# Evaluation of Confidence Measures for On-Line Handwriting Recognition

Anja Brakensiek[1], Andreas Kosmala[1], and Gerhard Rigoll[2]

[1] Dept. of Computer Science, Faculty of Electrical Engineering,
Gerhard-Mercator-University Duisburg, D-47057 Duisburg,
{anja,kosmala}@fb9-ti.uni-duisburg.de
[2] Inst. for Human-Machine Communication, Technical University of Munich,
D-80290 Munich, rigoll@ei.tum.de

**Abstract.** In this paper a writer-independent on-line handwriting recognition system is described comparing the effectiveness of several confidence measures. Our recognition system for single German words is based on Hidden Markov Models (HMMs) using a dictionary. We compare the ratio of rejected words to misrecognized words using four different confidence measures: One depends on the frame-normalized likelihood, the second on a garbage model, the third on a two-best list and the fourth on an unconstrained character recognition. The rating of recognition results is necessary for an unsupervised retraining or adaptation of recognition systems as well as for a user friendly human-computer interaction avoiding excessive call backs.

## 1 Introduction

Automatic recognition systems for unconstrained on-line handwritten words become more and more important, especially with respect to the use of pen based computers or electronic address books (PDA, compare also [6, 7]). In this field of research, HMM-based techniques (for a detailed introduction see [8]), which are well known in speech recognition, have been established because of their segmentation-free recognition approach and their automatic training capabilities.

Although the performance of recognition systems increases, the error rate of writer independent recognizers is still quite high. By computing confidence measures for recognition, a reliability assessment of the results becomes feasible. Thus, using these measures it is possible to decide whether the recognition result is uncertain or not. The goal is to reject most of the misrecognized words and as few as possible of the correct results. This problem seems to be easy, regarding the recognition of single characters (or pre-segmented words) using e.g. neural networks or distance measures (e.g. KNN-classifier) to prototypes. In general these techniques automatically compute a kind of posterior probability which can be used as a confidence measure. In contrast to this the HMM-based technique yields a likelihood, which has to be transformed resp. normalized, as it is described in Section 3. Here, just that word of a dictionary, which is the

From: DAGM 02, Springer "Lecture Notes in Computer Science"

most probable, is selected. And this probability cannot be regarded offhand for confidence.

A rating of correctness of a result is helpful in several applications for speech and handwriting recognition systems. The (re-) training of a system can be done in an unsupervised mode using automatically generated labels with a high confidence score. So the large amount of training data does not need to be labeled manually. The same applies to an unsupervised adaptation to a certain writer or speaker (compare [1, 10]). Another application is the detection of out-of-vocabulary (OOV) words (see [11]) or a more user friendly dialog between humans and automatic recognition systems (e.g. call backs in information systems).

In the following sections our baseline recognition system (Sec. 2), the theory (Sec. 3) and some results (Sec. 4) which are obtained by four investigated confidence measures are described.

## 2 System Architecture

Our handwriting recognition system (compare also [1]) consists of about 90 different linear HMMs, one for each character (upper- and lower-case letters, numbers and punctuation marks). In general the continuous density HMMs consist of 12 states (with up to 15 Gaussian mixtures per state depending on the amount of training data per HMM) for characters and numbers and fewer states for some special characters depending on their width. The presented results refer to a single word recognition rate using a dictionary of about 2200 German words (no out of vocabulary).

For our experiments we use a large on-line handwriting database of several writers (compare Fig.1), which is described in the following. The database consists of cursive script samples of 166 different writers, all writing several words or sentences on a digitizing surface. The training of the writer independent system is performed using about 24400 words of 145 writers. Testing is carried out with 2071 words of 21 different writers (about 100 words per writer).

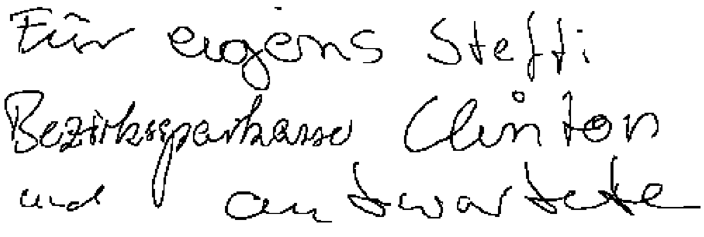

**Fig. 1.** Some examples of the handwritten database (7 different writers)

After the resampling of the pen trajectory in order to compensate different writing speeds the script samples are normalized. Normalization of the input-data implies the correction of slant and height. Slant normalization is performed by shearing the pen trajectory according to an entropy-criterion (see also [2]).

