A Question of Confidence: Using OCR Technology for Script analysis

Antonia Karaisl, Waseda University

Abstract

The following article proposes a method employing the Tesseract OCR engine to aid palaeographic analysis and scribal identification. Repurposing the so-called confidence score provided by the OCR engine, different methods of visualization are used to surface differences between font families, script types and manuscript hands.

1 Introduction

This paper introduces a simple method for conducting technology-assisted analysis of script and handwriting styles in printed books and manuscripts. The approach described uses a side product of a tried and tested technology: Optical Character Recognition (OCR). OCR software is traditionally employed for the automatic transcription of text from a digital image to machine-readable output. The method used here largely ignores the transcription but focuses on the so-called confidence scores. Confidence scores are usually employed in the process of OCR recognition to assess the probability that the output is correct. Normally, low-scoring results are undesirable as they signal a lower probability of accuracy. In this case, however, low scores will be used to identify pages, words or characters that could be of interest for a palaeographic analysis.

Digital methods, including Artificial Intelligence (AI), have been plied before to the field of palaeography, for example for the purpose of automatic transcription, the classification of writing styles or scribal identification (see e.g. Camps 2014; Castro Correa 2014; Christlein 2018; Cilia et al., 2019). A contest organized in 2017 by the Fifteenth International Conference on Frontiers in Handwriting Recognition in 2017, for example, solicited AI-based solutions for the classification of medieval script types, providing a set of labelled training material ("ICDAR2017 Competition on the Classification of Medieval Handwritings in Latin Script" 2017). A report submitted by Kestemont et al. discusses the efficacy of several submissions and remarks that the premise of the task itself builds on a simplified reality. Medieval script types do not always have firm boundaries, that is, definitive sets of features that reliably set one type apart from another; some hybrid forms are not easily described with one single label. The categorization of medieval script types often moves in grey zones and a technology trained on human-labelled script types is therefore not automatically free from human bias. Conversely, unsupervised learning, which does not rely on labelled data, does not necessarily sort material in ways that are meaningful to scholars and can also be hard to interpret (Kestemont, Christlein, and Stutzmann 2017, 104–7). Thus, whilst AI shows potential for palaeographic analysis, some caution is warranted: worst case scenario, human subjectivity is replaced by AI's accountability gap.

The approach introduced here is not meaning to replace human expertise with an automated solution. Rather, it attempts to re-purpose a pre-existing technology as a heuristic tool that can accelerate the palaeographer's, book historian's, bibliophile's quest for areas of interest in a book or on a page, with the help of OCR confidence scores. The paper mainly showcases different modes of visualizing confidence metrics to aid the discovery of palaeographic phenomena. The following argument will briefly introduce the metrics of word and character confidence and how they can be employed in script or scribal analysis. Proposed approach employs confidence scores to identify divergences in script style or abnormalities in letter shapes on book- or page level. The elements identified are then inspected with the help of statistics to establish whether they are of significance. Throughout, the open-source OCR

Proceedings of the Joint 3rd NLP4DH and 8th IWCLUL, pages 162–171
December 1-3, 2023 ©2023 Association for Computational Linguistics

engine Tesseract (LSTM version) is used with different OCR models.[1]

2 Visualizing confidence scores on printed and handwritten material

2.1 Confidence in Theory

In OCR (Optical Character Recognition, typically used for printed text) and HTR (Handwritten Text Recognition) technology, confidence scores are usually procured to aid decision-making calls on transcriptions during the recognition process. By definition, the confidence score marks the probability with which the OCR engine deems the transcription of a character or a word to be correct. Thus, lower confidence scores signal lower probability of accuracy. Confidence is not an absolute measure, and it is possible that low confidence scores can accompany accurate output, or high-confidence scores inaccurate transcriptions. For example, an OCR model trained on predominantly English texts will recognize the same font in a French-language document with fair accuracy but might produce low confidence scores on account of its English-language training. Conversely, a high confidence score does not always guarantee an accurate result – merely the engine's assessment that given the model's parameters the output is accurate.

