Rejection strategies for offline handwritten text line recognition

Roman Bertolami * Matthias Zimmermann [1] Horst Bunke

Institute of Computer Science and Applied Mathematics
University of Bern, Neubrückstrasse 10, CH-3012 Bern, Switzerland

Abstract

This paper investigates rejection strategies for unconstrained offline handwritten text line recognition. The rejection strategies depend on various confidence measures that are based on alternative word sequences. The alternative word sequences are derived from specific integration of a statistical language model in the hidden Markov model based recognition system. Extensive experiments on the IAM database validate the proposed schemes and show that the novel confidence measures clearly outperform two baseline systems which use normalised likelihoods and local n-best lists, respectively.

Key words: Handwritten Text Recognition - Rejection Strategies - Statistical Language Model

1 Introduction

After four decades of research, writer independent recognition of unconstrained offline handwritten text is still considered a very difficult problem. For this task, recognition rates between 50% and 80% are reported in literature depending on the experimental setup (Kim et al., 1999; Vinciarelli et al., 2004; Zimmermann and Bunke, 2004b). By implementing rejection strategies in a handwriting recognition system we are able to improve the reliability by rejecting certain parts of the input and increase the accuracy on the remaining

* Corresponding author.
Email addresses: bertolam@iam.unibe.ch (Roman Bertolami), zimmerma@icsi.berkeley.edu (Matthias Zimmermann), bunke@iam.unibe.ch (Horst Bunke).
[1] Present address: International Computer Science Institute (ICSI), Berkeley, USA

text. Furthermore, we are able to detect parts that may not have been recognised correctly and which we should either reject, classify with an additional recognition system, or submit to a human operator.

A common way to reject input units, such as letters, words, or text lines is to compute a confidence measure for each input unit. For such an approach rejection strategies can be formulated as simple thresholding operations. If the confidence measure of a letter, word, or text line exceeds a specific threshold, the recognition result is accepted. Otherwise, it is rejected.

Confidence measures cannot be used only for rejection. They can play an important role in classifier combination methods, as well (Oza et al., 2005). In general, however, confidence measures are domain specific, i.e. a confidence measure that performs well for rejection probably achieves only poor performance in a classifier combination task and vice versa. In this paper we exclusively focus on rejection.

A large number of confidence measures have been proposed in the literature. In contrast to previously published work in the domain of offline handwriting recognition that concentrated on isolated characters or words, we address the problem of rejecting words in the context of their surrounding text taking advantage of the fact that a statistical language model supports the recognition process. So far, confidence measures of this kind have only been applied in the domain of continuous speech recognition (Sanchis et al., 2000; Zeppenfeld et al., 1997). To the knowledge of the authors, it is the first time in this paper that confidence measures based on candidates which are derived from specific integration of a statistical language model are applied in handwriting recognition.

Statistical language models and lexicon driven approaches have shown to enable substantial improvements to text recognition (Bazzi et al., 1999; Brakensiek et al., 2000; Marti and Bunke, 2001; Shridhar et al., 1997; Vinciarelli et al., 2004). The contextual knowledge obtained from the language model helps to reduce the ambiguity of the segmentation. Furthermore, the search space can be reduce because this knowledge often allows one to prune unlikely hypotheses.

This paper builds upon some of our previous work (Zimmermann et al., 2004). Additional confidence measures are presented and experiments are conducted on a much larger scale. In contrast to (Zimmermann et al., 2004) we consider text lines instead of sentences in this paper, which is a more general approach.

The remaining part of this paper is organised as follows. In Sect. 2, related work is reviewed. The underlying recogniser is presented in Sect. 3.1. Next, the generation of the alternative candidates is described in Sect. 3.2, while Sect. 3.3 introduces novel confidence measures proposed in this paper. Experimental

results are provided in Sect. 4 and conclusions are drawn in the last section of this paper.

2 Related Work

In the literature, a large number of confidence measures have been proposed. They depend on the application and the underlying recogniser. In this section related work in the domain of offline and online handwriting recognition, and continuous speech recognition research is reviewed.

2.1 Offline Handwriting Recognition

In offline handwriting recognition confidence measures for address reading (Brakensiek and Rigoll, 2004), cheque processing (Gorski, 1997), character (Pitrelli and Perrone, 2003), and word (Koerich, 2004) recognition systems have been proposed.

