Many-Author Offline Handwriting Recognition Using a Warping-Based Approach

Douglas J. Kennard, William A. Barrett, and Thomas W. Sederberg
Department of Computer Science
Brigham Young University
Provo, Utah
{kennard,barrett,tom@cs.byu.edu}

Abstract

Optical Character Recognition (OCR) software allows computers to automatically convert scanned pages of typed or machine-printed text into searchable digital formats for use by humans. However, automatically transcribing or indexing materials that are handwritten (instead of machine-printed) is a much more difficult problem that is still not completely solved. Solving this problem will be of enormous benefit to family history research, historical data analysis, government digitization and archival projects, and many other application areas. We previously published a novel approach to handwriting recognition (HR) based on 2D geometric warping of word images. The approach showed encouraging initial results on two small, single-author datasets (1000 training examples and 1000 test examples each). In-vocabulary word recognition accuracy for those datasets was nearly 90%.

In this paper, we describe recent progress and improvements to our method, and we show results on a much larger, many-author dataset that is publicly available in the HR research community. We also compare our results to those of an existing state-of-the-art Hidden Markov Model (HMM) approach to HR. We show that the accuracy of our approach is either better than or comparable to the accuracy of the existing method in several comparisons.

1. Introduction

There are currently many governments, libraries, archives, companies, and other organizations undertaking massive digitization projects to convert books, documents, and microfilm into accessible digital formats. After scanning these materials into digital images, they must still be transcribed or indexed to make them truly searchable. For many books and other materials that are machine-printed, Optical Character Recognition (OCR) software can be used to automatically transcribe and index them (assuming the quality of the machine print is good enough). Examples of such efforts for machine-printed text include the book search efforts of Google and similar efforts by Internet Archive (http://archive.org).

Automatically indexing handwriting is a much more difficult task than OCR and has not yet been completely solved. As such, enormous amounts of time, money, and effort are spent to manually index or transcribe handwritten materials. Even so, it is really only practical to index a small fraction of the available materials within a reasonable time frame. A large portion of the available handwritten records will remain unindexed until automatic indexing methods are good enough to augment the manual indexing efforts.

We previously presented a warping-based approach to automatic handwriting recognition (HR) [1]. Our early results for single-author datasets were encouraging. We found that on two small datasets (the Washington and Smith datasets, with 1000 training images and 1000 test images each) our method achieved in-vocabulary word accuracy of 88.9% (Washington) and 89.38% (Smith).[1] The total accuracy — including errors caused by having no training example of a test word (i.e., including out-of-vocabulary test words) — on those two datasets was 66.5% (Washington) and 70.7% (Smith).

2. Improvements

Since our initial results, we have made a few minor improvements to our original method. Some of those improvements include:

- Better handling of the edges of images and distance maps
- Minor improvements to our word difference metric
- Better selection of algorithm parameters

Additional details of the implementation and parameter improvements will be available in [2].

The improvements to our algorithm result in slightly higher recognition rates. In-vocabulary accuracies are now 91.58% (Washington) and 90.64% (Smith). Total accuracies are 68.5% (Washington) and 71.7% (Smith). Since the test sets of these two datasets are used during development and parameter tuning (instead of using an independent validation and test set), they represent an upper-bound on the recognition accuracies our current method achieves for these two datasets.

3. Experiments and Results

In this paper, we expand our analysis to include results from a large dataset that has many authors. The dataset we use is the IAM database (IAMDB), made publicly available by Marti and Bunke at the University of Bern [3]. The database is split into a training set, two validation sets, and a test set, according to the pre-defined "Large Writer Independent Text Line Recognition Task." Each split of data is mutually exclusive and uses different authors than the other splits.

We create our own small parameter-tuning training and test splits from a subset of words from the training and validation sets. (We do not use any words from the test set for our tuning dataset.) We use our small tuning set to tune our parameters (vary each parameter over a reasonable range of values and select the value that maximizes recognition accuracy for the tuning test set). The tuned parameters are: $p = 0.28$, Sakoe-Chiba band width constraint = 15, and mesh size ratio denominator = 3.8. Using these tuned parameters, we test our accuracy for each of the two validation sets and the test set.

We report the accuracy of our method for each split of data in Table 1. For the two validation sets, total accuracy of our method is 55.3% and 56.88%, and in-vocabulary accuracy of our method is 67.14% and 67.45%. On the test set, our method achieves total accuracy of 55.53% and in-vocabulary accuracy of 65.93%. It is important to note that these numbers are raw recognition accuracies and do not reflect any improvements in accuracy that can be gained from word n-grams or other language models, multiple classifiers, or postprocessing.

---

1 The numbers reported here are slightly higher than the previously published results due to some corrections in the dataset ground truth and other minor corrections.

