Data and Structure adaptive Optical Character Reader

Sanjana M A[A], Usha K[B]

[A]Dept of Computer Science & Engineering, NSS College of Engineering, Palakkad ,Kerala
[B]Dept of Computer Science & Engineering, NSS College of Engineering, Palakkad ,Kerala

Abstract

Current Optical Character Reader (OCR) technologies can achieve high character recognition rates when applied on contemporary documents with good printing quality. However, these OCR systems achieve notably lower accuracy on degraded documents. This is especially the case for historical documents exhibiting degraded papers, low quality printing, ancient linguistic particularities, and old-style fonts which may be unknown to a standard OCR. For this purpose, adaptive hidden Markov model (HMM) OCR systems were designed that try to adapt to the specificities of the document to be recognized. We combine transition histogram, and structure and parameter adaptive HMM for printed character recognition. This work is an extension of paper [1].

Keywords: *Hidden Markov models, structure adaptation, parameter adaptation, printed text recognition, historical documents, and transition histograms.*

I. Introduction

To create new either nature is analyzed or existing knowledge is used. A part of knowledge which is carved ancient books and literatures seem to be abundant database for analytical and formative studies hence documents digitization projects comes to existent. The last two decades have seen the creation of paper documents digitization projects in libraries and archives. Some of the most representative projects are Google Books and Gallica. The final aim of these projects is to give remote access to digitized collections with a particular interest and emphasis on historical documents for which high resolution digital facsimiles provide an alternate media allowing the preservation of the paper documents. Digital access to these documents depends on the effectiveness of optical character recognition (OCR). Current OCR technologies can achieve high character recognition rates when applied on contemporary documents with good printing quality. However, these OCR systems achieve notably lower accuracy on degraded documents. This is especially the case for historical documents exhibiting degraded papers, low quality printing, ancient linguistic particularities, and old-style fonts which may be unknown to a standard OCR.

An OCR system first analyzes the layout of a document in order to extract the text blocks which are then segmented into lines. The success of this preliminary step is essential for the further recognition process. Present-day industrial OCR systems use a technology based on a segmentation stage that segments these text lines into characters which are then recognized in isolation using dedicated classifiers. This character segmentation stage is often ineffective on handwritten documents and also fails quite often when applied to some problematic printed documents such as highly degraded ones, or old documents (because of their degradation and their similarities in some aspects with handwritten documents3). The state-of-the-art systems try to correct these segmentation errors using information about character shapes, alternating segmentation and recognition steps in order to optimize the character recognition process [2]. Nevertheless, this segmentation problem remains an important limitation to the effectiveness of this technology.

Some segmentation-free recognition techniques have been proposed to achieve recognition and segmentation in conjunction. These methods use an implicit segmentation approach (also called “sliding window”), cutting the text lines at regular intervals (narrower than the width of the characters which are, therefore, over-segmented) and leaving it to the classifier to determine the boundaries of characters together with the best sequence of characters. The most commonly-used classifier falling into this framework is the hidden Markov model (HMM) [3]. The main advantage of HMMs is their ability to take into account all the information available to build a decision, by the combination of a data model (are presentation of the character shape information) and a model of the expected solution (the language model in a wider sense). The effectiveness of HMMs has been demonstrated in various domains such as speech processing [4], handwriting recognition (see [5] for a recent survey) and cursive printed scripts recognition [6], [7].

II. RELATED WORK

Statistical classifiers used for character recognition can cope with a large diversity of shapes. These generic (“multifont”) classifiers are trained on huge data sets [4]. In principle, these systems are able to recognize characters of any font but none of them can recognize with the same accuracy any type of font. It is generally agreed that a general-purpose system has much lower accuracy than dedicated (or ”monofont”) systems. Unfortunately, the large labeled data sets that are required to train such monofont systems are generally missing. One possible solution to the lack of labeled data is to specialize a generic system on new data using a limited amount of labeled data or without labeled data. For this purpose, adaptive systems were designed that try to adapt to the specificities of the document to be recognized. The first kind of adaptive systems consider training a new classifier with character images extracted from the document to be recognized, and labeled. Similar" prototyping" techniques have been introduced in most commercial OCRs [2]. Such methods can perform well but their effectiveness depends on the accuracy of the initial recognizer. A second kind of adaptive systems makes use of adaptation techniques. In practice, adaptation consists in tuning a general-purpose recognizer so as to better recognize new data, assuming the isogeny of the data (for example: the same font is used in the entire document). Thus, adaptation can be considered here as the specialization of an already trained generic classifier. The adaptation of an existing recognizer is particularly interesting when the available amount of labeled data is too small to train a new classifier from scratch [6]. Consequently, in these adaptation uses techniques that are similar to those used for training.

