Automatic Writer Identification in Medieval Papal Charters

By **Vincent Christlein** and **Elli Angelopoulou**

1. Introduction

Knowing the authors of handwritten texts can give new insights of life in the past. Questions such as: "Can we find out who is the author of a given handwritten section, cf. Figure 1 (top blue rectangle)?", (Fig. 1) "or can we relate other handwritings to specific individuals?" are typical examples of the issues tackled by the problem known as (offline[1]) writer[2] identification. Writer identification is the task of finding the correct author of a handwritten text of unknown authorship, based on a known set of handwritings with known authorship. In contrast, deciding if a handwriting is written by a specific scribe is called writer verification. An example of the latter would be the question whether the author of the context is the same as the author of the datum line or not, cf. blue rectangles of Figure 1 (Fig. 1).

2. Overview of Writer Identification Methods

Automatic writer identification makes use of visual features in the handwriting image, i. e., a linguistic analysis is not used. Writer identification (verification) methods can be grouped in two categories[3]: textural-based methods and allograph-based methods.

2.1. Textural-based methods

Textural-based methods measure statistics for the whole handwritten sample such as the angle distribution, or the width-height relation. For example, one could compute special attributes at the edges of the ink like the Hinge feature[4], or the Quill feature[5]. For the Hinge features the angles between the legs in each point are measured, cf. angles α and β in Figure 2. (Fig. 2) The Quill feature measures the width of the ink and the associated angle at each contour point, i. e., angle α and width w in Figure 2. (Fig. 2) It has been shown that such feature distributions can be very good indications for a handwriting[6].

---

[1] In offline writer identification only the handwriting images can be used. Unlike online writer identification, for which temporal information of the handwriting formation can be used, too, e. g., due to writing on a touchpad.

[2] Note: "writer" and "scribe" are used interchangeably throughout this paper.

[3] Marius BULACU / Lambert SCHOMAKER, "Text-Independent Writer Identification and Verification Using Textural and Allographic Features", in: Pattern Analysis and Machine Intelligence, IEEE Transactions on 29.4 (Apr. 2007), pp. 701–17. ISSN : 0162-8828.

[4] Ibid.

[5] A. A. BRINK et al., "Writer Identification Using Directional Ink-Trace Width Measurements", in: Pattern Recognition 45.1 (Jan. 2012), pp. 162–171. ISSN : 00313203.

[6] BULACU / SCHOMAKER, "Text-Independent Writer Identification and Verification Using Textural and Allographic Features"; BRINK et al., "Writer Identification Using Directional Ink-Trace Width Measurements".

2.2. Allograph-based methods

In allograph-based methods, local descriptors are computed at parts of letters (allographs). They are related to a background model of how such local descriptors occur in documents. The frequency of the occurrences of these descriptors in a text sample are used to describe a writer. More specifically, a global image descriptor is computed by encoding local descriptors. The encoding step is achieved by relating a universal background model[7] to the local descriptors. See Figure 3 (Fig. 3) for a schematic illustration of the encoding process. This process is also known as bag of (visual) words (BoW).

Local feature descriptors can be computed at different locations, cf. Figure 4. (Fig. 4) Typical allograph-based approaches compute descriptors at keypoints, e. g., SIFT keypoints as depicted in Figure 4a. (Fig. 4) However, they can also be used at different letter parts, such as sections computed at vertical cuts or by a connected component analysis as visualized in Figures 4b and 4c, (Fig. 4) respectively. Local descriptors often stem from the field of computer vision. For example, the use of gradient based descriptors like SIFT or SURF is quite common[8]. The allograph-based methods can also differ with respect to the encoding process. A variety of techniques can be employed, ranging from the simplest form of computing zeroth order statistics, which relates to vector quantization[9], to higher order statistics[10].

Allograph-based methods often generate very high dimensional global descriptors, which are difficult to visualize. In juxtaposition, the advantage of textural-based methods lies in their intuitive interpretation. However, the best methods in terms of accuracy stem from the allograph group. Current such methods achieve very high writer identification accuracy on contemporary datasets. For example on the well known ICDAR 2013 benchmark set, our method using GMM supervectors[11] achieves more than 97% TOP-1 accuracy, i. e., the scribe of the document, which is ranked as the most similar one, is with 97% probability also the scribe of the questioned document.

3. Writer Verification of Datum Lines

In this section different approaches for the verification of historical text documents are evaluated. Our dataset consists of 127 datum lines extracted from high medieval papal

---

[7] Also known as “vocabulary”, or “dictionary”.