Then the following features are derived from the trajectory of the pen input:

- the angle of the spatially resampled strokes ($\sin\alpha$, $\cos\alpha$)
- the difference angles ($\sin\Delta\alpha$, $\cos\Delta\alpha$)
- the pen pressure (binary)
- a sub-sampled bitmap slid along the pen trajectory (9-dimensional vector), containing the current image information in a $30 \times 30$ window

The baseline system and the feature extraction method is described in [1] in greater detail. To train the HMMs we use the Baum-Welch algorithm, whereas for recognition the Viterbi algorithm is used always presenting these 14-dimensional feature vectors $\underline{x}$ in one single stream.

The recognition problem using HMMs can be described by the following Eq. 1 using Bayes' rule (with $X$ is the sequence of feature vectors and $W$ represents the class resp. word):

$$P(W|X) = \frac{P(X|W) \cdot P(W)}{P(X)} \tag{1}$$

Here $P(W|X)$ is the posterior probability, $P(X|W)$ represents the feature model (the likelihood, which is computed by the HMM), $P(W)$ describes the a priori probability of the word $W$ (e.g. grammar or language model) and $P(X)$ represents the a priori probability of the feature vectors. The recognition of a single word $W^*$, which is defined in a dictionary (the same a priori probability for each entry) , leads to this equation

$$W^* \approx \underset{W}{\operatorname{argmax}}\; P(X|W) \cdot P(W) \approx \underset{W}{\operatorname{argmax}}\; P(X|W) \tag{2}$$

because $P(X)$ and $P(W)$ are the same for all classes $W$ ($P(X)$ is independent of $W$ and $P(W)$ is nonrelevant per definition). Disregarding these probabilities the relative order of the best recognition results will not change. Thus for recognition only, this simplification is permitted. However, the probability $P(X)$ is important to compute the probability of correctness – the confidence – of the recognition result (see Sec. 3).

## 3 Confidence Measures

The likelihood $P(X|W)$, which is used for recognition according to Eq. 2 is not an absolute measure of probability, but rather a relative measure. Thus, we just know which word of a given closed dictionary is the most likely, but we do not know the certainty of correctness of this recognition result. This certainty is described by confidence measures. If the confidence measure $Conf$ of a recognition result is below a threshold $\tau$, this data image resp. test-word is rejected. The consequence of that is a manual labeling or a call back in a human-machine interface, for example.

For our handwriting recognition problem of single words (no OOV) we compare four different confidence measures:

1. the frame normalized likelihood $P(X|W)$ (as a reference)
2. the posterior probability $P(W|X)$ by approximating $P(X)$ using a garbage-model $W_{garb}$
3. the posterior probability $P(W|X)$ by approximating $P(X)$ using a two-best recognition
4. the likelihood $P(X|W)$, which is normalized by the likelihood $P(X|C)$ obtained by a character decoding without dictionary

As first investigated measure, the likelihood $P(X|W)$, which is normalized by the number of frames (resp. corresponding feature vectors) is used as confidence measure (see e.g. [3]). Here the computational costs are very low, because this measure exists in any case. Because of the dynamic of the HMM-based decoding procedure, in general these measures are computed as log likelihoods. The higher the normalized likelihood, the higher the reliability.

The following confidence measures take Eq. 1 into account. The posterior probability $P(W|X)$ will be an optimal confidence measure, if it was possible to estimate $P(X)$:

$$Conf := \frac{P(X|W)}{P(X)} \qquad Conf \begin{cases} < \tau \rightarrow reject & (N_r) \\ \geq \tau \rightarrow classify & (N_c + N_e) \end{cases} \tag{3}$$

Thus, the second confidence measure, we tested, is based on a garbage- or filler-model (compare [9, 4]). The garbage-model $W_{garb}$ is trained on all features of the training-set (independent of the character-label), which leads to an unspecific average model. Often, such a garbage-model is used for OOV detection in speech recognition. The confidence measure can be calculated using the garbage-model as an approximation of $P(X)$:

$$P(X) \approx P(X|W_{garb}) \tag{4}$$

If this ratio of word and garbage dependent likelihood is large, the correctness is more likely. To determine $P(X|W_{garb})$ the decoding procedure has to be expanded, as it is shown in Fig. 2.