For HMM-based recognizers, efficient algorithms, namely maximum a posteriori (MAP) [7] and maximum likelihood linear regression (MLLR) [8] have been proposed for this adaptation task. To our knowledge, all these existing HMM adaptation algorithms only deal with the adaptation of the data model, leaving the HMM structure unchanged. However, in the literature, structure optimization has been addressed for training purpose. Indeed, HMM model structure has an important influence on its accuracy.

To train HMM models, there exist two well-established iterative training algorithms based on an Expectation - Maximization estimation: the Baum-Welch and the Viterbi algorithms. Given a predetermined structure, these algorithms estimate the model parameters using the maximum likelihood (ML) criterion. They are commonly used for HMM training but have some known defects. First, they require a large amount of training data to correctly estimate HMM parameters. Second, they are sensitive to the initialization as they do not necessarily converge to the global optimum. Finally, the HMM structure has to be determined prior to learning.

The use of HMM data model adaptation is an interesting approach to increase the accuracy of a generic system on a particular data set. But this technique remains limited when data become significantly different from the training data. In particular, we know that these methods modify the data model but retain the structure of the initial model unchanged, despite the fact that this structure, determined on the training data, is not necessarily the most appropriate to represent the new data set. Now, it is a known fact that choosing a good structure is important to build an effective HMM model. We have seen that great effort was made on the optimization of the HMM structure. Previous work includes heuristic as well as model selection methods, which are divided into " global" and " local" approaches (that differ in how they explore the HMM structure search-space). Despite the relatively large amount of studies devoted to HMM structure optimization, none of them have been developed in the context of adaptation. Therefore, we investigated how some of these techniques could be introduced in an adaptation task, since they appear to have potential merits for improving the existing HMM adaptation algorithms.

III. PROPOSED SYSTEM

The proposed system performs feature extraction and recognition in two phases: first phase segments the image into characters and uses transition histogram character image to find a match between it and the transition histogram of labeled data. So that it can recognize high confident characters. For degraded image segments structure and parameter adaptation of HMMs used as the final phase for recognizing characters.

3.1 Transition histogram

The goal of the feature extraction phase is to extract, in an ordered way a set of relevant features that reduce redundancy in the word image while preserving the discriminative information for recognition. Feature set is based on the analysis of the bi-dimensional contour transition histogram of each segment in the horizontal and vertical directions.

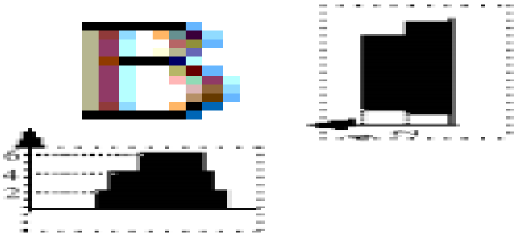

Figure 1: Transition histogram for character „B‟.

3.2 Combination of Structure and Parameter Adaptation of Sequential HMMs

In order to combine structure as well as parameter adaptation into a single framework, first derived an optimization scheme that can handle the main constraint of adaptation principle, which is the scarcity of labeled data. In targeted application, the parameters as well as the structure of a large set of models (89 characters HMM models) have to be re-estimated. The estimation of the criteria used in the literature for guiding the structure optimization requires a quite large amount of data. In this context, there is a possibility that the labeled adaptation data set does not comprise enough samples to adapt each model. To overcome this difficulty, a semi-supervised adaptation framework designed where the available new data is divided into two subsets: the small adaptation set contains only labeled data, and the larger validation set contains only unlabeled data. These two data sets are mutually exclusive. The adaptation data set is used for the re-estimation of the parameters of Gaussian mixtures while the unlabeled validation data set is used to optimize the structure modification operations, by estimating the criterion used by each algorithm.

