Predicting the scribe behind a page of medieval handwriting

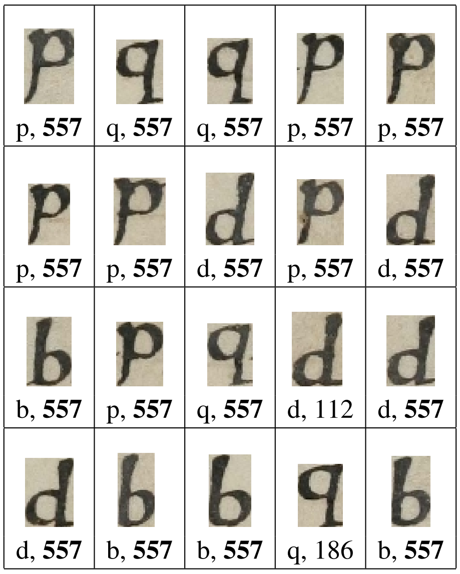
Mats Dahllöf

Dept of Linguistics and Philology and Dept of Information Technology, Uppsala University

mats.dahllof@lingfil.uu.se

Abstract

This paper addresses the issue of attributing pieces of medieval handwriting to scribes known from other examples of writing. The system is applied to manuscript page images and performs extraction and comparison of letter shapes. Letters and sequences of connected letters are identified by means of connected component labeling. This is followed by further splitting into letter-size pieces. The prediction process makes use of a dataset with instances of four letter types (*b*, *d*, *p*, and *q*), taken from manuscript pages with known scribes. Nearest neighbor classification is used for letter-level prediction of scribe (and grapheme). The image features capture the distribution of foreground, as it appears after a binarization step. Cosine similarity is used as the similarity metric. The system predicts the scribe behind a page by means of a voting procedure taking the highest-scoring letter-level hits for a page as its input. Evaluated on codicological units from five different scribes the system reached an accuracy above 99% for four of them and 87% for the fifth one.



p, **557**	q, **557**	q, **557**	p, **557**	p, **557**
p, **557**	p, **557**	d, **557**	p, **557**	d, **557**
b, **557**	p, **557**	q, **557**	d, 112	d, **557**
d, **557**	b, **557**	b, **557**	q, 186	b, **557**

Table 1: An example of the top 20 letter hits for a manuscript page (Cod. Sang. 557, p. 2). Scribe prediction gives two errors.

1. Introduction

The purpose of the work reported here is to identify which scribe (among a set of previously seen scribes) has produced a piece of handwriting using a procedure relying on a set of letter instances, manually cut out, for each scribe. It has also been reported in (Dahllöf, 2014).

The analysis of an unseen page (in the form of a high-quality jpeg image) starts with image binarization using the Otsu algorithm (Otsu, 1979), i.e. separation of foreground (ink) and background (parchment). Connected components of foreground, presumed to form letters and letter sequences are identified, and these components are split into pieces presumed to correspond to single letters. Vertical cuts are proposed where the horizontal pixel projection profile is thinnest.

The system predicts which scribal hand has produced a proposed letter instance by means of a nearest neighbor classification procedure using the dataset of cut-out letters. (Character prediction is a "by-product" here.)

The system was developed for and evaluated on early medieval Caroline minuscule manuscripts. The four characters *b*, *d*, *p*, and *q* were assumed to be useful here, as they have a tendency to occur unconnected to other letters. Only instances of these were used. This dataset contained a total of 436 items, 59–108 letter images for each hand, and 9–33 instances of each hand-character combination.

The hand behind a codex page is predicted by a voting procedure operating on the 29 toplisted letter-level proposals generated for that page. These proposals are ranked by the similarity score for the nearest neighbor hit. See Table 1 for an example. A Java implementation was used to develop and evaluate the method.

2. Feature model and similarity metric

Each letter image is represented by features, computed with reference to the minimal bounding box enclosing the foreground pixels. They characterize the image in the following size- and scale-invariant terms (where a and b are two constants, and w, h, f, are the width, height, and number of foreground pixels, respectively.):

- **Distribution of foreground pixels** as captured by a grid of $a \times b$ subrectangles over the bounding box, with the total number of foreground pixels as divisor: $r_n = f_n/f$, where f_n is the number of foreground pixels in the subrectangle n, $1 \leq n \leq ab$. See Figure 1.
- **Bounding box proportions:** $p = w/(w + h)$.
- **Foreground density:** $d = f/wh$.

