# Reviewing Performance Metrics for Handwriting Recognition: Must-Rejects and Recognition Graph Scores

Marc-Peter Schambach
*Siemens AG*
*Marc-Peter.Schambach@siemens.com*

## Abstract

*The performance of handwriting recognition systems is typically measured in terms of "recognition rate". Many academic competitions work this way. However, in practical applications, additional requirements may shift the view of recognition quality: Processing time and acceptable error rate may be limited; lexica may be missing, but are needed to unambiguously define result correctness. These aspects will be discussed in detail, and appropriate metrics will be proposed. A single-valued combination of these metrics may then be defined for specific application areas. It can be used in order to choose between recognition approaches or systems, and to optimize system parameters automatically.*

## 1. Introduction

Competitions of handwriting recognition systems have become increasingly popular in the last years [9, 5]. They define and standardize the image database, the recognition interface, and the evaluation method. Thus, they are an important step towards an objective comparison of recognition methodologies, a precondition to boost the progress in development of powerful recognition systems.

### 1.1. Standard evaluation

Typical competitions define an *image database* that consists of readable word images, labeled with literal transcription, and one or multiple lexica, which contain this image label. The image database is divided into published training and evaluation sets, and an unpublished test set on which the evaluation is performed.

The recognition *interface* takes word images and lexica as input, and requests one or more result word alternatives as output, accompanied and sorted by confidence values. The confidence values may be scaled to the interval $[0..1]$, and be interpreted as probability or similarity measure [8, 12].

*Evaluation* is typically performed on recognition rate only, sometimes added by $n$-best recognition, taking the first $n$ alternatives into account. It uses "forced recognition" mode, which ignores the difference between potentially rejected images and false recognition.[1]

Sometimes, the systems' run-time performance is also measured and provides additional information. This puts additional efforts to competition's organizers and participants, because running systems are evaluated instead of data only. But it is important information to rate the systems' practicality.

### 1.2. Additional aspects

In practice, the environment in which handwriting recognition is typically used, questions some aspects of the standard evaluation approach of competitions:

- Clients using handwriting recognition must balance recognition vs. error rate; or error rates are strictly limited. Realistic interfaces must allow rejects, realistic evaluations must rate reject-curves.
- Input data is always uncertain: Images may not contain writing, lexica may not contain the image's content: Word recognition is hypothesis testing. Realistic databases and evaluations must be able to handle these out-of-vocabulary (OOV) situations.
- Lexica may not be available, especially in spontaneous handwriting. Sayre's paradox [10], expanded from word to document level, would state the mutual dependence of lexicon-based recognition and recognition-based lexica. Realistic interfaces must allow missing lexica and provide ade-

[1] Theoretical considerations concerning lexicon-dependence, which lead to augmented interfaces (separate estimation of image quality and word similarity) and performance metrics (correct prediction of similarity), are not covered here.[12]

quate recognition graph output [13], realistic evaluations must rate the quality of these graphs.

### 1.3. Overview

The following sections 2 and 3 will discuss these aspects – reject handling, out-of-vocabulary and lexicon-free recognition – in detail. Section 4 supports the importance of the aspects with some data, and section 5 illustrates the proposed metrics with experimental results. Finally, section 6 draws some conclusions.

## 2. Reject

### 2.1. Error limitation

Many clients of word recognition have strict error limitations, which may vary even between successive tasks within the same application. For example, in postal address recognition, different address elements may have different accuracy requirements. They may correspond to the relative, application-specific costs of errors: Wrong recognition of country names may result in sending mail to wrong countries, being much more expensive than e. g. wrongly recognized recipient's names. Or they correspond to different amounts of redundancy within the address, which do or do not allow the subsequent correction of errors.

Thus, error limitations must be implemented in most word recognition systems. Uncertain results must be rejected, either explicitly or implicitly, by using confidence values. Confidences are normally calculated by combination of intermediate result values, specific to the selected recognition method (e. g. HMM log-probabilities, neural-net outputs), and may also include global features like image quality scores, joker-words and distances between alternatives [2, 3, 7].