The third evaluated confidence measure depends on a two-best recognition according to Eq. 5 (see also [3, 4]):

$$P(X) \approx \sum_{k=1}^{N} P(X|W_k) \cdot P(W_k) \quad \Rightarrow \quad P(X) \approx P(X|W_{1st}) + P(X|W_{2nd}) \tag{5}$$

This measure contains the difference of the log likelihoods between the best and the second best hypothesis for the same sequence of feature vectors. The approximation in Eq. 5 is valid under the assumption, that the likelihoods of the best and second best class are much higher than those of the other $(N-2)$ classes of the dictionary. Transforming this equation because of the dynamic range of the values, the new confidence measure $Conf^* = \frac{Conf}{1-Conf}$ can be defined as follows:

$$Conf = \frac{P(X|W_{1st})}{P(X|W_{1st}) + P(X|W_{2nd})} \quad \Rightarrow \quad Conf^* = \frac{P(X|W_{1st})}{P(X|W_{2nd})} \tag{6}$$

Again, for rejection the rule of Eq. 3 is applied (only the domain of $\tau$ has been changed).

The fourth method to obtain confidence measures is based on a character decoding without dictionary (compare e.g. [11, 5]), as it is shown in Fig. 2. The character-based likelihood $P(X|C)$ is used for normalization:

$$P(X|C) = P(X|c_1, ...c_k) = \prod P(X_{f_i}|c_i) \quad \Rightarrow \quad Conf := \frac{P(X|W)}{P(X|C)} \tag{7}$$

A recognition without vocabulary leads to an arbitrary sequence of characters $c_i : 1 \leq i \leq K$ ($K$ is the number of different character HMMs) which is the most likely without respect to the lexicon. In general $P(X|C)$ will be greater (or equal) than $P(X|W)$.

Additionally to this character-based likelihood, also the character-sequence itself can be regarded by calculating the Levenshtein-distance between $W$ and $C$. The Levenshtein-distance describes how many changes are necessary to transform one string into the other.

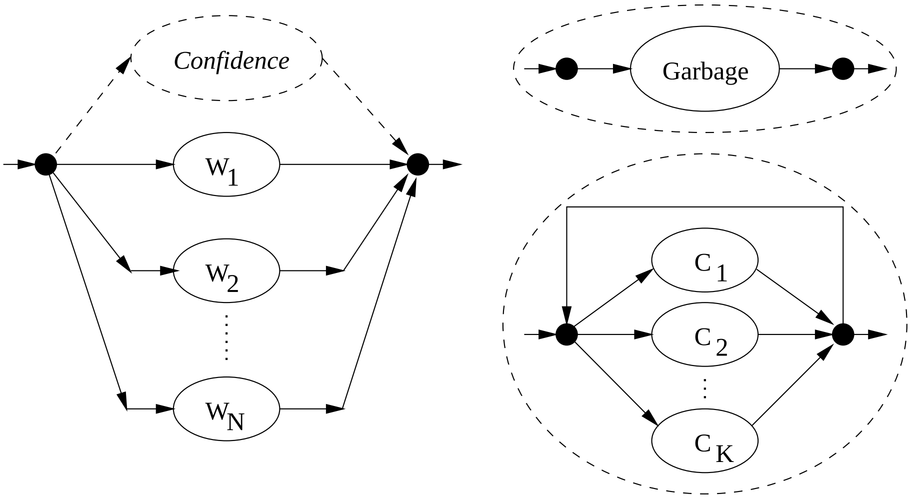

**Fig. 2.** Decoding configuration to obtain confidence measures: for recognition using a dictionary with the classes $W$ a further path to compute the confidence (garbage model or recognition of characters $c$) can be added

These confidence measures differ significantly regarding the computational costs, the effectiveness and the application. In the literature, there are described much more confidence measures especially for continuous speech recognition. But often, they take the grammar $P(W)$ of a sentence into account, which is not possible for single word recognition (such as in our experiments).

