Confidence Measures for an Address Reading System

Anja Brakensiek
University Duisburg
47057 Duisburg, Germany
now: anja.brakensiek@odt-oce.com

Jörg Rottland
Siemens Dematic AG
78467 Konstanz, Germany
Joerg.Rottland@siemens.com

Gerhard Rigoll
Technical University Munich
80290 Munich, Germany
rigoll@ei.tum.de

Abstract

In this paper the performance of different confidence measures used for an address recognition system are evaluated. The recognition system for cursive handwritten German address words is based on Hidden Markov Models (HMMs). It is essential, that the structure of the address (name, street, city, country) is known, so that a specific small but complete dictionary can be selected. Choosing a wrong dictionary (OOV: out-of-vocabulary) or misrecognize the word, the recognition result should be rejected by means of the confidence measure. This paper points out two aspects: the comparison of four confidence measures for single words – based on the likelihood, a garbage-model, a two-best recognition or a character decoding – and the comparison of using complete or wrong dictionaries. It is shown, that the best confidence measure – the two-best distance – has a quite different behavior using OOV.

1. Introduction

Automatic recognition systems for handwritten addresses become more and more important for postal automation [6, 10]. Usually, Hidden Markov Model-based techniques (HMM) are used for recognition of cursive handwritten words because of their segmentation-free recognition capability (see [8]).

Although the performance of recognition systems has increased since several years, the error rate of recognizers for unconstrained handwritten address words is still quite high. Thus, for practical usage, a reliability assessment of the recognized words will be necessary. It is cheaper to reject letters, which are recognized uncertain, and to label them manually, than to send them to a wrong address [3]. The decision, if a recognition is uncertain or not, is done by confidence measures. A recognition result obtained by a neural network or a KNN-classifier, for example, can be interpreted as a confidence measure immediately, because it is a kind of posterior probability.

In contrast to this, in an HMM-based system the confidence measures has to be computed additionally, because the recognition result itself is just a likelihood. Here, that word of the dictionary, which is the most probable, is selected. This leads to a further problem, if the selected dictionary does not contain the current word label (OOV: out-of-vocabulary). In postal applications the variety of words is so large, that a complete dictionary for all possible addresses is not really feasible. Once the structure of the address is known, i.e. the words have been categorized (city, street), the size of the dictionaries can be reduced to sizes of 1000 words or less depending on the zip code. If it is not possible to decide, which line of the address block belongs to which category, or if the zip code is misrecognized or wrong, the recognition may be carried out with an OOV-dictionary.

A rating of correctness is not only useful for the reject management in postal applications, but also for an unsupervised adaptation (e.g. PDA, mail streams) using automatically generated labels or a more user friendly dialog in a human-machine communication. In the following sections our address reading system, the theory and some results, which are obtained by four different confidence measures using complete or OOV dictionaries are described.

2. System architecture

Our handwriting recognition system consists of about 77 different linear HMMs (see [8]), using a semi-continuous (resp. tied-mixture) probability modeling structure with 300 Gaussian densities (full covariance matrix). We use one HMM for each character (upper- and lower-case letters, numbers and some special characters like '- / . ( )'), which consists mostly of three states (compare [1]), except for the special characters depending on their width. To estimate the HMM-parameters we use the Baum-Welch algorithm, whereas the recognition is performed by the Viterbi algorithm using city- or street-dictionaries.

In reality, the lexicon is determined and restricted in size by recognition of the zip code, which is easier to read than the rest of the address. To simulate this standard scenario,

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 IEEE

we use small complete dictionaries (no OOV), which are created artificially depending on the test set. Using these correct dictionaries, misrecognized words have to be rejected depending on the confidence measure. If it is not possible to decide, which part of the address belongs to the specification of the city and the street, a recognition using the wrong dictionary may become possible. This means a confusion of city and street dictionaries, which leads to an OOV recognition. In this case, all words have to be rejected using the same confidence measures and thresholds as for a correct specification.

The presented word recognition rates refer only to the recognition of cities (single words like 'Stuttgart' or short sequences like 'Frankfurt am Main') and streets and are independent of errors in zip code or street number recognition.

