR system, which uses a Hidden Markov Model (HMM) approach for training and recognition. To handle degraded documents, we present two approaches, which can be applied individually or jointly. In the first approach, we train the system on documents that exhibit the expected kind of degradation. For example, to perform OCR on fax documents, we train the system on fax data. In the second approach, the system performs unsupervised adaptation on each page to be recognized in such a way as to maximize a desired objective function. Several objective functions were attempted: Maximum Likelihood Linear Regression (MLLR), Maximum a Posteriori (MAP), and Leave-One-Out MAP. We report on results using the above approaches on fax text images generated from the University of Washington English Image Database I. Applying adaptation techniques, in addition to training on fax, we have reduced the character error rate by a factor of three from the base condition.

1 Introduction

In earlier papers [6,8], we presented an HMM-based OCR system referred to as the BBN BYBLOS OCR system in this paper, incorporating the BBN BYBLOS continuous speech recognition system. The BYBLOS OCR system uses a character model trained on a corpus of text images, a lexicon and a grammar. A brief review of the BYBLOS OCR system is provided in the following section.

While our earlier papers reported on the performance of the BYBLOS OCR system on data from the University of Washington English Image Database I (UW corpus) [9], in this paper we present techniques for dealing with degraded documents within the framework of the BYBLOS OCR system. The accuracy of OCR systems is fundamentally dependent upon the quality of the scanned text image [3]. Common document processing operations such as faxing induce significant degradations in image quality. Part of the degradation is due to the low resolution scanning devices in fax machines and part of the degradation is from the printing process. Sometimes transmission noise adds to the degradation. While existing literature contains examples of algorithms for processing the degraded image to enhance quality [2], this paper focuses on model-based techniques for handling the degradations within the OCR system. For our recognition experiments on degraded data we have used fax-degraded documents generated from *clean* documents in the UW corpus.

The paper is organized as follows. In section 2 we provide a brief review of the BYBLOS OCR system along with some background information. In section 3 we present recognition results using a model trained on degraded documents as well as the results obtained using a system trained on clean data. In Section 4 we discuss and demonstrate the use of adaptation to further improve recognition accuracy. A summary and conclusion in Section 5 follow this.

2 System Overview

This section gives a brief review of the BBN BYBLOS OCR system. For a more detailed description the reader is referred to [6]. A pictorial representation of the system is given in Fig. 1. In the figure, knowledge sources are depicted by ellipses and are dependent on the particular language or script. The OCR system components themselves are identified by rectangular boxes and are independent of the particular language or script. Thus, the same OCR system can be configured to perform recognition on any language.

At the top level, the OCR system can be subdivided into two basic functional components:

training and recognition. Both, training and recognition share a common pre-processing and feature extraction stage. The pre-processing and feature extraction stage starts off by first deskewing the scanned image and then locating the positions of the text lines on the deskewed image.

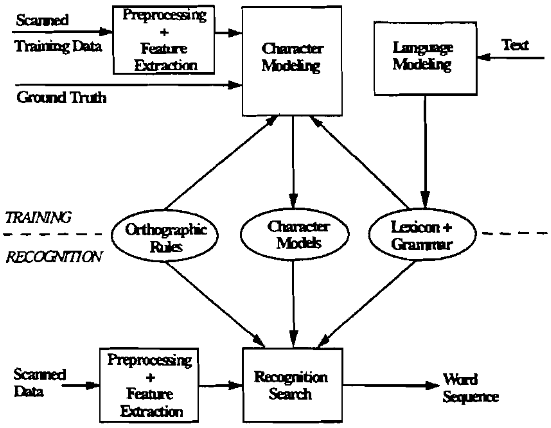

Figure 1: Block diagram of BBN BYBLOS OCR system

The feature extraction program computes a feature vector as a function of the horizontal position within a line, see Fig. 2. First, each line of text is horizontally segmented into a sequence of thin, overlapping, vertical strips called frames (one frame is shown in Fig. 2). For each frame we then compute a language-independent, feature vector that is a numerical representation of the frame.