Recognition performance is then often characterized as so-called "receiver operating characteristics", or simply ROC curves [15]. For the purpose of easy interpretation and viable competitions, single points at fixed error levels (e. g. 1%, 2% or 5%) may be sufficient. Arbitrary confidence values may be scaled to error levels with simple threshold histograms. The advantage is: The evaluation system itself can implement it, imposing no additional specifications to recognition systems.

### 2.2. Must-reject

Out-of-vocabulary (OOV) situations are those where the correct image transcription is not contained in the input lexicon for word recognition. It's synonymous to "must-reject", i. e. the *correct* answer of word recognition is "reject".

In practice, OOV situations happen quite often. For example, in address reading, recognition is guided hierarchically from postal code, city name, street name, street number to recipient's name. In each step, dynamic lexica implement tests of preceding hypotheses: Wrong hypotheses correspond to OOV-situations and must be rejected. Word recognition is *verification*.

There are two kinds of OOV situations which correspond to two types of wrong hypotheses: Images don't show legible words, or legible words are not represented in the lexicon: The image is "wrong", or the lexicon is "wrong". Both situations arise in practice and must be addressed by a word recognition system. See section 4.2 for a quantitative analysis.

However, OOV situations are rarely tested, maybe because it seems to be more difficult to construct realistic OOV situations compared to readable data. Realistic illegible "wrong" images are highly dependent on application and implementation. And "wrong" lexica need a useful definition, which addresses the problem of similar shapes, either real, as in e. g. *"Ireland"/"Iceland"*, or accidentally, as in spelling mistakes. See [12] for a discussion of similarity in word recognition.

To overcome this deficiency, the proposed procedure is to generate a special must-reject test, which consists of real-world recognition tasks from a client system. All tasks are logged, and must-rejects are marked. This can be done implicitly during standard generation of test and training data, which mostly is performed semi-automatically using a client system. Must-reject data can then be evaluated separately, but set into relation to standard evaluation.

## 3. Lexicon-free recognition

Frequently, especially early in the recognition process, no useful lexicon is available for word recognition. Lexicon-free recognition is then performed — as in "standard" segmentation based OCR — and results are provided as *"recognition graph"*, a probabilistic character graph containing all segmentation and classification alternatives above a given cost threshold [13].

Reduction of time and memory costs may be accomplished by the use of character $n$-gram language models, but besides performance improvements have been reported in isolated word recognition [1], no positive effects on higher-level text recognition systems could be observed [13]. Instead, external lexical knowledge should only be used in later recognition and interpretation stages.

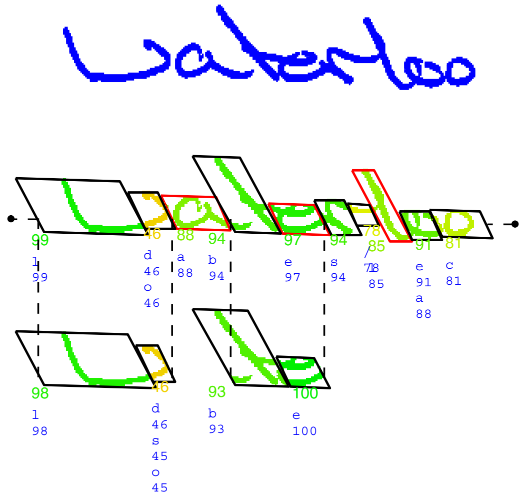


**Figure 1. The recognition graph: Quadrangles with color-coded confidence values denote segmentation alternatives, blue characters and confidences denote classification alternatives. Path costs limit the graph size. The best-matching sub-sequence "ael" is marked in red.**

