IS&T Intl Symp. on Electronic Imaging 2016, Document Recognition and Retrieval XXIII, San Francisco, CA, Feb. 15–18, 2016, paper 056.

Cuckoos among Your Data: A Quality Control Method to Retrieve Mislabeled Writer Identities from Handwriting Datasets

Vlad Atanasiu, Department of Informatics, University of Fribourg, Bd de Pérolles 98, 1700 Fribourg, Switzerland

Abstract

Motivation: *Handwriting datasets may contain specimens assigned to the wrong writer. A little discussed problem, such misclassifications, "cuckoos", can bias recognition, retrieval, identification, and other expertise systems, with serious consequences in biometric and forensic applications. Indeed, misclassification research has been purported as the most important topic in pattern recognition.* Objective: *We describe the design of a generic semi-automatic method for detecting possible misclassifications and illustrate it by way of an exemplary classification criteria (writer identity), measurement feature (contour orientation), and document distance metrics combination.* Method: *The core of the method consists in automated ranking of writer classes by stylistic variability, using the open source software Alphonse, followed by visual inspection of a limited number of top ranking classes, using an interactive handwriting datasets visualization tool, Rex. The method is independent from dataset producers and does not necessitate training. It is the result of empirical and theoretical research, and its performance demonstrated on the Swiss IAM offline handwriting dataset.* Findings: *We show that to evaluate the performance of a quality control it is necessary to consider the interdependency between system sensitivity and task difficulty. We propose a dataset-independent measure of the scrambling severity of a dataset and its proneness to misclassification. We find that in a broad writer population the variability of the contour orientation approaches a log-normal distribution, increasing the amount of genuine outliers.*

1. Introduction

Proposition 0: Rationale. A document recognition and retrieval system is only as good as its training and live data.

Corollary: Implication. The user does not care if the wrong document was retrieved because of the low quality of the system or that of the data, it only sees a wrong answer to his or her request, thus possibly spelling the demise of an otherwise good product.

Proposition 1: Solution. A writer verification system turns into a dataset quality control method by verifying the identity of every sample in a dataset.

Proposition 2: Generalization. Any system verifying the membership of an item to a class becomes a method for controlling the classification quality, once applied to every data item.

Proposition 3: Specificity. Writer identity is just one among many instances of handwriting classification criteria to which Proposition 2 applies, such as classification by style, handness, sex, age, or country.

Proposition 4: Constraints. The method given here can be applied indistinctly to any of the above classification criteria, beyond writer identity, insofar as the classes are stylistically homogeneous and the measuring instrument is sensitive to the classification criteria.

Definition: Terminology. We call a mislabeled data item a "cuckoo", and a class containing cuckoos "cuckold". The task of retrieving misclassified data is "cuckoo retrieval".

*

Paper objective — The objective of this paper is to describe the implementation of the generic principles of quality control expounded in the above Propositions, by way of an exemplary application to the specific case of writer identity. To this end, a semi-automatic software system, Alphonse [6], was created to experiment with an equally exemplary, yet effective, combination of image processing techniques using contour orientation as measuring instrument of the handwriting style, a cocktail of statistical metrics for distances between documents, and an interactive visualization tool for handwriting datasets, Rex [8].

Quality control in light of document retrieval — Quality control can be viewed as a special task in document retrieval, insofar as it means retrieving all mislabeled documents in a dataset. To the difference of document verification, often performed on a single document, quality control is not mere verification repeated as many times as there are documents in a dataset. The challenge of quality control is to group all misclassification candidates at the top of a ranking list, have an estimate of how many there are, and have the tools for visual confirmation by the human users of the system. In a previous publication we minted the concept of "writer retrieval" – the retrieval of all documents in a dataset produced by the same writer –, which is itself an extension of writer identification [7]; here we present the retrieval of misclassifications as an extension of the verification task.

Topic relevance — Misclassification research has been purported as the most important topic in pattern recognition [82, 38]. Specifically, data quality control contributes to document recognition and retrieval by improving the performance and security of systems [74, 14, 36]. In parallel, it serves other document processing tasks, such as identification and verification, by ensuring a healthy data basis. As for the relevance of writer identity, it is essential to the broadest spectrum of fields in which handwriting plays a role: biometrics, forensics, paleography, art dealership, medicine, education, digital libraries, and so forth. Indeed, it is the availability of writer identities what makes handwriting-based biometrics and forensics possible. When assignments of handwriting samples to the wrong writer occur, its impact varies in nature and intensity. As an example, the performance of writer identification systems evaluated on the dataset tested in this paper, IAM OffLine Handwriting database 3.0 [53], popular with academics, might be better than reported, or worse, due to a small amount of misclassifications. Quality control is also time consuming and tedious: it took about six months for one person to manually label the 382,200 characters of NIST SD3, another widely used handwriting dataset [86: 209; 35: 1; 87]. Thus, it seems desirable to have at disposal methods for the quality control of handwriting datasets.

Paper contribution and originality — The paper points toward an important, but little discussed topic. Its *theoretical* contribution is the insight that dataset quality control can be cast as an extension of item verification, which leads to a generic framework for the quality control of handwriting datasets, valid for various application domains, tasks, and classification types. The *operative* novelty of the paper's method is to create a quality control layer independent of the dataset production workflow and software, where usually quality control takes place. At the level of *implementation*, we propose an original system utilizing the synergy between machine and human capabilities on one hand, and image processing, statistics, and interactive visualization on the other hand. Our choices are based on empirical and theoretical research, in the favor of which arguments their good performance. Interviews with the dataset producers were conducted to understand its design process and history. We also refine the *evaluation concept* to account for the interdependence between system performance, data complication, and use case requirements. From a *practical* point of view, the system doesn't demand training on preexisting data models, hence its simplicity and generality. The selected handwriting feature is versatile, robust, and its characteristics understood. Method implementation and reusability are facilitated by the publishing of the handwriting analysis software and visualization tool. An *unexpected* contribution of the paper is the discovery of a log-normal type distribution of variability in a writer population, a topic that begs for further research.

Paper structure — Having clarified the object of the paper, stated the principles guiding us, and stressed the need for quality control of handwriting datasets, we will dwell into the state of the art of quality control (Section 2) and explore the typology, factors, and implications of misclassification (Section 3). We examine the proposed detection method (Section 4) and evaluate experimentally its performance (Section 5) and task difficulty (Section 6). We conclude with perspectives on future work (Section 7) and a retrospective of the accomplishments (Section 8).

2. Research on quality control

Fields — Quality control is addressed in document processing by *normed workflow protocols* for data acquisition [86, 75], *data cleaning* of digital repositories [51], *ground truthing* of training data [86, 68, 12], *error-rate estimation* of classification systems [1], and results *evaluation* [86, 42]. More generally, the subject relates to a number of broader scientific fields: *statistical process control*, devoted to quality control in industrial and management settings, a field emerged in the United States before World War Two and later contributed to the success of the Japanese industry [64, 48, 65]; *data quality*, crystallized during the "dotcom boom" of the late 1990s and specifically concerned with digital data [74, 40, 57]; *outliers theory*, elaborated in statistics [10, 62]; and *misclassification research*, fundamental to pattern recognition [82, 38]. Ad hoc solutions to misclassification have been developed in a wide range of areas, beyond the dedicated fields just mentioned, providing interesting inspirational material: medicine (diagnosis reliability) [50], industrial inspection (surfaces defects) [59], geographical information sciences (imprecise coordinates) [76], biometry [52], surveillance (abnormal behavior) [43], satellite imagery (feature detection) [17, 33], and many more.

Challenges — Regarding writer identity, most quality controls are consubstantial with the data production process or classification algorithm [86, 75], creating a need for independent control capabilities once datasets and classification outputs reach end users. Also, control methods usually employ machine learning, which depends on the availability of training data and implementation efforts [37, 35, 14]. Finally, some datasets are more prone to errors than others [87], an information that could improve control systems and evaluation methods (like in scanning software, where the use of image degradation models are common [9]). These three issues – *independence from dataset producer*, *learning-free quality control*, and *prediction of misclassification proneness* – are addressed in these pages.

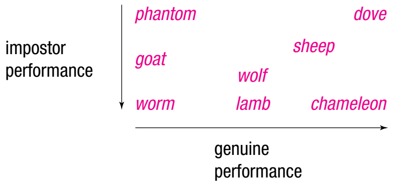
Prior work — Despite our best efforts we discovered only two sets of publications on the identification of misclassifications in handwritten documents, both from the 1990's [56, 36; 37]. (As a testimony of the topic's pedigree, a historical note: The first publications originate in the AT&T Bell Laboratories research group who invented the Support Vector Machines and, later, Deep Learning [21, 78]. The second publication was part of a research that sparked a fruitful era in computational handwriting processing at the Institute of Applied Mathematics, Bern, Switzerland, of which the IAM handwriting dataset used in this paper was one outcrop.) The difference between our method and the two others starts with a difference in *goals*: writer identification vs handwriting recognition. It results in a *structural difference of the input data*: on one hand classes with good *homogeneity*, due to the relative stability of an individual's handwriting, and on the other hand a high intra-class *heterogeneity*, generated by mixing allographic characters written by many writers. A pure *statistical* approach involving no learning is sufficient in the first case, while *learning* is necessary in the second. An ensuing requirement is that of preexisting *training data*. Both methods contend, however, partially with the *same problems*, such as whether to include a *manual check* of the machine's output (the answer from all researchers was "yes") and how to determine when the check should *stop* (here, the solutions are framed by the general approach of the respective quality control methods: using measures of statistical variance, of information theoretical surprise, and of Bayesian probabilities of error rates, respectively). In conclusion, *when intra-class variability is low, frequentist statistics are sufficient to detect misclassifications, otherwise machine learning is necessary*.

Alternative — Anticipating the method description, we mention that the underlying principle of the proposed handwriting dataset quality control consists in ranking writer classes by their internal stylistic variability, ideally resulting in an accumulation of the misclassifications in the top ranks. There exists a straightforward alternative to our *class variability ranking* method: reclassify anew the dataset, using a concurrent number of classifiers and compare their results with the original classification. *Concurrent classification* is a well-known technique, e.g. Web search engines use it for labeling images (sometimes by human volunteers) and it is even embodied in hardware (multi-viewer microscopes used by biologists, medical doctors, and others, to classify specimens by consensus [88]). The idea sounds attractive for automating handwriting quality control, were it not for the lack of information on the amount of cuckoos, which makes it a task of classification with unknown number of classes, a much harder problem than ranking existing classes by variability [29, 80, 26]. At the same time one has to be sure of the quality of individual components and the fine tuning of their interaction in this system of systems. Last, there is the practical difficulty in acquiring multiple systems: do they exist? how much do they cost? what resources do they demand? The interest of the class variability ranking method appears, in these circumstances, as a compelling solution.

3. Fundamentals of misclassification

Genesis — Cuckoos are typically created when labeling data with metadata (*labeling errors*) and when clustering the resulting entities in classes (*categorization errors*). Responsibility lies principally with the dataset producers, but software, the document writers themselves, and even environmental factors can play a role in misclassification. Cuckoos are mostly *natural*, emerging by accident, due to *attentional* and *procedural issues* (e.g. typos, merging of incompatible data, natural language or script ambiguities). Another cuckoo specie is *artificial*,

Writing variability factors				
writer	consistency	individual	weak · *hen* · **272**	(1)
			normal · *sparrow*	(2)
			strong · *swan* · **231**	(3)
		population	weak · *metropolis*	
			normal · *city*	
			strong · *village*	
	style	single · *flamingo*		(4)
		multiple · *myriapod*	shape · *lesser* · **81**	(5)
			structure · *higher* · **555**	
context	demographics	foreigner · *stork* · **159**		(6)
		child · *rabbit*		
		senior · *tortoise*		
		sick · *octopus*	*etc.*	(7)
	text	combinatorics · *mutant* · **297**, **208**		

Misclassification detectability				
contrast	endogenes · *locals*	cluster · *commoners*	● ● ●	
		outlier · *jester*	● ● ●	
	exogenes · *foreigners*	none · *spy*	●●● ●	(8)
		weak · *missionary*	● ● ● ●	(9)
		strong · *ambassador*	● ● ● ●	(10)
observer	humans only detect it · *dog*			(11)
	algorithms only · *mouse*			(12)
	both humans and algorithms · *cat* · **577**			(13)
	neither humans, nor algorithms · *fairy*			(14)
proneness	 impostor performance phantom dove sheep goat wolf worm lamb chameleon genuine performance			

Classification quality							
	model			admit · *pigeon* · **231**	(17)	reject · *chicken* · **208**, **577**	(18)
	ground truth	correct · *dove* · **231**, **208**	(15)	true no alarm · *white dove* · **231**		false alarm · *black dove* · **208**	(19)
		wrong · *cuckoo* · **577**	(16)	false no alarm · *black cuckoo*		true alarm · *white cuckoo* · **577**	

(1) "Hen writing" stands for bad, illegible handwriting. Expression attributed to the Latin playwright Plautus, 2nd c. BC, still in use in contemporary Italy [84, 85]. — (2) An epitome of commonness. — (3) Geese were mascots of Taoist calligraphers in Ancient China [47: 27]; swans emblems of grace in Western cultures. — (4) Stands on one foot / script style. — (5) Caterpillars master multiple "hands" / styles. — (6) Celebrated migratory bird. — (7) It drank ink. — (8) Hides well. — (9) Adopts local customs / script styles, but ends modifying them / the data distribution. — (10) Stands out. — (11) Man's best friend. — (12) Of the electronic pointing device sort. — (13) Even algorithms watch lolcatz on the Internet. — (14) We known domestic fairies are there, but can't see them. — (15) Symbol of peace. — (16) Lays its eggs in other birds' nests. — (17) Admit: unremarkable sight. — (18) Reject: is a bird, but doesn't fly. — (19) Alarming, although harmless, like a black sheep.