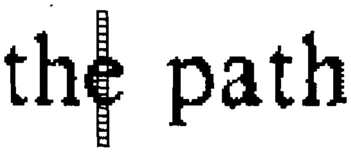

Figure 2: Feature extraction on a line of English text

The OCR system models each character with a multi-state, left-to-right HMM. Each state has an associated output probability distribution over the features. The number of states and the allowable transitions are system parameters that can be set. For our experiments we have used 14-state, left-to-right HMMs with the topology shown in Fig. 3. Training is performed using the Baum-Welch or Forward-Backward algorithm, which aligns the feature vectors with the character-models to obtain maximum likelihood estimates of HMM parameters.

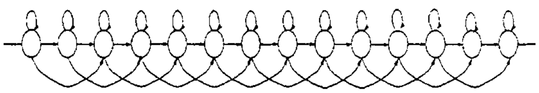

Figure 3: Figure shows a 14-state, left-to-right HMM with self-loops and skips

For our system the HMM parameters are the means and variances of the component gaussians in the gaussian mixture model of the state output probabilities, the mixture component weights and the state transition probabilities. On the other hand during recognition we search for the sequence of characters that is most likely given the feature-vector sequence and the trained character-models, in accordance with the constraints imposed by a lexicon and/or a statistical grammar. The use of a lexicon during recognition is optional but its use generally results in a lower Character Error Rate (CER). The lexicon is estimated from a suitably large text corpus. Typically the grammar (language model), which provides the probability of any character or word sequence, is also estimated from the same corpus.

A significant advantage of HMM-based systems is that they provide a language-independent framework for training and recognition. At the same time, they do not require the training data to be segmented into words or characters, i.e., they automatically train themselves on non-segmented data.

3 Training on Fax

The most straightforward way to improve recognition performance on degraded data is to train on data that has been subjected to a similar degradation process. In this section we present our results with fax-degraded data.

3.1 Parallel Fax Corpus

For our English OCR experiments we used data from the UW corpus. The UW corpus consists of 958 pages scanned from technical articles containing more than 11000 zones of text. For our experiments we randomly selected 95 zones for training and 36 zones for testing. To generate the fax-degraded documents, the selected documents from the UW corpus were first printed on paper. The printed images were then faxed from one plain paper fax machine to another and the faxed images were scanned into bitmaps on the computer. The procedure used for generating the faxed data resulted in fax-degraded training and test corpora that mirrors the clean data from the UW English database. Also, this design allows us to compare the recognition results on the degraded documents with the results on corresponding *clean* documents from the UW corpus.

3.2 Fax Training Results

In our first set of experiments we trained the system using three different training data sets: clean data alone, fax data alone and a mixture of the clean and fax data. For each training condition we tested the system on both clean and faxed data. The output of the recognizer was compared with the reference transcriptions and the average character error rate (CER) was measured by adding the number of substitutions, deletions and insertions to obtain the total number of errors, and then dividing the total number of errors by the total number of characters in the reference transcriptions. The CER's for different training conditions are listed in Table 1.

Table 1: Character Error Rates Under Different Training Conditions

Training	CER % (Clean Test)	CER % (Fax Test)
Clean Only	0.6	5.3
Clean + Fax	0.6	2.7
Fax Only	1.0	2.2

For the model trained on clean data alone, the CER on the fax data, 5.3%, is about nine times higher than the CER on clean data, 0.6%. By training the system on the fax training data we were able to bring down the error rate on the fax test data from 5.3% to 2.2%. At the same time the CER on clean data increased from 0.6% to 1.0%. With the aim of restoring the performance on clean data while maintaining the improved accuracy on the fax data we trained our system on a mix of the clean and fax training. Using this system we achieved a CER of 0.6% on clean data and 2.7% on faxed data.

The fact that a model trained on fax data alone yields a CER of 2.2% on the fax test set while the model trained on clean data alone yields a CER of 0.6% on the clean test set indicates that the recognition of fax documents is an inherently more difficult problem than the problem of recognizing clean documents. In the next section we discuss the use of powerful adaptation techniques to further improve the accuracy of recognition on fax-degraded documents.