Confidence measures for an HMM based handwriting recognition system for German address reading are introduced in (Brakensiek and Rigoll, 2004). In order to reject isolated handwritten street and city names, four different strategies based on normalised likelihoods and the estimation of posterior probabilities are described. For likelihood normalisation the number of frames is used, while for the estimation of posterior probabilities the normalisation is performed using a garbage model, a two-best recognition strategy, and a character-based recogniser.

Rejection strategies for cheque processing systems are presented in (Gorski, 1997), where 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 produced by the recogniser, for example, the log of the best score.

Several confidence measures for an offline handwritten character recognition system are investigated in (Pitrelli and Perrone, 2003). The measures of recognition confidence are recognition score, likelihood ratio, estimated posterior probability, and exponentiated probability. An additional confidence measure is built by using a Multi-Layer Perceptron to combine the individual confidence measures mentioned before.

Various rejection strategies for offline handwritten word recognition are proposed in (Koerich, 2004). Class-dependent, hypothesis-dependent, as well as

a class-independent and hypothesis-independent confidence measures are presented.

2.2 Online Handwriting Recognition

In (Pitrelli and Perrone, 2002) confidence measures are evaluated in the field of online handwriting recognition. These confidence measures are similar to those investigated in (Pitrelli and Perrone, 2003) for offline recognition. An artificial neural network, combining different confidence measures, is used to decide when to reject isolated digits or words.

Various confidence measures for online handwriting recognition are investigated in (Marukatat et al., 2002). The confidence measures are integrated in an isolated word recognition system as well as in a sentence recognition system. Four different letter-level confidence measures based on different implicit anti-models are applied. Anti-models are used to normalise the likelihood of an unknown observation sequence by calculating the ratio between the probability of the hypothesised word and its anti-model.

2.3 Speech Recognition

In the field of continuous speech recognition additional confidence measures based on the integration of a statistical language model are used. The integration of the language model in the recognition process can be controlled by two factors: the *Grammar Scale Factor* (GSF) and the *Word Insertion Penalty* (WIP) (Zimmermann and Bunke, 2004b). The GSF is used to weight the influence of the language model against the optical recogniser, while the WIP helps to control over- and undersegmentation, i.e. the insertion and deletion of words.

In (Sanchis et al., 2000) the GSF is used to classify incorrect words in a speech recognition system. Two models based on acoustic stability are presented. The study additionally investigates the reduction of computational costs of the reject models.

Not only the GSF, but also the WIP is used in (Zeppenfeld et al., 1997) in the field of conversational telephone speech recognition. Multiple candidate sentences derived from GSF and WIP variations are used to determine the confidence measure.

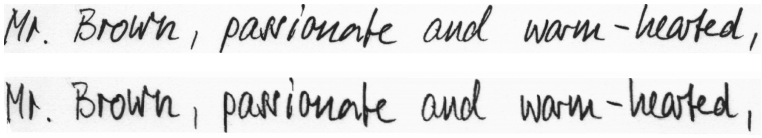


Fig. 1. Preprocessing of the handwritten text line image. The first line shows the original image, while the normalised image is shown on the second line.

3 Methodology

3.1 HMM Based Recognition System

The offline handwriting recognition system we used is based on the system described in detail in (Marti and Bunke, 2001). It can be divided into three major parts: preprocessing and feature extraction, *Hidden Markov Model* (HMM) based recognition, and postprocessing.

In the preprocessing part skew, slant, and baseline position are normalised. This normalisation is necessary to reduce the impact of the different writing styles. An example of these normalisation steps is shown in Fig. 1. For any further details we refer to (Marti and Bunke, 2001).

After preprocessing, a handwritten text line is converted into a sequence of feature vectors. For this purpose, a sliding window is used. The window has a width of one pixel and is moved from the left to the right, one pixel per step, over the image, which was scanned with a resolution of 300 dpi. At each position of the window, nine geometrical features are extracted. The first three features contain the number of foreground pixels in the window as well as the first and the second order moment of the foreground pixels. Features four to seven contain the position of the upper and the lower contour, and the first order derivative from the upper and the lower contour, respectively. The last two features contain the number of vertical black-white transitions and the pixel density between the upper and the lower contour. Again, we refer to (Marti and Bunke, 2001) for further details.

