Rejection Strategies for Handwritten Word Recognition

Alessandro L. Koerich
Pontifical Catholic University of Paraná
Faculdades Integradas Curitiba
Curitiba, PR, Brazil
alekoe@computer.org

Abstract

In this paper, we investigate different rejection strategies to verify the output of a handwriting recognition system. We evaluate a variety of novel rejection thresholds including global, class–dependent and hypothesis–dependent thresholds to improve the reliability in recognizing unconstrained handwritten words. The rejection thresholds are applied in a post–processing mode to either reject or accept the output of the handwriting recognition system which consists of a list with the N–best word hypotheses. Experimental results show that the best rejection strategy is able to improve the reliability of the handwriting recognition system from about 78% to 94% while rejecting 30% of the word hypotheses.

1 Introduction

Handwriting recognition has been an intensive research field in the last decade [5, 10]. Most of the efforts have been devoted to build systems that are able to recognize handwriting in constrained environments. Besides that, the focus has primarily been in building handwriting recognition systems and improving their recognition rate [5, 8, 10]. Nevertheless, in the overall recognition process, high recognition rates is not the final goal. Recognition rate is a valid measure to characterize the quality of a recognition system, but for practical applications, it is also important to look at the reliability [3, 4, 6, 7]. Reliability is related to the capability of a recognition system to not accept false word hypotheses and to not reject true word hypotheses. The question is not only to find a word hypothesis, but most importantly find out how trustworthy is the hypothesis provided by a handwriting recognition system. However, this problem may be regarded as difficult as the recognition itself is. For such an aim, rejection mechanisms are usually used to reject word hypotheses according to an established threshold [3, 4, 6, 7]. Garbage models and anti–models have also been used to establish rejection criteria [1, 6].

Pitrelli and Perrone [7] compare several confidence scores for the verification of the output of an hidden Markov model based on–line handwriting recognizer. Better rejection performance is achieved by a multilayer perceptron neural network classifier that combines seven different confidence measures. Marukatat et al. [6] have shown an efficient measure of confidence for an on–line handwriting recognizer based on anti–model measures which improves accuracy from 80% to 95% at 30% rejection level. Gorski [4] presents several confidence measures and a neural network to either accept or reject word hypothesis lists. Such a rejection mechanism is applied to the recognition of courtesy check amount to find suitable error/rejection tradeoff. Gloger et al. [3] presented two different rejection mechanisms, one based on the relative frequencies of reject feature values and another based on a statistical model of normal distributions to find a best tradeoff between rejection and error rate for a handwritten word recognition system. El–Yacoubi et al. [2] proposed a rejection mechanism to account for the cases where the input word image is not guaranteed to belong to the lexicon. For such an aim two other terms are considered in the computation of the *a posteriori* word probability: the a priori probability that a word belongs to the lexicon and the probability of a observation sequence given a word out of the lexicon.

In this paper, we present novel rejection strategies for a hidden Markov model based off–line handwritten word recognition. Different from previous works, three types of rejection strategies applied at post–processing level are investigated with the aim of improving the reliability of a handwriting recognition system: global, class–dependent and hypothesis–dependent rejection strategies.

This paper is organized as follows. Section 2 presents some definitions of important measures used through the paper. In order to motivate the work described in this paper it is important to provide some minimal understanding of the context in which the rejection techniques are applied. Section 3 presents a brief overview of the handwriting recognition system. Section 4 presents the details of the

Proceedings of the 9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004)
0-7695-2187-8/04 $20.00 © 2004 IEEE

rejection strategies proposed in this paper. Experimental results are presented in Section 5. The conclusions of this paper are presented in the last section.

2 Definitions

The task is to recognize an unknown handwritten word which can belong to L classes, where L coincides with the number of lexicon entries. Therefore, there are $N \leq L$ possible answers which are called hypotheses, each of which is associated with a confidence score. In our case, such confidence scores are *a posteriori* probabilities. The handwriting recognition system classify correctly an input word when it assigns the correct lexicon entry to the word since it is a lexicon–driven approach.

To evaluate the results of the rejection strategies proposed in this paper, the following measures are employed: recognition rate, error rate, rejection rate, and reliability, which are defined as follows:

$$RecognitionRate = \frac{N_{recog}}{N_{test}} \times 100 \quad (1)$$

$$ErrorRate = \frac{N_{err}}{N_{test}} \times 100 \quad (2)$$

$$RejectionRate = \frac{N_{rej}}{N_{test}} \times 100 \quad (3)$$

$$Reliability = \frac{N_{recog}}{N_{recog} + N_{err}} \times 100 \quad (4)$$

where N_{recog} is defined as the number of words correctly classified, N_{err} is defined as the number of words misclassified, N_{rej} is defined as the number of input words rejected after classification, and N_{test} is the number of input words tested.