Table 1. *The handwriting and quality control bestiary — In Roman characters the formal terms, in Italics the informal, and in bold writer ids of illustrative samples from the IAM dataset.*

Writing variability factors — *Handwriting variability results from complex interactions of a profusion of writer specific and contextual factors [5: 55–57, 41]. "Bad" writing is just the most familiar cause, the number of distinct "hands" mastered by an individual is another: polystylistic writers can vary both formal features, such as slant, and structural features, like allographs. Of particular interest for misclassification are within and inbetween writer variability, the consistency of the demographic make-up of writer population and dataset determining the amount and detectability of misclassifications. A case in point is the presence of foreigners, remarked by both this author and author Ha [26] in their respective datasets. Authorship verification poses peculiar problems when dealing with data obtained over long periods of time, such as in biometric, medical, or historical applications, because the transformation of script with biological and cognitive development. Conversely, short-term events, such as substance intake, stress, and illness, create outlier samples prone to misclassification. Little, if anything, in the metadata of handwriting datasets records, however, such events, complicating quality control. · Given that, by design, our method identifies writers with high script variability, it can double as a tool for retrieving specific demographics. Though, the capability raises ethical and legal questions. For instance, our system has thrust out of the mass of documents items ostensibly written by two foreigners, non-native Latin script writers. Would this have been a real life system, its potential for indiscriminate use would be blatant.*

Misclassification detectability — *Besides within writer variability, the second most important factor for the detectability of misclassifications is the disparity, or "contrast", between the endogenous and exogenous samples in a writer class. Cuckoos are more difficult to detect if genuine within writer variability is low and endogenous-to-exogenous contrast is weak. Similar difficulties arise for populations or datasets with low between writer variability. Different observers have also different detection capabilities. Documents, individuals, and populations can be further categorized according to their propensity at being correctly identified in terms of authorship. The schema, called the "zoo plot" and based on the "biometric menagerie" typology [27: 166–167], is useful to predict the performance of identification systems or take particular workflow and managerial measures, such as avoiding some demographics or submitting them to human expertise.*

Classification quality — *By introducing the term "cuckoo" for a misclassified item, one might feel curious, or in need for, the nomenclature for what is not cuckoo and the different sorts of cuckoos. This typology is provided here in the form of a confusion matrix for misclassification types.*

produced intentionally, for legitimate reasons (scientific experiments, such as in this paper) or out of malice (e.g. forgeries, vandalism). For other origins of cuckoos see also [57: 5–22, 48: 157, 10: 26–28].

Typology — A typology of writer misclassifications enriches our understanding on their genesis, prevention, and rectification. Table 1 provides such a classification schema and Fig. 3 supplies examples. To explore the dataset used in developing the typology, the reader is encouraged to use the online handwriting features browser Rex, created by the author for navigation within handwriting datasets. In regard to the adopted nomenclature, we were guided by the principle of keeping terms short, clear, and memorable. Insofar as many concepts have no established names, we shadowed the formal, cumbersome terms, with names of (mostly) animals, following in this a custom from the early days of speech processing [25], recently adopted in biometrics [27: 161–180] and with echoes in criminalistics [71: 150].

Detectability — The detectability of cuckoos depends on the coupling effect of the sensitivity of the control *instrument* to misclassifications, the *scrambling* pattern of which and how many documents are misclassified to which classes, the *dataset* structure, the *producer* workmanship [11: 35], the *sample* representativeness, the variability of the underlying demographics of the writer *population*, and the *text combinatorics*, i.e. the visual pattern resulting from the interaction of language, script, orthography, and content [2]. Section 6 will further discuss how these layers affect performance evaluation.

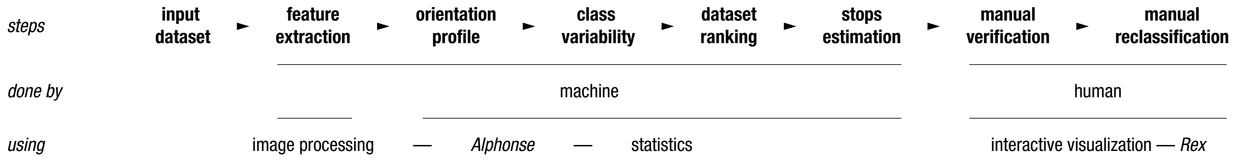


Figure 1. *Process diagram, elements, actuators, and capabilities of the Alphonse Rex quality control system.*

Severity — The primary aspects to consider when assessing the implications of identity misclassification are the *stringency* to detect them, the *tolerance* range of accepted failure, the *affordance* to meet these requirements with the available resources, the *proneness* of populations and producers for misclassifications, and the *prevalence* of the phenomenon [83]. The importance of the issue is area and task dependent, ranging from critical for biometric, forensic, and medical applications, to benign for art historical studies.

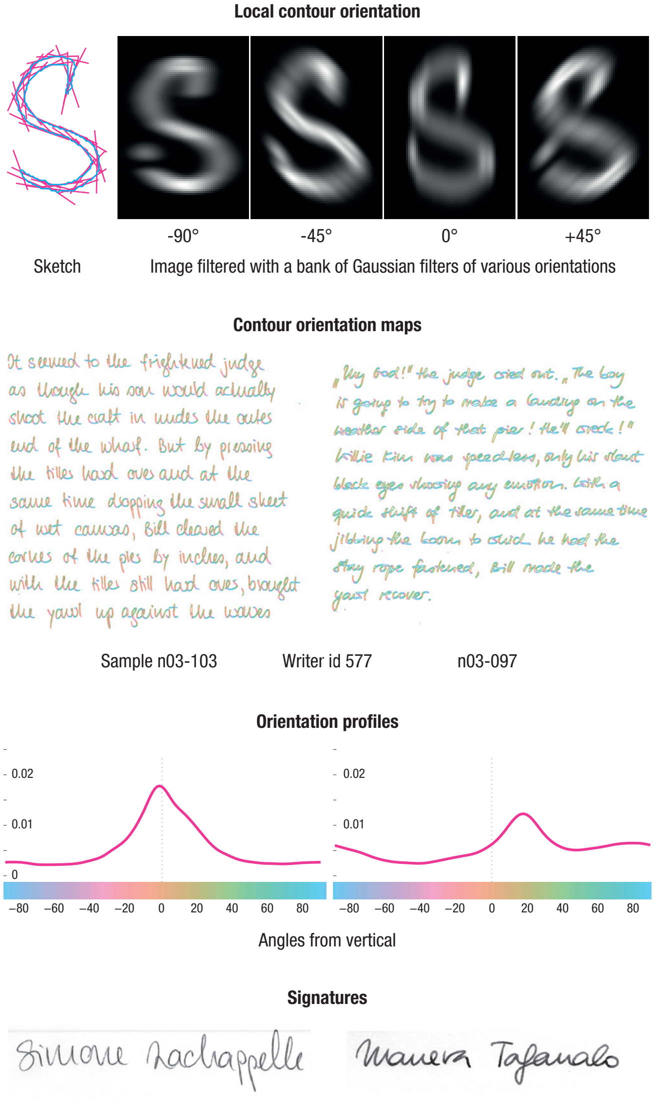
Workflow — Data quality control is managed at the *organizational* level of data producers or consumers, *architectural* level of the adopted control strategy, and *computational* level of software implementation [74: 8–10]. By combining the various references in the quality control literature, a generic procedure for the control of misclassifications emerges, comprising five steps: (*1*) *assessment* of the errors' impact for the given use case, (*2*) *procurement* of necessary control methods and tools, (*3*) *preventive* action by integrating quality control in the dataset production process, (*4*) *post-production* error detection and correction, and (*5*) *reporting* the quality control procedure and performance. Steps 4 and 5 are relevant to this paper and are expanded upon hereafter.

4. Method description

Method summary — The backbone of our quality control method consists in using image processing and statistical analysis to automatically rank writer classes according to within writer variability, then manually inspect a number of the top-most classes for misclassified items with an interactive visualization tool. Fig. 1 gives a schematic overview of the quality control process: extraction of a writing feature, measurement of class variability, estimation of stops, check for misclassifications within stop limits of those classes with high variability, and reassignment of misclassified samples to new writer classes.

Feature measurement — The feature is the local orientation along the writing contour, an angular value, given for each pixel of the contour, in respect to the handwriting's horizontal baseline, assumed parallel to the image abscissa (Fig. 2). It is obtained by convolving the binarized contour image with a bank of gaussian filters of size 30×30 pixels, sigma 2 and 0.5, and 1 degree orientation step [7]. The probability density function of the dominant orientation angles is a vector, the *contour orientation profile* of a script. It tells us how long a writer moved the writing instrument in a given direction. Fig. 2 provides an example of two documents attributed in the IAM dataset to the same writer, a classification to which we were alerted by the difference of their orientation profiles, leading to the discovery of distinct signatures on the original documents, and thus to a cuckold writer class.

The rationale for having selected orientation as feature is as follows. The scope of this paper is to prove a concept – quality control by exhaustive writer verification –, not find the best possible feature for this task. So from this point of view any well performing feature would do. As evident from Table 2, there are anyways to many features



It seemed to the frightened judge
as though his son would actually
shoot the craft in under the outer
end of the wharf. But by pressing
the tiller hard over and at the
same time dropping the small sheet
of wet canvas, Bill cleared the
corner of the pier by inches, and
with the tiller still hard over, brought
the yawl up against the waves

„My God!" the judge cried out. „The boy
is going to try to make a landing on the
weather side of that pier! He'll wreck!"
Willie Kim was speechless, only his slant
black eyes showing any emotion. With a
quick shift of tiller, and at the same time
jibbing the boom to which he had the
stay rope fastened, Bill made the
yawl recover.