In the HMM based recogniser an HMM is provided for each character. For all HMMs a linear topology is used. This means that there are only two transitions per state, one to itself and one to the next state. The number of states of a character HMM is chosen depending on the individual character following the procedure described in (Zimmermann and Bunke, 2002). A mixture of twelve Gaussians is used to model the output distribution in each state. The character HMMs are concatenated to word models. There is exactly one word model for each word in the underlying lexicon.

For the training of the character HMMs the Baum-Welch algorithm (Rabiner,

1989) is used. The recognition is performed by Viterbi decoding (Viterbi, 1967) supported by a statistical n-gram language model (Jelinek, 1990) with backoff (Katz, 1987). N-gram language models are based on the observation that we are often able to guess the next word when we are reading a given text. In other words, the probability of a word is highly depending on the previous text.

For the HMM based recogniser of this paper a word bigram language model is used. In the case of bigram language models the previous text is approximated by the last word and the dependency is modelled by the probability $p(w_i|w_{i-1})$, where w_i represents the considered word and w_{i-1} stands for the previous word. The probability $p(W)$ of a text line $W = (w_1, \ldots, w_n)$ can then be computed as follows:

$$p(W) = p(w_1) \prod_{i=2}^{n} p(w_i|w_{i-1}) \tag{1}$$

Bigram language models seem to be a good trade-off between model accuracy and generalisation. Unigrams are usually not able to describe the language accuratly, whereas for trigrams more text is required to estimate the language model probabilities reliably (Vinciarelli et al., 2004; Zimmermann and Bunke, 2004a).

For this paper the bigram language model is obtained from the LOB corpus (Johansson et al., 1986). Upper and lower case words are distinguished and punctuation marks are modelled as separate words.

The rejection strategies proposed in this paper are implemented as a postprocessing step. Given the output of the HMM based recogniser, we first generate K alternative candidates. Based on these candidates a confidence measure is then calculated for each recognised word w_i. Only if this confidence measure exceeds a given threshold t the word w_i is accepted.

The generation of the K candidates is explained in the following subsection while the confidence measures are described in Sect. 3.3.

3.2 Generation of Alternative Candidates

The rejection strategies introduced in this paper are based on confidence measures derived from alternative candidate word sequences. This means that the recogniser not only produces the top ranked word sequence, but also a list of alternative candidate sequences which are used to compute the confidence measures. The quality of these alternative candidates is a key aspect for a good performance of the proposed confidence measures. In the ideal case, an

alternative candidate sequence should distinguish itself from the top ranked output word sequence exactly at the positions where top ranked output words have been recognised incorrectly. Of course, in practice, this is rarely the case, as alternative candidates sometimes differ in words that have been recognised correctly or coincide with wrongly recognised words.

A common way to produce alternative candidates is the extraction of an n-best list, containing the n highest ranked transcriptions of a given image of handwritten text. However, it has been shown in the speech recognition literature (Zeppenfeld et al., 1997) as well as in the handwriting literature (Zimmermann et al., 2004) that candidates based on language model variations have the potential to provide better rejection performance than n-best lists. Therefore, we use language model variations to obtain the alternative candidates.

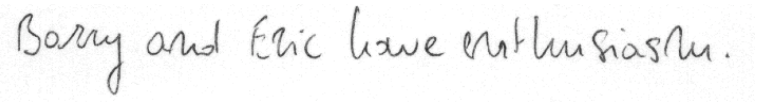
For an HMM based recognition system with integrated language model, such as the one used in this paper, the most likely word sequence $\hat{W} = (w_1, \ldots, w_m)$ for a given observation sequence X is computed in the following way:

$$\hat{W} = \underset{W}{\operatorname{argmax}} \;\; \log p(X|W) + \alpha \;\; \log p(W) + \beta \; m \tag{2}$$

According to Eq. 2 the optical model $p(X|W)$, which is the result of the HMM decoding, is combined with the likelihood of a text line $p(W)$ obtained from the language model. Because the HMM system and the language model merely produce approximations of probabilities, two additional parameters α and β are necessary to compensate the deficiencies and to control the integration of the language model. The parameter α is called *Grammar Scale Factor* (GSF) and weights the impact of the statistical language model. The term *Word Insertion Penalty* (WIP) is used for the parameter β. Multiplied with m, the number of words in W, parameter β controls the segmentation rate of the recogniser.

