ZELIG: An Unconstrained Offline Handwriting Identifier

Mark O'Meara
Stanford University
omeara13@stanford.edu

Daniel Koning
Stanford University
dkoning@stanford.edu

Abstract

For all the human qualities involved, the problem of handwriting classification — matching written samples to their writers — is a surprisingly natural fit for machine learning techniques. The traits that most strongly characterize a writer's unique style are difficult for an untrained human observer to pick out, but easy for a computer to glean from a scan. We present a classifier, ZELIG[1]*, that learns to recognize handwriting style from a set of tagged samples. Our initial selection of features was inspired by an earlier model [Hertel 2003] that achieved a never-before-seen rate of 90% accuracy on a large set of authors. By including a set of features original to our project, we improve upon that precursor by a significant margin.*

Introduction

Handwriting classification is a problem of great interest to biometric analysts, for reasons both theoretical and practical. A person's handwriting is an extremely convenient identifying characteristic: easy to collect, yet difficult for another individual to imitate, and always varying from person to person in ways both obvious and subtle. Historically, signatures have often been used as the standard legal confirmation of identity for exactly these reasons.

There are also strong practical incentives to develop handwriting recognition software. The most obvious (and likely the most lucrative) would be to automate the process of fraud detection by allowing receipt signatures to be automatically authenticated. But there are other possible applications that demand an approach that deals with the whole text. Forensic analysis of Many databases of historical texts are prepared from unorganized handwritten scans by many authors, and automated recognition would greatly expedite the process of sorting the material. The same process could even be applied for original historical work, providing strong supporting evidence for the authorship of an anonymous text. [4]

We conceived of our project by analogy to other . As a "starting point", we

Resources

While we initially considered collecting and preprocessing our own samples, this would have greatly limited the number of authors our classifier would have had to distinguish; the problem would have been correspondingly less interesting. Instead, we made use of the substantial IAM Handwriting Database [6], which offers a collection of some 1,539 pages of text by 657 writers.

Methods

Feature selection

Figure 1: *Cursive vs. print analysis. The first style, with one contiguous component per word, is purely cursive. The second has as many components as letters and is purely print. The third is in an intermediate style.*

[1] "ZELIG Efficiently Learns to Identify Graphology." Inspired by the 1983 film *Zelig*, featuring a human chameleon and habitual forger as its protagonist, whose theatrical poster consists of the title written in a large variety of styles.

Figure 2: *When we draw diagonal lines at angle θ across this sample, the density of the pixels intersected by the lines is maximized when θ is about 60° from the x-axis. This is the vertical slant of the lettering.*

IAM provides a comprehensive set of pre-processed, tagged metadata associated with its sample images. Some of these metadata have an obvious relation to the problem. Of the included features that reflected individual style, we identified several as particularly relevant to our analysis. Ultimately, after performing ablative analysis (q.v.), we continued to use some features and eliminated other, less helpful features.

Feature	IAM variable	Description	Used?
Fractal dimensions	`fd0, fd1, fd2`	The "jaggedness" of the text, as measured by sampling it at progressively smaller resolutions and observing the change in pixel density. The parameters `fd0, fd1, fd2` characterize the curve produced. [1]	✓
Slant	`slant`	Angle between baseline and average vertical pen stroke	✓
Stroke width	`stroke-width`	Pixel width of average pen stroke	✓
Ascender and descender slopes	`ass, dss`	Angles of extenders for letters like lowercase *f* and *y*	✗
Baseline slopes	`ubs, lbs`	Rise or fall of the baseline	✗

In addition, we derived the following novel measures using included features as primitives:

Feature	Description	Used?
Letter height	Average height of the main body of a letter; measured by distance from upper to lower baseline	✓
Word gap	Average horizontal whitespace between words	✓
Character width	Width of the average single character	✓
Fragmentation	Average number of connected components per word	✓

The first three of these are very simple measures that bear obviously on
Generally, the

Training

After comparing several alternatives (including softmax regression and SVMs), we chose a modified naive Bayes classfier as our program's technique. Naive Bayes, in its simplest form, demands discrete distributions and all the attributes we could use were continuous-valued. The underlying probability model justified this approach. Each of our features is independent of most others. And each is the physical outcome of a physiological process, and one which depends on the interplay of several bodily systems — therefore, the density curve of its values likely approximates the Gaussian distribution.

Culling features

We performed ablative analysis in `MATLAB` to

Results

Let's take a look...