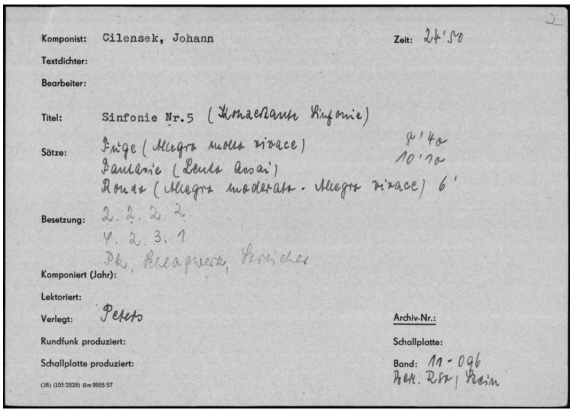
### 3.1. Recognition graph

The recognition graph represents the results of lexicon-free recognition. It is a probabilistic, edge-based character graph. Nodes represent segmentation cuts, edges are character segments, each containing a set of character recognition alternatives.

Fig. 1 shows the visualization of a recognition graph. It has been created by an HMM-based system which will be described in section 5.1. It is based on a sliding-window approach with constant slant. The size of the graph is limited by maximum path cost – each segment must belong to at least one path with costs below the given threshold. Limitation by number of characters is implemented by setting the cost threshold appropriately.

### 3.2. Lexicon-free performance

When no lexicon is applicable, recognition performance is often given by *character error rate*. It is defined – in terms of the recognition graph – by the normalized string distance to the *single best path* in the graph. However, no segmentation and classification alternatives are taken into account this way. They may be cheap compared to character substitution, which justified the use of recognition graphs in the beginning. Obviously, all parts of a recognition graph shall be used by clients. Thus, it's better to calculate the string distance to the *best-matching path* in the recognition graph.



Komponist: Cilensek, Johann
Zeit: 24'50
Textdichter:
Bearbeiter:
Titel: Sinfonie Nr.5 (Konzertante Sinfonie)
Sätze: Fuge (Allegro molto vivace) 8'40
Pantasie (Lento assai) 10'10
Rondo (Allegro moderato - Allegro vivace) 6'
Besetzung: 2. 2. 2. 2
4. 2. 3. 1
Pk, Schlagwerk, Streicher
Komponiert (Jahr):
Lektoriert:
Verlegt: Peters
Archiv-Nr.:
Rundfunk produziert:
Schallplatte:
Schallplatte produziert:
Band: 11-096
Rek. RSO / Klein
(36) (105/2528) Bm 9955/57

**Figure 2. File card from the Music Information Center of the German National Library**

A recognition system's performance is characterized by a function which relates fixed graph sizes (or costs) to average string distances. A practical metric is defined by the average string distance at a reasonable, predefined maximum graph size

Because a client system considers all alternatives in the graph, it works best with the smallest graph which still contains the correct result. These graphs vary in size with the readability of the image data. Thus, the best metric is to predefine an *average* graph size, and determine the average distance between graph and transcription label.

## 4. Data analysis

### 4.1. Application domains

Experiments are conducted with data taken from two domains.

**Postal addresses** Lexicon based script word recognition has been used since the mid nineties in the postal domain for address reading systems [14]. Postal address databases give a solid foundation for the hierarchical creation of dynamical lexica for word recognition. Address reading is well understood. Sorting systems for German and Canadian postal addresses are used to generate statistics of real-world usage of script word recognition.

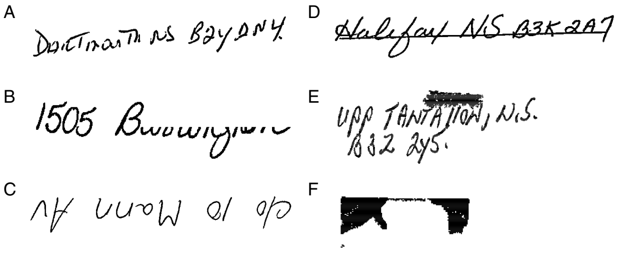


**Figure 3. Samples of word images from must-reject situations: Difficult writing, or wrong lexica (A), occlusions (B), upside-down (C), writing lines (D), distortions (E), or non-word images (F).**

**Music file cards** Lexicon-free recognition is used in domains where only little context information is available. The Music Information Center (MIZ) of the German National Library (DNB) has a large collection of file cards containing information on pieces of music available in the library. The file cards (Fig. 2) are partly hand-written and shall be digitized. Relevant data fields are, among others: composer, style, date and title. Music titles are in many languages and even contain artificial words, thus not allowing the usage of lexica, providing a good example for lexicon-free recognition.