3 Handwriting Recognition System

Our system is a large vocabulary off–line handwritten word recognition based on discrete hidden Markov models. The recognition system was designed to deal with unconstrained handwriting (handprinted, cursive and mixed styles), multiple writers (writer–independent), and dynamically generated lexicons. Each character is modeled by a ten–state left–right transition–based HMM with no self–transitions. Intra–word and inter–word spaces are modeled by a two–state left–right transition–based HMM [1]. Words are formed using standard concatenation techniques.

The general problem of recognizing a handwritten word w, or equivalently a character sequence constrained to spellings in a lexicon $\mathcal{L}$, is framed from a statistical perspective, where the goal is to find the sequence of labels $c_1^L = (c_1 c_2 \dots c_L)$ (e.g. characters) that is most likely given the sequence of T observations $o_1^T = (o_1 o_2 \dots o_T)$:

$$\hat{w} \ni P(\hat{w}|o_1^T) = \max_{w \in \mathcal{L}} P(w|o_1^T) \quad (5)$$

The posteriori probability of a word w can be rewritten using Bayes' rule:

$$P(w|o_1^T) = \frac{P(o_1^T|w)P(w)}{P(o_1^T)} \quad (6)$$

where $P(w)$ is the prior probability of the word occurring. The probability of data occurring $P(o_1^T)$ is unknown, but assuming that the word is in the lexicon $\mathcal{L}$ and that the decoder computes the likelihoods of the entire set of possible hypotheses, then the probabilities must sum to one, and can be normalized:

$$\sum_{w \in \mathcal{L}} P(w|o_1^T) = 1 \quad (7)$$

In such a way, estimated *a posteriori* probability can be used as confidence estimates which is obtained as:

$$P(w|o_1^T) = \frac{P(o_1^T|w)P(w)}{\sum_{w \in \mathcal{L}} P(o_1^T|w)P(w)} \quad (8)$$

At the output, the handwriting recognition system provides a list with the *N*–best word hypotheses ranked accordingly to the *a posteriori* probability assigned to each word hypothesis.

4 Rejection Strategies

The concept of rejection admits the potential refusal of a word hypothesis if the classifier is not certain enough about the hypothesis. In our case, an evidence about the certainty is given by the probabilities assigned to the word hypotheses (Equation 8). Assuming that all words are present in the lexicon, the refusal of a word hypothesis may have two different reasons:

- there is not enough evidence to come to a unique decision since more than one word hypothesis among the *N*–best word hypotheses appears adequate;
- there is not enough evidence to come to a decision since no word hypothesis among the *N*–best word hypotheses appears adequate;

In the first case, it may happen that the probabilities do not indicate a unique decision in the sense that there is not just one probability exhibiting a value close to one. In the second case, it may happen that there is no probability exhibiting a value close to one. Therefore, the probabilities

assigned to the word hypotheses in the *N*–best word hypothesis list should be used as a guide to establish a rejection criterion.

Bayes decision rule embodies already a rejection rule, namely, find the maximum of $P(w|o)$ but check whether the maximum found exceeds a certain threshold value or not. Due to the decision–theoretic conceptions this reject rule is optimum for the case of insufficient evidence if the closed–world assumption holds and if the *a posteriori* probabilities are know [9]. Therefore, this suggests rejecting a word hypothesis if the probability for that hypothesis is less than a threshold.

In the context of the handwriting recognition system, the task of a rejection mechanism is to decide on whether the best word hypothesis in the *N*–best word hypothesis list can be accepted or not. For such an aim, we have investigated different rejection strategies: class–dependent rejection where the rejection threshold depends on the class of the word; hypothesis–dependent rejection where the rejection threshold depends on the probabilities of the word hypotheses at the *N*–best list; global threshold that depends neither on the class nor on the hypotheses. The details of the rejection strategies are presented as follows.

Class–Dependent Rejection Threshold

- Average probability of recognizing the class correctly (*avg_class*): given k samples of a word w in the training dataset, we average the *a posteriori* probability provided by the handwriting recognition system to the samples when they are recognized as the best word hypotheses. Accordingly, the rejection threshold R_{avg_class} is defined as:

$$R_{avg\_class} = \frac{1}{K} \sum_{k=1}^{K} P(w_k|o_1^t(k)) \tag{9}$$

  where K is the number of times the word w_k appears in the training dataset.