Figure 2. *The orientation feature, its profile, and its analysis — A sketch illustrates the concept of local contour orientation for a single character. The anisotropic filtered images offer a peek into the process of extracting the orientation angles. Once they are obtained for each pixel along the contour, the result can be visualized by color coding the document. The orientations distribution gives the probability density function shown by a graph, where the zero angle is relative to the vertical axis of the image. — These specific handwriting samples are an example of cuckoos. The quality control system has detected a substantial difference in the measured signal, the orientation profile; visual comparison of the two samples reveals, indeed, a stylistic distinctiveness. A verification of the signatures apposed on the original forms proves the machine and human assumption: despite being labeled with the same writer id in the dataset metadata, the samples belong to different authors.*

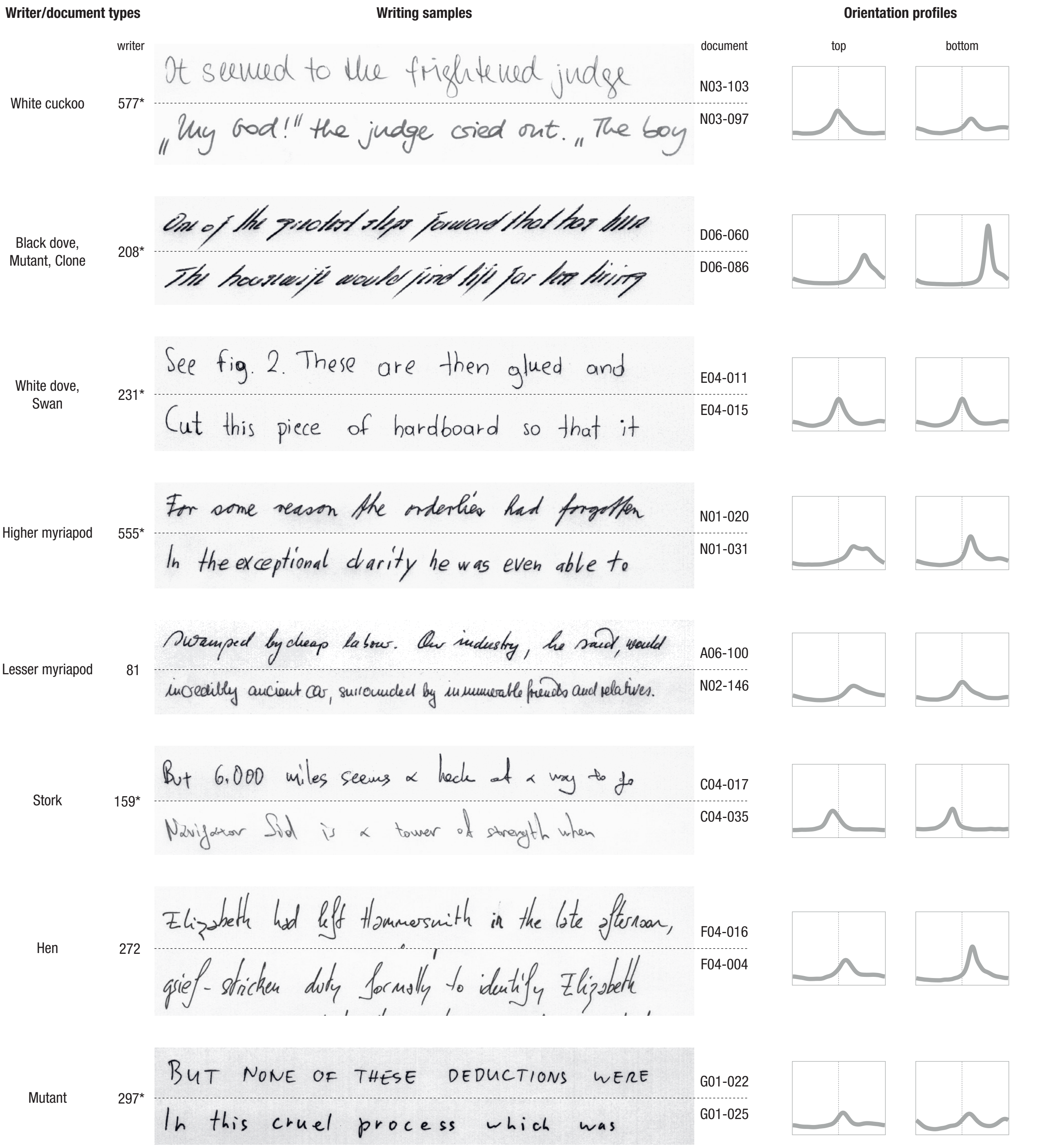


Figure 3. *Typology samples — The dataset is IAM 3.0; samples are cut-out of larger paragraphs; writer class id left of the samples, document id to the right; on the left hand is the nomenclature defined in Table 1; on the right hand the orientation profiles of the paragraphs to which the top and bottom samples belong. An asterisk marks documents signed by their respective writers. E.g., the documents of class 577 are signed with two different names, proving that this is a cuckold class; while class 555 exhibits the same name in its items, comforting the graphonomical expertise hypothesis of a polystylistic writer class. When writer 555 makes structural style changes, such as using different allographs, writer 81, also a myriapod, morphs superficial parameters, in this case the slant. Class 208 appears to the algorithm to be written by two writers, possibly due to a difference in text combinatorics. Class 272 has one of the largest natural within writer variability in the dataset (rank seven in Fig. 6, with method M2), while class 231 has one of the smallest variability. Class 159 is signed with a Greek name – note the Latin "a" written as a Greek "α" alpha. The disparity registered by the instrument in class 297 is due to a change in writing case, not writer.*

to implement, not a few beyond present capabilities in computational graphonomics. Nevertheless, contour orientation was not chosen arbitrary. First, it has good descriptive performance, being one of the top writer identification and verification instruments of the state of the art [16, 77]; it is a widely used handwriting expertise feature and its properties studied [4]. Second, it is versatile: various statistical features of the orientation profile have perceptual correlates. The mode, for example, is indicative of the script slant, the profile's entropy relates to the script roundness, and the spread around the mode reflects the writer's consistency [7]. Under statistical scrutiny, contour orientation is more than a numerical vector, it is a prism that splits an abstract numeric datum into a variety of meaningful perceptual concepts. In this sense, there are more than one handwriting feature which are employed in this paper. The advantage of analytical–perceptual correlates is to make the quality control system usable beyond the computing community, by users who don't necessarily have a mathematical background.

Within writer variability — The orientation profiles allow us to estimate the stylistic variability (or consistency) of a writer, by using some measure of distance between the handwriting samples. An example of profiles for a single writer is illustrated in Fig. 4a. In the following we explain the process by which we chose our measures and outline the algorithm for their use.

Choice of methods — Our goal being to find exogenous items in a collection of distribution vectors, we need a measure that emphasizes the variability within classes. This can be variously approached as a problem in outlier detection [10], distance measurement [23, 18], or shape description [30, 58, 28]. From the aforementioned literature we selected three basic descriptive statistics of dispersion (*summation*, *standard deviation*, *l2-norm*), four well known pair metrics (*chi-square distance*, *cosine distance*, *earth mover's distance*, and *dynamic time warping*), and a data preparation technique (*Fourier transform*).

The choices are motivated by their adequacy to the nature of our data and the goal mentioned above, and their profusion is a sign of complementarity. Observing that orientation profiles are circular, suggests the use of the frequency domain, where regularities are easily analyzed. Given that the data represents handwriting and handwriting is deformed in an "elastic" fashion, makes dynamic time warping and earth mover's distance natural choices to measure variability, the very reason for which the two methods were created. The sensitivity to outliers of summation, standard deviation, and l2-norm is why they are convenient in quality control. The chi-square and cosine distances are chosen for the more prosaically reason of being standard methods of vector distance measurement.

To design an integrated method we proceed along two ways. On one hand we find the best methods by experimentation. On the other hand we use ensemble methods to fuse the results of the base methods (with the *Borda count* for sum of ranks and normalized magnitudes [13, 49, 69, 70, 81, 46]). This allows to both circumvent the dilemma of method selection and improve performance.

Algorithm (Fig. 4 bottom) — (*1*) *Domain selection*: Use as starting point the probability density function orientation vectors or the magnitude of their Fourier transform. (*2*) *Matrix to vector reduction*: *Point-wise variant*: Make a matrix whose rows are the orientation vectors; then apply to each column the standard deviation or l2-norm statistics to obtain a vector. *Pair-wise variant*: Compute the pair distance between orientation vectors using chi-square, cosine, earth mover's, or dynamic time warping. (*3*) *Vector to scalar reduction*: Apply summation, standard deviation, or l2-norm to the result of step 2, to get the writer class variability.

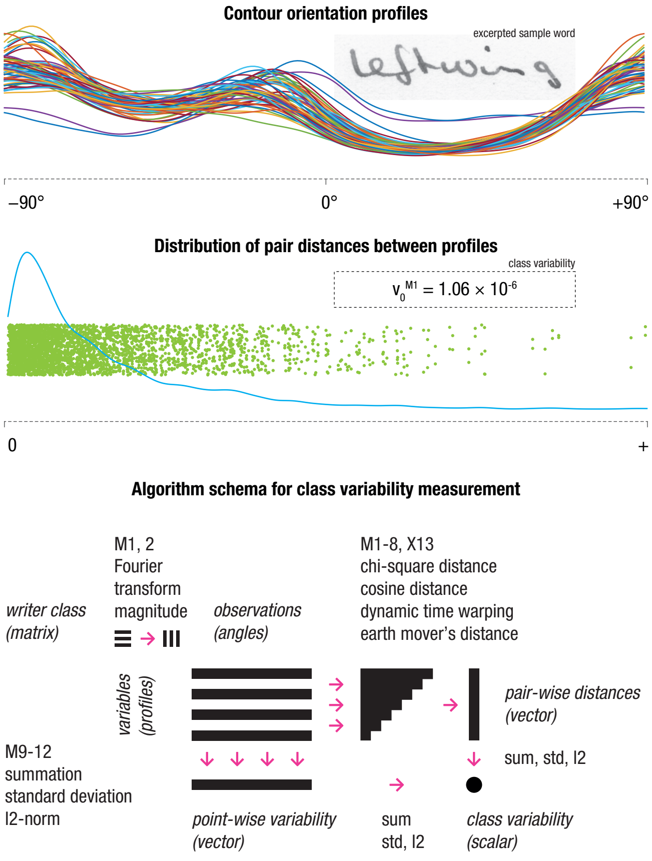


Figure 4. *Within writer variability* — Top: *Contour orientation profiles of the 59 handwriting samples of writer id 0 of the IAM dataset, an unusual case of left-slanting.* — Middle: *Distribution of pair distances between the profiles obtained with method M1 of Fig. 6; values as green dots, density as blue line, numerical value of class variability.* — Bottom: *Graphical representation of the algorithm for class variability measurement, consisting in a matrix to vector to scalar reduction.*

Inbetween writer variability — Once we got the variability for each writer using the above methods, we simply rank them and obtain the writer variability profile of the entire dataset (Fig. 5). We expect the cuckoos to lie in the top classes, on the presumption that writers have consistent writing styles, i.e. low variability.

Ranking stops — Before we start to manually check for misclassifications, there is a last automated step: determine when to stop the manual check of top ranking classes. We devised a generic rule, highlighted the role of interactivity, and sketched a model-based method.

Generic stops — Looking at the dataset writer variability distributions (Fig. 5) we observe that, except for the low performing method M12, few writer exhibit very high handwriting variability. By taking as stops analytically meaningful values, e.g. multiples of the standard deviation, we hope to isolate high variability writers from the rest of the population. Fig. 6 confirms that cuckoos in the IAM dataset fall within the range of the first two statistical stops, again, with the exception of M8 and M12.

Depending on the available amount of time and desired quality control precision, the user can chose at which level to stop the manual check – knowing that one should ideally check as many classes as possible. The stops suggested here are an approximation of the 3 sigma rule used in process control applications [20: 69–70].

Subjective stops — It is critical that the user be not blindly guided by inflexible thresholds, but takes advantage of the information on the whole dataset variability distribution revealed by its visualization. Note the distinction between "when" and "were" to stop. The more a class has members, the more time it takes time to verify them and less long will be the affordable total number of checked ranks.

Model based stops — Two types of information could help us refine setting the stops: knowledge of the natural writer variability in the sampled population and knowledge of the misclassification propensity of the human or machine classifier having produced the dataset.

The former demands sufficient data for inferring a model. Sampling theory could be used to estimate how much exactly (one study estimated at 1,200 the minimal sampling number representative of the US writer population for forensic writer identification [79]). The real problem, however, is not one of quantity, as of sampling method, given the rudimentary knowledge of handwriting demographics [5: 185].

This issue leaves us unsure about if and when a variability model based on one dataset realistically models any other. For instance, the qualifier "Swiss" for the IAM dataset has to be understood in the light of what it really samples. The unsuspecting user might think it as representative of the Swiss population, while someone aware of the diversity characteristic of this tiny country, will approach it with more watchfulness. Indeed, the styles of scripts taught to children in Switzerland varies with canton, school, and even individual teacher [22]. To clarify the matter, the author conducted interviews with the dataset producers [54]. It appears that specimens were collected mostly from residents of canton Bern, friends and colleagues, their extended families (id 472–475), some foreigners, including one "Gorbatchev" (id 51) and "Zorro" himself (*sic*, id 607). Furthermore, sizable batches were written outside Switzerland, in Greece, France, and Germany, as well as a piece "written on the train" (id d05-021). The language of the text to copy, English, was with few exceptions not the first language of the participants, and, for some, even Latin was not their everyday script. Production was guided by a ruled page positioned under the writing form, adding to the graphonomical constraints. In conclusion, the dataset is heterogeneous in terms of participants origins, motivations, and acquisition settings. The question is then what value has its comparison with other existing handwriting collections, Dutch, Greek, American, Arabic, or Chinese?