## 4 Experimental Results

In the presented experiments we examine the influence of four different confidence measures for on-line handwriting recognition of single words (no OOV).

The recognition results, which are shown in Fig. 3, are determined using an increasing threshold $\tau$ (which differs using different confidence measures). Using the baseline system without rejection ($N_r$ =0, $r$ =0%) a word recognition rate of 87.0% ($N_c$ =1801 words are recognized correctly) is achieved testing the entire test-set of $N_a$= 2071 words. For testing of confidence measures always the ratio of error rate $e$ and rejection rate $r$ is calculated using the following definition ($N_r$ is the number of rejected examples, $N_e$ is the number of errors within the not rejected set). The term $f$ denotes the quota of false classified examples, which are not rejected.

$$r = \frac{N_r}{N_a} \,, \quad e = \frac{N_e}{N_a} \,, \quad f = \frac{N_e}{N_a - N_r} \qquad \text{with:} \quad N_r + N_e + N_c = N_a \quad (8)$$

As can be seen in Fig. 3 the frame-based likelihood (1) is the worst confidence measure. Even if 52.1% of the test words are rejected, the error rate decreases only from 13.0% to 3.3% (Tab. 1). That implies 7.0% false classified words (referring to the number of words, which are not rejected). The confidence measures based on the garbage model (2) or the character decoding (4) are only slightly more effective. The best method relies on the two-best list (3). For example, rejecting 44.8% of the data, the error is reduced by about 93% relative and the quota of false classified words decreases to 1.7% (Tab. 1: last row).

Additionally, in Fig. 3 two significant points (5,6) are marked depending on the Levenshtein-distance. For this a character recognition without dictionary is implemented (compare confidence measure (4)) and the Levenshtein-distance between the dictionary-based word recognition and the recognized character sequence is computed. In contrast to (4) and (5), in (6) this character recognition is performed using a statistical language model (backoff 3-gram) to enhance the recognition performance. (5) and (6) denote the ratio of rejected and error when the dictionary-based recognition and the character-based recognition leads to the same word-class (Levenshtein-distance=0). Regarding (6) the word error rate is not smaller than using the frame-normalized likelihood, but here most of the errors are caused by a substitution of a single character only ('wo' – 'Wo', 'der' – 'dar') depending on the lexicon.

The recognition rate is highly dependent on the underlying vocabulary. Here, not only the size of the dictionary but also the similarity between distinct entries is essential. This fact is considered by the two-best confidence measure. If the best and second-best hypothesis are quite similar (small Levenshtein-distance), because they differ only in one character, the corresponding likelihoods will be

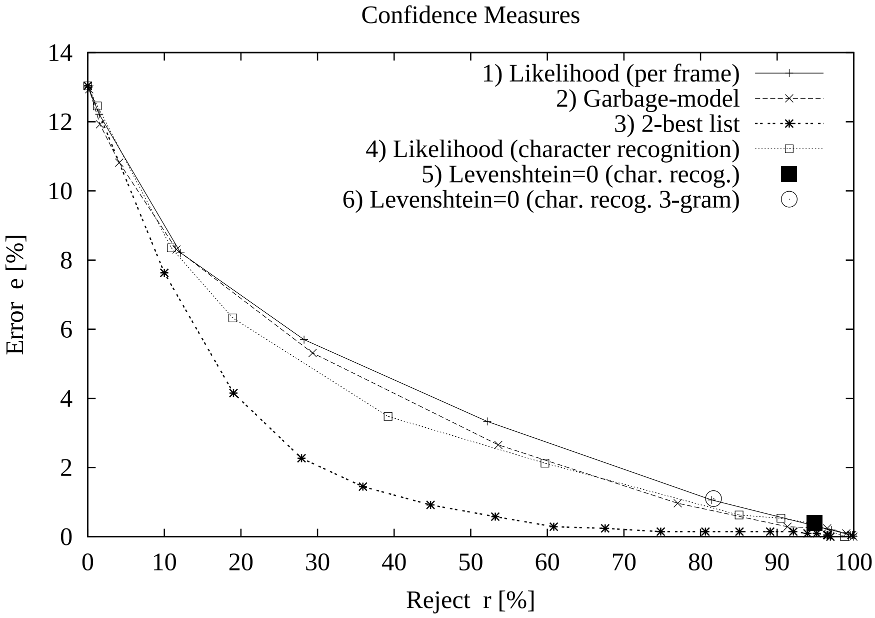

**Fig. 3.** Recognition results for 21 writers using different confidence measures

quite similar, too. Thus, such an uncertain recognition result will be rejected. Using another dictionary the ratio of rejection to error could be completely different. The other evaluated confidence measures are independent of the dictionary, and so just these errors cannot be avoided (compare confidence (5,6)).

