Rejection Strategies for Offline Handwritten Sentence Recognition

Matthias Zimmermann, Roman Bertolami, Horst Bunke
Institute of Informatics and Applied Mathematics
University of Bern, Neubrückstrasse 10, CH-3012 Bern, Switzerland
{zimmerma, bertolam, bunke}@iam.unibe.ch

Abstract

This paper investigates three different rejection strategies for offline handwritten sentence recognition. The rejection strategies are implemented as a postprocessing step of a Hidden Markov Model based text recognition system and are based on confidence measures derived from a list of candidate sentences produced by the recognizer. The better performing confidence measures make use of the fact that the recognizer integrates a word bigram language model. Experimental results on extracted sentences from the IAM database validate the effectiveness of the proposed rejection strategies.

1. Introduction

Writer independent recognition of general handwritten text is still considered a very difficult problem. Depending on the experimental setup word recognition rates between 50% and 80% are reported in the literature [6, 9, 12]. For many application such low recognition rates are not acceptable. If a complete automation of the transcription process is not required, rejection strategies may be used to reject certain parts of the handwritten text to achieve the required level of accuracies on the remaining input.

The rejection of input (e.g. letters, words, sentences) is typically based on a confidence measure. If the confidence measure exceeds a specific threshold, the recognition result is accepted. Otherwise, it is rejected. In the literature a large number of confidence measures are proposed depending on the application and the nature of the underlying recognizer. Examples of such confidence measures are provided in the next section which briefly surveys works in the domain of online and offline handwriting recognition. In contrast to previously published works in the domain of handwriting recognition our rejection strategies are based on the fact that a statistical language model supports the recognition process. So far, such rejection strategies have only been applied in the domain of continuous speech recognition.

The remaining part of this paper is organized as follows. Sec. 3 describes the investigated confidence measures. Experimental results are provided in Sec. 4 and conclusions are drawn in the last section of this paper.

2. Related Work

In offline handwriting recognition rejection strategies for both address reading [1] and check processing [2] systems have been presented. In [1] four different strategies to reject isolated handwritten street and city names are described which are based on normalized likelihoods and the estimation of posterior probabilities. For the likelihood normalization the number of frames is used. In the case of estimation of the posterior probabilities the normalization is performed using a garbage model, a two-best recognition and a character based recognizer. In [2] an artificial neural network computes a confidence measure from a set of 10-20 features. Most features represent quantities derived from the scores of the n-best candidate list which is produced by the recognizer (e.g. the log of the best score or the estimated posterior probabilities).

For the case of online handwriting recognition similar confidence measures were used. In [7] an artificial neural network analogously to [2] has been used to decide when to reject isolated characters or words. Four different letter level confidence measures were applied in [5]. The letter confidence measures are defined using different types of anti-models. Word confidence measures are then derived from the letter confidence measures.

Additional confidence measures are frequently used in the field of continuous speech recognition [8, 10] which are based on the integration of a statistical language model in the recognition process. The integration of the language model can be controlled by two factors. The *Grammar Scale Factor* (GSF) which weights the impact of the statistical language model against the acoustic recognition of the utterance (sentence) and the *Word Insertion Penalty* (WIP[1])

1 Both the GSF and the WIP are defined formally in Sec. 3 and explained in more detail in [12].

0-7695-2128-2/04 $20.00 (C) 2004 IEEE

Transcription:	Mr.	Lisbon	has	it	taped
Hypothesis:	Mr.	Lisbon	had		escaped
Alternative 1:	Mr.	Lisbon	has	it	taped
Alternative 2:	Mr.	Lisbon	has	it	taped
Alternative 3:	Mr.	Lisbon	had		escaped
n:	3	3	1		1

Figure 1. Counting the number of times *n* a hypothesized word occurs in alternative candidate sentences.

which controls the segmentation rate of the recognizer. In continuous speech recognition it has been observed that those words of the hypothesized sentence that are very sensitive to a specific integration of the statistical language model are frequently recognized incorrectly. Such words are therefore to be rejected. In [8] only the GSF is varied to identify words to be rejected while both the GSF and the WIP are varied in [10].