2.1. Address database

The database consists of single handwritten address words (cursive or block letters) of several post offices in Germany (see Fig. 1: 'BERLIN, Rostock, WEINMEISTERHORNWEG, Wildungerstr.', compare also [1]). The baseline system is trained with about 20000 words, the test set consists of about 2000 words ($N = 935$ cities and $N = 1092$ streets). The complete dictionaries for recognition of cities or streets contain all corresponding test-labels. This leads to a dictionary size of 421 for the cities and of 901 for the streets. The evaluated OOV-dictionaries consist of 100% out-of-vocabulary. Here, the street dictionary is used for the recognition of cities and vice versa.

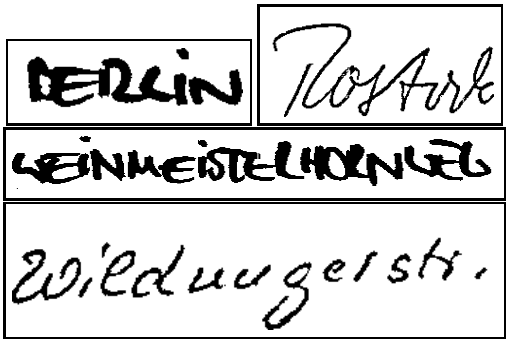

Figure 1. Examples of the database

After localization, segmentation and preprocessing of the address words, the script samples are normalized according to skew, slant and height. The following feature extraction is based on the estimation of the baseline and the line above the lowercase letters, the ruler lines. Therefore, a sliding window technique is used. Every sliding window is divided horizontally by the ruler lines in five overlapping areas, in which dashes, dots, cusp, upstrokes, curves and horizontal or vertical lines are detected to determine 20 features. At last, a linear discriminant analysis (LDA) is performed on always three neighboring frames of these features and the resulting feature vectors X are reduced to 30 dimensions [1].

2.2. HMM based recognition

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):

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

Here $P(W|X)$ represents the posterior probability, $P(X|W)$ is the likelihood, which is computed by the HMMs, $P(W)$ describes the a priori probability of the word W 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 (here: the same a priori probability for each entry), leads to Eq. 2, because $P(X)$ is independent of W:

$$W^* \approx \underset{W}{\operatorname{argmax}}\, P(X|W)\, P(W) \approx \underset{W}{\operatorname{argmax}}\, P(X|W) \quad (2)$$

The probability $P(W)$ is nonrelevant per definition. Disregarding the probability $P(X)$ 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.

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 – the confidence measure $Conf$ – of this recognition result. For our handwriting recognition problem we compare four different confidence measures (compare [5, 7]):

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

The first investigated confidence measure, the likelihood $P(X|W)$, which is normalized by the number of corresponding feature frames is used as a reference. Because of the dynamic of the HMM-based decoding procedure, in general these measures are computed as log likelihoods. Thus, in practice, the logarithm of the confidence measures are examined, which leads to a simple subtraction of the computed log likelihood values. The higher the normalized likelihood resp. the ratio of the likelihoods, the higher the reliability. If the confidence measure is below a threshold t, this test-word has to be rejected.

The following confidence measures take Eq. 1 into account. The posterior probability $P(W|X)$ will be an optimal confidence measure, if it was be possible to estimate $P(X)$ (see also [11]):

$$Conf := \frac{P(X|W)}{P(X)} \qquad Conf \begin{cases} < t \rightarrow reject \\ \geq t \rightarrow accept \end{cases} \tag{3}$$

Thus, the second confidence measure, we tested, is based on a garbage-model (compare [9]). 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. The confidence measure can be calculated using the garbage-model as an approximation of $P(X)$:

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

To determine $P(X|W_{garb})$ the decoding procedure has to be expanded by an additional HMM, as it is shown in Fig. 2.

