Writer identification through information retrieval: the allograph weight vector

Ralph Niels, Franc Grootjen, and Louis Vuurpijl
Nijmegen Institute for Cognition and Information, Radboud University,
The Netherlands
{r.niels, f.grootjen, l.vuurpijl}@nici.ru.nl

Abstract

We show a number of promising results in writer identification, by recasting the traditional information retrieval (IR) problem of finding documents based on the frequency of occurrence of their terms. In IR, the tf-idf is a well-known statistical measure that weighs the importance of certain terms occurring in a database of documents. Here, writers are searched on the basis of the frequency of occurrence of particular character shapes: the allographs. The results show a high retrieval score. Moreover, by using the af-iwf (allograph frequency - inverse writer frequency) measure, qualitative and quantitative analyses can be made that elaborate on the particular allograph shapes that lead to a successful writer identification. In this paper, we sketch the application of these techniques in forensic science.

Keywords: Writer identification, information retrieval, allograph frequencies

1. Introduction

In forensic writer identification, the task is to establish the identity of the writer of a questioned handwritten document, by comparing the questioned handwriting to handwritten samples with known identities which are stored in a database [3, 12]. The first to address this challenging problem using methods from the well-established research area of information retrieval were Bensefia, Pacquet, and Heutte [1, 2]. In information retrieval, text-based documents are indexed and stored in a database. Based on a query document, documents are retrieved from the database by computing a query index and retrieving the documents that most resemble the query. As shown in [1, 2], the information retrieval approach can be recast such that the query is specified through a questioned handwritten document and writer identification boils down to retrieving the pre-indexed handwritten documents stored in the database.

In this paper, we proceed with this approach. As part of our ongoing investigations within the Trigraph project [7], we are exploring new methods and tools for forensic document examination. Our investigations focus on the use of allograph-based information. We argued recently in [9] that a person's handwriting can be described by a vector containing the frequence of occurrence of prototypical characters, the so-called *allograph membership vector*. We will investigate in this paper how membership vectors can be used as a mechanism to index a person's handwriting information, such that it can be employed directly for information retrieval. As we will show, our methods yield a high writer identification performance on a moderately sized database (43 writers).

The organization of this paper is as follows. Below, in Section 2, we explain the concept of allograph membership (or allograph frequency) vectors. In Section 3, the information retrieval system used in this paper is discussed. The results of our experiments are presented in Section 4. We conclude this paper with a discussion in which we sketch that the outcomes of our experiments can be used for qualitative and quantitative assessments of the importance of the particular allograph shapes for writer identification.

2. Membership vectors and the allograph frequency vector

In [9], we introduced a method to use a collection of prototypical character shapes (allographs) for describing a person's handwriting. A prototype set was generated through semi-automatic clustering of the characters in the UNIPEN database [6]. The list of resulting prototypes can be used to generate so-called membership vectors, by counting the frequency of occurrence of each prototype, given a number of handwritten characters from a certain writer. Nearest-neighbor search was employed for determining whether a certain prototype matched a given input character. Matching was performed using the dynamic time warping (DTW) distance function [8].

The use of allograph memberships is a well-established method in forensic science [9, 10]. When investigating pieces of handwriting, experts often compile lists of allographs that occur in it. They try to distinguish between common allographs (allographs that can be found in the handwriting of a large writer population people) and less common allographs, which are characteristic for only a limited number of writers. If the same uncommon allograph appears in two documents, chances are higher that they are produced by the same writer, than when two common allographs occur in both documents. Subsequently, allograph lists of different documents can be combined and closely inspected to see how well two pieces of handwriting match.

In this paper, we formalize the generation of allograph frequency vectors as follows. We assume that for each writer w, a number of pre-segmented handwritten character samples $C(w)$ are available. Given a set of allograph prototypes (the set P), the allograph frequency vector *af(w)* can be computed using a so-called *allograph membership function* [9]:

$$\mathcal{M}(P, C(w)) = \textit{af(w)}$$

The membership function uses DTW-matching to determine the best matching prototype from P for each character sample from *C(w)*. The number of times each prototype is the best match for a sample in the handwriting of w is $\textit{af(w)}_p$, which represents the allograph frequency of prototype p for writer w.