- *A priori* class probability (*pri_class*): a simple rejection threshold which is based on the *a priori* probability of a word w to appear in the training dataset.

$$R_{pri\_class} = P(w) \tag{10}$$

Hypothesis–Dependent Rejection Threshold

- Average probability of the *N*–best word hypotheses (*avg_top*): given N word hypotheses provided by the handwriting recognition system, we average the *a posteriori* probabilities assigned to the word hypotheses. The rejection threshold R_{avg_top} is defined as:

$$R_{avg\_top} = \frac{1}{N} \sum_{n=1}^{N} P(H_n) \tag{11}$$

  where H_n denotes the n–th word hypothesis.

- Difference between the *a posteriori* probabilities of the best word hypothesis and the second best word hypothesis (*dif_12*). It is defined as:

$$R_{dif\_12} = P(H_1) - P(H_2) \tag{12}$$

Fixed Rejection Threshold

- fixed threshold (*fixed*): a global rejection threshold that is class–independent and hypothesis–independent.

$$R_{fixed} = P \tag{13}$$

where P is a probability obtained experimentally on a validation dataset, according to the rejection level expected.

Given the rejection thresholds defined as before and denoted as $R_{(.)}$, the rejection rule will be given as:

1) The best word hypothesis is accepted whenever

$$P(H_1) \geq \gamma R_{(.)} \tag{14}$$

2) The best word hypothesis is rejected whenever

$$P(H_1) < \gamma R_{(.)} \tag{15}$$

where $P(H_1)$ is the *a posteriori* probability of best word hypothesis provided by the handwriting recognition system, and $\gamma \in [0, 1]$ is a parameter that indicates the amount of variation of the probability between the best word hypothesis and the rejection threshold. The value of γ is set according to the rejection level required.

5 Experiments and Results

For the experiments, a proprietary database containing more than 20,000 real postal envelops was used. Three datasets that contain city names manually located on postal envelopes were used in the experiments, as well as a very–large vocabulary of 85,092 city names. The training dataset contains 12,023 unconstrained handwritten words, the validation dataset contains 3,475 unconstrained handwritten words and the test dataset contains 4,674 unconstrained handwritten words.

We have applied the rejection strategies on the word hypotheses produced by the handwriting recognition system. Figure 1 shows the word error rates on the test dataset as a function of rejection rate for the different rejection criteria and considering an 80,000–word lexicon. Among the

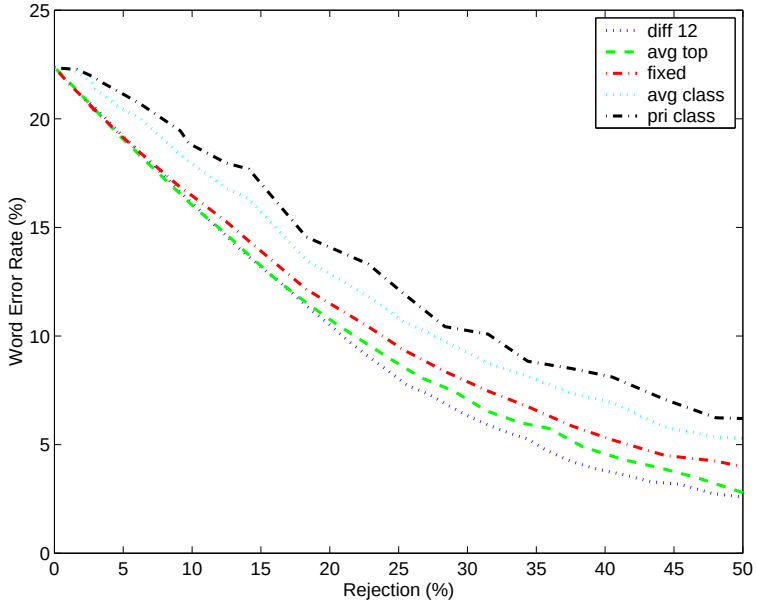

Figure 1. Word error rates versus the rejection rate for the different rejection thresholds and an 80,000–word lexicon

different rejection criteria, the criterion based on the difference between the probabilities of the first best word hypothesis (H_1) and the second best word hypothesis (H_2) performs the best. A similar performance was observed in different lexicon sizes. Surprisingly, the class–dependent rejection thresholds did not provide good results. This is due to the reduced number of samples (or even the absence) in the training dataset for some word classes.

Figure 2 shows the word error rates on the test dataset as a function of the rejection rate for different lexicon sizes and using the R_{dif_12} rejection threshold. It is clear that such a rejection strategy provides an interesting error–rejection tradeoff for all lexicon sizes. For instance, at a 40% rejection level, the word error rate on small and large lexicons is less than 1%, while on very large lexicons, the word error rate is less than 4% .