Tesseract can produce confidence scores at word level and character level. The calculation of the respective confidence scores is complex but understanding the context of their generation can help gauging the underlying parameters. Tesseract's documentation for the current, neural-network-based version (4.0.0 and higher) does not contain any information on the calculation of the confidence score. The documentation for Tesseract's Legacy engine (version 3.0.0 and lower), however, specifies the circumstances and formula for calculating character and word confidence in the context of character classification. When processing new input with the Legacy Engine, Tesseract performs the segmentation of a text image into lines and words down to individual characters. Each segmented character is then classified by mapping it to the closest-matching prototype. The shape of the character is described by a visual feature vector combining a number of 3-dimensional features mapping the character's outline; the distance between the recognized character's visual feature vector and that of the closest matching prototype is then used to calculate the character's confidence (Perveen, n.d.; Smith 2007). On a word-level, the confidence of the lowest-scoring character doubles up as the confidence score for the entire word. (Tesseract Documentation FAQ). If the word formed from the recognized characters turns out highly improbable or linguistically implausible, the Legacy engine tries to re-segment the characters in different ways to see whether a more satisfactory solution can be found (Smith 2007). Output of the Legacy engine expresses character confidence as a percentage; the percentages for all character suggestions for one symbol stand independently and do not necessarily add up to 100%.[2]

In 2016, Tesseract's system was upgraded to include recurrent neural networks with LSTM (Long-Short-Term-Memory). The advantage of such a network is its capacity for context-aware processing, particularly with LSTM, resulting in lower error rates (Ul-Hasan et al 2013). Tesseract's LSTM implementation was adapted from OCRopus, an OCR system based on convolutional neural networks (Smith 2016). The novelty of OCRopus' initial design vis a vis Tesseract's concurrent version was documented at its inception in 2008. In contrast to Tesseract's Legacy engine, OCRopus' recognition process does not segment a text image into separate, single characters, but takes words as base units: proceeding sequentially across an identified word in so-called timesteps, the string is oversegmented – meaning, the word is not chopped into a discrete set of characters, but each timestep presents a separate segmentation attempt for part of the word. Each of these segments presents a character hypothesis; each character hypothesis is assigned a probability that it presents a valid character, and assuming that it is, the "posterior probability" (or confidence) for each character class, based on image features. Recognition results and the relationships between each potential character are expressed in a graph structure, from which the best sequence

[1] Tesseract models and test books used in this study are listed in the appendix.

[2] As noted in the commentary to Tesseract's code, see https://github.com/tesseract-ocr/tesseract/blob/7c178276d78fc4d2e55d531563275fd9631a72fb/src/ccmain/ltrresultiterator.cpp#L458

representing the whole word is then identified, using statistical language modelling (Breuel 2008).

OCRopus' LSTM system is integrated in Tesseract 4.0.0 and later. The last layer of Tesseract's network moreover contains a so-called softmax classifier which normalizes the confidence score for each character before the final output. (Smith 2016). Where Tesseract's Legacy engine yields the confidence as the original percentage in the final output, the LSTM-based Tesseract engine normalizes the confidence score for the chosen character so that the confidences for all classification attempts for one character approximately add up to 1.[3] The final confidence score is amplified through this normalization process, with the result that for the LSTM engine, the character confidences usually move within a band between 1 and 0.90. The softmax normalization only applies to the character confidence, not the word confidence, which fluctuates between 0 and 100%.

Tesseract's OCR models utilize neural networks trained on a pool of so-called ground truth, that is a quantity of labelled material. In the case of OCR, this material consists of image files of text lines, matched with transcriptions saved in simple text files. Throughout the training process, the neural network of the OCR engine iterates over this ground truth, producing transcriptions and evaluating their accuracy against the ground truth labels. Each iteration produces a model which is assessed on its word and character error rate with a separate pool of ground truth. At the end of the process, the model with the highest accuracy is chosen and can then be used with the OCR engine for the automatic transcription of related image material.