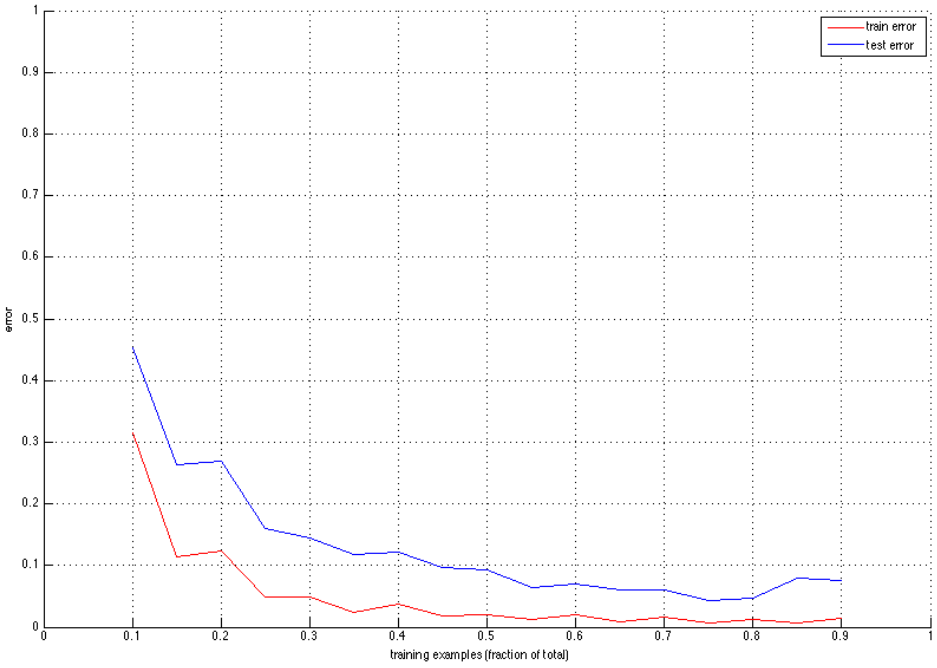


Figure 3: *Learning curve for guessing 5 close matches for each of 100 authors*

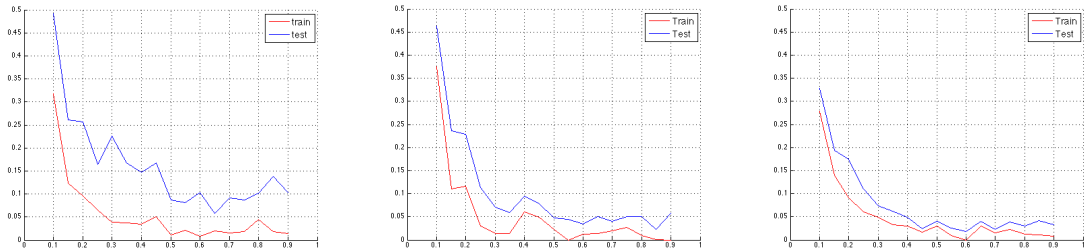
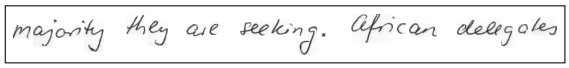
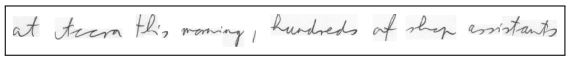


Figure 4: *Learning curves for just 20 authors, guessing 1 match, 3 matches, and 5 matches.*
etc. etc. etc.

Discussion

(1) 

(2) 

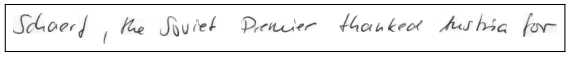
(3) 

Table 1: *Samples of three styles found by our model to be very similar.*

Hertel [4], the paper that provided the foundation for our project, demonstrated an accuracy of
Despite this substantial improvement from

References

[1] Bouletreau, V., Vincent, N., Sabourin, R. and Emptoz, H. (1997). Synthetic parameters for handwriting classification. *Proceedings of the 4th International Conference on Document Analysis and Recognition*, 102–106.

[2] Cheriet, M., et al. (2009). Handwriting recognition research: Twenty years of achievement . . . and beyond. *Pattern Recognition*, 42:3131–3135.

[3] Epshtein, B., Ofek, E. and Wexler, Y. (2010). Detecting text in natural scenes with stroke width transform. *Computer Vision and Pattern Recognition*, 2963-2970.

[4] Hertel, C. and Bunke, H. (2003). A set of novel features for writer identification. *Audio- and Video-Based Biometric Person Authentication*, 679–687.

[5] Huber, R. A. and Headrick, A. M. (1999). *Handwriting Identification: Facts and Fundamentals.*

[6] Marti, U. V. and Bunke, H. (2002). The IAM-database: an English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5:39–46.

[7] Sahu, V. L. and Kubde, B. (2013). Offline Handwritten Character Recognition Techniques using Neural Network: A Review. *International Journal of Science and Research (IJSR), India*, 2:87–94.