[8] Vincent CHRISTLEIN et al., “Writer Identification and Verification Using GMM Supervectors”, in: Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on. Mar. 2014, pp. 998–1005; Stefan FIEL / Robert SABLATNIG, “Writer Identification and Writer Retrieval Using the Fisher Vector on Visual Vocabularies”, in: Document Analysis and Recognition (ICDAR), 2013 12th International Conference on. Washington DC, NY, Aug. 2013, pp. 545–549; Rajiv JAIN / David DOERMANN, “Combining Local Features for Offline Writer Identification”, in: Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on. Heraklion, Sept. 2014, pp. 583–588.

[9] BULACU / SCHOMAKER, “Text-Independent Writer Identification and Verification Using Textural and Allographic Features”; Rajiv JAIN / David DOERMANN, “Writer Identification Using an Alphabet of Contour Gradient Descriptors”, in: Document Analysis and Recognition (ICDAR), International Conference on. Buffalo, Aug. 2013, pp. 550–554. ISBN : 978-0-7695-4999-6.

[10] CHRISTLEIN et al., “Writer Identification and Verification Using GMM Supervectors”; FIEL / SABLATNIG, “Writer Identification and Writer Retrieval Using the Fisher Vector on Visual Vocabularies”.

[11] CHRISTLEIN et al., “Writer Identification and Verification Using GMM Supervectors”.

charters. Typically, the authorship of a datum line is known since it contains the name of the author (exceptions: wrong authorship or forgeries). Seventy-four of the datum lines were written by John of Gaeta. The goal was to automatically classify each datum line as either belonging to the class "written by John of Gaeta" or to the class "not written by John of Gaeta". Sample datum lines are shown in Figure 5 (Fig. 5).

3.1. Letter-based approach

First, we were interested in investigating whether it is possible to find out the correct class base on some given letters of a datum line. Four different letters ("a", "e","o", "c") were extracted from the datum lines. Local binary pattern (LBP) descriptors were computed from these letters and classified using Support Vector Machines (SVM). However, maximally an accuracy of about 60% was achieved. Possible reasons are: a) The quality of the letters is often bad, since most of the letters were extracted from digitalizations of analog photographs[12]; b) The chosen letters may not be individual enough, i. e., they do not vary significantly across distinct scribes. Other letters which carry more individuality like uppercase letters or letters with ascenders and descenders may perform better; and c) The extracted features might not be discriminative enough. LBPs work well for image retrieval cases such as face recognition. However, they might not carry enough information for writer identification.

3.2. Texture-based approach

Subsequently, another allograph-based method was evaluated on the complete extracted datum line (texture[13]). Hereby, SIFT descriptors were sampled at SIFT keypoints. SIFT descriptors are 128-dimensional rotational and scale-invariant descriptors. They encode the gradient distribution of a local neighborhood around its associated keypoint. k-means with 1000 clusters was employed for generating the background model. The local SIFT descriptors were encoded using vector quantization, i. e., for each cluster the number of the nearest descriptors was counted. In other words, each datum line was represented by the frequency of its SIFT descriptors. This histogram was then normalized by its l_1 norm. The correlation distance was used to compare two histograms with each other. An accuracy of 82.2% was achieved by using the correlation distance for comparing two normalized histograms with each other. This method is favorable in comparison to the letter-based approach. It achieves significantly higher accuracy and the annotation process is much faster, since only the datum line (instead of multiple individual letters) is extracted. However, an accuracy of about 80% seems to be low in comparison to the much higher rates of writer identification using contemporary documents. The reason for that is that the background is sometimes be along with the script. For instance, artifacts such as rips, folds or water marks are often treated as parts of the datum lines. Additionally, they often contain sections of symbols of the whole charter such as the "rota" or "benevalete" symbols, cf. Figures 5 and 6 (Fig. 5, Fig. 6).

[12] This is also known as retro-digitalization.

[13] Not to be confused with textural-based method.

3.3. Word-based approach

To improve the recognition accuracy, we took advantage of the typically limited vocabulary contained in datum lines. For example the five words that appear most often in these lines are “indic(t)” (127), “anno” (126), “dat / datum” (125), the abbreviation symbol “ m̊” (117), and “dnice” (117). Each word was individually annotated. Subsequently, the words were encoded using vector quantization in a similar manner to the previously described texture-based method (see Section 3.2). Each word is compared with all the instances of the same word. The most similar instance is found and the author of this word is kept in a list, i. e., in our case: “John of Gaeta” or “other”. If the occurrences of “John of Gaeta” is higher than of “other”, then the datum line is presumably written by him and vice-versa. With this method, an accuracy of 96.6% is achieved using the $\chi 2$ -distance for comparing each word histogram.