The ground truth pool used for training is the key parameter determining the model's capacity to interpret real material – and strategizing on its size and composition is crucial for an effective OCR strategy. In most cases, the aim is to train a model that is specific enough to perform well within its context but capable of generalizing beyond the ground truth pool. Within certain boundaries – say, a language, an alphabet, or a font group – diversity within a ground truth pool can improve this the model's ability to generalize. Too little specialization means the OCR model does not work well within its context. Yet if the ground truth pool is too small or not sufficiently diversified, so-called overfitting occurs: the neural networks' recognition capacity is over-adjusted to its training material. (Kestemont et al. 2017, 97). Since the approach introduced here is less interested in transcription than analysis, some of the models used are deliberately overfitted in order to understand how they can signal affinity or difference between a very small and specific training pool and the test material.

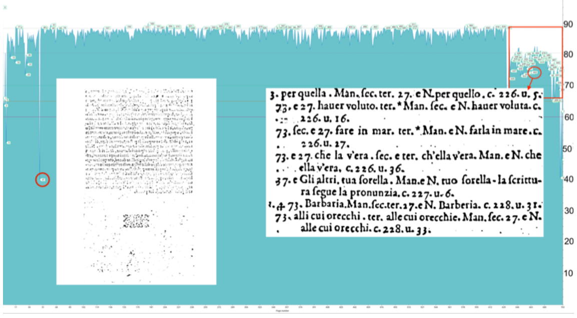
As said, the recognition process employs confidence scores to assess the probability that an output is correct. The confidence scores not only highlight potential good or bad transcriptions; they communicate the efficacy of the chosen OCR model with regard to the input material – and by extension, the affinity of the material it was trained on with the material it is used on. As a consequence, characters underrepresented in a LSTM training set tend to be misclassified in transcription (Ul-Hasan et al 2013). We would expect, therefore, that atypical glyphs or letter shapes not included in the training material are likely to be badly transcribed and consequently flagged up by low confidence scores when running the OCR model over a text image. This is the assumption that the following experiment is seeking to corroborate and utilize.

The caveat to this approach is that the confidence score is an uncertain metric, which is not only affected by the character's shape, but also by factors such as image quality, skewing, or discolorations on the page. With an LSTM-based system, moreover, the impact of context on the confidence score remains difficult to gauge. Overall, therefore, confidence is too complex a metric to serve as an absolute indicator; what the following experiment means to show empirically, however, is that using tools to visualize confidence metrics can still help to identify areas of palaeographic interest on book or page level.

2.2 Confidence in Practice

When analysing the OCR output across a whole book, the word confidence score can help identifying problem areas across the full range using the the average word confidence score for each page in a whole book. Sections where the

[3] See the link to the code commentary in Footnote 2.



3. per quella . Man. ſec. ter. 27. e N. per quello . c. 226. u. 5.
73. e 27. hauer voluto. ter.* Man. ſec. e N. hauer voluta. c. 226. u. 16.
73. ſec. e 27. fare in mar. ter.* Man. e N. farla in mare. c. 226. u. 27.
73. e 27. che la v'era . ſec. e ter. ch'ella v'era. Man. e N. che ella v'era. c. 226. u. 36.
37. e Gli altri. tua ſorella . Man. e N. tuo ſorella - la ſcrittura ſegue la pronunzia. c. 227. u. 6.
4. 73. Barbaria. Man. ſec. ter. 27. e N. Barberia. c. 228. u. 31.
73. alli cui orecchi . ter. alle cui orecchie. Man. ſec. 27. e N. alle cui orecchi. c. 228. u. 33.

Figure 1: Word confidence graph, book-level

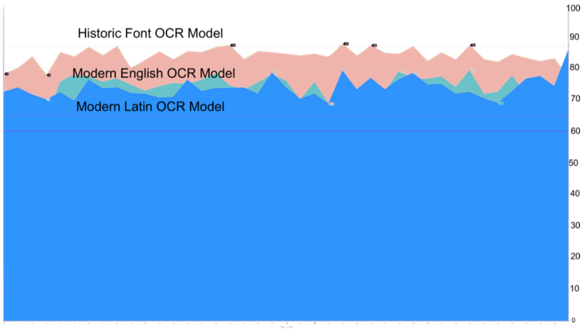


Figure 2: Word confidence graph for different Tesseract models, run over historic printed book