Table 1: Recognition Accuracy on IAMDB Dataset (large, many writers)

IAMDB Dataset	# writers	# words	Total Accuracy (# correct/# possible)	In-Vocab Accuracy (# correct/# possible)
Train	283	46,947	—	—
Validation 1	46	7,894	55.30% (4,365/7,894)	67.14% (4,365/6,501)
Validation 2	43	8,556	56.88% (4,867/8,556)	67.45% (4,867/7,216)
Test	128	17,584	55.53% (9,764/17,584)	65.93% (9,764/14,809)

4. Comparisons to HMM methods

Over the past several years, most of the large improvements in HR accuracy have not been due to completely new recognition methods or major improvements in existing recognition methods, themselves. Most of the improvements have come through incorporating language models (such as word n-grams to provide word context), combining multiple recognition methods into one system, or other such strategies that often have little to do with the core recognition portion of HR systems. In fact, most current HR systems use Hidden Markov Model (HMM) approaches that have been around for many years as the core recognition portion.

Our method would also benefit from incorporating n-gram language models, combining it with other methods, etc. However, we have not yet implemented those portions of our HR system. As such, we compare the raw accuracy of our method with the raw recognition accuracies (before language models and multiple classifiers are added) of some existing HMM-based systems reported in the literature. Some of the methods build off of others, incorporating minor improvements. The comparisons are not direct comparisons because of some differences in versions and splits of the IAMDB data, as well as some minor differences in how accuracy is measured. However, by making reasonable assumptions of consistency, we are able to estimate the raw accuracies of these HMM-based systems to indirectly compare them with our method. We explain our assumptions and our estimate calculations in great detail in [2], but do not include them here due to space constraints.

We estimate that our total accuracy of 55.53% (test set) is approximately equivalent to a slightly higher reported accuracy of 56.75% when using the metric reported by some of the papers we use for comparison. Marti and Bunke [4] report recognition accuracies ranging from 40.47% to 51.44%, depending on vocabulary size. Our method is clearly more accurate. Vinciarelli, Bengio, and Bunke [5] report accuracies from about 29% to 35%. Even with trigrams, their best accuracy is reported as about 46.5% — still much lower than ours. Depending on which assumptions we make, we estimate the accuracy of Zimmermann, Chappelier, and Bunke [6] to be anywhere from as low as 51.6% to as high as 58.48%. Since several assumptions must be made for this comparison, we cannot be completely confident that our method is more accurate than their base HMM method. However, it appears that our accuracy is at least close to theirs, if not better. Finally, we estimate that Bertolami and Bunke [7] would have accuracies from 39.28% to 46.16% if language models are ignored (they report accuracies as high as 67.17% when language models are used). Again, several assumptions have to be made for this comparison because they do not report any results without language models. As such, we cannot be completely sure that our estimates are correct. However, it appears that our method is significantly more accurate on this dataset when language models are discounted.

All of the comparisons in the previous paragraph are done on raw recognition accuracies (without using language models, multiple classifier/ensemble approaches, etc.). We anticipate that adding language models to our method and using multiple classifiers will increase our accuracy just as it has increased the accuracy of HMM-based methods. As such, we believe that a complete recognition system using our method as the core recognition portion of the system will be more accurate than (or at least nearly as accurate as) the state-of-the-art HMM methods reported in the literature.

5. Conclusions

In this paper, we have mentioned minor improvements to our novel handwriting recognition algorithm and provided a reference for more details [2]. We have reported experimental results for our method on a large, many-writer dataset. Our in-vocabulary accuracies are above 65% and total accuracies are above 55% for all splits of the data. We have compared our results to those of several papers that use HMM-based recognition methods. We used estimates where necessary to compare how those systems would perform without the benefit of language models and multiple classifiers. We found that our method compares favorably, achieving accuracies higher than (or at least close to) the other methods. Since our method compares favorably, we believe that it will also compare favorably once we take advantage of language models and multiple recognition methods like the HMM-based methods do.

References

[1] Douglas J. Kennard, William A. Barrett, and Thomas W. Sederberg. Word Warping for Offline Handwriting Recogntion. In *International Conference on Document Analysis and Recognition (ICDAR)*, pages 1349–1353, Beijing, China, Sep. 2011.

[2] Douglas J. Kennard, "Warping-Based Approach to Offline Handwriting Recognition," Ph.D. dissertation, Department of Computer Science, Brigham Young University, Provo, UT, 2013. Electronic version will be archived at http://etd.lib.byu.edu/

[3] Urs-Victor Marti and Horst Bunke. The IAM-database: an English Sentence Database of Offline Handwriting Recognition. *International Journal on Document Analysis and Recognition (IJDAR)*, 5(1):39–46, Jul. 2002.

[4] Urs-Viktor Marti and Horst Bunke. Using a Statistical Language Model to Improve the Performance of an HMM-based Cursive Handwriting Recognition System. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 15(1):65–90, 2001.

[5] Alessandro Vinciarelli, Samy Bengio, and Horst Bunke. Offline recognition of unconstrained handwritten texts using HMMs and statistical language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(6):709–720, Jun. 2004.

[6] Matthias Zimmermann, Jean-Cédric Chappelier, and Horst Bunke. Offline Grammar-based Recognition of Handrwitten Sentences. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(5):818–821, May 2006.

[7] Roman Bertolami and Horst Bunke. Hidden Markov Model-based Ensemble Methods for Offline Handwritten Text Line Recognition. *Pattern Recognition (PR)*, 41(11):3452–3460, 2008.