By varying the two parameters α and β, multiple candidates can be produced from the same image of a handwritten text. To obtain K alternative candidates $\hat{W}_i$, we choose K different parameter pairs (α_i, β_i) $i \in \{1, \ldots, K\}$.

An example of candidates based on language model variation is shown in Fig. 2. Multiple recognition results are produced for the handwritten text *"Barry and Eric have enthusiasm."* The obtained candidates provide an illustration of the impact of parameter β_i on the segmentation of $\hat{W}_i$. The average amount of words for $\beta_i = -100$ is 4.33, while for $\beta_i = 150$ there are seven words on average (including punctuation marks). Furthermore, we observe that if we increase parameter α, nonsense word sequences, such as *we m run rush*, are usually eliminated. Even though all the candidate text lines differ in the example of Fig. 2, in general, the candidates may differ or not.



i	α_i	β_i	$\hat{W}_i$
1	0	-100	Barry arm inch we enthusiasm
2	0	25	Barry arm inch we m run rush
3	0	150	B my arm inch we m run rush :
4	30	-100	Barry and include enthusiasm
5	30	25	Barry and Eric have enthusiasm
6	30	150	Barry and Eric have enthusiasm .
7	60	-100	Barry and include enthusiasm
8	60	25	Barry and include enthusiasm .
9	60	150	Barry and in have enthusiasm .

Fig. 2. Candidate text lines resulting from language model variation.

W	Barry	and	Eric	have	enthusiasm	.
$\hat{W}_1$	Barry	arm	inch	we	enthusiasm	
$\hat{W}_4$	Barry	and		include	enthusiasm	
$\hat{W}_9$	Barry	and	in	have	enthusiasm	.

Fig. 3. Example of aligning alternative candidates ($\hat{W}_1, \hat{W}_4, \hat{W}_9$) with the top ranked output W.

3.3 Confidence Measures

The confidence measures proposed in this paper are derived from a list of candidates. As described in Sect. 3.2, in addition to the recogniser's top ranked output $W = (w_1, \ldots, w_m)$, the list contains K alternative candidates $\hat{W}_1, \ldots, \hat{W}_K$, where $\hat{W}_i = (w_1^i, \ldots, w_m^i)$. The alternative candidates are aligned with the top ranked output W using dynamic string alignment (Wagner and Fischer, 1974). See Fig. 3 for an example. Based on the alignment a confidence measure $p(c|w_i, n)$ is computed for each word w_i of W in order to decide whether to accept w_i or to reject it. The quantity $p(c|w, n)$ represents the probability of a word w of the top ranked output beeing recognised correctly, where $c \in \{0, 1\}$ (0 stands for incorrect and 1 for correct) and $n \in 0, \ldots, K$ corresponds to the number of times a word w is observed in the K alternative candidates. In the text below we describe the three different confidence measures that are different approximations of the probability $p(c|w, n)$. The resulting rejection strategies will be called Strategy 1, Strategy 2, and Strategy 3.

Strategy 1 is the simplest of the three strategies where $p(c|n, w)$ is estimated by $p(c|n)$. The underlying assumption is that the probability of being correctly recognised is independent of the considered word w. This assumption allows a straightforward and robust estimation of $p(c|n)$. The probability $p(c|n)$ is

then used as a confidence measure ρ_1 for Strategy 1.

$$\rho_1 = p(c|n) \tag{3}$$

During the training phase, the quantities $p(c|n)$ are estimated for every $n = 0, \ldots, K$ using the relative frequencies obtained from the training set. Although Strategy 1 is simple to train and to use, the following two limitations will possibly lead to a limited performance of confidence measure ρ_1.

- The assumption that the probability of a correct recognition is independent of the considered word may be too strong. There are words that are easy to recognise, while others are more difficult.
- Because Strategy 1 assigns the same weight to all alternative candidates just summing up the number of identical word instances among the K alternative candidates may lead to some information loss. This procedure seems reasonable as long as all alternatives are of more or less of the same quality and reliability. In general this condition is not true.