Some results using selected confidence thresholds are described in Tab. 1 in greater detail. For comparison, the usual false acceptance rate $FAR$ and the false rejection rate $FRR$ are computed.

**Table 1.** Selected values of word recognition results (%) for 21 writers

| confidence measure | reject | error | false classified | | $FAR$ | $FRR$ |
|---|---|---|---|---|---|---|
| | word ($r$) | word ($e$) | word ($f$) | char | word | word |
| baseline system | 0.0 | 13.0 | 13.0 | 6.4 | 100 | 0.0 |
| (1) likelihood (per frame) | 52.1 | 3.3 | 7.0 | 2.4 | 25.6 | 48.8 |
| (3) 2-best list | 44.8 | 0.9 | 1.7 | 0.6 | 7.0 | 37.5 |

For recognition purpose with high certainty (e.g. human-computer interface without enquiry call or address recognition for postal automation as an off-line application) the confidence measure based on the two-best list will be reasonable. For a retraining of a recognition system or an unsupervised writer adaptation the other confidence measures can be sufficient, too. In this application the character error rate (see Tab. 1) is much more important than the word error rate and, in general, the amount of unlabeled training data is such high that the rejection rate can be disregarded. These aspects will be evaluated in the future work.

## 5 Summary and Conclusions

In this paper we presented the comparison of four different confidence measures for an on-line handwriting recognition system for single words, which is based on HMMs. It has been shown that a confidence measure depending on a two-best list performs significant better than those which are based on the frame-normalized likelihood, a garbage model or an unconstrained character recognition. Some extensions to this work, we want to examine in the future, are the combination of different confidence measures and the usage for unsupervised writer adaptation.

## References

[1] A. Brakensiek, A. Kosmala, and G. Rigoll. Writer Adaptation for On-Line Handwriting Recognition. In *23. DAGM-Symposium, Tagungsband Springer-Verlag*, pages 32–37, Munich, Germany, Sept. 2001.

[2] A. Brakensiek, A. Kosmala, and G. Rigoll. Comparing Normalization and Adaptation Techniques for On-Line Handwriting Recognition. In *16th Int. Conference on Pattern Recognition (ICPR), to appear*, Quebec, Canada, Aug. 2002.

[3] J. Dolfing and A. Wendemuth. Combination of Confidence Measures in Isolated Word Recognition. In *5th Int. Conference on Spoken Language Processsing (IC-SLP)*, pages 3237–3240, Sydney, Australia, Dec. 1998.

[4] S. Eickeler, M. Jabs, and G. Rigoll. Comparison of Confidence Measures for Face Recognition. In *IEEE Int. Conference on Automatic Face and Gesture Recognition*, pages 257–262, Grenoble, France, Mar. 2000.

[5] T. Hazen and I. Bazzi. A Comparison and Combination of Methods for OOV Word Detection and Word Confidence Scoring. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, Utah, May 2001.

[6] J. Hu, S. Lim, and M. Brown. HMM Based Writer Independent On-line Handwritten Character and Word Recognition. In *6th Int. Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pages 143–155, Taejon, Korea, 1998.

[7] R. Plamondon and S. Srihari. On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(1):63–84, Jan. 2000.

[8] L. Rabiner and B. Juang. An Introduction to Hidden Markov Models. *IEEE ASSP Magazine*, pages 4–16, 1986.

[9] R. Rose and D. Paul. A Hidden Markov Model based Keyword Recognition System. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 129–132, Albuquerque, New Mexico, 1990.

[10] F. Wallhoff, D. Willett, and G. Rigoll. Frame Discriminative and Confidence-Driven Adaptation for LVCSR. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1835–1838, Istanbul, Turkey, June 2000.

[11] S. Young. Detecting Misrecognitions and Out-Of Vocabulary Words. In *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 21–24, Adelaide, Australia, Apr. 1994.