3. IR and the allograph weight vector

As we will explain in this section, the writer identification process may be seen as a standard Information Retrieval task: finding document(s) using a query. Although the use of Information Retrieval (IR) techniques for writer identification is not new (see [2]) the approach presented here differs on an essential point: instead of using graphemes we map the input to prototypes which, besides the normal ranking, facilitates visual feedback to a supervising forensic expert. Instead of just presenting the top 5 possible writers, the system can justify *why* it finds a writer relevant.

3.1. The model

Where IR normally recognizes a set of *documents* and a set of *terminals*, in writer identification we are confronted with *writers* (the set W) and their written text. We will briefly describe how we use a standard IR model for writer identification (see Figure 1). As explained in Section 2, the writer input will be segmented into sets of characters *C(w)*, mapped on prototypes P using the membership function $\mathcal{M}$ and represented by allograph frequency vectors *af(w)*.

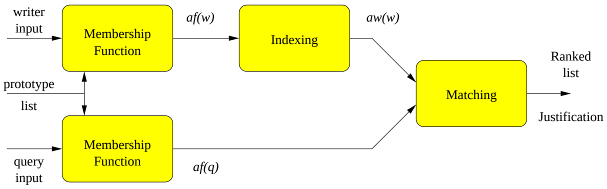

Figure 1. The IR model for writer identification.

3.1.1. Indexing

In IR retrieval applications it is a common practice to adjust the weight of terminals: after all some terminals appear so frequently that their contribution to the retrieval process is minimal or even negative (cf.stopwords). One of the most used methods is called *tf·idf*, which multiplies the terminal frequency component with a factor which is called *inverse document frequency*. For terminals with an high occurrence in the collection this factor is small, for rare terminals it is high.

For writer identification we follow the same reasoning: we will multiply the allograph frequency with the inverse writer frequency *af · iwf*. The allograph frequency is the value found in the previous step, the inverse writer frequency is defined as follows:

$$\mathit{iwf}(p) = {}^{2}\mathit{log}(\frac{n}{\mathit{wf}(p)})$$

where $n = |W|$, and $\mathit{wf}(p)$ is the number of writers that used allograph p. Obviously, the *iwf* factor will be close to zero for allographs with high occurrences. If we use $\mathit{af}(w)$ for our *af* component we can determine an *allograph weight vector* $\mathit{aw}(w)$ for each writer:

$$\mathit{aw}(w)_p = \mathit{af}(w) \cdot \mathit{iwf}(p)$$

However, most retrieval tools based on Okapi's BM25 [11], use a different formula which enables them to tune the system with parameters k_1 and b:

$$\mathit{aw}(w)_p = \frac{\mathit{af}(w)}{\mathit{af}(w) + k_1((1-b) + b \cdot \theta(w))} \cdot \mathit{iwf}(p)$$

where $\theta(w)$ represents the ratio between the number of used letters by w, and the average number of used letters (over all writers). For our experiment we selected the default values $k_1 = 1.2$ and $b = 1$.

3.1.2. Matching

In the matching phase, a query is used to generate a ranked list of possible writers. The ranking score should reflect some kind of similarity between the writer and the query.

Just as we did for the writers, we can segment the query into characters $C(q)$ and map them to prototypes.

Subsequently we can determine an allograph frequency vector for that query: *af*(q). Although it is possible to reweight the query allograph frequency vectors, it is common practice to leave them untouched. As we will see later, linear normalization of the query does not influence the ranking.

We now can define the similarity measure between a query q and a writer w:

$$\mathit{sim}(q, w) = \mathit{af}(q) \cdot \mathit{aw}(w) \tag{1}$$

The rationale behind this measure is linked to the (cosine of) the angle between the two vectors, and their relation with the vector in-product: if the writer and the query are highly related, their angle is close to 0 and their cosine is near to 1. If they are unrelated, their angle will be larger, and their cosine smaller.

3.2. Output

For each query the model will produce:

- A ranked list of writers, ordered by similarity.
- For each entry of the ranked list a justification of the result. This justification shows the letters in the query which were mapped to the ones in the writer's document together with the impact they had on the retrieval result.