In Strategy 2 and Strategy 3 we try to overcome these two potential weaknesses.

Strategy 2 takes into account that some words are more likely to be recognised correctly than others. In its confidence measure ρ_2, Strategy 2 explicitly considers the current word w, instead of assuming that the recognition result is independent of word w, as it was supposed in Strategy 1. For Strategy 2 the Bayes' rule is used to reformulate $p(c|n, w)$:

$$p(c|n, w) = \frac{p(n|c, w) \cdot p(c|w)}{\sum_{x=0,1} p(n|x, w) \cdot p(x|w)} \tag{4}$$

We then simplify the right hand side of Eq. 4 using the assumption that $p(n|c, w) \simeq p(n|c)$. By means of this approximation the resulting confidence measure ρ_2 is defined as follows:

$$\rho_2 = \frac{p(n|c) \cdot p(c|w)}{\sum_{x=0,1} p(n|x) \cdot p(x|w)} \tag{5}$$

Both $p(n|c)$ and $p(c|w)$ are estimated using relative frequencies obtained from the training set during the training phase. If there are no or not enough training samples for a word w to estimate $p(c|w)$, confidence measure ρ_1 is used instead of ρ_2.

Strategy 3 addresses the problem that some sources of candidates are more reliable than others. The sources which produce better results should have

a larger impact on the confidence measure than sources of weaker quality. Confidence measure ρ_3 of Strategy 3 is based on a *Multi-Layer Perceptron (MLP)* with a single hidden layer that conceives the rejection strategy as a two-class classification problem. Based on a feature vector extracted from the K alternative candidates, the system must decide whether to accept a word, or to reject it. As an additional benefit the MLP is able to consider relations between different sources of alternative candidates, as these sources are typically not independent.

For the MLP architecture we choose K input neurons, one hidden layer with l neurons, and two output neurons. The feature vectors $(x_1, \ldots, x_K)$ are acquired from the alternative candidates. For every word in the input, each alternative candidate $\hat{W}_i$ contributes one element x_i to the feature vector where $x_i = 1$ if the word of $\hat{W}_i$ matches the word in the top ranked output, and $x_i = 0$ otherwise. The output neurons y_0 and y_1 represent the score for acceptance (y_1) and the score for rejection (y_0), respectively. The score for acceptance y_1 is used as the confidence measure ρ_3 of Strategy 3.

$$\rho_3 = y_1 \quad \text{(acceptance score of the MLP)} \tag{6}$$

To illustrate the performance of the proposed confidence measures we implement two additional confidence measure which act as baseline systems against which the previously described confidence measures are compared.

The first baseline system uses confidence measures based on normalised likelihoods. The HMM based recogniser accumulates likelihoods for each frame, i.e. each position of the sliding window. The resulting likelihood score for a word is used in the decoding step. Because the raw recognition score is influenced by the length of the handwritten word it is normalised by the number of frames. The result is an average likelihood which is then used as confidence measure.

The confidence measure of the second baseline system is derived from local n-best lists. These n-best list are obtained from the recognition lattice. The recognition score of the second best word hypothesis score is divided by the score of best word hypothesis. One minus this ratio is then used as a confidence measure.

4 Experiments and Results

All experiments reported in this paper make use of the Hidden Markov Model (HMM) based handwritten text recognition system described in Sect. 3.1.

4.1 Experimental Setup

A writer independent text line recognition task is considered where no handwritten samples of the writers in the test set are available for the training or the validation of the recognition system. The text lines originate from the IAM database (Marti and Bunke, 2002). The recognition system is trained on 6166 text lines (283 writers). The rejection strategies are trained on 3686 text lines (201 writers). The MLP strategy is validated on another 941 text lines (43 writers). Finally, the test set consists of 1863 text lines written by 128 writers. All these data sets are disjoint, and no writer has contributed to more than one set. The underlying lexicon includes all those 12,502 words classes that occur in the union of the training, validation, and test sets. To determine (α, β) of the top ranked candidate the two parameters are optimised globally. This optimisation is performed on part of the rejection training set (900 text lines written by 46 writers). The value of K for the number of alternative candidates is set to 64. Eight different values for each of the parameters α and β are used. Parameter α is equally varied between 0 and 60 (step size: 8.5), while β is equally varied between -100 and 150 (step size: 36). The optimised value for (α, β) is found at $(20, 20)$.

