The snowball principle for handwritten word-image retrieval
van Oosten, Jean-Paul

DOI:
10.33612/diss.160750597

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2021

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
SUMMARY

Handwriting recognition is an active field of research, even though today our writing is mostly done digitally. There is a large number of archives that contain vast collections of handwritten documents, often in script styles that are hard to read for most people. Searching, and finding relevant pages, is a manual and tedious process.

In the field of handwriting recognition, many researchers use standardized benchmarking data sets to develop the machine learning and pattern recognition techniques and to compare their results to others. However, these datasets are usually very clean and are not comparable to the noisy quality of the historical handwritten document collections in archives and national libraries. When applying the techniques from research to such problematic collections, a number of hidden assumptions that are usually entertained by researchers in Machine Learning become apparent. Such assumptions are discussed in this thesis, along with a number of issues that were encountered in the application of machine learning techniques within a large-scale search engine for historical documents: Monk.

One of the assumptions is that the handwriting recognition process is usually considered a linear pipeline consisting of feature extraction and machine learning\(^1\). A ground truth for the data is generally presumed by researchers, and therefore not part of these pipelines. When designing a search engine, the challenging part was integrating the labelling into the process. Line-by-line editors of annotations are not ideal: A text line is not a natural object for search and starting from the first page moving down line by line does not use the full potential an integrated model has to offer. Therefore, we propose a more data-mining oriented

\(^1\) Segmentation and other pre-processing steps are outside the scope of this thesis
approach that uses a hit-list interface to gather labels per word image.

A data-mining approach to labelling images allows a human user to have a direct impact on the performance of the entire handwriting recognition process (as depicted in Figure 1.6, on page 10). There are several aspects where humans have an impact on the process: (a) On the machine learning methods, (b) on the feature engineering and (c) on the labelling. These aspects were studied in more depth in each of the chapters of this thesis. For each aspect, we looked at the issues that come up during the design of a search engine in order to answer the main question: Where can one have the most impact on the quality of the results of a search engine for historical handwritten documents: By improving the machine learning methods, the feature engineering methods or the labelling methods?

Chapter 2

Chapter based on
van Oosten, J.-P. and Schomaker, L. (Submitted). Examining common assumptions about the convergence of the Baum-Welch training algorithm for hidden Markov models. *Journal of Machine Learning Research*

In Chapter 2, we examined a number of assumptions related to the machine learning aspect of the handwriting recognition process. We were especially interested in assumptions about convergence in the training algorithm for Hidden Markov Models (HMMs), since HMMs have played such an important role in handwriting recognition. The main assumptions that were studied in this thesis are related to the fact that the Baum-Welch training method converges to a local optimum.

The first assumption is that the closer a model is to the global optimum, the better the method will perform. This was studied by generating data with a known global optimum, and training many models on this generated data. We could then measure
the distance between the models and the global optimum and measure the performance in terms of log-likelihood. Surprisingly, this experiment showed us that the \( \chi^2 \) distance to the global optimum is not a good predictor of likelihood of a trained model. One would expect that models closer to the global optimum would also have a better performance.

The other main assumption that we tested in Chapter 2, is that models that are already close to the global optimum (Baum-Welch starts by picking a random starting point and trains from there) will most likely end up close to the global optimum. However, we found that it is hard for models to converge to a point close to the global optimum without guidance.

**Chapter 3**

Chapter based on

In Chapter 3, we continued studying HMMs, but zoomed in on the essential elements of the models. We were mostly interested in the relation between the state-transition probabilities, that model the temporal structure of the data, and the observation probabilities, that model the feature representation of the data per state. The main assumption that is tested in this chapter is that the temporal structure is as important as the feature representation.

We studied the relation between the two parts by generating data with a particular temporal structure and trained a model on this data. This structure should be present in the state-transition probabilities in a trained model. However, the experiments in Chapter 3 showed that there is no clear indication that the original structure could be found. Another experiment that was performed in Chapter 3 is removing the temporal relation between
states and observe whether or not classification performance dropped in these models. Surprisingly, the performance did not drop as drastically as expected.

The main conclusion that we can draw from these experiments is that the observation probabilities seem to have a larger impact on the model performance than the transition probabilities. This means, related to the general research question, that special attention should be given to the feature representation.

Chapter 4

Chapter based on

Finally, in Chapter 4 we turned our attention to the labelling part of the handwriting recognition process. We consider it an essential part of the process and explicitly integrate it into a continuous loop. The hit-list interface is introduced in this chapter and it helps gather a large amount of training data by achieving a snow-ball effect (i.e., an initially small number of labels can accumulate more and more labels over time). A hit-list is constructed by classifying words into the different lists and then ranking each list. We found that one cannot assume that a good recognizer will also be good at ranking.

Related to labelling, we found that it is important to consider the construction of your dataset as part of the process and integrate human annotators in a continuous learning cycle. An implication from Chapter 4 is that one should alternate between classification and ranking and use different methods that are optimized for each subtask. Also, the specific classification and ranking methods should not be considered as fixed. It is necessary to alternate between these, e.g., if the current method does not yield enough new labels anymore to keep the momentum going. The final conclusion from this is that the handwriting recognition process
is not a static process, a single training event, but needs constant maintenance.

**Discussion**

To conclude, this thesis discusses human involvement in the handwriting process from three different angles: In the design of machine learning methods, design of feature extraction methods and representations, and labelling. Chapters 2 and 3 mainly deal with assumptions around the machine learning and feature extraction methods, while the main concern in Chapter 4 is about how to deal with a changing dataset, especially with continuous additions of labels.

The main method of examining the assumptions in the use of HMMs is the generation of data from known models, and to study what happens in the models during training. Taking a global perspective to study what happens in local (gradient descent) processes is a method that can be used to study other machine learning methods as well, such as neural networks. If the models themselves can be used to generate data, it is relatively easy to compare the trained model versus the global optimum.

The bigger theme in the thesis is about the idea that we need to regard the handwriting recognition process as a dynamic process. In the Monk system, this is expressed in a flexible hit list interface. While the hit list method is only applied to handwritten words in this thesis, we believe that this form of active learning is relevant for machine learning in general. Ideally, the beneficial effect of a labelling action is experienced as soon as possible by the user. This creates a snow-ball effect in the feedback loop and leads to broadly labelled datasets. Another benefit of the hit list interface is that it allows exploration. Using different classification and ranking methods is useful when the addition of labels is stagnating.

It is crucial for all machine learning methods to have proper, labelled data. Therefore, the framework described in this thesis is relevant for all applications of machine learning across fields
and industries. The advice in this thesis is therefore to invest in a system of getting more and better labels and to incorporate this framework into any application of machine learning.