3. Rejection Strategies

The three rejection strategies investigated in this paper are based on confidence measures derived from a list of candidate sentences. In addition to the recognizer's top ranked output which is the hypothesized sentence $W = (w_1, \ldots, w_m)$, the list contains K alternative candidate sentences produced by the recognition process. The probability of word w being recognized correctly can then be defined by $P(C = c \,|\, n, w)$ where $c \in \{correct, incorrect\}$ and $n \in [0..K]$ represents the number of times word w is observed in the K alternative candidate sentences. The presence of word w_i in a given candidate sentence is determined with dynamic string alignment as shown in the example provided in Fig. 1.

If our training set would be large enough, we could estimate this probability for every value of n and all the words w contained in the dictionary. Since most words appear only a few times (many words appear not at all in the training set) expression $P(C = c \,|\, n, w)$ is approximated by $P(C = c \,|\, n)$ as described in [8]. I.e. the assumption is made that this probability does not depend on the hypothesized word w but only on the value of n. The probability $P(C = c \,|\, n)$ is then directly used as the confidence measure. Consequently, a word w is rejected if the confidence measure is less than a specified threshold τ. For the example provided in Fig. 1 both mis-recognized words *had* and *escaped* are rejected if we use n directly as confidence measure and set $\tau = 2$. Note that the word *it* will be ignored since it is not present in the hypothesized candidate sentence.

Let us now consider some methods to generate the K alternative candidate sentences. For this paper we have examined the following three methods:

- n-best list extraction
- variation of grammar scale factor
- variation of grammar scale factor and word insertion penalty

N-Best List Extraction. The most obvious way to get multiple candidate sentences from a recognition process is to generate a n-best list which contains the n most likely sentences for a given image of a handwritten sentence. In our case n is set to $K + 1$ since K alternative candidate sentences in addition to the hypothesized sentence are needed.

Variation of Grammar Scale Factor. It has been shown that words with a high stability concerning the integration of a statistical language model are relatively error-free compared to words that rapidly change when this integration is varied [10]. This means that words which are observed less frequently in the alternative candidate sentences are more likely to be incorrect than words which appear in all or most candidate sentences.

For the integration of a statistical language model in an HMM based recognition system the most likely sentence $\hat{W} = (w_1, \ldots, w_m)$ for a given observation sequence X is computed in the following way [12] .

$$\hat{W} = \underset{W}{argmax} \;\; \log p(X|W) + \alpha \;\log p(W) + m \;\beta$$

where parameter α stands for the GSF and parameter β corresponds to the WIP.

For the method described in this paragraph we only vary the GSF as described in [8]. Consequently, K values in the range $[\alpha - x, \alpha + y]$ are chosen where $x, y > 0$ and α represents the GSF of the hypothesized sentence.

Variation of Grammar Scale Factor and Word Insertion Penalty. The third method to produce multiple candidate sentences is an extension of the method described above. Not only the GSF, but also the WIP is varied, as mentioned in [10]. Therefore K parameter pairs (GSF, WIP) are selected where $GSF \in [\alpha - x, \alpha + y]$ and $WIP \in [\beta - s, \beta + t]$. The value of β corresponds to the WIP of the hypothesized sentence and $s, t > 0$.

4. Experiments and Results

All experiments reported in this paper make use of a Hidden Markov Model (HMM) based handwritten sentence recognition system which uses word bigrams as a statistical language model. The recognition system is based on individual character models with a linear topology and multi-Gaussian output densities (see [4, 12] for details). Furthermore the same experimental setup as described in [12] was used. The training and the test set contain 200 complete English sentences extracted from the segmented IAM