Manual check — The stop ranking allows, in signal theory terms, to *detect* the presence of misclassified documents in writer classes; the visual inspection carried out in this step, *confirms* and *identifies* specific documents as such. Sometimes, clues demanding no handwriting expertise are enough to take a decision: signatures on the original documents of the IAM dataset weighted strongly in favor of handwriting samples being considered cuckoos, or not. Formal handwriting expertise makes, however, better informed results. In this case, the author is a trained paleographer and calligrapher. Table 2 lists handwriting features examined during the manual check, with their use frequency. Inspection was greatly facilitated by the tool at our disposal, the handwriting dataset visualizer Rex.

Reclassification — Reassigning discovered cuckoos to their genuine membership class could be considered an activity beyond quality control and, insofar as it is a identification task, a new start in the classification–verification loop. For automated methods, see the appropriate writer identification literature, for example [15].

Here, reclassification was performed semi-automatically, using Rex. We first ranked the IAM dataset by slant, which corresponds to the statistical mode of the orientation profile and is the most discriminant among the profile's features [7: 631]. Then, we searched in the neighborhood of each detected cuckoo for possible identical writers. None were found and we conclude that there are, in fact, 664 writers in the IAM dataset, instead of 657 as suggested by its metadata.

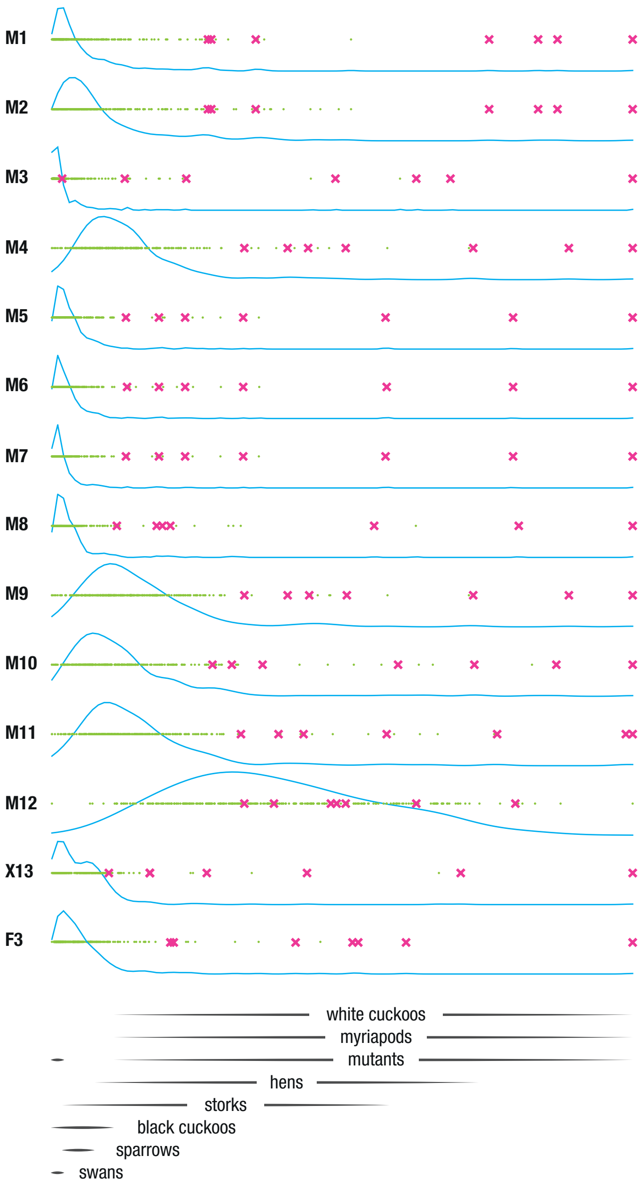


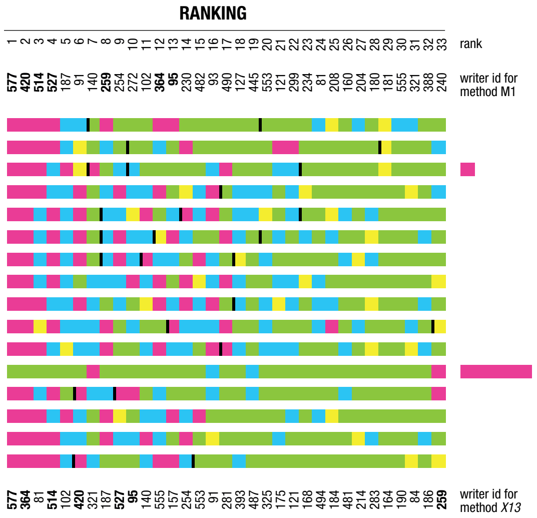
Figure 5. *The handwriting variability of the IAM dataset writers — The graphs represent the distribution of within writer variability according to the experimental methods listed in Fig. 6. The x-axis gives the variability magnitude for the 301 classes with at least two handwriting samples (green dots); the blue line is the probability density function of the distributions, normalized to fit the same area for illustrative purposes. Cuckoos are marked by red crosses. — At the bottom the approximate location of various demographic strata, according to the nomenclature of Table 1, which concur to give the distributions their particular shapes.*

Software — Image processing and statistical analysis implemented in the Alphonse system are custom functions written in the Matlab programming language. The interactive visualizer Rex is a web application written server-side in PHP and utilizing a MySQL database, and client-side in HTML, CSS, and JavaScript, with hyperlinks to the online IAM dataset.

METHODS

RANKING

PERFORMANCE



id	representa-tion domain	documents distance	class variability	run ranks			spread	precision	
				top	soft	hard	variance	absolute	stop-1
M1	**frequency**	**dynamic time warping**	**l2-norm**	**1**	**4**	**13**	2.90	.57	.57
M2	frequency	dynamic time warping	summation	**1**	**4**	22	8.22	.57	.57
M3	spatial	dynamic time warping	l2-norm	**1**	3	93	64.47	**.71**	.71
M4	spatial	earth mover's distance	summation	**1**	3	16	3.79	.57	**1.00**
M5	spatial	chi-square distance	summation	**1**	2	16	4.50	.57	.57
M6	spatial	chi-square distance	standard deviation	**1**	2	17	4.25	.57	.57
M7	spatial	chi-square distance	l2-norm	**1**	2	15	3.45	.57	.57
M8	spatial	cosine distance	summation	**1**	2	17	5.35	.43	**1.00**
M9	spatial	standard deviation	summation	**1**	3	16	4.61	.57	**1.00**
M10	spatial	standard deviation	standard deviation	**1**	2	25	8.34	.43	.57
M11	spatial	standard deviation	l2-norm	**1**	3	17	5.47	.43	.86
M12	spatial	l2-norm	standard deviation	7	1	169	488.30	.14	.71
X13	*spatial*	*chi-square*	*standard deviation*	**1**	2	33	9.47	.57	.43
F1	fusion	M1–11	Borda rank sum	**1**	3	15	3.55	.57	—
F2	fusion	M1–11, X13	Borda rank sum	**1**	**4**	14	2.93	.57	—
F3	**fusion**	**M1, X13**	**Borda magnitude sum**	**1**	**4**	**13**	**2.70**	**.71**	.57

● cuckoo ● myriapod ● mutant ● hen ❙ stop

Figure 6. *System performance — The figure shows the methods tested in this paper, with the data representation domain and metric names for measuring the distance between documents and the class variability; the best performing methods are in bold. Method X13 represents a third party system. The writer ids are given for method M1; note that they change across methods; cuckoos beyond rank 33 are aggregated left of the main bars. The writer typology is indicated by color codes; refer to Table 1 for definitions. Proposed check stops are marked by bars between ranks. In the performance table are given the extremities of the soft and hard runs in term of ranks. The spread is the ratio of observed and optimal rank variances of cuckold classes. The absolute precision indicates the number of cuckold classes observable within the top-7 classes, where the value "7" is given by the amount of genuine cuckold classes. The precision for the 1st stop is an estimation of how well the stops were defined, i.e. the ratio of cuckold classes found if stopping the manual verification at the proposed stop.*

5. Performance evaluation and insights

Performance criteria and indicators — A quality controller has to know in which order to check writer classes for misclassification and how many of them to check. The better a system responds to these questions, the better is its performance, hence the choice of the following evaluation criteria. (*1*) *Rank* of genuine cuckold writer classes in the system output, expressed as rank of the first genuine cuckold class (*top*), rank of the last class in the first consecutive run (*soft stop*), and rank of the last ever genuine cuckold class (*hard stop*). (*2*) *Rank spread* of cuckold classes $RS = \sum_{i=1}^{n} r_i^2 / \sum_{i=1}^{n} i^2$, were n is the number of genuine cuckold classes and r their rank. The closer RS is to 1, the better the system performs. The denominator normalizes the actual system performance by the maximal performance, attained when all cuckoos are contingent at the top of the ranking. The measure of rank spread, a ratio of variances with mean zero, palliates to the independence of the precision measures from the distribution of classes within the considered range. In a single value it makes a synthesis of the other performance measures given here. Note its sensitivity to outliers, which favors control methods yielding shorter hard stops and faster completion of the quality control task. (*3*) *Absolute precision*, indicating the number of cuckold classes observable within the top-7 classes, where the value "7" is given by the amount of genuine cuckold classes. (*4*) *Distinctiveness* of cuckold classes, conveyed by the distance between ranks (Fig. 5) and measurable by the *precision for the first stop*, an estimation of how well the stops were defined, i.e. the ratio of cuckold classes found if stopping the manual verification process at the proposed stop. (*5*) *Separability* between explainable and non explainable class variability (i.e. cuckoos, myriapods, and mutants vs hen writers). It is preferable to have explainable handwriting variability among classes in the top ranks of the output, because such classes take less resources to discard during verification. (*6*) *Difficulty* of the quality control task, the special topic to which Section 6 is devoted.

Ground truthing — Seven cuckoos were manually identified in the IAM dataset (either of the two documents of writer classes id 95, 259, 364, 420, 514, 527, and 577). The hypothesis that classes reported here are cuckolds was confirmed by two police forensic experts and two graphonomics computer scientists collaborating with the author (but see [31] for the reliability of human experts).

Inquiries with the dataset producers revealed how the errors came about. Initially, the IAM dataset was meant to serve the training of a handwriting recognition system, so recording the writer identity was secondary. After it became of interest, it was realized that writers do not systematically sign their samples, making necessary a manual post-hoc classification for part of the dataset. At that point, we surmise, appeared the cuckoos. Forms were distributed to participants according to the id sequence and the distance between cuckoos of same class in that system is 0, 0, 2, 3, 6, 16, and 268. This logical and, possibly physical proximity in the document stack, might have been a confusing factor during classification.

The success of our system at detecting misclassifications demonstrates the utility of an additional quality control layer, independent of the dataset creators. Likewise, it portrays the advantages of combining human and machine verification, supported by interactive visualization.

Results — Fig. 6 supplies numerical performance indicators and permits comparative visual analysis of the experimental methods. The best method is F3, which gives the longest soft stops, shortest hard stops, and lowest spread, closely followed by M1. As it can be seen, the system ranks classes containing misclassifications in reasonable stop ranges, given the amount of writers with high variability in this dataset. The quality of the performance can be even better apprehended when looking not at the rankings (Fig. 6), but at the magnitude distribution (Fig. 5), where the cuckoos clearly appear as outliers.

Insights — (a) *In a handwriting expertise system the script feature is not to always the most important factor.* Some features have, both for human and machine experts, stronger discriminative power than others, e.g. allographs vs run lengths [15, 16, 61, 3]. So we compared our system, which uses a good but not the best feature, contour orientation, with an award winning system (X13), based on the multiscale distribution of interest points in a handwriting sample, such as corners, junctions, and extremities ([32], best student paper at the International Graphonomics Society conference 2015). The fact that we fare better in performance, prompted the ensuing two remarks.

(b) *All elements of a handwriting expertise system are important.* Our experiments underscore the importance of the document distance metric to the expense of handwriting features. Additionally, the experiments M11 and M12 show that just by choosing different permutations of the same metric, performance can be significantly affected. Besides ranking, the order of classes is also modified.