3.3. Software

All retrieval runs were performed on a standard PC running BRIGHT, a retrieval engine designed for indexing and retrieving large databases (see [5]).

4. Allograph-based information retrieval

In this section, the first results of information retrieval based on the allograph weight vector *aw(w)* are presented. For our experiments, we used a database available in our department, the *plucoll* set. This database contains lowercase handwritten words, segmented into characters, and written by 43 writers. Each writer was requested to generate 5 sets of the same 210 words. From these sets, we selected 2 sets as query data $\mathcal{Q}$ and 2 sets as database data $\mathcal{D}$. For each of the 26 lowercase characters, at least 10 samples were available in both $\mathcal{Q}$ and $\mathcal{D}$, for each of the writers w.

The prototype set P used in the experiment is the *Mergesamples* collection described in [10]. This set of 1583 prototypes consists of actual character shapes that were produced by hierarchical clustering a large database of segmented lowercase letters (the UNIPEN trainset [6]), and merging the members of each cluster into an 'average' shape, using a variation of learning vector quantization.

4.1. Performance experiment

This experiment was performed to (i) determine how well the proposed IR-based technique (see Section 3) performs on the task of writer identification, and (ii) to investigate how much character samples are required to achieve a suitable level of performance. We varied the sizes of both query and database "documents" by manipulating the number of available characters per writer for formulating a query (*Nq*), and for indexing the database (*Nd*).

Indexed database documents were constructed by randomly selecting an equal number of *Nd* characters from the database data $\mathcal{D}$, for each writer w in the population. Writer identification of a writer q was performed by randomly selecting a number of *Nq* query characters from $C(q)$ and returning the ranked list of writers, ordered on Eq 1. Neither the query documents nor the database documents were balanced over alphabet letters, to simulate the real forensic practice, where in most cases the distribution of available letters is not equal over the alphabet.

The correct identification performance was assessed by counting for which relative amount of the 43 query documents (one for each writer) the system was able to find the corresponding database document. Both query and database documents were compiled randomly 10 times for each size combination $(\mathit{Nq}, \mathit{Nd})$, such that each experiment was re-run 100 times. The performance values reported in this section are the average performances of these 100 re-runs. The value of *Nq* was varied between 10 and 100 characters, while the value of *Nd* was varied between 100 and 1000 characters. These values correspond to the daily forensic practice: an average sentence in English contains about 75-100 letters, and the amount of available material is often limited to a few handwritten sentences.

Two measures were used to evaluate the performance of the system, given specific query and database sizes. In the first method (*top-1*), the relative amount of writers that were correctly identified by the system (i.e., the database document corresponding to the writer of the query document was ranked by the system at the first position) is reported. In the second method (*top-4*), a ranking within the first 4 positions (10%) was considered to be correct.

4.2. Results

Both the *top-1* and the *top-4* writer identification performances were calculated on each combination of query sizes *Nq* and database sizes *Np* that were tested. Below, in Figure 2, a graphical representation of the *top-1* performances is depicted. To explore these results in more detail, Table 1 below shows a selection of the measured performances.

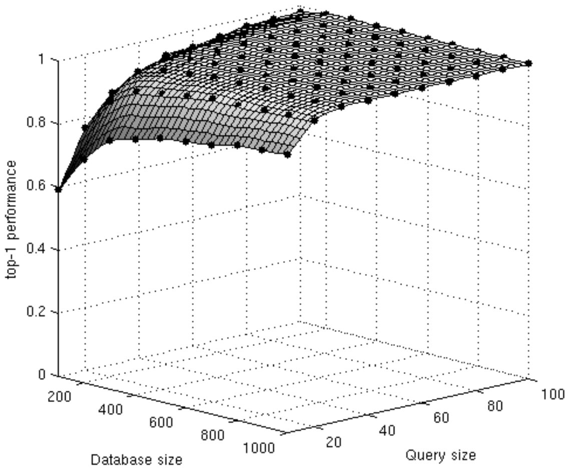

Figure 2. The average *top 1* performances given different sizes of query and database documents.