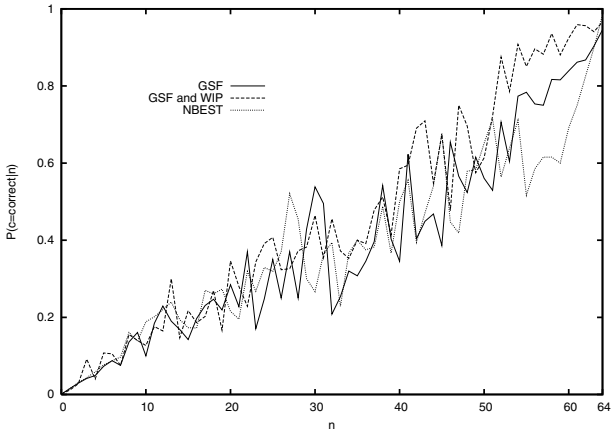

Figure 2. Estimated probability of being correct as a function of *n*

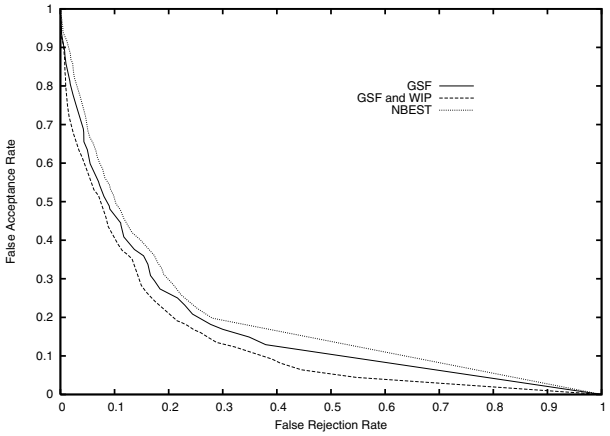

Figure 3. ROC Curves of the different reject models.

database [3, 11]. The 400 sentences with an average length of 23.1 words have been written by 200 individual writers where the first 100 writers are represented in the training set and the second 100 writers contributed to the test set. The lexicon has been closed over the test (training) set and included 8,819 (8,825) words[2]. The training set was used to estimate $P(C=c\,|\,n)$ and the proposed rejection strategies were evaluated on the sentences from the test set.

For the generation of the hypothesized sentence we used $\alpha = 30$ and $\beta = 50$ which has been shown to lead to a maximum recognition rate for this experimental setup [12]. The value of K for the number of alternative candidate sentences was set to 64. For the variation of both the GSF and WIP eight different values for each of the parameters were used. Parameters x and y were set to 30, s to 150 and t to 100, thus $GSF \in [0, 60]$ and $WIP \in [-100, 150]$.

The quantities $P(C=c\,|\,n)$ are estimated using the relative frequencies obtained from the training set as follows

$$P(C=correct\,|\,n) \simeq \frac{\#correct_n}{\#correct_n + \#incorrect_n}$$

where $\#correct_n$ counts correctly recognized words in which n is the number of times a hypothesized word appears in the alternative candidate sentences. The quantity $\#incorrect_n$ is used to count the cases of the words not correctly recognized. Fig. 2 shows the estimated probabilities for each of the three considered models and each of the 65 values of n.

To evaluate the efficiency of the rejection strategies a confusion matrix as defined in [5] is used. A word can either be recognized correctly or incorrectly. In both cases the recognition result may be accepted or rejected by the rejection mechanism which results in one of the following possible outcomes:

- *Correct Accept* (CA) - A correctly recognized word has been accepted by the postprocessor.
- *False Accept* (FA) - A word has not been recognized correctly but has been accepted by the postprocessor.
- *Correct Reject* (CR) - A incorrectly recognized word has been rejected by the postprocessor.
- *False Reject* (FR) - A word that has been recognized correctly has been rejected by the postprocessor.

A *Receiver Operating Characteristic* (ROC) curve can then be constructed by plotting the *False Acceptance Rate* (FAR) against the *False Rejection Rate* (FRR). These measures are defined as follows:

$$FAR = \frac{FA}{FA + CR} \qquad FRR = \frac{FR}{FR + CA}$$

.

The experimental results for the three confidence measures are shown in Fig. 3. We observe that the models based on language model variation perform better than the n-best list approach. The inclusion of the WIP in language model variations results in a clearly superior performance over the confidence measure based on the variation of the GSF alone. Where the confidence measure varying both the GSF and WIP requires 20% false reject rate to attain a false accept rate of 20% a 28% false reject rate is required with the confidence measure based on the n-best lists.

In terms of error-reject statistics, the best confidence measure (which varies both the GSF and the WIP) performs as follows: for a 0% reject rate the word error rate lies at 19.9%. To achieve a word error rate of 5% a reject rate of 29% is required and to get the word error rate below 2% we

2 The closing the lexicon over the test (training) set ensures that all words of the test (training) set are contained in the task lexicon.