4 Adaptation

Adaptation is the process of adjusting the parameters of an initial trained model so as to improve performance on a particular document. Adaptation techniques for HMMs have been used earlier by researchers in the speech community [4,5]. For example in speech recognition systems, a speaker-independent (SI) model is first trained on speech data from many speakers. At recognition time, the SI model is then adapted to each speaker to better model the finer variations for that particular speaker. Similarly for OCR, we can first train a Document Independent (DI) model on data from many documents. We may then adapt the parameters of this DI model using adaptation data for a particular document in such a way as to improve the recognition accuracy for that document.

Adaptation techniques can be broadly divided into two categories: supervised adaptation and unsupervised adaptation. In supervised adaptation the character transcriptions for the adaptation data are provided whereas in unsupervised adaptation we first use the DI model to recognize the document and then use the errorful, recognized text as the transcriptions for the adaptation data. Using the adapted model we can then recognize the document again, typically with higher accuracy.

In the following we present a brief description of unsupervised adaptation using two popular objective functions, the Likelihood function and the Posterior probability. The technique based on maximizing the Likelihood function of the adaptation data is referred to as the Maximum Likelihood Linear Regression (MLLR) technique while the technique based on maximizing the posterior probability of the adaptation data is referred to as the Maximum A Posteriori (MAP) technique. In both cases, the technique must deal with the fact that we do not have sufficient data to re-estimate all the parameters.

4.1 MLLR Adaptation

The MLLR technique handles the data-insufficiency problem by first segmenting the gaussians into a few distinctive sets and then inferring a shared transformation for all the gaussians in each set. In our MLLR adaptation program we only re-estimate the means of the gaussians. Thus, we do not re-estimate transition probabilities, mixture component weights or mixture component covariance's and these parameters take their values from the original model set. Mathematically the MLLR method is described as follows.

$$\lambda_{mllr} = \text{argmax}_{\lambda}\ P(X|\lambda)$$

where λ is the model parameter vector, X is the observation vector and λ_{mllr} is the model parameter vector that maximizes the likelihood function, $P(X|\lambda)$.

4.2 MAP Adaptation

Unlike the MLLR technique, the MAP approach deals with the lack of data by incorporating prior information into the adaptation process. Usually the prior knowledge is obtained from a prior training stage where prior distributions are estimated for the parameters that are to be adapted. The MAP technique infers a separate transformation for each gaussian in the model. Since a larger number of transformations are estimated, MAP adaptation typically requires more data than the MLLR technique. In addition the MAP technique also imposes the additional burden of training the prior distributions for the model parameters to be re-estimated.

Mathematically the MAP procedure is described as follows,

$$\lambda_{map} = \text{argmax}_{\lambda}\ P(\lambda| X)$$

where λ is the unknown model parameter vector, X is the observation vector and λ_{map} is the model parameter vector that maximizes the posterior probability $P(\lambda| X)$. The Posterior probability, $P(\lambda|X)$, may be computed using Bayes' Law as follows:

$$P(\lambda|X)=P(X|\lambda)\ P(\lambda)/ P(X)$$

and since the observation probability, P (X), is independent of λ, we may rewrite the MAP equation as,

$$\lambda_{map} = \text{argmax}_{\lambda}\ [P(X|\lambda)\ P(\lambda)]$$

For any document, the resulting MAP adapted model is an interpolation between a DI model trained on data from many documents and the document dependent (DD) model trained on data from that particular document. The more adaptation data available for the particular document, the closer the MAP model is to the DD model. It is typical to have more adaptation data available in OCR than in speech. For example, while using OCR on a whole book or newspaper, a large amount of adaptation data becomes available in an incremental fashion. For our experiments we implemented adaptation for each text zone separately.

4.3 Results with Adaptation

For our first set of adaptation experiments we started off by using the model trained on the mix of clean and fax data as our Document Independent (DI) model. We performed a first pass of recognition using the DI model and used the recognition results of the first pass as input to the adaptation program. The results of the experiments are listed in Table 2. As can be seen from the figures in Table 2, MLLR is seen to be better than MAP adaptation; a conclusion that is counter-intuitive. A detailed analysis of the errors showed that the MAP adapted model tended to repeat the recognition errors made in the first pass of recognition, indicating that the MAP adaptation was overfitting the models to the adaptation data thereby *memorizing* the errors. To get over this problem we devised a different strategy, called the **Leave-One Out MAP** technique, for implementing MAP adaptation. In the Leave-One Out technique adaptation is done at the sentence (just one line of text, in this context) level and the adaptation data consists of all the sentences except the one to be recognized.