The third evaluated confidence measure depends on a two-best recognition according to Eq. 5 (see also [2]):

$$P(X) \approx \sum_{k=1}^{N} P(X|W_k) \cdot P(W_k)$$

$$\Rightarrow \quad P(X) \approx \frac{1}{N} \cdot (P(X|W_{1st}) + P(X|W_{2nd})) \tag{5}$$

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 $Conf^* = \frac{Conf}{N-Conf}$ can be defined as follows:

$$Conf = \frac{N \cdot P(X|W_{1st})}{P(X|W_{1st}) + P(X|W_{2nd})}$$

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

Thus, this term is easy to compute, only the domain of $Conf$ resp. t has been changed.

The fourth method to obtain confidence measures is based on an unconstrained character decoding without dictionary (compare e.g. [4]), as it is shown in Fig. 2. The character-based likelihood $P(X|C)$ is used for normalization (again, for rejection the rule of Eq. 3 is applied):

$$P(X) \approx P(X|C) = P(X|c_1, ...c_k) = \prod P(X_{f_i}|c_i) \tag{7}$$

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 $P(X|W)$.

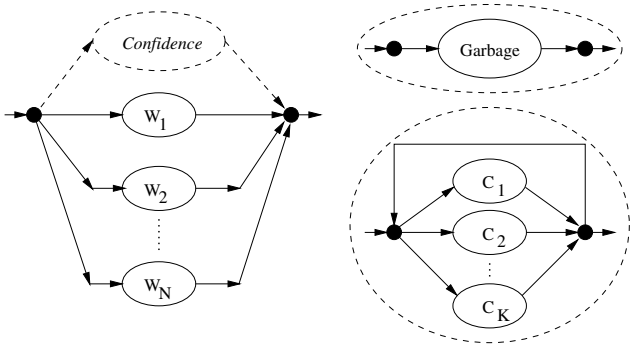

Figure 2. Decoding configuration to obtain confidence measures

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. A further error reduction in address recognition would be possible by taken the context of the entire address into account (holistic approach).

4. Experimental results

In the presented experiments we examine the influence of four different confidence measures for handwriting recognition of single address words (compare also [1]). Additionally, we compare the effect of confidence measures using a complete or an OOV dictionary.

Without rejection, a recognition accuracy of 86.8% for the cities and 91.0% for the streets is obtained by the respective complete dictionary. Fig. 3 shows the false acceptance rate FAR against the false rejection rate FRR for city and street recognition using different confidence measures and increasing thresholds. The usual terms FAR and FRR are

defined as follows (Eq. 8):

$$P(\text{type I error}) = FRR = \frac{FR}{FR + CA}$$

$$P(\text{type II error}) = FAR = \frac{FA}{FA + CR} \quad (8)$$

Here, FR is the number of false rejected words (the result is correct but rejected), CA is the number of correct accepted words, FA the number of false accepted (the result is wrong but accepted) and CR the number of correct rejected examples with $FR + CA + FA + CR = N$.