Table 1. Correct writer identification performance given different sizes of query and database documents (*top-1* and *top-4*. The reported values are averages over 100 random re-runs.

		Database size *Nd*			
		100		300	
top		*1*	*4*	*1*	*4*
Query size *Nq*	10	59.3	84.1	78.7	95.6
	30	86.0	97.7	97.2	99.9
	50	94.2	99.5	99.2	99.9
	70	96.3	99.9	99.8	100.0
	100	98.3	99.9	100.0	100.0

		Database size *Nd*			
		500		1000	
top		*1*	*4*	*1*	*4*
Query size *Nq*	10	83.3	97.3	88.2	98.9
	30	98.8	99.9	99.6	100.0
	50	99.8	100.0	99.8	100.0
	70	99.9	100.0	99.9	100.0
	100	100.0	100.0	100.0	100.0

As can be observed from this table, the *top-1* performance varies between 59.3% (for *Nq* $= 10$) and 100.0% (for large queries and a relatively large database). For considering more entries in the hit-list and examining whether the correct writer is contained in the list, we compute the *top-4* performance, which varies between 84.1% (for *Nq* $= 10$) and 100.0% (for larger database sizes).

5. Applications of IR for forensics

As shown in the previous section, the allograph weight measure *aw(w)* which we inspired on the well-known *tf·idf* measure from information retrieval, provides a promising means to perform writer identification. Please note that these findings sustain the outcomes of the experiments performed by Bensefia *et al* [1, 2], who achieved high recognition rates for databases of a larger size than our Plucoll collection. As we will argue in this section, the IR-measures computed based on allograph memberships can also be used for a more elaborate exploration of the information retrieval process, which may be of use for the forensic scientist.

5.1. Most distinctive characters for IR

As a first case, we would like to show how the inverse writer frequency can be used to explore which letters are most important for writer identification — for our current Plucoll database. Recall that the value *iwf*(p) corresponds to the 'importance' of a prototype p in our writer identification process. By averaging the values of all *iwf*(p) for each prototype belonging to a certain alphabet letter, the importance of that letter for writer identification can be computed. Note that this average is computed for the complete writer population. Figure 3 shows the average value of *iwf*(p), for the prototypes belonging to each alphabet letter. From this figure, it can concluded that, given our database, the letters 'q' and 'k' are the letters that are, on average, the most suitable letters for distinguishing between writers, while the letters 'e' and 'o' are the least distinguishing ones, on average.

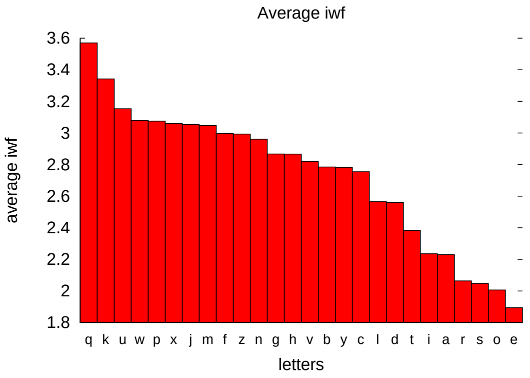

Figure 3. The importance of each alphabet letter for writer identification according to our technique. The letters 'q' and 'k' are the most distinctive for our dataset.

It should be noted that the frequency of occurrence of characters in the database is a very important factor for the *iwf*(p), which makes it plausible that, e.g., an 'e' or 'o' are least informative. Similarly, the letter 'q' occurs less frequently in the Plucoll collection, which enhances its distinctive properties. Other investigations on the distinctiveness of characters [4, 9, 13] indicate that indeed, the letters 'q' and 'k' are very informative letters when considering shape information.

5.2. IR revisited: under the hood

To further illustrate a number of potential applications, we first sketch a typical use scenario of this technique. Given handwritten material from a certain writer population, indexing the database results in an allograph weight vector *aw(w)* for each writer w. Search in this database is performed by formulating a query $C(q)$, a collection of questioned handwritten characters. The resulting ranked list of writers from the database is found through Equation 1.