3.4. Failure cases

Figure 6 (Fig. 6) shows three datum lines which were falsely marked by our method as being written by John of Gaeta. The first and third one contain many artifacts. Thus, they could possibly be correctly identified by preprocessing the image to decrease the presence of such artifacts. The second datum line stems from a forgery. It is falsely misclassified by chance since all the other datum lines are very dissimilar to it. Another issue lies in authenticity. There are datum lines which were not written by John of Gaeta himself. In those cases, the best match might not be the correct one. By examining the sum of word distances, one could make a manual inspection of in-between cases.

4. Conclusion

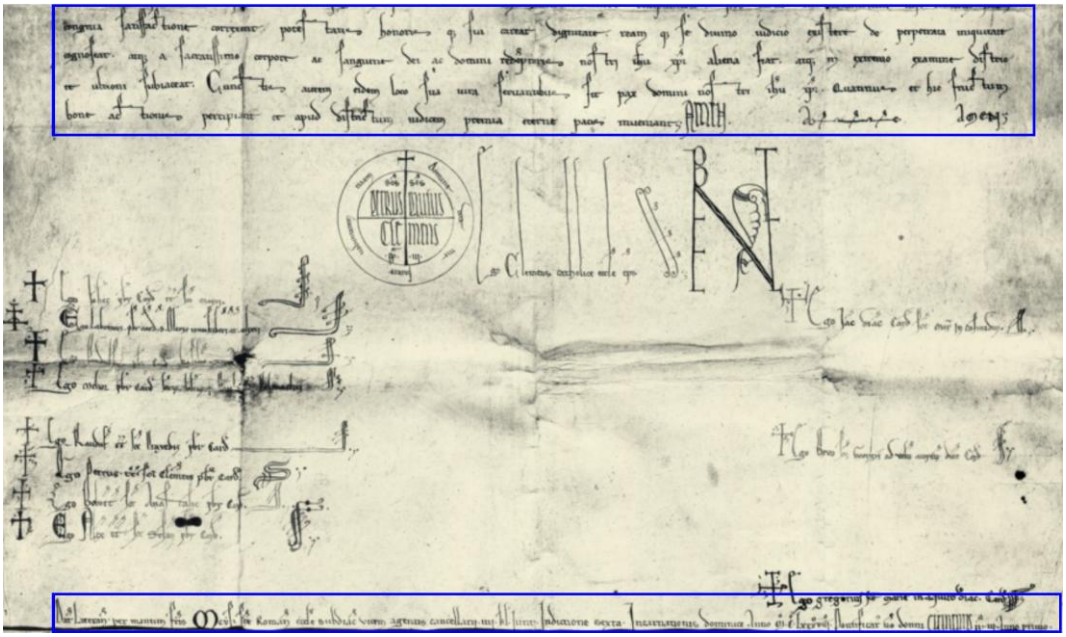
Several methods for writer identification (verification) were presented. While they achieve very high recognition results on contemporary documents, they are not directly applicable to historical data. Our experiments show that artifacts such as rips, holes, or water marks, negatively influence the recognition accuracy. Thus, we employed a verification mode on the word-level. The additional effort in annotating the words comes with an increased accuracy. Nearly all datum lines could be assigned correctly.

However, nobody should blindly trust the output of automatic methods. The algorithms can only give indications for handwritings being written by a specific person. But, an automatic system for writer identification can provide a confidence measure. Furthermore, it enables one to presort large amounts of data, and thus can drastically reduce the time for manual inspection. We believe that automatic methodscan already be used for finding similar writers in large amounts of data. However, in-between cases need to be carefully reviewed. Also note that in the case of deciding whether a document is written by a specific person or not, a human still outperforms current approaches. Technology-wise, other encoding methods would possibly perform better than vector quantization. Furthermore, distinct approaches for writer identification could be combined to achieve a higher accuracy.

Abstract

Automatic writer identification and writer verification has recently received significant attention in the field of historical analysis. In this work a short overview of current approaches for writer identification is given. Current state-of-the-art results on contemporary data are related to different approaches for writer verification on a small dataset of datum lines extracted from papal charters of the high middle ages. In the case of these datum lines, a word-based approach is superior to texture-based approaches.