To determine a strategy to explore the HMM structure search space, two structure adaptation algorithms designed that are inspired by the two main families of HMM structure optimization methods. The first explores the search-space using a local model selection approach, and uses the likelihood as a guiding criterion for structure adaptation. The second derives its guiding criterion from data heuristics.

Based on these initial choices, the general adaptation framework is proceeding by alternatively adapting the parameters and the structure of the model. The parameter adaptation stage is based on supervised MAP or MLLR, while the structure adaptation stage involves two basic merging and splitting operators of HMM states.

3.2.1 Basic Operations Used for HMM Structure Adaptation

Let consider an initial model with a given number of states. In order to converge to an optimal number of states with regard to the new data (greater or lower than the initial number), both state splitting and state merging operations are allowed. The algorithms are restricted to the case of left/ right models, which allows us to take advantage of the left/ right a priori when performing an operation of structure modification. This is clearly an advantage because it reduces the number of candidates for state merging operations (only two successive states can be merged, contrarily to fully-connected HMM models). Also assume that all HMM states have the same number of Gaussians as required by the use of MLLR where matrix operations are possible only if this constraint is satisfied.

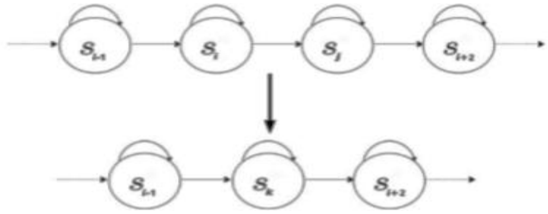

Figure 2: State deletion

State deletion is performed by merging two consecutive states of the model (see Figure 2). Let G be the number of Gaussian components associated to each state, the data model of the new state resulting from the merging process is originally associated to 2G components. In order to avoid a growth phenomenon, the number of components of this new state is kept constant. For this purpose, the application of a mixture collapsing algorithm that uses a particular clustering strategy is introduced. This algorithm iteratively merges the two closest Gaussian components in the mixture until the desired number of components is reached. The self-transition probability a_{kk} of this new state is computed in the same way as in so that its length l_k is half the sum of the lengths l_i and l_j of the two initial states giving rise to the update Equations (6):

$$l_k = \frac{l_i + l_j}{2}; \; a_{kk} = \frac{a_{ii} + a_{jj} - 2a_{ii}a_{jj}}{2 - a_{ii} - a_{jj}} \qquad (6)$$

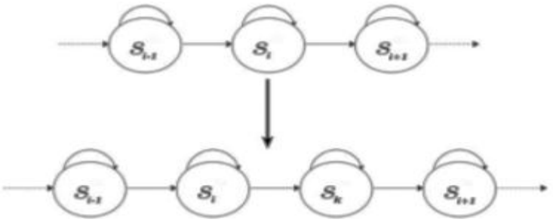

Figure 3: State insertion

Regarding the state insertion, an original procedure quite different from existing methods is developed. Existing approaches usually perform state insertion (see Figure 3) by duplicating one state s_i, thus creating two new states s_i and s_k identical to the initial state. The self-transition probabilities of these states a_{kk} and a_{ii} are updated so that the sum of the lengths of the two states is twice the length of the initial splitted state l_{init} (Eq. (7)):

$$l_k + l_i = 2l_i = 2l_{init}; \;\; a_{kk} = a_{ii} = a_{init} \qquad (7)$$

The re-estimation procedure following this duplication should result in differentiating these two states. However this duplication rule is inappropriate for MLLR adaptation because the two states share the same emission probability and thus they would be modified in the same manner and would remain identical all along the MLLR adaptation process. Therefore, the splitting procedure is modified as follows. A new state is inserted between the state to split s_{split} and s_{adj} (the closest adjacent neighbour of s_{split} according to the Kullback-Leibler divergence between the emission distributions of the two states). This new state is the result of merging s_{split} and s_{adj}. This procedure preserves the left/right topology of the model, as depicted in Figure 3.