Whereas in general a writer identification system merely yields such a ranked list of retrieved writers sorted on *sim*(q, w), we envisage an interactive workbench in which a forensic document examiner can interrogate the system for revealing more details on the decisions underlying the writer identification process. A typical starting point for the interaction between the forensic expert and such a system could be the chart depicted in Figure 4. The query $C(q)$ consisted of $Nq = 10$ randomly drawn samples from $\mathcal{Q}(paulus)$.

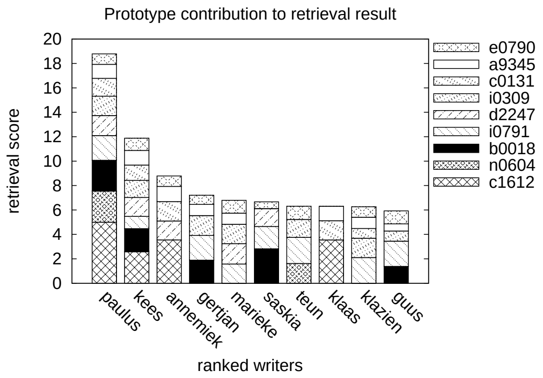

Figure 4. Chart showing how much each prototype contributed to the retrieval result, for each writer. The numbers depicted in the legenda on the right are prototype id's.

In this chart, the probability that the query was written by a certain writer is indicated by the height of the bar corresponding to that writer. Writer *paulus* has the highest rank, so the top-1 result is correct. Furthermore, each bar corresponding to a writer w consists of blocks that represent the value *af(q)p* · *aw(w)p* of the corresponding prototype p, i.e., the relative weight of that allograph for this particular query q. The height of each "prototype block" reflects the amplitude of the corresponding relative allograph weight. So, apparently the most important prototype that ruled this particular IR outcome is *c1612*, which is shared by almost all writers. On the other hand, a very distinctive prototype seems to be *n0604*, which distinguishes writer *paulus* from almost all other writers.

If we use this chart as an example, typical questions the expert could ask are: (i) "Since the prototype *c1612* plays such an important role in the handwritings of both writer *paulus* and *kees*. Let me see the characters of these writers that matched with the prototype", or (ii) "Prototype *n0604* is important in the handwriting of writer *paulus*, but not at all in that of writer *kees*. Let me see the 'n' characters from *paulus*' corresponding to *n0604*, and all the 'n's written by *kees*." For both questions, it is obvious that the expert would want to visualize the prototypes that influenced the writer identification results.

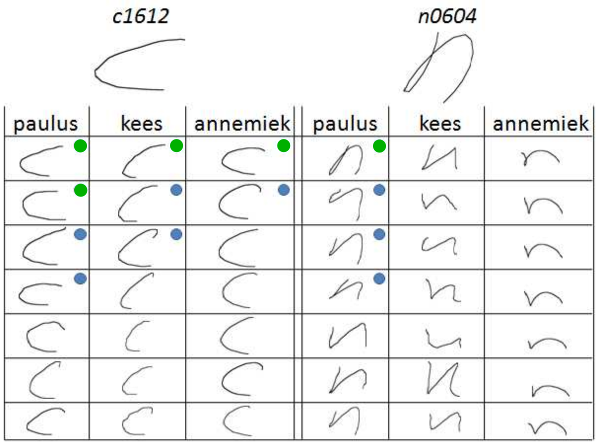

Figure 5. Possible system output that reveals the prototype shapes and character samples that rule the writer identification process. For the two prototypes *c1612* and *n0604*, the characters from the indexed database of the top-3 writers that matched to these prototypes are marked with a green dot. Also depicted are some character samples from these writers which were not selected for this query. Note that some of these samples have a DTW-match to these prototypes (marked with a blue dot) and some other characters do not match (no dot).

As can be observed in Figure 5, the option to inspect the prototypes and corresponding characters seems a very informative tool for understanding why a "black box" like a writer identification system yields a certain outcome. For example, it becomes clear why writers *kees* and *annemiek* have no samples that match to *n0604*.