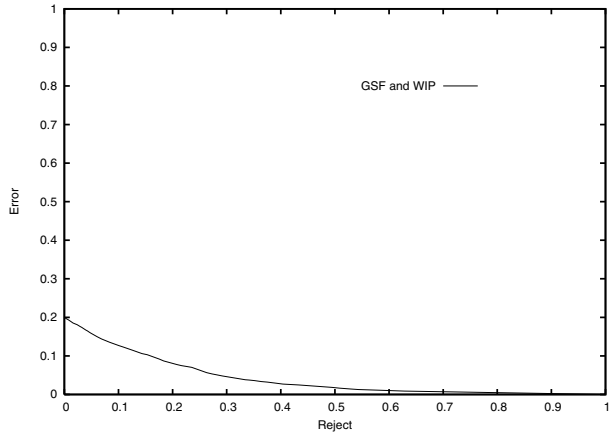

Figure 4. Error-reject plot

need to reject 49% of the words. The full error-reject rate of the best confidence measure is shown in Fig. 4.

5. Conclusions

We have investigated three different confidence measures to implement a rejection strategy for an HMM based offline handwritten sentence recognition system. In contrast to the other works in the domain of handwriting recognition the confidence measures used in this paper are not derived from normalized likelihoods or estimated posterior probabilities but make use of the fact that a statistical language model supports the recognition process.

Based on experiments using complete English sentences extracted from the IAM database we have found that the good results reported in the domain of continuous speech recognition can be confirmed for the case of offline handwritten text recognition. Furthermore, the results achieved with the confidence measure based on the variation of the grammar scale factor and the word insertion penalty compare favorably with previously published works in the domain of handwriting recognition.

Future research will include the combination of the confidence measure proposed in this paper with "traditional" confidence measures derived from the recognition scores provided by the HMM recognition system.

References

[1] A. Brakensiek, J. Rottland, and G. Rigoll. Confidence measures for an address reading system. In *7th Int. Conf. on Document Analysis and Recognition, Edinburgh, Scotland*, volume 1, pages 294–298, 2003.

[2] N. Gorski. Optimizing error-reject trade off in recognition systems. In *4th Int. Conf. on Document Analysis and Recognition*, pages 1092–1096, Ulm, Germany, 1997.

[3] U.-V. Marti and H. Bunke. A full English sentence database for off-line handwriting recognition. In *5th Int. Conf. on Document Analysis and Recognition 99, Bangalore, India*, pages 705–708, 1999.

[4] U.-V. Marti and H. Bunke. Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 15:65–90, 2001.

[5] S. Marukatat, T. Artières, and P. Gallinari. Rejection measures for handwriting sentence recognition. In *8th Int. Workshop on Frontiers in Handwriting Recognition*, pages 24–29, Niagra-on-the-Lake, Canada, 2002.

[6] F. Perraud, C. Viard-Gaudin, E. Morin, and P.-M. Lallican. N-gram and n-class models for on line handwriting recognition. In *7th Int. Conf. on Document Analysis and Recognition, Edinburgh, Scotland*, volume 2, pages 1053–1057, 2003.

[7] J.F. Pitrelli and M.P. Perrone. Confidence modeling for verification post-processing for handwriting recognition. In *8th Int. Workshop on Frontiers in Handwriting Recognition*, pages 30–35, Niagra-on-the-Lake, Canada, August 2002.

[8] A Sanchis, V. Jiménez, and E. Vidal. Efficient use of the grammar scale factor to classify incorrect words in speech recognition verification. In *Proc. 15th Int. Conf. on Pattern Recognition*, pages 278–281, Barcelona, Spain, 2000.

[9] A. Vinciarelli, S. Bengio, and H. Bunke. Offline recognition of unconstrained handwritten texts using HMM and statistical language models. IDIAP-RR 03-22, Dalle Molle Institute for Perceptual Artificial Intelligence, 2003.

[10] T. Zeppenfeld, M. Finke, K. Ries, M. Westphal, and A. Waibel. Recognition of conversational telephone speech using the janus speech engine. In *Proc. ICASSP '97*, pages 1815–1818, Munich, Germany, 1997.

[11] M. Zimmermann and H. Bunke. Automatic segmentation of the IAM off-line handwritten English text database. In *16th Int. Conf. on Pattern Recognition*, volume 4, pages 35–39, Quebec, Canada, August 2002.

[12] M. Zimmermann and H. Bunke. Optimizing the integration of statistical language models in HMM based offline handwritten text recognition. In *Submitted*, 2004.