### 4.2. Must-reject statistics

In order to get an impression of real-world usage of script word recognition, program calls to script street name recognition in an address recognition system for German and Canadian addresses have been logged and analyzed. 10000 mail pieces have been recognized, of which 2350 have been labeled as hand written. The total number of calls to script word recognition has been 8753. All 987 street name image hypotheses have been labeled with transcription, or classified "unreadable". See Fig. 3 for some examples of unreadable images. All calls have been lexicon-based, so each image has been matched with its lexicon in order to detect OOV situations. Table 1 shows the statistic of the calls.

It has to be noted that this data is only a snapshot during development, because tuning of word recognition to the calling pattern changes the pattern itself again. Although data evaluation is performed semi-automatically – reference data and recognition results propose most data correctly – it still needs manual interaction which prevents multiple iterations. However, Table 1 shows quite clearly the scale of typical calls of word recognition.

| | | | |
|---|---|---|---|
| Total | | 987 | |
| Image readable | | 801 | 81% |
| Lexicon matches | | 415 | 42% |
| .. exact match | *"Main Street"* | 334 | 34% |
| .. extension | *"53, Main Street"* | 79 | 8% |
| .. abbreviation | *"Main St"* | 2 | 0% |
| Out-of-vocabulary | | 386 | 39% |
| .. wrong street | *"Side Road"* | 237 | 24% |
| .. not a street | *"Corporation"* | 149 | 15% |
| Image unreadable | | 186 | 19% |

**Table 1. Analysis of calls to script street name recognition in a postal address reading system**

The most outstanding result is the high total must-reject rate (unreadable and OOV) of 58%, and the low number of exact matches with 34%.

Another aspect has been noticed during the acquisition of script handwriting training data. Script data from Canadian postal address reading has been collected: 7000 calls to script word recognition have been logged, and images have been labeled semi-automatically. The amount of unreadable images has been 20%, but another 17% have been readable, but machine printed. Thus they were not useful for training and evaluation of *script* word recognition. We can conclude, it is also important to define the scope and follow the responsibility of recognition modules.

## 5. Experiments

### 5.1. Recognition systems

The first word recognition system tested here is a fairly standard HMM-based system. It extracts the image contour, writing lines, uses a sliding window, performs features extraction, vector quantization, creates a dynamic lexicon HMM built from letter-HMMs, and performs Viterbi evaluation, and finally normalizes the confidence. It has been described in detail in [6, 11]. It has been expanded to lexicon-free and $n$-gram recognition with recurrent HMMs, producing recognition graphs as output [13].

In order to compare different performance characteristics, a second system using a different approach, based on multi-dimensional recurrent neural nets and long short-term memory [4] has been evaluated. However, no experiments on lexicon-free, graph based recognition have been performed, because this feature has not been available to the author at the point of writing. The

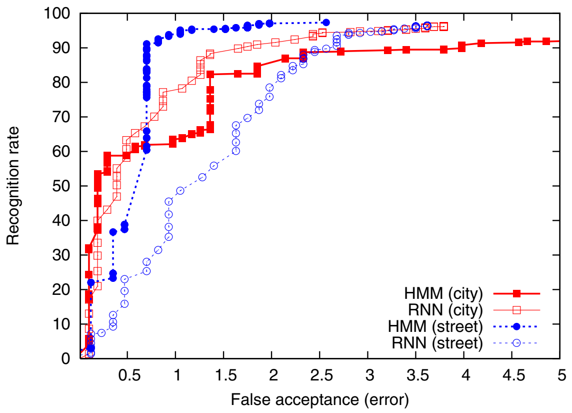


**Figure 4. Recognition rates at different error levels for Canadian city and street names; based on hidden Markov models (HMM) and recurrent neural nets (RNN).**