Besides the reduction in word error rates afforded by the different rejection strategies, it is also interesting to look at another rejection statistics, such as the false–rejection rate (Type II Error) and the false–acceptance rate (Type I Error). Figure 3 shows the detection and tradeoff curve. This figure shows again that the R_{dif_12} rejection threshold provides the best results among all rejection strategies. For instance, if a low false–acceptance rate is the goal, the handwriting recognition system requires only a 25% false–rejection rate to achieve a false–acceptance rate below 10%.

Finally, the last aspect that is interesting to analyze is the improvement in reliability afforded by the rejection strategies. Reliability is an interesting performance measure because it takes into account both the error rate and the rejection rate. Figure 4 shows the evolution of the recognition rate, error rate and reliability as a function of the rejection rate which is based on the R_{dif_12} rejection threshold. We can observe from this figure that for low rejection rates, the rejection strategy based on the R_{dif_12} rejection threshold produces interesting error–reject tradeoff. Reliability is a more suitable measure to assess the performance of a classifier in real applications because it gives an impression of the classifier behavior in several different situations, that is, at different rejection and error levels.

6 Discussion and Conclusion

In this paper we have presented different rejection strategies for the problem of off–line handwritten word recognition. Three different rejection strategies were investigated: class–dependent, hypothesis–dependent and global. The experimental results have shown that the hypothesis–dependent is the best rejection strategy in combination with the R_{dif_12}. Notice that only this strategy is in accordance with the reasons for rejecting a word hypothesis stated in Section 4.

In this way, incorporating a rejection mechanism to the handwriting recognition system is a powerful method for reducing error rate and improving reliability. As we have seen, the word error rates can be reduced in more than 10% for very–large vocabularies (>40,000 words) at the cost of

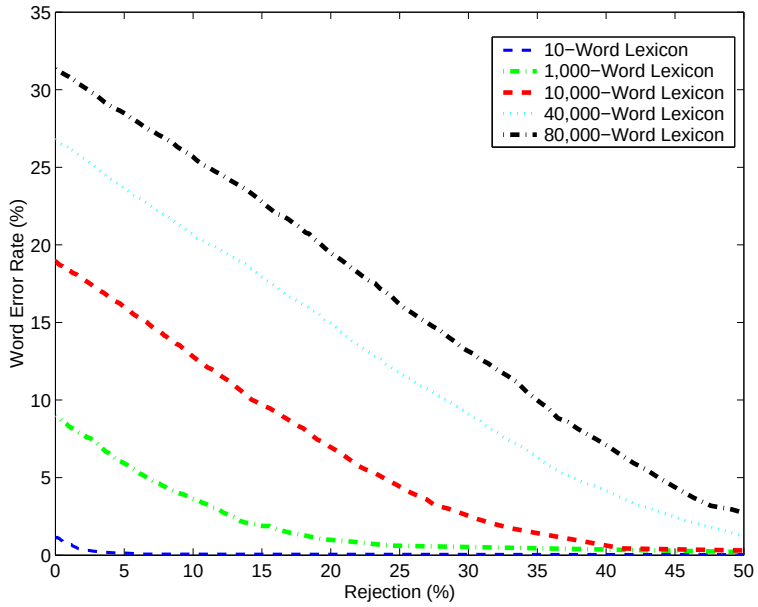

Figure 2. Word error rates versus the rejection rate for different lexicon sizes using the R_{dif_12} rejection criterion

rejecting 20% of the input word images.

In spite of the differences in the experimental environment, the Type I and Type II error rates are close to the results presented in [7] which uses a combination of seven confidence measures, a database of 1,157 words, and a 30,000–word lexicon. The performance of the proposed rejection mechanism is also very similar to the rejection mechanism proposed in [6] which uses anti–model confidence measures, a database of 2,000 words, and a 3,000–word lexicon.

In the future, given the individual rejection thresholds, we plan to study the combination of hypothesis–dependent and class–dependent rejection strategy as a means to improve rejection performance. We also plan to evaluate the proposed rejection strategies on other databases.

References

[1] A. El-Yacoubi, M. Gilloux, R. Sabourin, and C. Y. Suen. Unconstrained handwritten word recognition using hidden markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):752–760, 1999.

[2] A. El-Yacoubi, R. Sabourin, C. Y. Suen, and M. Gilloux. Improved model architecture and training phase in a off–line hmm-based word recognition system. In *Proc. 14th International Conference on Pattern Recognition*, pages 17–20, Brisbaine, Australia, 1998.