In accordance with some model selection approaches, rely on the Gaussian mixtures of the states for the choice of the two states to merge (in case of state deletion) or the state to split (in case of state insertion). The two successive states in the model with the closest emission probability densities (in term of Kullback-Leibler divergence) are chose to merge. Also assume that the greater similarity between these two GMMs minimizes the risk of information loss resulting from state merging.

In the same way, temporal split on the state with the GMM of higher variance proceeded, because by allowing one more state, expecting a reduction in the variance and thus a better model with a higher likelihood on the adaptation data. Temporal split is not suitable when training generic models. This defect is ineffective at creating models that are specialized on a unique font (i.e. the developing system).

Based on the previous operations, two adaptation algorithms Model selection based structure adaptation (MS-SA)[1] and State duration based structure adaptation (SD-SA)[1] used for recognizing characters.

IV. CONCLUSION AND FUTURE WORK

The proposed system performs feature extraction and recognition by using transition histogram for high confident characters and for degraded image segments structure and parameter adaptation of HMMs.

The adaptation algorithms MA-SA, SD-SA improve hidden Markov model adaptation by a combination of structure optimization procedures with data model adaptation. These adaptation algorithms can be improved in many respects. First, a more refined modelling of the data distribution in the representation space can be envisaged in order to reflect the new data statistics, either by adapting the number of components of each Gaussian mixture to take into account the statistics of the new data, or by including contextual state splitting procedures. Second, these algorithms have only been applied to left- right HMMs. It would be interesting to evaluate a generalization of these algorithms to other types of HMM topology.

HMMs can also specified for transition histograms of labeled data and adaptation to the same improve accuracy and can avoid the complexity of using MA-SA, SD-SA for recognition purpose.

Regarding the printed text recognition area further improvements of the system can be by the integration of linguistic constraints such as lexicon and language models. This word recognition system could easily be trained to deal with different printed cursive scripts such as Malayalam. The evaluation of the adaptation algorithms on handwritten documents, for example, on a writer adaptation task, would be of interest. The approach, based on hidden Markov models, is easily transferable to this type of data.

ACKNOWLEDGMENT

The authors would like to thank the professors of NSSCE, Palakkad for the suggestions and support on this paper.

REFERENCES

[1] Kamel Ait Mohand, ThierryPaquet, Nicolas Ragot, “Combining Structure and Parameter Adaptation of HMMs for Printed Text Recognition. ” IEEE Trans. Pattern Anal. Mach. Intell., 2014

[2] Marosi, “Industrial OCR Approaches: Architecture, Algorithms, and Adaptation Techniques,” Proc. SPIE, vol. 6500, pp. 650002.1-650002.10, Jan. 2007.

[3] L.R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” Proc. IEEE, vol. 77, no. 2, pp. 257-286, http://dx.doi.org/10.1109/5.18626, Feb. 1989.

[4] H.S. Baird and R. Fossey, “A 100-Font Classifier,” Proc. First Int‟l Conf. Document Analysis and Recognition (ICDAR ‟91), pp. 332-340, 1991.

[5] Y. Xu and G. Nsagy, “Prototype Extraction and Adaptive OCR,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 12, pp. 1280-1296, Dec. 1999.

[6] H.S. Baird and G. Nagy, “Self-Correcting 100-Font Classifier,” Proc. Document Recognition, vol. 2181 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conf. Series, pp. 106-115, Mar. 1994.

[7] J. Gauvain and C. Lee, “Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,” IEEE Trans. Speech and Audio Processing, vol. 2, no. 2, pp. 291-299, Apr. 1994.

[8] C.J. Leggetter and P.C. Woodland, “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hid- den Markov Models,” Computer Speech & Language, vol. 9, no. 2, pp. 171-185, http://www.sciencedirect.com/science/article/ B6WCW-45PTX5N-4/2/becda3edade8 783a68e5875f1645ce3e, 1995.