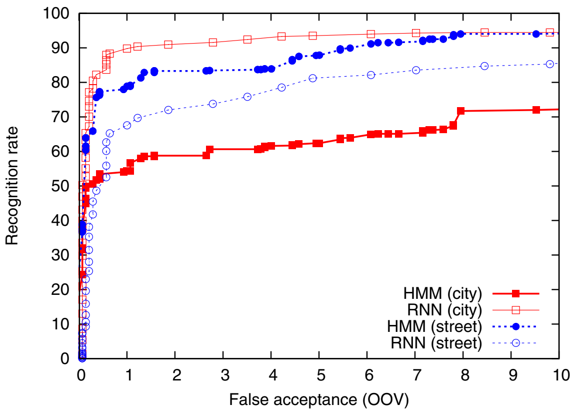


**Figure 5. Recognition rates at different levels of OOV errors for Canadian city and street names.**

system has not been adapted specifically to the data.

### 5.2. Reject performance

**Error limitation** In order to illustrate the significance of reject behavior on the evaluation of recognition engines, the two system, based on hidden Markov models (HMM) and recurrent neural nets (RNN), have been tested with the same data. Canadian city and street images with typical lexica had to be recognized.

The ROC curves on in Fig. 4 show a remarkable result: Pure recognition performance and correct confidence estimation must not necessarily correlate. While for street name recognition the ranking of the systems is constant at all error levels, the curves are crossing for city name recognition. For maximum performance the RNN based system is best, while for low error rates the HMMs are best. So the choice of the best system depends on the client's requirements.

**Must-reject** For the evaluation of the OOV-behavior of the systems, a realistic mix of the Canadian must-reject images and lexica (non-images, distorted images, wrong extracts and wrong lexica) of section 4.2 has been chosen for the experiment. By adjusting the confidence threshold, (correct) reject and (false) acceptance rates are set. In the ROC curve of Fig. 5, the latter is set into relation to the correct recognition rates on readable data of the last experiment.

Having separate tests with readable and must-reject data makes it easy adjust the evaluation dynamically to the real distribution of recognition tasks, as given in Table 1. Additionally, a cost model with separate parameters for recognition rate, false recognition, and false OOV acceptance can be incorporated. It provides the basis for the optimal estimation of confidence reject thresholds, and the combination of different recognition systems.

### 5.3. Recognition graph scores

For the evaluation of recognition graphs, word line images from the music file cards as shown in Fig. 2 are used. It is very difficult to determine a useful lexicon for title words, so lexicon-free recognition is reasonable.

Fig. 6 correlates the average graph size to the the string similarity of the best matching path in the graph. Graphs are limited to fixed sizes, or by maximum path costs, which results in no differences in the average similarity. String similarities around 90% are reached for graph sizes around 1000; this result is acceptable, because a huge reduction of hypotheses can be used for inexact string matching in further processing steps. Complete strings are contained in 40% of the graphs, which is surprisingly high for an average word line length of 26 characters.

Surprisingly, the string similarity is not influenced by the graph-limiting method: Fixed graph sizes yield the same results as dynamic, cost limited graphs. They are even better for complete matches.

