Training, Adaptation, and Semi-Supervised Learning in a Real-World OCR System

Thomas M. Breuel
DFKI & U. Kaiserslautern
Kaiserslautern, Germany

February 19, 2009

OCR and handwriting recognition have been key applications and drivers for research in machine learning and pattern recognition. Printed OCR has a reputation for being an "easy" or "solved" problem because, compared to many other classification problems, it is comparatively easy to reach nominally low error rates for many printed OCR tasks when classifiers are presented with isolated characters from a finite character set extracted from cleanly scanned documents. For unsegmented text lines, degraded documents, and handwritten input, the problem is already considerably harder. In addition to the text recognition itself, OCR also involves a number of other tasks that have a strong, and often dominant, influence on overall error rates of the OCR system as a whole. These tasks include page orientation detection, noise removal, 2D page segmentation (into text, images, text blocks), logical layout analysis, and language modeling.

OCRopus (`www.ocropus.org`) is an open-source OCR system that has the goal of replacing many of the heuristic or rule-based components of "traditional" OCR systems with adaptive and machine learning algorithms, as well as to deliver state-of-the-art OCR performance. This presentation is about work towards the upcoming OCRopus 0.5 release, and preliminary work towards the 1.0 release at the end of 2009. We are also designing OCRopus to make it easy for researchers in machine learning to benchmark and compare character classifiers and integrated segmentation-and-recognition methods.

The goal of the presentation is to inform about the state of the art of OCR systems in a number of areas relevant to machine learning, to point out open and challenging problems for machine learning theory, and provide information about the availability of tools and data sets for evaluating machine learning algorithms in the context of a large real-world learning task.

Some of the specific points, challenges, and solutions I will address are:

- **Layout analysis and page segmentation** often dominates the error rate of real-world recognition systems, but attempts to create high performance, trainable layout analysis methods have had limited success. We have developed a statistically justifiable method for layout analysis called *structural mixture models* [4] that is trainable, fast, yields competitive performance on layout analysis tasks, and is generalizable to other 2D segmentation problems.

- **Automatic parameter selection and training** is essential for real-world applications of OCR; OCRopus provides a novel algorithm for parameter selection for arbitrary classifiers, and permits *total benchmarking* of classifiers, taking into account parameter selection costs.

- **Adaptation** to variations in input resolution, geometry, and imaging parameters, strongly influence OCR system performance. Most OCR systems use non-adaptive ad-hoc methods for such adaptation. OCRopus provides unsupervised, trainable adaptation based on generative geometric models and language models [3].

- **Semi-Supervised Learning** is important for OCR and handwriting recognition systems and has been used extensively in the past (e.g., [1]). Although usually considered separately, **style modeling**, document-level adaptation, and multi-task learning are closely related techniques. OCRopus is routinely trained using semi-supervised learning, using a sequence of EM-steps, alternating between classification, language modeling, stochastic gradient descent, and clustering (equivalent to hierarchical Bayesian methods, [2]).

- **Very Large Datasets** Because of the availability of semi-supervised learning techniques, we can construct (and, if necessary, manually correct) very large training and test sets. We are preparing the release of a new standard OCR database in MNIST format consisting 60 million training samples and 10 million test samples, including style information and rare classes.

- **Parallel and Distributed Training** become important because of the large training and test sets used in OCR. OCRopus takes advantage of multicore CPUs during training, classification, and language modeling. Furthermore, it implements distributed training based on ensenbles; unlike other ensemble methods, in our approach, training set splitting is not under the control of the ensemble learning algorithm, permitting users to contribute independently trained models into a common pool that can then be combined into a "super-classifier".

References

[1] Thomas M. Breuel. Design and implementation of a system for the off-line recognition of handwritten responses on us census forms. In *Proceedings of the IAPR Workshop on Document Analysis Systems*, 1994.

[2] Charles Mathis and Thomas M. Breuel. Classification using a hierarchical bayesian approach. In *Proc. International Conference on Pattern Recognition (ICPR)*, 2002.

[3] Yves Rangoni, Faisal Shafait, and Thomas M. Breuel. Trainable orientation detection for multiple scripts using character similarity. Submitted to ICDAR 2009.

[4] Faisal Shafait, Joost van Beusekom, Daniel Keysers, and Thomas M. Breuel. Structural mixtures for statistical layout analysis. In *Proc. 8th Int. Workshop on Document Analysis Systems (DAS)*, 2008.