4.2 Evaluation Methodology

To evaluate the rejection strategies a confusion matrix is used. A word can either be recognised correctly or incorrectly. In both cases the recognition result may be accepted or rejected by the postprocessing procedure which results in one of the four following outcomes:

- *Correct Acceptance* (CA) - A correctly recognised word has been accepted by the postprocessor.
- *False Acceptance* (FA) - A word has not been recognised correctly but has been accepted by the postprocessor.
- *Correct Rejection* (CR) - A incorrectly recognised word has been rejected by the postprocessor.
- *False Rejection* (FR) - A word that has been recognised 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) (Maltoni et al., 2003). These measures are defined as follows:

$$FAR = \frac{FA}{FA + CR} \tag{7}$$

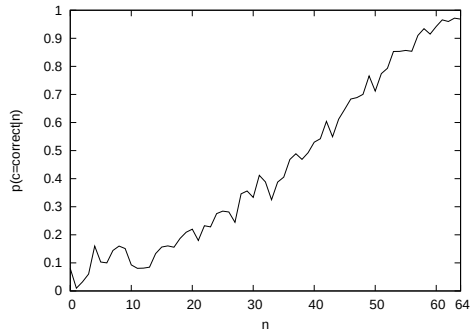


Fig. 4. Estimated probability $p(c|n)$ of being correct as a function of n.

$$FRR = \frac{FR}{FR + CA} \tag{8}$$

A second characteristic curve is the *Error-Reject Plot* where the *Error Rate* (ERR) is plotted against the *Rejection Rate* (REJ). ERR and REJ are defined as follows:

$$ERR = \frac{FA}{CA + FA} \tag{9}$$

$$REJ = \frac{CR + FR}{CA + FA + CR + FR} \tag{10}$$

4.3 Training and Validation

The quantities $p(c|n)$, $p(c|w)$, and $p(n|c)$ are estimated on the rejection strategy training set using relative frequencies. The MLP of Strategy 3 is trained using standard back-propagation. Figure 4 shows the resulting probabilities for $p(c|n)$. As expected, the probability of a word being correctly recognised is usually higher if the word appears more often among the alternative candidates. A few examples of the probabilities $p(c|w)$ are listed in Fig. 5 illustrating the fact that short words are often more difficult to recognise correctly than longer words. The estimations of the probabilities $p(n|c)$ are shown in Fig. 6.

To conduct experiments with Strategy 3, the number of hidden neurons l of the MLP has to be determined on the validation set. On the validation set we evaluated each value of $l \in (1, \ldots, 50)$ using the *Equal Error Rate*

w	$p(correct\|w)$
do	0.14
get	0.65
his	0.89
other	0.86
which	0.99

Fig. 5. Extract of the estimated probabilities $p(c|w)$ from training set.

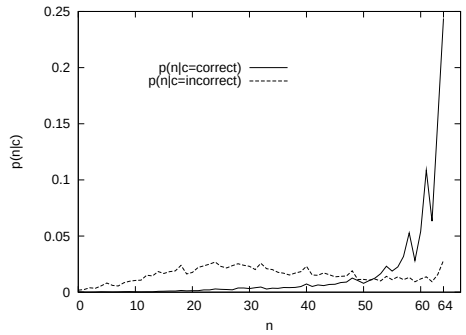


Fig. 6. Estimated probability $p(n|c)$ of appearing n times in the alternative candidates.

(EER) (Maltoni et al., 2003). For $l = 6$ hidden neurons the system performed best.

4.4 Test Set Results

The experimental results on the test set are shown in the ROC curve plot of Fig. 7. The three proposed confidence measures as well as the two baseline measures are shown. The proposed confidence measures clearly outperform the baseline measures. Furthermore, the more complicated confidence measures of Strategy 2 and Strategy 3 perform better than Strategy 1. The best performing confidence measure is Strategy 2, indicating that the considered word delivers more information than the consideration of the source of the alternative candidates performed in Strategy 3. However, because Strategy 2 is based on the quantities $p(c|w)$ the training of Strategy 2 is dependent on the underlying lexicon. If this lexicon is extended or changed, in general, Strategy 2 has to be adapted as well. Thus, if more flexibility concerning the lexicon is required, Strategy 3 has to preferred.