## 6. Conclusions

Competitions of recognition systems mostly use $n$-best *(forced)* recognition rates as measure to rank sys-

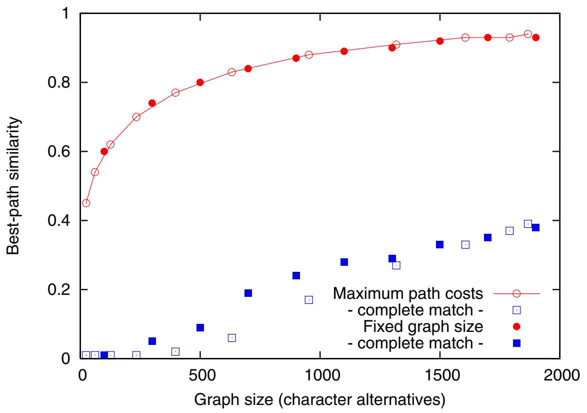


**Figure 6. Graph matching for different graph sizes. The average image word line length is 26 characters.**

tem performance. However, these results may be of limited significance for the decision to choose a recognition system for a real-life application. It has been shown, that the following aspects should be regarded in addition to simple recognition rates:

**Reject handling** Controlling error rates is a must for every recognition system, because there always exists a cost model for error and reject, even if it is unknown. Furthermore, OOV situations arise in most recognition systems; they may be even the majority of usage. Word recognition systems may differ greatly in the ability to recognize, or to reject. Thus, OOV test cases should be integrated in every test set.

**Lexicon-free performance** can be reasonably measured in terms of graph sizes. However, there are no principal differences to evaluation of lexicon-based recognition. Graph size evaluations may give a good impression of the influence on the higher-level recognition system.

Computation time is a limiting factor for many applications and should always be measured. Although, in the context of feasibility studies, it may play an inferior role.

Typically, systems are ranked similarly if evaluated with different metrics [5]. However, this may be true for similar technical solutions only. To truly compare different approaches, e. g. HMM-based, RNN-based and holistic systems, metrics have to be chosen which reflect the requirements of typical applications. The more orthogonal the approaches, the more necessary are the advanced metrics proposed here. This may give hints and proposals for organizers of future competitions.

## Acknowledgments

The work presented here was partially funded by the German Federal Ministry of Economy and Technology (BMWi) under the THESEUS project.

## References

[1] A. Brakensiek and G. Rigoll. A comparison of character N-grams and dictionaries used for script recognition. In *6th ICDAR*, pages 241–245, Seattle, WA, 2001.

[2] A. Brakensiek, J. Rottland, and G. Rigoll. Confidence measures for an address reading system. In *7th ICDAR*, pages 294–298, Edinburgh, Scotland, 2003.

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

[4] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):855–868, 2009.

[5] E. Grosicki and H. E. Abed. ICDAR 2009 handwriting recognition competition. In *10th ICDAR*, Barcelona, Spain, 2009.

[6] A. Kaltenmeier, T. Caesar, J. Gloger, and E. Mandler. Sophisticated topology of hidden Markov models for cursive script recognition. In *2nd ICDAR*, pages 139–142, Tsukuba Science City, Japan, 1993.

[7] A. L. Koerich. Rejection strategies for handwritten word recognition. In *9th IWFHR*, Tokyo, Japan, 2004.

[8] 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.

[9] V. Märgner, M. Pechwitz, and H. E. Abed. ICDAR 2005 — Arabic handwriting recognition competition. In *8th ICDAR*, pages 70–74, Seoul, Korea, 2005.

[10] K. M. Sayre. Machine recognition of handwritten words: A project report. *Pattern Recognition*, 5(3):213–228, 1973.

[11] M.-P. Schambach. *Automatische Modellierung gebundener Handschrift in einem HMM-basierten Erkennungssystem.* Dissertation, Universität Ulm, 2004.

[12] M.-P. Schambach. A new view of the output from word recognition. In *9th IWFHR*, Tokyo, Japan, 2004.

[13] M.-P. Schambach. Recurrent HMMs and cursive handwriting recognition graphs. In *10th ICDAR*, Barcelona, Spain, 2009.

[14] S. N. Srihari and E. J. Kuebert. Integration of handwritten address interpretation technology into the united states postal service remote computer reader system. In *4th ICDAR*, pages 892–896, Ulm, Germany, 1997.

[15] Wikipedia. Receiver operating characteristic — Wikipedia, The Free Encyclopedia, 2010.