Table 2: Results of Using Adaptation Techniques on Fax Data

Training Data/ AdaptationCondition	CER % (Fax)
Clean + Fax / No Adaptation	2.7
Clean + Fax / MAP	2.5
Clean + Fax / MLLR	2.2
Clean + Fax / Leave-One-Out MAP	2.1

From Table 2, it can be seen that the Leave-One Out MAP technique is indeed much better than the basic MAP technique and slightly better than the MLLR adaptation technique. The MLLR technique, on the other hand, is easier to implement and computationally much less expensive than the MAP. Based on this we have settled on the use of the MLLR technique as the algorithm of choice for our adaptation system.

For our final adaptation experiment we used the model trained on the fax data alone as the initial DI model and used the MLLR adaptation technique to adapt the models. As before, we ran one pass of recognition using the DI models and used the recognized text as the adaptation data. After adaptation the error rate decreased from 2.2% to 1.7%. The results of this experiment along are listed in Table 3.

Table 3: Summary of Results with and Without Adaptation

Training Data / Adaptation Method	CER% (Fax)
Fax only / No Adaptation	2.2
Fax only / MLLR Adaptation	1.7

5 Conclusions

In this paper we have addressed the problem of performing character recognition on fax degraded documents. The techniques presented are general and may be applied to other kinds of degradations or noise environments. We have demonstrated that HMM-based systems can easily be trained to model different degradations by using a properly chosen training data set. As indicated by experimental results, a properly chosen training set can significantly reduce the recognition error rate.

Another useful aspect of HMM-based systems is the possibility of adapting the models to a particular document. Experimental results presented in the paper indicate the substantial improvement that adaptation can offer. A comparative analysis of two basic adaptation techniques, MAP adaptation and MLLR adaptation, is provided. The Leave-One Out MAP technique is marginally superior to the MLLR method but the small gain in accuracy does not justify the added complexity and computational expense. As such, the MLLR method is the adaptation technique of choice for our system.

6 References

[1] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, Vol. 77, No. 2, pp. 257-286, Feb. 1989

[2] J.D. Hobby and H.S. Baird, "Degraded Character Image Restoration," *Fifth Annual Symposium on Document Analysis and Information Retrieval*, Alexis Park Resort, Las Vegas, Nevada, pp. 233-245, April 15-17, 1996

[3] S.V. Rice, J. Kanai, T.A. Nartker, "An Evaluation of OCR Accuracy,"In *Information Science Research Institute, 1993 Annual Research Report*, University of Nevada, Las Vegas, pp. 9-20, 1993

[4] J.L. Gauvain and C.H. Lee, "Maximum-a-posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,"In *IEEE Transactions on Speech and Audio Processing*, Vol. 2, pp. 291-298, 1994

[5] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs," In *Computer Speech and Language*, Vol. 9, pp. 171-186, 1995

[6] R. Schwartz, C. LaPre, J. Makhoul, C. Raphael, and Y Zhao, "Language-Independent OCR Using a Continuous Speech Recognition System," *Proc. Int. Conf. on Pattern Recognition*, Vienna, Austria, pp. 99-103, August 1996

[7] I. Bazzi, C. LaPre, J. Makhoul, C. Rapahel, R. Schwartz, "Omnifont and Unlimited Vocabulary OCR for English and Arabic," *Proc. Int. Conf. Doc. Analysis and Recognition*, Ulm, Germany, pp. 842-845, Aug. 1997

[8] L. Nguyen, T. Anastasakos, F. Kubala, C. LaPre, J. Makhoul, R. Schwartz, N. Yuan, G. Zavaliagkos, and Y. Zhao, "The 1994 BBN/BYBLOS Speech Recognition System," *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, Morgan Kaufmann Publishers, pp. 77-81, January 1995

[9] I.T. Phillips, S. Chen, and R.M. Haralick, "CD-ROM document database standard," *Proc. Int. Conf. Document Analysis and Recognition*, Tsukuba City, Japan, pp. 478-483, Oct. 1993