Cosine similarity (Jones and Furnas, 1987) is used as the similarity metric. The features are given equal weight.

Different choices for a and b were evaluated during the development stage using the letter dataset. The accuracy of the letter classifier (based on the feature scheme and the similarity metric) with respect to hand and character prediction can be computed by applying it in a leave-one-out manner, i.e. by predicting the type of each letter instance

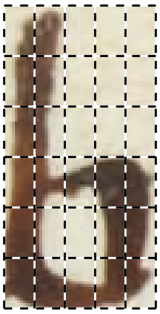


Figure 1: The grid of $a \times b$ rectangles corresponding to the features (here $5 \times 6 = 30$) used to capture the distribution of foreground (ink) in the bounding boxes (enclosing letters). From (Dahllöf, 2014) © 2014 IEEE.

Scribe/unit	Corr.	Incorr.	Accuracy	Dev.
CS 112, 1–322	287	2	99.3%	33
CS 186, 3–146	112	16	87.5%	15
CS 557, 2–274	241	0	100.0%	29
CS 562, 3–93	79	0	100.0%	12
CS 565, 3–222	196	0	100.0%	24
All five	915	18	98.1%	113

Table 2: The performance of the hand prediction system applied to single pages of the five codicological units/hands, evaluated on the unseen subset of the data. Number of pages correctly and incorrectly classified, accuracy, and number of pages seen in the development process. CS – Codex Sangallensis. From (Dahllöf, 2014) © 2014 IEEE.

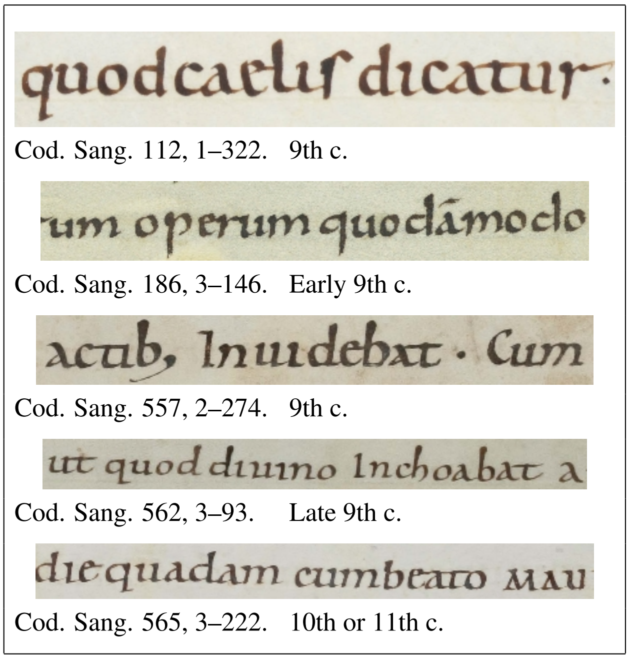
by comparing it to every other letter in the letter dataset (containing 436 items).

The choices for $a \times b$ were thus optimized with regard to hand prediction to 5×6 for feature models using only $(r_1, \ldots, r_{ab})$. The addition of the p and d features jointly further improved the performance, thus justifying the feature model $(r_1, \ldots, r_{30}, p, d)$, which was used in the further experiments.

This model yields an accuracy rate of 92.4% (403 correct) for hand prediction and 99.8% (435 correct) for character (*b*, *d*, *p*, or *q*) prediction, when evaluated in the leave-one-out manner on the manually produced letter dataset.

3. Evaluation

The method proposed here was evaluated on five well-preserved manuscripts written in the style known as the Caroline (or Carolingian) minuscule, see Table 3. The books were all written at the Abbey of St. Gall (Switzerland) and still reside there. They are found in high-quality digitized form at www.e-codices.unifr.ch. The scribe prediction method was evaluated on unseen pages from five different scribes and reached an accuracy above 99% for four of them and 87% for a fifth significantly more difficult one. About a tenth of the pages had been used for parameter refinement during the development process. Of these, 1 or 2 provided letters for the letter dataset. Table 2 gives an overview of the performance.