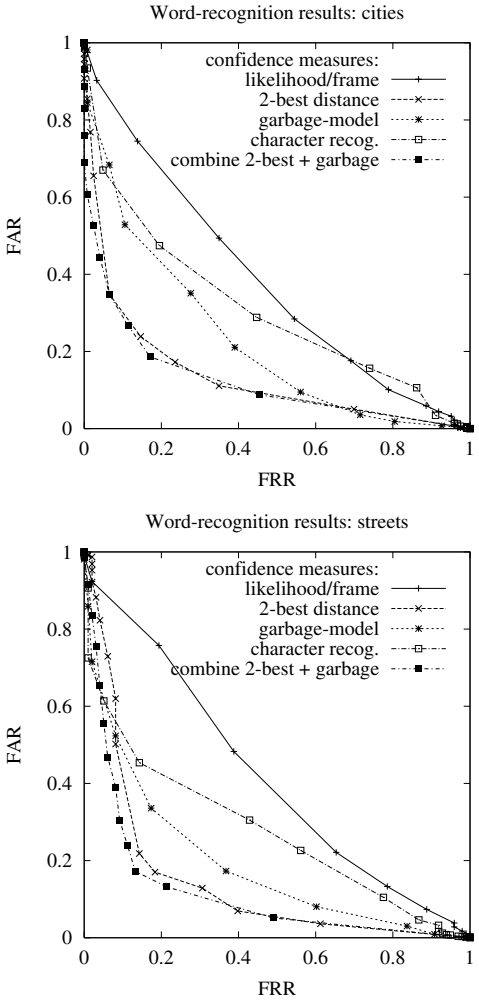

Figure 3. ROC using complete dictionaries

As can be seen in Fig. 3, the best confidence measure is based on the 2-best distance of the likelihoods. And the frame normalized likelihood is the worst confidence measure, as it is expected. The analysis of the 2-best recognition considers the similarity of the entries of the dictionary, thus potential errors can be avoided. This means, that this confidence is highly dependent on the kind of dictionary (e.g. size, similar words like 'Hamburg' and 'Homberg', OOV). The 2-best confidence for the same test-word can be quite different using a dictionary without the second best hypothesis. Additionally, the ROC-curve using a simple linear combination (weighted sum) of the two best measures – the 2-best distance and the garbage-model – is shown. Here, the performance is only little better and this result is obtained by optimizing the parameters manually. Several other combination methods (other weights, maximum decision, AND-conjunction), which perform well on the city test-set, flop on the street-data and reverse. For evaluation of (automatic, more complex) combination methods (see e.g. [7]: MLP) a larger database has to be used to be more independent of the specific test-set.

The second aspect considered in this paper is the behavior of confidence measures using OOV dictionaries. The presented results refer to the city-database, whereas in principle the results are transferable to street recognition. Using the baseline system without rejection ($r = 0\%$) an error rate of $e = 13.2\%$ is achieved testing the city names. Sure, rejecting 100% of the test-data, the error will become zero (compare Fig. 4: left side). For comparison of confidence measures using complete and OOV-dictionaries another presentation – the ratio of error rate e and rejection rate r resp. threshold t – is shown using the following definition (N is the total number of words).

$$r = \frac{CR + FR}{N} \qquad e = \frac{FA}{N} \quad (9)$$

These terms are used, because in a recognition using the wrong dictionary, it is $FR = CA = 0$. The recognition results, which are shown in Fig. 4, are determined using an increasing threshold $t = t0, ..., tn$. On the right side of Fig. 4 the corresponding error rates using a correct or wrong (OOV) dictionary at the same confidence threshold are presented. The left curve presents the rejection rate r referring to this error rate e using complete dictionaries (OOV: $r + e = 100\%$).

First, regarding the 2-best confidence: if any threshold t_k (or t_l) is chosen, such that $e_k = 1.9\%$ ($e_l = 0.8\%$) using the complete dictionary, the corresponding error rate for the OOV-dictionary is $e_k = 21.1\%$ ($e_l = 15.2\%$). Then the rejection rate for the complete dictionary is $r_k = 32.0\%$ ($r_l = 42.6\%$). Now some selected values obtained by the garbage-model (see Fig. 4): if a threshold t_u (t_v) is chosen, such that $e_u = 3.6\%$ ($e_v = 1.4\%$) using the complete dictionary, the corresponding error rate for the OOV-dictionary is $e_u = 4.8\%$ ($e_v = 1.1\%$). Then the rejection rate for the complete dictionary is $r_u = 40.0\%$ ($r_v = 57.6\%$).

The 2-best confidence leads to curves for complete and OOV dictionaries, which run nearly parallel. Thus, for detection resp. rejection of OOV-words the garbage-model performs better than the 2-best confidence.