(c) *Understanding the properties of the elements of a handwriting expertise system helps fine tuning its performance.* This truism is explained when we consider that even such a lowly feature as run lengths has its place in the expert's arsenal: it can model word spacing, what the mighty allographs can not (for examples, select "run lengths" in Rex). In respect to our experiments, the good performance of method M1 was achieved after realizing (α) that the circular nature of the orientation pdf allows for vector comparison in the frequency domain, which we knew to be a competitive shape distance measurement technique (the MPEG-7 multimedia standard incorporates frequency domain based shape description [58]), and (β) that translation invariance, a property of the frequency domain, has no adverse impact. A translation in the orientation profile corresponds to a change in handwriting slant [4: Fig. 5.1–5.2]. Due to the biomechanics of handwriting, slant is correlated to shear, which keeps profiles distinct (e.g., **O** → ***O***). Only mirrored styles will look the same to the frequency domain distance measures (**O** = ***O***), but most Latin script is right slanted [4: Fig. 4].

(d) *Importance of manual check.* Long tailed distributions, as exhibited by handwriting variation, have natural outliers [60, 10: 37–38], due to polystylistic writers and random text combinatorics. A more imaginative handwriting expertise than usual is required to establish their genuineness and separate them from cuckoos (Fig. 6), strengthening thereby the rationale for manual check.

(e) *System overfit.* Studying the similarities and disparities of the performance indicators, the question arises whether the system is not overfitting the data, and what is the part of randomness in the apparent fluctuation. One way to approach the question is *brute force* method evaluation with many systems, on many datasets. The comparison with a third party system, X13, alleviates overfitting apprehensions; while multiplying the tested datasets has the drawback of using handwriting samples about which it is unknown what exactly they represent within the wide variety of writer populations and writing conditions. Another way, presented in the next section, is by *simulation* of misclassifications in a controlled setting in a single dataset and studying the system's response.

(f) *Limits.* A "masking effect" between a cuckoo and members of its host class occurs when the stylistic *contrast* between handwritings is too low for the misclassification to be detected [10: 40] (Table 1). Such cases are difficult to sort out, but the issue is mitigated by the fact that it is more stringent to detect outlier misclassifications, since they reduce more strongly the accuracy of expertise systems, than exogenous items with low within class contrast.

(g) *Handwriting variability for both a single writer and a population appears to follow roughly a log-normal distribution.* It is well known that many types of rapid human motions, such as produced during handwriting and signature, can be modeled with log-normal distributions [66, 67, 24, 73]. Looking at Fig. 4b and 5, it is interesting to observe that, for entirely other reasons, this distribution type manifestly applies equally to within and inbetween writer variability of contour orientation. To test the hypothesis, we fitted the data to a log-normal and an exponential distribution [55], respectively, and performed an Anderson-Darling goodness-of-fit test between the empirical and fitted data [63]. The null hypothesis is supported for the log-normal, with the exception of methods M8 and X13, and rejected for the exponential. Visual confirmation is obtained from probability plots (Fig. 7). How can this phenomenon be explained and what are its implications?

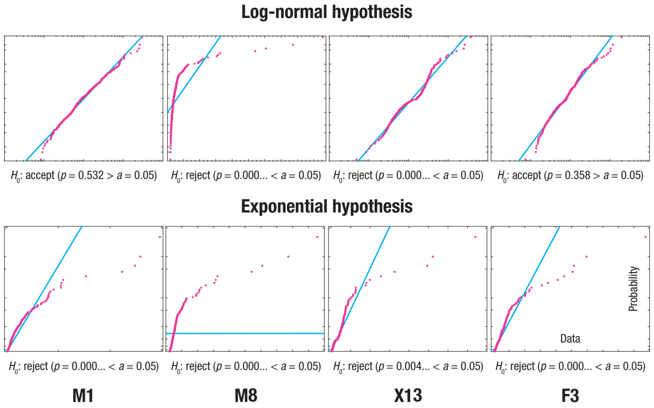


Figure 7. *The nature of the inbetween writer distribution — Probability plots of data from Fig. 5 tested against fitted log-normal and exponential distributions. The more the data dots fit the line, the better the adequacy between distributions.*

Writing by hand is similar to physical sports in that the writer strives to achieve and maintain biomechanical consistency. Assuming a statistical normal performance distribution, its left side is positively skewed away from zero mean variability because of the difficulty of near-perfect results, reserved to the – nowadays vanishing – elite of penmanship and calligraphers. Because it is easy to loose concentration on the quality of script while writing, especially long texts, a process of decay stretches the right side of the distribution. The two writer specific forces are supplemented by external ones. The left-hand skew is amplified because identical results can't be expected when the written text, the writing implements, and the contextual distractions perpetually change. Similarly, rare events such as deletions, introduce outliers on the right of the distribution. When considering now the handwriting variability in a population, its distribution shape is influenced by the mixture of demographic niches. While the mass of writers maintain a certain regularity in writing, there are as well individuals with extreme consistency, masters of multiple styles, adult learners of foreign scripts, or simply sloppy writers, sometimes involuntary subjects of random events, like writing produced on a train.

A practical implication of the log-normal handwriting variability distribution among a broad writer population is that in expertise applications the human or machine expert should be aware of the real possibility of genuine outliers [60]. At a theoretical level, the above insights allow the creation of models of handwriting variability. In respect to quality control, the log-normal distribution explains some aspects of our system performance, namely the ease of cuckoo detection (note in Fig. 5 that with little exception cuckoos are outliers) and their coexistence with genuine outliers in the top ranks (see the mix of color coded writer types in Fig. 6).

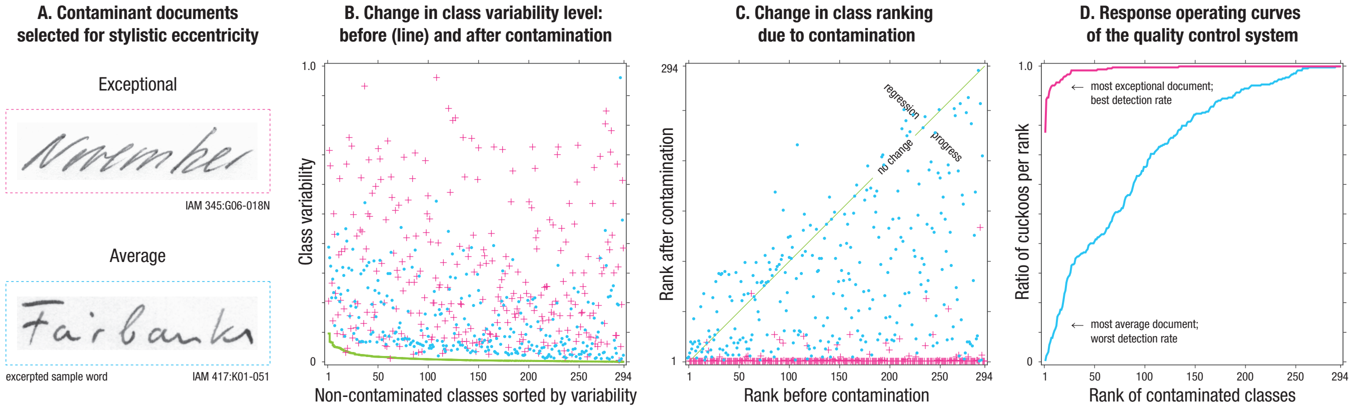


Figure 8. *Impact of dataset contamination with high and low contrast documents on quality control performance — Unusual documents (a, top) induce a high class variability (b) and push up class ranks (c), improving quality control performance (d) more than average documents (a, bottom). — The continuous line in the (b) diagram represents writer classes with more than two samples in the IAM dataset corrected for misclassifications and ranked by level of variability. The markers give the amount of within class variability in the same ranking order, after contamination with the documents shown in (a). Rearranging the classes according to the new variability, lets us compare the two distributions (c). When a class has "progressed" in rank, it means that the cuckoo it contains can be found more quickly. The degree of improvement in quality control performance can be judged by the vertical distance of markers form the diagonal. Clearly, contamination with highly individual samples produces far more detectable cuckoos (the markers are close to the ordinate). The analysis is confirmed by the response operating curves of the system, which compute the cumulative amount of cuckoos per rank using method M1 (d).*

6. Task difficulty and prediction

In the previous section we evaluated the quality control system Alphonse relative to an a priori unknown number of "wild cuckoos" occurring naturally in the IAM dataset. Would the performance be the same if the misclassified documents were other? In other words, *how difficult was the task*? To answer the question, controlled experiments of "artificial insemination" with known "laboratory cuckoos" will be carried out in the following. They will allow to investigate the possibility of a more generic performance measure, that proves the soundness of our concept beyond its realization in the specifics of a system and a dataset.

Performance in a coupled environment — We postulate that the performance of a quality control system can be fully understood only in reference to the task difficulty. To exemplify this, let us look at the dependence of quality control on dataset structure, by performing an *endogenous contamination* experiment with selected "Eastern Eggs" (Fig. 8). First, we retrieve the two stylistically most exceptional and average, respectively, handwriting samples, by finding, with method M1, the documents with the maximal and minimal sum of distances to all other documents in the dataset. They represent the best and the worst performance we can expect from our system on this dataset. Subsequently, these two Eastern Eggs are hidden in turn in each class except their own and the change in class variability and rank measured. We build a response operating curve (ROC) that for each rank k gives the precision p_k of the quality control as the number c of Eastern Eggs appearing in the top-k ranks weighted by the total number n of writer classes minus one: $p_k = (\sum_{i=1}^{k} c_i) / (n - 1)$. As expected, the performance curve of the average document is inferior to that of the exceptional document, confirming that contamination with high contrast documents is easier to detect than with low contrast items (Fig. 8d).

Please consider that the "exceptional document" is so from the vantage point of the measurement instrument. A human observer might have selected, for example, writer id 122 as the most idiosyncratic: its stretched intercharacter ligatures (included) are typical of Arabic script, not Latin.

Layered performance evaluation — Given the interdependence demonstrated above, the question arises about what exactly is evaluated: the system sensitivity, the data complication, or something else? If we consider the quality control system to be the articulation between data and user, then it is surrounded by a sequence of layers modulating its performance (Fig. 9). Data-side, the performance depends on the *dataset "complication"* (in the horology sense of the term), which is a result of a number of factors contributing to the emergence of misclassifications: dataset *producers excellence*, dataset *scrambling severity*, *writer variability*, and *text combinatorics*. User-side, performance is framed by the *stringency* of the use case, its *tolerance* range, and the *affordance* of the system and user to meet these requirements. (These factors determine the selection of check stops, i.e. how many classes are to be manually checked for cuckoos.) Finally, at the center of data–user interaction lies the *sensitivity* of the quality control system to detect misclassifications. The observed quality control performance fluctuates with all these parameters. In conclusion, *it seems necessary to distinguish between system sensitivity, data complication, and use case constraints, when addressing quality control performance*.