[3] J. Gloger, A. Kaltenmeier, E. Mandler, and L. Andrews. Reject management in a handwriting recognition system. In *Proc. 4th International Conference Document Analysis and Recognition*, pages 556–559, Ulm, Germany, 1997.

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

[5] A. L. Koerich, R. Sabourin, and C. Y. Suen. Large vocabulary off–line handwriting recognition: A survey. *Pattern Analysis and Applications*, 6(2):97–121, 2003.

[6] S. Marukatat, T. Artires, P. Gallinari, and B. Dorizzi. Rejection measures for handwriting sentence recognition. In *Proc. 8th International Workshop on Frontiers in Handwriting Recognition*, pages 24–29, Niagara-on-the-Lake, Canada, 2002.

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

[8] R. K. Powalka, N. Sherkat, and R. J. Whitrow. Word shape analysis for a hybrid recognition system. *Pattern Recognition*, 30(3):412–445, 1997.

[9] J. Schurmann. *Pattern Classification: A Unified View of Statistical and Neural Approaches*. John Wiley and Sons, 1996.

[10] T. Steinherz, E. Rivlin, and N. Intrator. Offline cursive script word recognition – a survey. *International Journal on Document Analysis and Recognition*, 2:90–110, 1999.

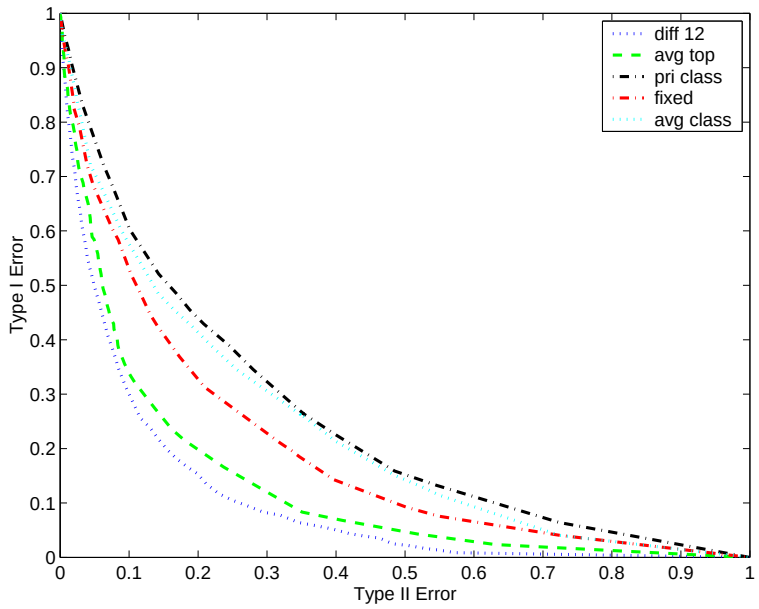

Figure 3. False–rejection rate on correctly–recognized words (Type II error) versus false–acceptance rate on incorrectly–recognized words (Type I error rate) for the different rejection thresholds

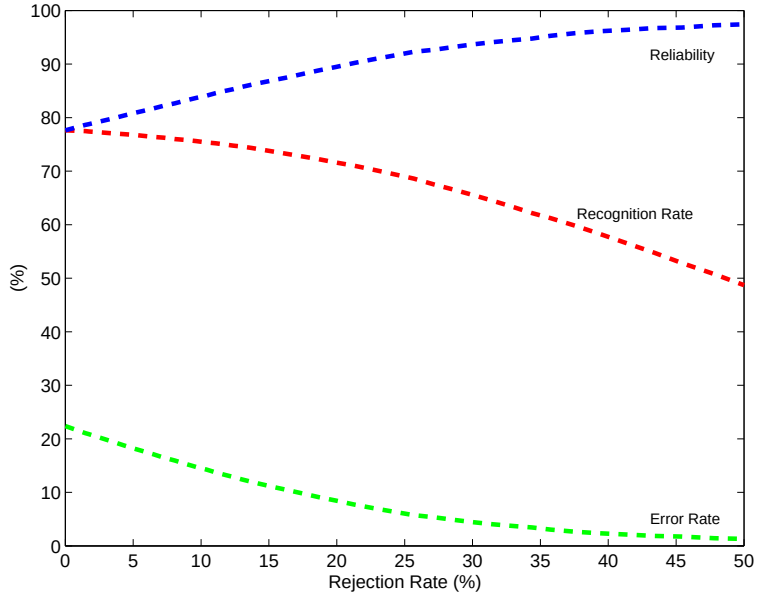

Figure 4. Recognition rate, error rate and reliability as a function of rejection rate for the R_{dif_12} rejection threshold