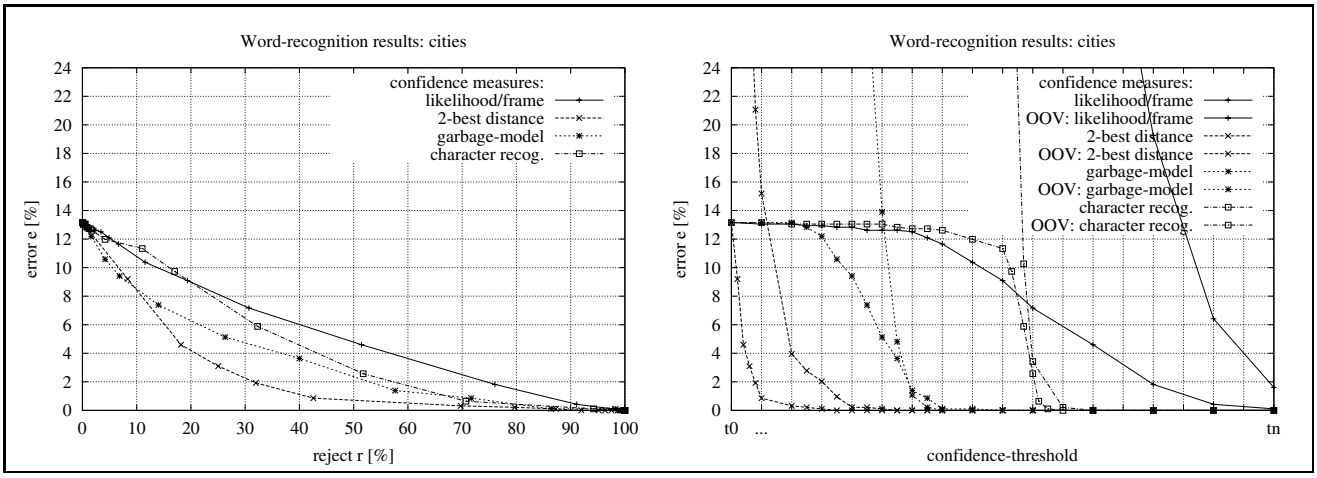

Figure 4. Rejection management using a complete dictionary (left) and corresponding comparison of complete and OOV dictionaries using the same threshold (right)

5. Summary and outlook

We presented in this paper an HMM based handwriting recognition system for German address words with focus on confidence measures to reject uncertain results. Comparing four investigated confidence measures – based on the frame normalized likelihood, a 2-best distance, a garbage-model or an unconstrained character decoding – the best performance is obtained by the 2-best recognition. Regarding the same problem when using OOV-dictionaries the garbage-model based confidence and also the character decoding lead to some better results.

Acknowledgments

This work has been (partially) funded by the BMBF - German Federal Ministry of Education and Research (project: Adaptive READ).

References

[1] A. Brakensiek, J. Rottland, F. Wallhoff, and G. Rigoll. Adaptation of an Address Reading System to Local Mail Streams. In *6th Int. Conference on Document Analysis and Recognition (ICDAR)*, pages 872–876, Seattle, USA, Sept. 2001.

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

[3] J. Gloger, A. Kaltenmaier, E. Mandler, and L. Andrews. Reject Management in a Handwriting Recognition System. In *Int. Conference on Document Analysis and Recognition (ICDAR)*, pages 556–559, Ulm, Germany, Aug. 1997.

[4] 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.

[5] S. Marukatat, T. Artieres, and P. Gallinari. Rejection measures for Handwriting sentence Recognition. In *8th Int. Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pages 24–29, Niagara-on-the-Lake, Canada, Aug. 2002.

[6] U. Miletzki, T. Bayer, and H. Schäfer. Continuous Learning Systems: Postal Address Readers with built-in learning capability. In *5th Int. Conference on Document Analysis and Recognition (ICDAR)*, pages 329–332, Bangalore, India, 1999.

[7] J. Pitrelli and M. Perrone. Confidence Modeling for Verification Post-Processing for Handwriting Recognition. In *8th Int. Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pages 30–35, Niagara-on-the-Lake, Canada, Aug. 2002.

[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] M. Shridhar, F. Kimura, B. Truijen, and G. Houle. Impact of Lexicon Completeness on City Name Recognition. In *8th Int. Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pages 513–518, Niagara-on-the-Lake, Canada, Aug. 2002.

[11] G. Williams and S. Renals. Confidence measures from local posterior probability estimates. *Computer Speech and Language*, 13:395–411, 1999.