Relative grading — To grade a quality control system we would need a set of reference criteria. However, the class variability can vary considerably given the number of layers affecting it – especially due to the randomness of scrambling – and there is no ready list of criteria telling why a handwriting is more easy to expertise than another. A solution to the *first problem* proceeds by "peeling" away the contribution of each layer. After performing quality control we obtain a version of the dataset hopefully unbiased by scrambling. The amount of scrambling is indicative of the producers excellence, an information also derivable from the quality of other datasets of the same producer. The level of masking in the unbiased dataset characterizes the difficulty of any system to deal with this specific set of handwriting samples, and, in extenso, the difficulty created by a writer population, assuming that the sample is statistically representative. A solution to the *second problem* – grading performance without a reference scale – consists in comparing the performance

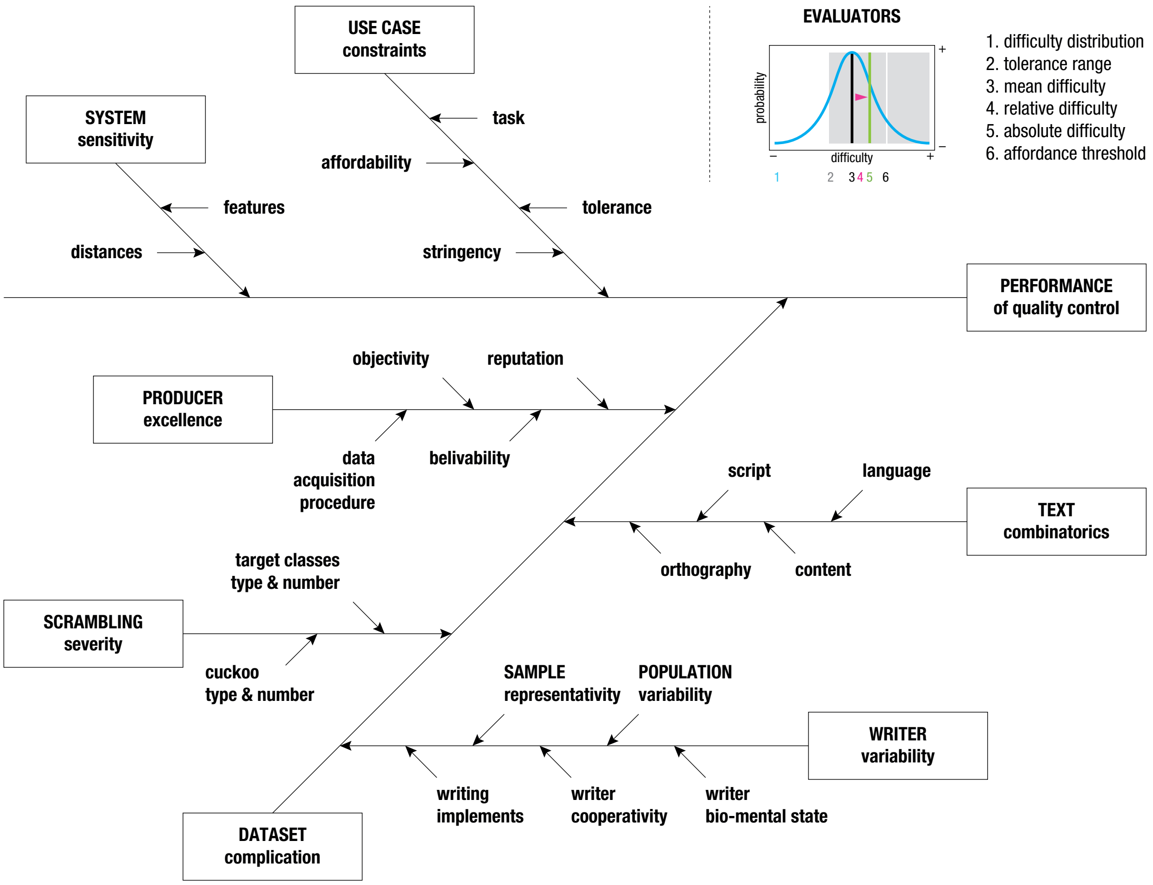


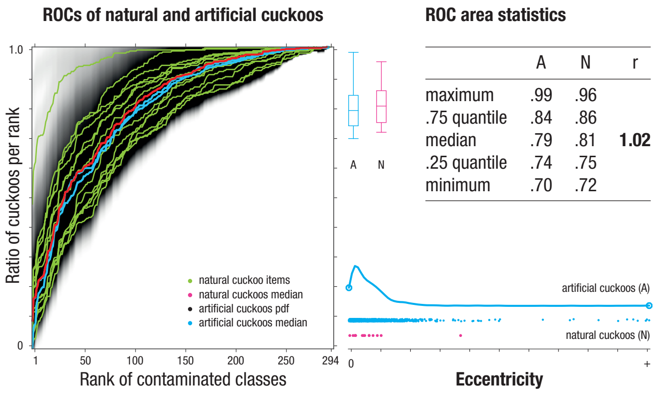
Figure 9. *Performance evaluation factors — This cause-and-effect diagram shows how quality control performance is affected by a multitude of factors. The concept of the inset diagram can be applied to each factor to evaluate its degree of "difficulty" (the data is fictitious). To evaluate, for example, the relative performance of a quality control system, we estimate the distribution of the performance of other systems (blue curve) and compute the distance (red arrow), in standard deviation units, between its mean (black line) and the absolute performance value of the evaluated system (green line). Its performance can also be evaluated in regard to whether it is within the use case tolerance range (gray area) and below the affordance threshold (white line).*

of the evaluated system with that of many others: a well-performing system is one that is at least as good as the majority of other systems, within the frame of the same set of data and use case parameters. Fig. 9 shows, for fictitious data, how the relative grading could be visually represented. Next we turn to obtaining numerical indicators for the scrambling severity and sample difficulty.

Prediction — What does it mean that a population or dataset is "prone to misclassifications"? When there is high in-class variability and low between-class variability, then cuckoos are easily masked by genuine outliers, humans and machines can easier make classification mistakes, and it is more difficult to establish good quality control stops. These characteristics create conditions favorable to the emergence of cuckoos. The question to consider is how to predict the misclassification proneness. We also have to keep in mind that this is a distinct question from the measurement and prediction of the dataset scrambling severity.

In order to evaluate the proneness for misclassification of a dataset unbiased by cuckoos, it would have to be scrambled in all possible ways [10: 37–38]. As an approximation, a sample could be drawn from the totality of permutations. For instance, we could extend the Eastern Eggs experiment by systematically contaminating *each* writer class with *each* item of the dataset, not just the most exceptional and most common (Fig. 10). The precision of quality control after this operation is inverse-proportional to the *masking potential of the sampled population* and represents the *absolute quality control performance* to expect from this system–dataset pairing. Because the method scrambles only one cuckoo egg at any given time, it produces an *optimistic* estimation: its value would decrease with an increase of contaminant documents and contaminated classes. If the performance of other systems is known, these values let us compute the *relative challenge to quality control*. Since outcomes depend on the selected script features and variability measurements, the same instruments have to be used for benchmarking the performance of different systems and difficulties of different dataset.

To compare two datasets *X* and *Y* we rely on the ratio *r* of the medians m_x and m_y, respectively, of the set of areas A_x and A_y under the ROCs: $r = m_x / m_y$. Depending on the nature of the compared datasets, a result greater than unit indicates (*1*) *a low scrambling severity* or (*2*) *a lower difficulty for quality control of dataset* X *against* Y. The IAM dataset appears to have both a medium scrambling severity (r = 1.02) and a medium structural difficulty (large spread of ROC curves) (Fig. 10). A look at the location of cuckoos on the centrality distribution in the same figure suggests that, with one exception, they are not outliers, that is, they could have occurred with relative ease.



	A	N	r
maximum	.99	.96	
.75 quantile	.84	.86	
median	.79	.81	**1.02**
.25 quantile	.74	.75	
minimum	.70	.72	

Figure 10. *Scrambling severity and dataset difficulty* — Left: *The grayscale bitmap represents the distribution of 1539 × 294 ROCs for the exhaustive scrambling of the IAM dataset (contamination by each document of all writer classes having more than two samples). It is obtained from pdfs of ROC values for each rank and informs us about the dataset difficulty. Superposed is the median ROC (blue) and the 14 ROCs of natural cuckoos (green) and their own median (red). By comparing the ROC distribution of artificial (A) and natural (N) cuckoos we can estimate the scrambling severity.* – Right top: *Statistics of the area under the ROCs.* — Right bottom: *Eccentricity of individual documents in respect to all others, as the sum of pairwise distances measured with method M1 (blue). The two documents in Fig. 9 are the most central and eccentric and coincide with the graph extremities. Cuckoos appear in the densest part of the distribution (red).*

SCRIPT	general	feature variability — feature use context — likelihood: converging indices are mutually reinforcing
	shape	allographs of characters — allographic sets (e.g. upper- & lower case) — size — vertical alignement lines (e.g. baseline, x-height): variability, proportions between lines — affine transform: slant, stretch — stroke: shape, weight, contrast, chroma — inline brokenness — outline irregularity — extremities shape — roundness — trace fluidity — ligaturing — spacing: characters, words; kerning, tracking — surface: texture, color — 3D imprint of writing implement — paragraph: margin alignment, indentation, interline spacing, columns amount & spacing — rivers — crossing-outs — hyphenation (rules, number of consecutive hyphens) — regularity — quality
	layout	shape outline — location — dynamism
	text	an abundance of personal names increases the number of upper case letters, modifying the script texture
CONTEXT	physical semantic	paper quality explains script quality word meaning explains its shape
LOGIC	heuristics likelihood	demographics explain script style converging indices are mutually reinforcing

Table 2. *Elements of human handwriting expertise — Lists of graphonomic expertise elements [41, 5: 128]; in red those effectively used by us for groundtruthing.*

7. Future research

Multiple features systems — What would it take to fully automate quality control? For handwriting recognition, the creators of the NIST dataset have convincingly argued in favor of semiautomatic solutions as being faster and more reliable. Their work brought them to the startling conclusion that "a quoted accuracy rate for a set of segmented characters is meaningless without reference to human performance on the same set of characters" [86: 1, 6–10]. To address the question for writer verification, we looked at the elements that went into the human expertise deployed for cleaning datasets infested with cuckoos.

We found that all decisions were based on a limited set of handwriting features and our decisions were strengthened when several indices concurred (Table 2). Only a single document could not have been possible to attribute to the correct writer given only sample dissimilarities, were it not for an identical signature on two forms (Fig. 3, id 555; a fact check revealed two homonymous students by that name at IAM, but from different points in time). Logic and contextual knowledge (e.g. on the dataset demographics) didn't play any role in our decisions (but it might do for other datasets and use cases). There were instances a writer using exclusively uppercase letters in one sample and mixed cases in another (Fig. 3, id 297), or a prevalence of some letters due to the content that made two samples by the same writer look dissimilar to our measuring instrument (Fig. 3, id 208). By reasoning we understood from where the disparity came, but the factors where not something not detectable by state of the art algorithms. *We conclude that it seems possible to substantially improve the quality of authorship control using a multidimensional handwriting features space.* Indeed, multiple classifier systems gave good results in other handwriting expertise domains, such as writer identification [44, 39], and had in this very paper positive impact on performance.

Probabilistic approach — The layered performance factors discussed in the last section and graphically represented in Fig. 9, strike as apt to be modeled by a probabilistic network. The probabilistic approach to quality control – following the Bayesian [45] or likelihood paradigms [72] – would indeed allow the evaluator to take into account, in a flexible way, the richness of factors that shape performance. Presently, however, the paucity of data and theory precludes such promises. For example, our knowledge of handwriting demographics is too limited to define priors and likelihoods, as is our ability to quantify the difficulty of various expertise tasks, such as quality control, verification, identification, classification, or retrieval.

In truth, this work does have a Bayesian twist. Our belief in the quality of the system's performance was updated to include information ("para-data") not present in the data measured by the contour orientation instrument or in the dataset metadata. First, the signatures revealed non-native Latin writers, whose high handwriting variability masked true cuckoo classes. This contributed to the necessity of developing a writer typology and prompted a reconsideration of how performance could be measured. Second, interviewing the dataset producers illuminated the unexplained complication of the IAM dataset and what to expect from it in terms of sample representativity. Last, the development of the quality control method itself, predicated by the potentialities of the various technical choices tributary to the subjective approach of the author and feeding a lengthy stream of preprint versions of this paper, was an iterative Bayesian optimization process, reflected by the performance of the proposed method. The above remarks are meant to recall that the psychology and sociology of scientific research are also part of the solution to an engineering problem.

8. Conclusion

We presented a generic method to detect mislabeled script samples in handwriting datasets and an exemplary software system for quality control of writer identities. The method is appropriate when no machine learning is desired or possible. The procedure's essence consists in an automated ranking of writer classes by within-class handwriting variability, given some statistical measure of spread, followed by interactive manual inspection of a number of classes deemed potential harborers of misclassifications. Experiments were conducted and misclassifications successfully detected; the performance of our system compares well with that of other. Performance is best evaluated considering system sensitivity, dataset complication, and use case requirements. Methods are discussed to measure and predict the difficulty of performing quality control on arbitrary datasets.

Acknowledgments

The dedication of the IAM dataset producers and their generosity in making it publicly available is what made our study possible. The few classification errors that were found are only a tribute to the quality of their work. — We wish to thank Marcel Würsch for sharing the document distance data and Andreas Fischer for fruitful discussions.

References

[1] J. Arlandis, J.C. Perez-Cortes, and J. Cano, "Rejection Strategies and Confidence Measures for a k-NN Classifier in an OCR Task," in *Proc. 16th Intl. Conf. on Pattern Recognition, Quebec, Canada, Aug. 11–15, 2002*, vol. 1, pp. 576–579.