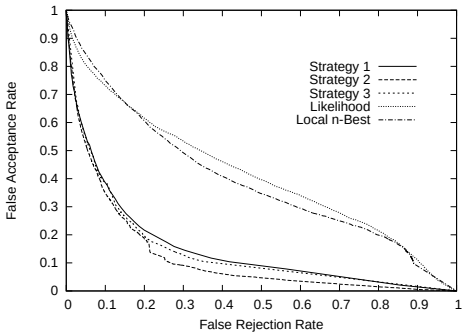


Fig. 7. ROC curves of the different reject strategies.

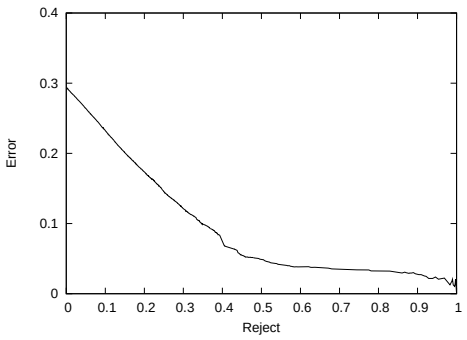


Fig. 8. Error-reject plot of Strategy 2.

The tradeoff between remaining errors and rejects is depicted in Fig. 8. In terms of error-reject statistics, the best performing confidence measure (Strategy 2) performs as follows: without any rejection the word error rate is equal to 29.3%. To attain a word error rate of 10% a rejection rate of 34.8% is required.

5 Conclusions

This paper investigated various rejection strategies for an HMM based off-line handwritten text recognition system supported by a statistical n-gram language model. The rejection strategies depend on different confidence measures that are used in a postprocessing step to decide whether to accept or to reject a recognised word in a given line of handwritten text.

The proposed confidence measures are based on a set of alternative text line candidates. To generate these alternative candidates we make use of the fact that the inclusion of a statistical language model in the recognition process can be controlled by the two parameters grammar scale factor and word insertion penalty. By varying these two parameters multiple candidates are produced.

The first proposed confidence measure, Strategy 1, is only based on the number of times a recognised word appears among the alternative candidates. The confidence measure of Strategy 2 also takes into account the considered word class, as some words are more likely to be correctly recognised than others. A Multi-Layer Perceptron is used in Strategy 3 to combine the results from the various alternative candidate sources.

Experiments have been conducted on a large set of text lines from the IAM database. Confidence measures based on normalised likelihoods and on local n-best lists were used as benchmarks in the evaluation of the performance of the proposed confidence measures. Each of the proposed confidence measures substantially outperforms the confidence measures of the baseline systems. The best performing confidence measure of Strategy 2 takes into account the considered word class and attained a false acceptance rate of 20% at a false rejection rate of less than 19%.

Acknowledgement

This research was supported by the Swiss National Science Foundation (Nr. 20-52087.97). Additional funding was provided by the Swiss National Science Foundation NCCR program "Interactive Multimodal Information Management (IM)2" in the Individual Project "Scene Analysis".

References

Bazzi, I., Schwartz, R. M., Makhoul, J., 1999. An omnifont open-vocabulary OCR system for English and Arabic. IEEE Transactions on Pattern Analysis