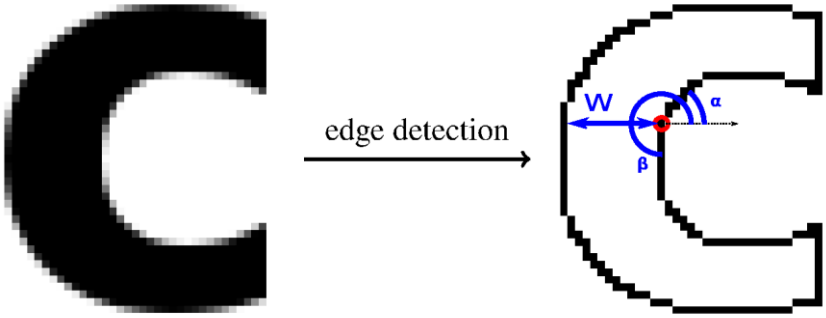
Fig. 1



Writer verification: Are the scribes of the context (top blue rectangle) and the datum line (bottomblue rectangle) the same individual? [Image source: Göttingen Academy of Sciences and Humanities]

Back to text

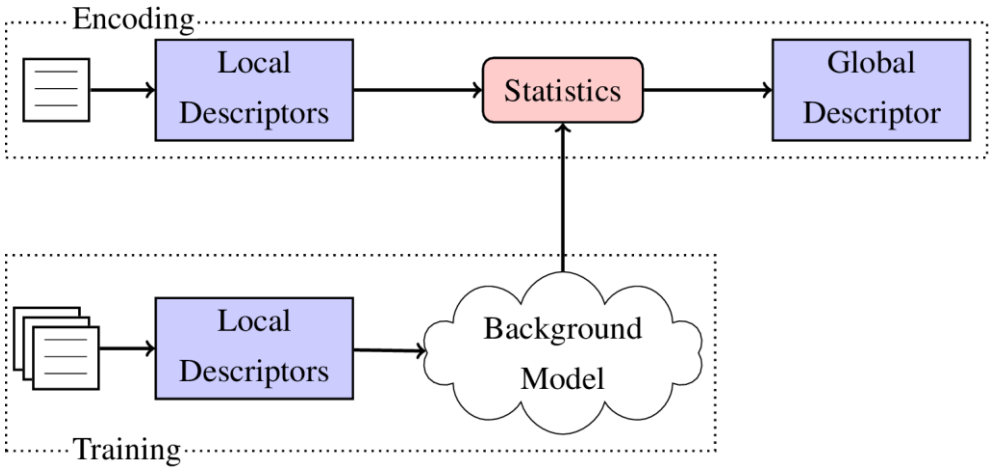
Fig. 2



Textural-based features for writer identification: At each edge point features like the width ordifferent angles are computed and aggregated to form a scribe-dependent descriptor.

Back to text

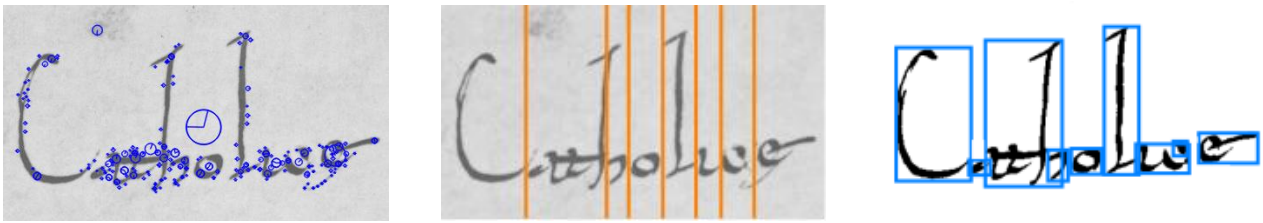
Fig. 3



Schematic illustration of the mode of operation of allograph-based methods.