[2] V. Atanasiu, *On letter frequencies and their influence on Arabic calligraphy* [*De la fréquence des lettres et de son influence en calligraphie arabe*], Paris: L'Harmattan, 1999.

[3] V. Atanasiu, "Allographic biometrics and behavior synthesis", *TUGboat* [*Proc. 14th European TEX Conference, Brest, France, June 24–7, 2003*], vol. 24 (3), pp. 998–1002, 2003.

[4] V. Atanasiu, "Forensic vs. Computing writing features as seen by Rex, the intuitive document retriever," in *Proc. First Intl. Workshop on Automated Forensic Handwriting Analysis, Beijing, China, September 17–18, 2011*, pp. 16–20.

[5] V. Atanasiu, *Expert Bytes. Computer Expertise in Forensic Documents: Players, Needs, Resources, and Image*, Boca Raton, FL: CRC Press, 2014.

[6] V. Atanasiu, "Alphonse: Handwriting Expertise Tools", *Matlab Central File Exchange*, software, 2015.11.25, http://www.mathworks.com/matlabcentral/fileexchange/?term=vlad+atanasiu.

[7] V. Atanasiu, L. Likforman-Sulem, and N. Vincent, "Writer Retrieval–Exploration of a Novel Biometric Scenario Using Perceptual Features Derived from Script Orientation," in *Proc. 11th Intl. Conf. on Document Analysis and Recognition, Beijing, China, Sep. 18–21, 2011*, pp. 628–632.

[8] V. Atanasiu, L. Likforman-Sulem, and N. Vincent, *Rex, a description-based retriever for written documents*, web application, April 2011, http://glyph.telecom-paristech.fr.

[9] H.S. Baird, "The State of the Art of Document Image Degradation Modeling," in B.B. Chaudhuri (ed.), *Digital Document Processing: Major Directions and Recent Advances*, New York: Springer, pp. 261–279, 2007.

[10] V. Barnett and T. Lewis, *Outliers in Statistical Data*, Chichester: John Wiley & Sons, 1978.

[11] C. Batini and M. Scannapieca, *Data Quality: Concepts, Methodologies and Techniques*, Berlin: Springer, 2006.

[12] J. van Beusekom, F. Shafait, and Th.M. Breuel, "Automated OCR Ground Truth Generation," in *Proc. 8th IAPR Workshop on Document Analysis Systems, Nara, Japan, September 16–19, 2008*, pp. 119–125.

[13] J.-Ch. Borda, "Mémoire sur les élections au scrutin," *Mémoires de l'Académie Royale des Sciences*, pp. 657–658, 1781.

[14] C.E. Brodley and M.A. Friedl, "Identifying Mislabeled Training Data", *J. of Artificial Intelligence Research*, vol. 11, pp. 133–169, 1999.

[15] M.L. Bulacu, "Statistical pattern recognition for automatic writer identification and verification," Ph.D. dissertation, Artificial Intelligence Institute, Univ. of Groningen, The Netherlands, 2007.

[16] M.L. Bulacu and L.R.B. Schomaker, "Text-independent writer identification and verification using textural and allographic features," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29 (4), pp. 701–717, 2007.

[17] M.C. Burl, U.M. Fayyad, P. Perona, and P. Smyth, "Automated Analysis of Radar Imagery of Venus: Handling Lack of Ground Truth," in *Proc. IEEE Intl. Conf. on Image Proc., Austin, TX, Nov. 13–16, 1994*, pp. 236–240.

[18] S.-H. Cha, "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions," *Intl. J. of Mathematical Models and Methods in Applied Sciences*, vol. 1 (4), pp. 300–307, 2007.

[19] S.-H. Cha and S.N. Srihari, "Multiple Feature Integration for Writer Verification," in *Proc. 7th Intl. Workshop on Frontiers in Handwriting Recognition, Amsterdam, The Netherlands, September 11–13, 2000*, pp. 333–342.

[20] Committee E-11 on Quality and Statistics, *Manual on Presentation of Data and Control Charts Analysis*, West Conshohocken, PA: ASTM. 2002.

[21] C. Cortes and V. Vapnik, "Support-Vector Networks", *Machine Learning*, vol. 20 (3), pp. 273–297, 1995.

[22] Deutschschweizer Erziehungsdirektoren-Konferenz, Arbeitsgruppe Schrift, *Entscheidungsgrundlagen zur Zukunft der Schweizer Schulschrift*, Luzern, 2013, http://www.basisschrift.ch/kantonale-informationen.

[23] M.M. Deza and E. Deza, *Encyclopedia of Distances*, New York: Springer, 2009.

[24] M. Djioua and R. Plamondon, "Studying the variability of handwriting patterns using the Kinematic Theory", *Human Movement Science*, vol. 28, pp. 588–601, 2009.

[25] G. Doddington, W. Ligget, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, Goats, Lambs and Wolves. A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation," in *Proc. Intl. Conf. Spoken Language Processing, Sydney, Australia, Nov. 30 – Dec. 4, 1998*, paper 0608.

[26] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, Hoboken (NJ): John Wiley & Sons, 2000.

[27] T. Dunstone and N. Yager, *Biometric System and data Analysis. Design, Evaluation, and Data Mining*, New York: Springer, 2009.

[28] I.L. Dryden and K.V. Mardia, *Statistical Shape Analysis*, Hoboken (NJ): John Wiley & Sons, 1998.

[29] G.A. Fink and Th. Plötz, "Unsupervised Estimation of Writing Style Models for Improved Unconstrained Off-line Handwriting Recognition," in *Proc. 10th Intl. Workshop on Frontiers in Handwriting Recognition, La Baule, France, Oct 23–26, 2006*, pp. 429–434.

[30] C.L. da Fontoura and R.M. Cesar Jr., *Shape Analysis and Classification: Theory and Practice*, Boca Raton, FL: CRC Press, 2000.

[31] B. Found and D. Rogers, "The Probative Character of Forensic Document Examiners' Identification and Elimination Opinions on Questioned Signatures," in *Proc. of the 13th Conf. of the Intl. Graphonomics Society, Melbourne, Australia, November 11–14, 2007*, pp. 171–174.

[32] A. Garz, M. Würsch, and R. Ingold. "Training- and Segmentation-Free Intuitive Writer Identification with Task-Adapted Interest Points," in *Proc. 17th Conf. of the International Graphonomics Society, Pointe-à-Pitre, Guadeloupe, June 21–24, 2015*, pp. 109–112.

[33] N. Ghoggali and F. Melgani, "Automatic Ground-Truth Validation With Genetic Algorithms for Multispectral Image Classification," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 47 (7), pp. 2172–2181, 2009.

[34] M. Gomez-Barrero, J. Galbally, J. Fierrez, J. Ortega-Garcia, and R. Plamondon, "Variations of handwritten signatures with time: A sigma-lognormal analysis", in *Proc. 6th Intl Conf. on Biometrics, Madrid, Spain, June 4–7, 201*3, pp. 1–6.

[35] P.J. Grother, *NIST Special Database 19. NIST Handprinted Forms and Characters Database*, NIST, technical report, Gaithersburg, MD, 1995.

[36] I. Guyon, N. Matić, and V. Vapnik, "Discovering Informative Patterns and Data Cleaning", in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.), *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: The MIT Press, pp. 181–203, 1996.

[37] Th.M. Ha, "Efficient Detection of Abnormalities in Large OCR Databases," in *Proc. 4th Intl. Conf. on Document Analysis and Recognition, Ulm, Germany, Aug. 18–20, 1997*, pp. 1006–1010.

[38] D.J. Hand, "Recent advances in error rate estimation," *Pattern Recognition Letters*, vol. 4, pp. 335–346, 1986.

[39] C. Hertel and H. Bunke, "A set of novel features for writer identification," in *Proc. 4th Intl. Conf. on Audio- and Video-Based Biometric Person Authentication, UK, Guildford, June 9–11, 2003*, pp. 679–687.

[40] K. Hildebrand, M. Gebauer, H. Hinrichs, and M. Mielke, *Daten- und Informationsqualität*, Wiesbaden: Vieweg+Teubner, 2011.

[41] R.A. Huber and A.M. Headrick, *Handwriting Identification: Facts and Fundamentals*, Boca Raton, FL: CRC Press.

[42] J.J. Hull, “Performance Evaluation for Document Analysis,” *Intl. J. of Imaging Systems and Technology*, vol. 7, pp. 357–362, 1996.

[43] I. Ivanov, F. Dufaux, Th.M. Ha, and T. Ebrahimi, “Towards Generic Detection of Unusual Events in Video Surveillance,” in *Proc. 6th IEEE Intl. Conf. on Advanced Video and Signal Based Surveillance, Genoa, Italy, Sep. 2–4, 2009*, pp. 69–74.

[44] R. Jain and D. Doermann, “Combining Local Features for Offline Writer Identification,” in *Proc. 14th Intl. Conf. on Frontiers in Handwriting Recognition,” Heraklion, Greece, Sept. 1–4, 2014*, pp. 583–588.

[45] F.V. Jensen and Th.D. Nielsen, *Bayesian Networks and Decision Graphs*, Berlin: Springer, 2007.

[46] J. Kittler *et al.* (eds.), *Multiple Classifier Systems Intl. Workshop*, Berlin: Springer, 2001–2015.

[47] R.C. Kraus, *Brushes with Power: Modern Politics and the Chinese Art of Calligraphy*, Los Angeles: University of California Press, 1991.

[48] H. Kume, *Statistical Methods for Quality Improvement*, Tokyo: The Association for Overseas Technical Scholarship, 1985.

[49] L. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, Hoboken, N.J.: John Wiley & Sons, 2014.

[50] N. Liyanage, “Misclassification bias in epidemiologic studies,” MS dissertation, Dept. of Mathematics and Statistics, McGill Univ., Montreal, QC, 1995.

[51] D. Lopresti, “Models and Algorithms for Duplicate Document Detection,” in *Proc. 5th Intl. Conf. on Document Analysis and Recognition, Bangalore, India, Sep. 20–22, 1999*, pp. 1006–1010.

[52] E. Marasco, A. Ross, and C. Sansone, “Predicting Identification Errors in a Multibiometric System Based on Ranks and Scores,” in *Proc. of 4th IEEE Intl. Conf. on Biometrics: Theory, Applications and Systems, Washington DC, Sep. 27–29, 2010*, pp. 1–6.

[53] U.-V. Marti and H. Bunke. “The IAM-database: an English sentence database for off-line handwriting recognition,” *Intl. J. on Document Analysis and Recognition*, vol. 5, pp. 39–46, 2002, http://www.iam.unibe.ch/fki/databases/iam-handwriting-database.

[54] U.-V. Marti, Bern, Switzerland; H. Bunke and V. Frinken, Institute of Computer Science and Applied Mathematics, Bern; M. Lewicki, German Research Center for Artificial Intelligence, Kaiserslautern, Germany; personal communications, March–April 2011, August, October 2015.

[55] MathWorks, “fitdist”, *Matlab Documentation*, 2015.11.24, http://ch.mathworks.com/help/stats/fitdist.html.

[56] N. Matić, I. Guyon, L. Bottou, J. Denker, and V. Vapnik, “Computer aided cleaning of large databases for character recognition”, in *Proc. 11th Intl. Conf. on Pattern Recognition, The Hague, The Netherlands, Aug. 30 – Sep. 3, 1992*, vol. 2, pp. 330–333.

[57] A. Maydanchik, *Data Quality Assessment*, Bradley Beach, NJ: Technics Publications, 2007.

[58] F. Mokhtarian and M. Bober, *Curvature Scale Space Representation: Theory, Applications and MPEG-7 Standardization*, Dordrecht: Kluwer Academic Publisher, 2003.

[59] A. Monadjemi, “Towards Efficient Texture Classification and Abnormality Detection,” Ph.D. dissertation, Univ. of Bristol, Bristol, UK, 2004.

[60] J. Neyman and E.L. Scott, “Outlier Proneness of Phenomena and of Related Distributions”, in J.S. Rustagi (ed.), *Optimizing Methods in Statistics*, New York: Academic Press, pp. 413–430, 1971.

[61] R. Niels, “Allograph based writer identification, handwriting analysis and character recognition,” PhD dissertation, Donders Inst. for Brain, Cognition and Behaviour, Radboud Univ. Nijmegen, The Netherlands, 2010.