Cod. Sang. 112, 1–322. 9th c.

Cod. Sang. 186, 3–146. Early 9th c.

Cod. Sang. 557, 2–274. 9th c.

Cod. Sang. 562, 3–93. Late 9th c.

Cod. Sang. 565, 3–222. 10th or 11th c.

Table 3: Text examples from the five codicological units (each being the work of one scribe) providing data for the present study. St. Gallen, Stiftsbibliothek. Digitized at www.e-codices.unifr.ch.

4. Discussion

According to an overview of nine approaches (Brink et al., 2012) hand identification systems achieve accuracy rates between 62% and 97% for modern handwriting, when several hundred hands are included in the data. The authors add that “the numbers cannot be well compared because of differences in dataset material, required level of human interference, and number of writers”. The results reported for medieval manuscript data amount to lower accuracy rates for smaller sets of scribes, suggesting that hand prediction for medieval handwriting is a more difficult task. From this perspective, the present results are promising, even if they leave open the question of how well the present method would perform with larger sets of scribes, or on other styles of writing.

Seen from the perspective of language technology, images are challenging in the sense that they represent two-dimensional continuous signals. As can be expected, segmentation of an image into letters or other useful units is one of the most difficult steps in the pipeline of handwriting analysis. The present system uses a simple letter segmentation method. Its good performance is due to the clearly separated lines and good contrast between ink and parchment which are typical of the Caroline manuscripts studied here.

A weakness of the present kind of approach is that it requires a manually produced dataset of letters. This suggests that a letter-based approach for hand classification should be combined with a procedure for automatic extraction of reference letters. One way of doing this could be to use some clustering method for finding typical and recurring image elements.

Manuscripts

Manuscripts from *e-codices – Virtual Manuscript Library of Switzerland*: "The use of our digitally reproduced images is regulated by a Creative Commons License Creative Commons License. This permits the use of individual images for non-commercial purposes (particularly in the areas of education and research), as long as proper source citation is used."

St. Gallen, Stiftsbibliothek, Cod. Sang. 112, 3–146.
Hieronymus, Commentarii in Esaiam, libri I–V.
http://www.e-codices.unifr.ch/en/list/one/csg/0112

St. Gallen, Stiftsbibliothek, Cod. Sang. 186, 1–322.
Prosper de activa et contemplativa vita libri III.
http://www.e-codices.unifr.ch/en/list/one/csg/0186

St. Gallen, Stiftsbibliothek, Cod. Sang. 557, 2–274.
Vita sancti Martini, Dialogi de orientalibus patribus.
http://www.e-codices.unifr.ch/en/list/one/csg/0557

St. Gallen, Stiftsbibliothek, Cod. Sang. 562, 3-93.
Vitae sancti Galli et Otmari.
http://www.e-codices.unifr.ch/en/list/one/csg/0562

St. Gallen, Stiftsbibliothek, Cod. Sang. 565, 3-222.
Lives of the Benedictine Saints.
http://www.e-codices.unifr.ch/en/list/one/csg/0565

References

A. A. Brink, J. Smit, M. L. Bulacu, and L. R. B. Schomaker. 2012. Writer identification using directional ink-trace width measurements. *Pattern Recognition* 45, 162–171.

M. Dahllöf. 2014. Scribe Attribution for Early Medieval Handwriting by Means of Letter Extraction and Classification and a Voting Procedure for Larger Pieces. *22nd International Conference on Pattern Recognition, August 24–28, 2014, Stockholm Waterfront, Stockholm, Sweden.*

W. P. Jones and G. W. Furnas. 1987. Pictures of Relevance: A Geometric Analysis of Similarity Measures. *Journal of the American Society for Information Science and Technology* 38(6), 420–442.

N. Otsu. 1979. A threshold selection method from gray-level histograms, *IEEE Transactions on Systems, Man and Cybernetics* 9(1), 62–66.

Note