6. Conclusions

Within the Trigraph project, we are exploring novel methods and tools for forensic document examination. Inspired by the works from Bensefia *et al*, we explored techniques from information retrieval for the task of writer identification. In our work, we have used the concept of allograph membership functions as introduced in [9] to compute allograph frequency vectors that provide a straight-forward input representation for IR. We have shown that the resulting allograph weight vectors have promising potential for effective writer identification tasks, which sustains the findings from Bensefia. Top-1 performances of almost 60% were achieved for small query documents containing only 10 characters. For larger databases, perfect top-1 writer identification rates were achieved.

We have sketched a number of applications of information retrieval techniques for forensic science. By ranking the inverse writer frequency, a quantitative assessment of the distinctive properties of each of the letters in the alphabet can be computed. Furthermore, we have shown examples of how the results of a writer identification system can be inspected such that a forensic document examiner would be able to better understand and justify why a specific retrieval result was generated.

We encourage the reader to proceed with the relatively unexplored use of IR techniques for writer identification. Our current efforts are directed on (i) performing more elaborate experiments on larger datasets, (ii) implementing the envisaged interactive workbench for examining and visualizing the results of a writer identification query, and (iii) exploring the possibilities of using information retrieval for probabilistic reasoning, something which is particularly important for forensic science, where the outcomes of case studies have to be adorned with real chances. The vast body of literature in information retrieval provides sufficient pointers to this latter issue.

References

[1] A. Bensefia, T. Paquet, and L. Heutte. Information retrieval based writer identification. In *ICDAR '03: Proceedings of the Seventh International Conference on Document Analysis and Recognition*, pages 946–950, Washington, DC, USA, 2003. IEEE Computer Society.

[2] A. Bensefia, T. Paquet, and L. Heutte. A writer identification and verification system. *Pattern Recogn. Lett.*, 26(13):2080–2092, 2005.

[3] M. Bulacu and L. Schomaker. Text-independent writer identification and verification using textural and allographic features. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), Special Issue - Biometrics: Progress and Directions*, 29(4):701–717, April 2007.

[4] Sung-Hyuk Cha, Sungsoo Yoon, and Charles C. Tappert. Handwriting copybook style identification for questioned document examination. *Journal of Forensic Document Examination*, 17:1–16, 2006.

[5] F.A. Grootjen and Th. P. van der Weide. The Bright side of information retrieval. Technical Report NIII, Radboud University of Nijmegen, 2004.

[6] Isabelle Guyon, Lambert Schomaker, Rejean Plamondon, Mark Liberman, and Stan Janet. UNIPEN project of on-line data exchange and recognizer benchmarks. In *Proceedings of the 12th International Conference on Pattern Recognition (ICPR'94)*, pages 29–33, Jerusalem, Israel, October 1994.

[7] R. Niels, L. Vuurpijl, and L.R.B. Schomaker. Introducing Trigraph - trimodal writer identification. In *Proc. European Network of Forensic Handwr. Experts*, Budapest, Hungary, 2005.

[8] Ralph Niels and Louis Vuurpijl. Using Dynamic Time Warping for intuitive handwriting recognition. In A. Marcellli and C. De Stefano, editors, *Advances in Graphonomics, Proceedings of the 12th Conference of the International Graphonomics Society (IGS2005)*, pages 217–221, Salerno, Italy, June 2005.

[9] Ralph Niels and Louis Vuurpijl. Generating copybooks from consistent handwriting styles. In *Proc. ICDAR 2007*, pages 1009–1013, Curitiba, Brazil, September 2007.

[10] Ralph Niels, Louis Vuurpijl, and Lambert Schomaker. Automatic allograph matching in forensic writer identification. *International Journal of Pattern Recognition and Artificial Intelligence*, 21(1):61–81, February 2006.

[11] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau. Okapi at TREC. In *Text REtrieval Conference*, pages 21–30, 1992.

[12] Sargur N. Srihari, Sung-Hyuk Cha, and Sangjik Lee. Establishing handwriting individuality using pattern recognition techniques. In *Proc. 6th ICDAR*, pages 1195–1204, Seattle, USA, 2001.

[13] Sungsoo Yoon, Seungseok Choi, Sung-Hyuk Cha, and Charles C. Tappert. Writer profiling using handwriting copybook styles. In *Proc. ICDAR 2005*, pages 600–604, Seoul, Korea, September 2005.