Back to text

Fig. 4

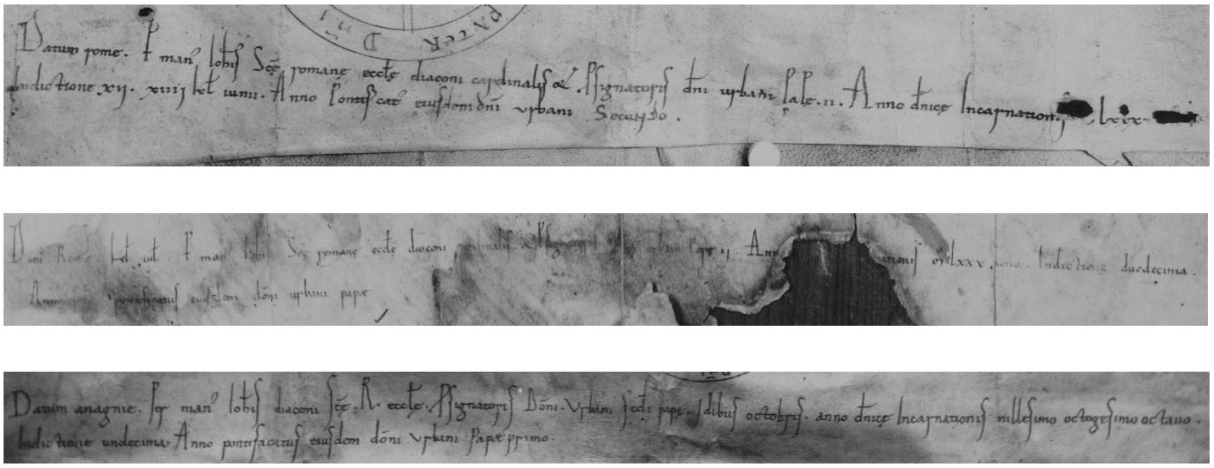


(a) SIFT keypoints *(b) Vertical cuts* *(c) Connected Components*

Different local feature partitions or positions.

Back to text

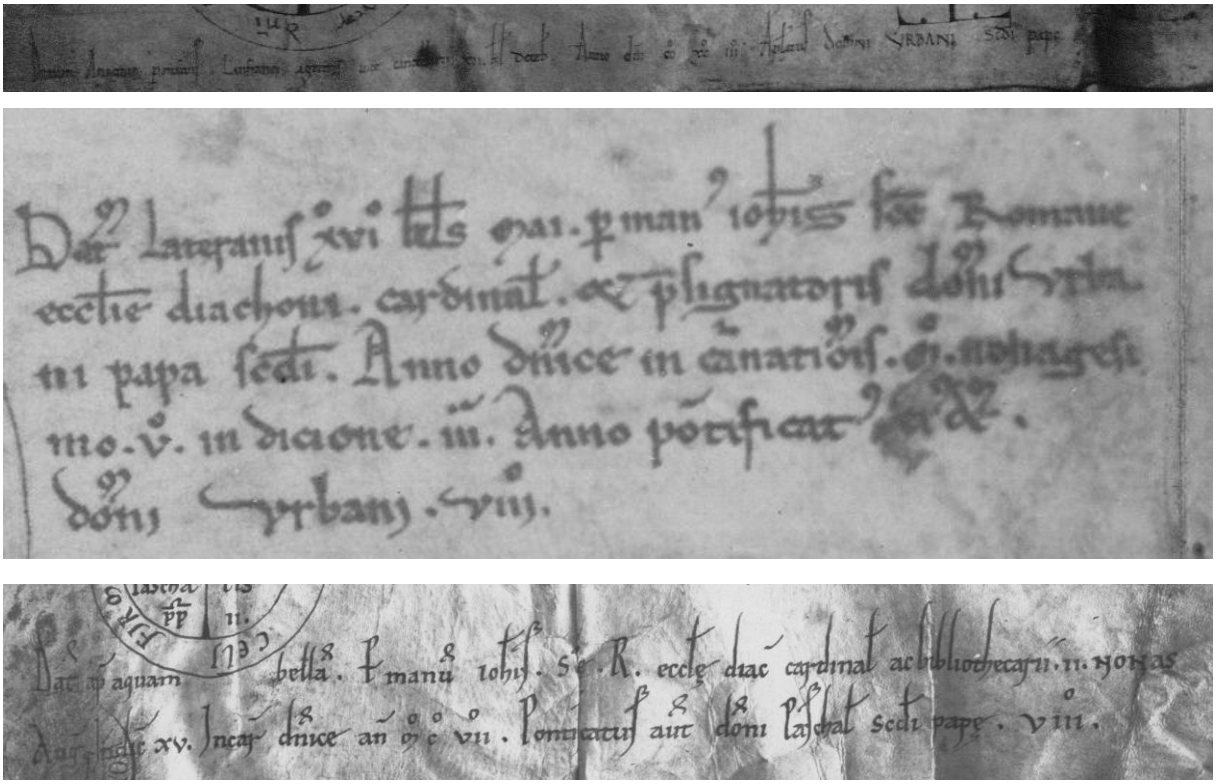
Fig. 5



Different datum lines extracted from high medieval papal charters. These datum lines were written by John of Gaeta. [Image source: Göttingen Academy of Sciences and Humanities]

Back to text

Fig. 6



Failure cases of the word-based approach. [Image source: Göttingen Academy of Sciences and Humanities]

Back to text