[62] V. Niennattrakul, Ch.A. Ratanamahatana, and E. Keogh, “Data Editing Techniques to Allow the Application of Distance-Based Outlier Detection to Streams,” in *Proc. Intl. Conf. on Data Mining, Sydney, Australia, Dec. 13–17, 2010*, pp. 947–952.

[63] NIST/SEMATECH, “Anderson-Darling Test”, *e-Handbook of Statistical Methods*, 2015.11.24, http://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm.

[64] J. Oakland, *Statistical Process Control*, Amsterdam: Butterworth-Heineman, 2008.

[65] W.S. Peters, *Counting for Something. Statistical Principles and Personalities*, New York: Springer, pp. 159–170, 1987.

[66] R. Plamondon, “A kinematic theory of rapid human movements: Part I. Movement representation and generation”, *Biological Cybernetics*, vol. 72, pp. 295–307, 1995.

[67] R. Plamondon, Ch. O’Reilly, C. Rémi, and T. Duval, “The lognormal handwriter: learning, performing, and declining”, *Frontiers in Psychology*, vol. 4, art. 945, 2013.

[68] S. Pletschacher and A. Antonacopoulos, “The PAGE (Page Analysis and Ground-truth Elements) Format Framework,” in *Proc. 20th Intl Conf. on Pattern Recognition, Istanbul, Turkey, Aug. 23–26, 2010*, pp. 257–260.

[69] L. Rokach, *Pattern classification using ensemble methods*, Singapore: World Publishing, 2010.

[70] A.R. Ross, K. Nandakumar, and A.K. Jain, *Handbook of Multibiometrics*, Berlin: Springer, 2006.

[71] D.K. Rossmo, *Geographic Profiling*, Boca Raton, FL: CRC Press, 2000.

[72] R. Royall, *Statistical Evidence: A likelihood paradigm*, Boca Raton, FL: Chapman & Hall / CRC Press, 1997.

[73] V.I. Rupasov, M.A. Lebedev, J.S. Erlichman, M. Linderman, “Neuronal Variability during Handwriting: Lognormal Distribution”, *PLoS One*, vol. 7 (4), art. e34759, 2012.

[74] Sh. Sadiq (ed.), *Handbook of Data Quality Research and Practice*, Berlin: Springer, 2013.

[75] L. Schomaker and L. Vuurpijl, *iUF Firemaker: A benchmark dataset for writer identification*, technical report, Nijmegen: Nijmegen Institute for Cognition and Information, Univ. of Nijmegen, 2000.

[76] W. Shi, *Principles of Modeling Uncertainties in Spatial Data and Spatial Analyses*, Boca Raton, FL: CRC Press, 2010.

[77] I. Siddiqi and N. Vincent, “Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features,” *Pattern Recognition*, vol. 43, pp. 3853–3865, 2010.

[78] T. Simonite, “Teaching Machines to Understand Us”, *MIT Technology Review*, vol. 120 (5), pp. 72–79, 2015.

[79] S.N. Srihari *et al.*, “Individuality of Handwriting,” *J. of Forensic Sciences*, vol. 47, pp. 856–872, 2002.

[80] J. Subrahmonia, “Similarity Measures For Writer Clustering,” in *Proc. 7th Intl. Workshop on Frontiers in Handwriting Recognition, Amsterdam, The Netherlands, September 11–13, 2000*, pp. 541–546.

[81] V. Torra and Y. Narukawa, *Modeling Decisions: Information Fusion and Aggregation Operators*, Berlin: Springer, 2007.

[82] G.T. Toussaint, “Bibliography on Estimation of Misclassification,” *IEEE Trans. on Information Theory*, vol. 20 (4), pp. 472–479, 1974.

[83] R.Y. Wang and D.M. Strong, “Beyond Accuracy: What Data Quality Means to Data Consumers,” *J. of Management Information Systems*, vol. 12 (4), pp. 5–34, 1996.

[84] Wikipedia contributors, “Roman cursive”, *Wikipedia*, 2015.11.24, https://en.wikipedia.org/wiki/Roman_cursive.

[85] Wiktionary contributors, “zampa di gallina”, *Wiktionary*, 2015.11.28, https://en.wiktionary.org/wiki/zampa_di_gallina.

[86] R.A. Wilkinson, M.D. Garris, and J. Geist, “Machine-Assisted Human Classification of Segmented Characters for OCR Testing and Training,” in *Proc. SPIE Conf. Character Recognition Technologies, San Jose, CA, 1 Feb. 1993*, vol. 1906, pp. 208–217.

[87] R.A. Wilkinson *et al.*, *The First Census Optical Character Recognition Systems Conference*, NIST, technical report, Gaithersburg, MD, 1992.

[88] Zeiss, *Did You Know... that up to 21 People Can Look Through a Microscope at the Same Time?*, webpage, http://www.zeiss.com/corporate/en_de/about-zeiss/news/did-you-know/58nd-edition-of--did-you-know----.html.

Author Biography

Vlad Atanasiu is doctoral candidate in computer science at the University of Fribourg, Switzerland. He holds a PhD in paleography (ÉPHÉ, Paris, 2003) and an MA in Arabic linguistics (University of Provence, 1995). He was postdoc in the Architecture and Cognitive Science Departments at MIT, worked on applications to Digital Humanities of image processing, visualization, and geographical information systems at the Austrian Academy of Sciences, and on handwriting analysis at Telecom ParisTech. atanasiu@alum.mit.edu, http://alum.mit.edu/www/atanasiu/

Supplementum

On *data complexity* in pattern recognition, including handwriting analysis, the reader might want to consult the rich collection of articles in M. Basu and T.K. Ho (eds.), *Data Complexity in Pattern Recognition*, London: Springer, 2006. [Many thanks to George Nagy for pointing out this reference.]

A further *contribution to misclassification research*, in the context of data streams, is M.-R. Bouguelia, Y. Belaïd, and A. Belaïd, "Stream-based Active Learning in the Presence of Label Noise", in *Proc. 4th Intl Conf. on Pattern Recognition Applications and Methods, Lisbon, Portugal, Jan. 10–12, 2015*, pp. 25–34. Inspiration on the study of misclassification can be gathered also from the topic of *rare categories*: Jingrui He, *Analysis of Rare Categories*, Berlin: Springer, 2012.

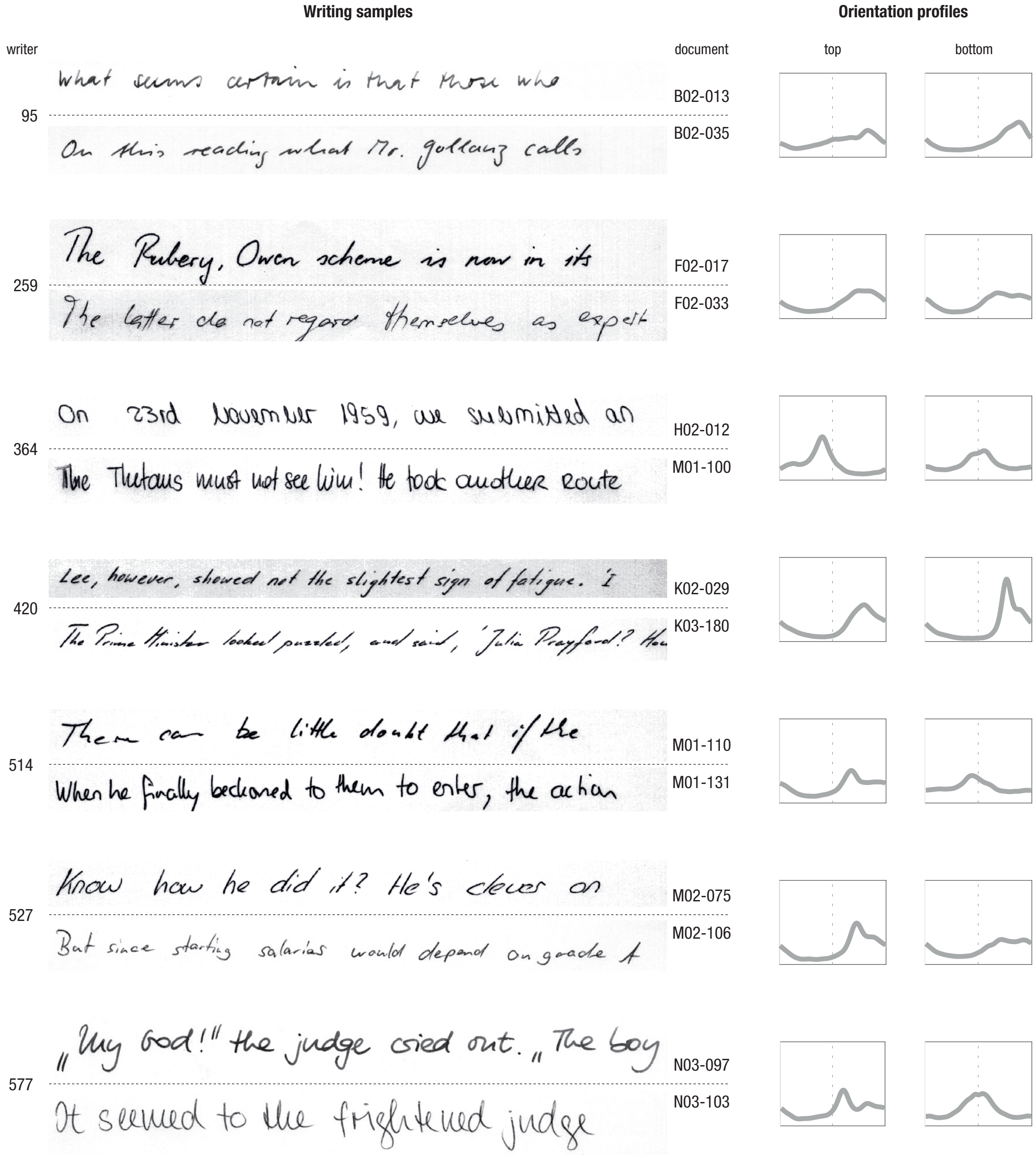
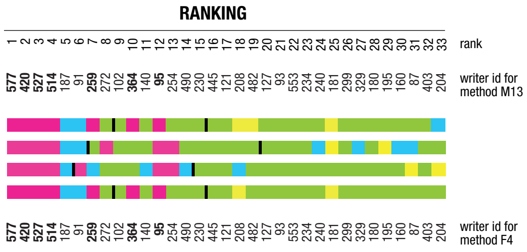


Figure S1. *The Seven IAM Cuckoos — Complete list of the cuckoos discovered in the IAM dataset, with sample handwritings of each.*

METHODS				PERFORMANCE					
				run ranks			spread	precision	
id	representa-tion domain	documents distance	class variability	top	soft	hard	variance	absolute	stop-1
M13	**frequency** [2]	**dynamic time warping**	**l2-norm**	**1**	**5**	**12**	**2.30**	**.71**	.71
M1	frequency [1]	dynamic time warping	l2-norm	1	4	13	2.90	.57	.57
F3	fusion	M1, X13	Borda magnitude sum	1	4	13	2.70	**.71**	.57
F4	**fusion**	**M13, X13**	**Borda magnitude sum**	**1**	**5**	**12**	**2.30**	**.71**	.71



● cuckoo ● myriapod ● mutant ● hen ▎ stop

Figure S2. *Improved system performance — (1) Magnitude of Fourier components. (2) Magnitudes weighted inverse proportionally to their frequency.*

The best method to measure the distance between two vectors was based on the magnitude of these vectors in the frequency domain (Fig. 6). Fig. S2 shows an improved method (M13), implementing a low-pass filter in which the magnitude is weighted by 1/*f*, were *f* represents the frequency associated to the indices of the magnitude vector. Let us note that the phase distortion between vectors produced much worse results (calculated as sum of the arclengths of the circular distance between phase vectors).

A recent addition to metaphoric nomenclatures comes from political sciences and concerns the voters bestiary (Jason Brennan, *Against Democracy*, Princeton: Princeton University Press, 2016, pp. 4–5); older ones concern typologies of museum visitors, teacher attitudes towards high school dropouts, and hikers' attitudes towards outdoors travel (Michael Q. Patton, *Qualitative Evaluation Methods*, Beverly Hills, CA: Sage, 1980, p. 460, 469, 458). A host of individual metaphorical terms related to statistics exist, some having become popular idioms: "butterfly effect", "black swan", "gray rhinoceros", "dragon kings", "the ostrich effect", even "the elephant in the room", "the skeleton in the closet".