and Machine Intelligence 21 (6), 495–504.
Brakensiek, A., Rigoll, G., 2004. Handwritten address recognition using hidden Markov models. In: Dengel, A., Junker, M., Weisbecker, A. (Eds.), Reading and Learning. Springer, pp. 103–122.
Brakensiek, A., Rottland, J., Kosmala, A., Rigoll, G., 2000. Off-line handwriting recognition using various hybrid modeling techniques and character n-grams. In: 7th International Workshop on Frontiers in Handwriting Recognition, Amsterdam, The Netherlands. pp. 343–352.
Gorski, N., 1997. Optimizing error-reject trade off in recognition systems. In: 4th International Conference on Document Analysis and Recognition, Ulm, Germany. Vol. 2. pp. 1092–1096.
Jelinek, F., 1990. Self-organized language modeling for speech recognition. Readings in Speech Recognition, 450–506.
Johansson, S., Atwell, E., Garside, R., Leech, G., 1986. The Tagged LOB Corpus, User's Manual. Norwegian Computing Center for the Humanities, Bergen, Norway.
Katz, S. M., 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Transactions on Acoustics, Speech and Signal Processing 35 (3), 400–401.
Kim, G., Govindaraju, V., Srihari, S., 1999. Architecture for handwritten text recognition systems. In: Lee, S.-W. (Ed.), Advances in Handwriting Recognition. World Scientific Publ. Co., pp. 163–172.
Koerich, A. L., 2004. Rejection strategies for handwritten word recognition. In: 9th International Workshop on Frontiers in Handwriting Recognition, Tokyo, Japan. pp. 479–484.
Maltoni, D., Maio, D., Jain, A. K., Prabhakar, S., 2003. Handbook of Fingerprint Recognition. Springer Professional Computing, New York.
Marti, U.-V., Bunke, H., 2001. Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system. International Journal of Pattern Recognition and Artificial Intelligence 15, 65–90.
Marti, U.-V., Bunke, H., 2002. The IAM-database: an English sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition 5, 39 – 46.
Marukatat, S., Artieres, T., Gallinari, P., 2002. Rejection measures for handwriting sentence recognition. In: 8th International Workshop on Frontiers in Handwriting Recognition, Niagara-on-the-Lake, Canada. pp. 24–29.
Oza, N., Polikar, R., Kittler, J., Roli, F. (Eds.), 2005. Multiple Classifier Systems, 6th International Workshop. Springer LNCS 3541.
Pitrelli, J., Perrone, M. P., 2002. Confidence modeling for verification post-processing for handwriting recognition. In: 8th International Workshop on Frontiers in Handwriting Recognition, Niagara-on-the-Lake, Canada. pp. 30–35.
Pitrelli, J., Perrone, M. P., 2003. Confidence-scoring post-processing for off-line handwritten-character recognition verification. In: 7th International Confer-

ence on Document Analysis and Recognition, Edinburgh, Scotland. Vol. 1. pp. 278–282.

Rabiner, L., 1989. A tutorial on hidden Markov models and selected application in speech recognition. Proc. of the IEEE 77 (2), 257–286.

Sanchis, A., Jimenez, V., Vidal, E., September 2000. Efficient use of the grammar scale factor to classify incorrect words in speech recognition verification. In: International Conference on Pattern Recognition, Barcelona, Spain. Vol. 3. pp. 278–281.

Shridhar, M., Houle, G., Kimura, F., 1997. Handwritten word recognition using lexicon free and lexicon directed word recognition algorithms. In: 4th International Conference on Document Analysis and Recognition, Ulm, Germany. pp. 861–865.

Vinciarelli, A., Bengio, S., Bunke, H., 2004. Offline recognition of unconstrained handwritten texts using HMMs and statistical language models. IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (6), 709–720.

Viterbi, A., 1967. Error bounds for convolutional codes and an asimptotically optimal decoding algorithm. IEEE Transactions on Information Theory 13 (2), 260–269.

Wagner, R., Fischer, M., 1974. The string-to-string correction problem. Journal of the ACM 21 (1), 168–173.

Zeppenfeld, T., Finke, M., Ries, K., Westphal, M., Waibel, A., 1997. Recognition of conversational telephone speech using the janus speech engine. In: International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany. Munich, Germany, pp. 1815–1818.

Zimmermann, M., Bertolami, R., Bunke, H., 2004. Rejection strategies for offline handwritten sentence recognition. In: 17th International Conference on Pattern Recognition, Cambridge, England. Vol. 2. pp. 550–553.

Zimmermann, M., Bunke, H., 2002. Hidden Markov model length optimization for handwriting recognition systems. In: 8th International Workshop on Frontiers in Handwriting Recognition, Niagara-on-the-Lake, Canada. pp. 369–374.

Zimmermann, M., Bunke, H., 2004a. N-gram language models for offline handwritten text recognition. In: 9th International Workshop on Frontiers in Handwriting Recognition, Tokyo, Japan. pp. 203 – 208.

Zimmermann, M., Bunke, H., 2004b. Optimizing the integration of a statistical language model in HMM based offline handwriting text recognition. In: 17th International Conference on Pattern Recognition, Cambridge, England. Vol. 2. pp. 541 – 544.