scores move in a narrow, consistent band tend to be transcribed with fair accuracy. Strong fluctuations or exceptionally low confidence scores typically denote pages that differ in some way from the rest. In the word confidence graph for Giovanni Boccaccio's Decameron (see Figure 1), for example, the egregiously low score in the beginning section (circled in red) corresponds with a blank page with ink bleed-through misinterpreted by the OCR model as writing. The finishing section of the same graph, too, shows up with consistently lower scores. Looking at a sample page from this section, it turns out that it contains the book's appendix, which features uncommon symbols, more numbers than usual and truncated words scoring low in confidence, regardless of whether they are transcribed correctly or not (see Figure 1). These particular examples might or might not be of concern. Rather, the point is that outlier pages in OCR output can be identified from the top level with the help of a confidence graph. Beyond the identification of outlier pages, however, there is no further indication what the issue could be in each case – the page image itself needs to be considered to understand what the cause of the low confidence score might be.

Most of Tesseract's standard OCR models are ostensibly trained for specific languages, on modern alphabets (see available list on Tesseract Github). When using these standard models, we would expect low confidence scores (independently from actual accuracy) where an OCR model is plied outside its "comfort zone", so to speak, e.g. to a text containing unknown or underrepresented characters, written in a different language or in a radically unusual font. Granted there is rarely much information about the training material used to train the model (and granted it tends to be too copious for a close review), it is not immediately obvious where such a "comfort zone" starts or ends. Running the model over different kinds of materials and comparing the confidence score, however, can help to get a bearing of the model's capacities. Conversely, running several models over the same material provides some context within which to judge each model's "comfort" and "discomfort zone".

Tesseract's standard models are very effective for modern printed text, and particularly the English language model shows a high performance for many different types of fonts, including typewritten texts. Viewed from the point of methodology, therefore, we should assume that a digitized text's language would be the main parameter to affect confidence scores when it comes to OCR processing. Manuscripts, modern or old, tend to fare badly with these standard models. This is not particularly surprising, granted handwriting tends to be much more irregular than print. Yet experiments with historic printed text, too, show that the OCR models trained for modern languages struggle with such material – as opposed to OCR models trained on a mix of languages, but on historic printed text.

In the following experiment, a set of pages from a Latin publication printed in 1475 were processed with three Tesseract models: the first one trained on English language material printed in modern fonts; the second on Latin language material printed in modern fonts; and the last one trained on material printed between 1500-1800, in Latin, English and French. The accuracy of the transcription was not considered in this experiment; only the confidence scores were assessed in order to understand each model's "comfort" or "discomfort" with historic material printed in Latin. Comparing the graph mapping the average page confidence (see Figure 2), we see that the models for modern English and modern Latin roughly play out in the same

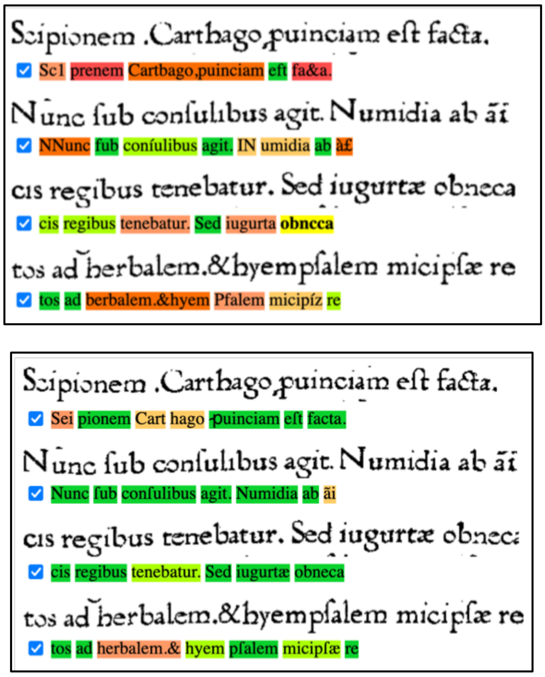

Sc1 prenem Cartbago,puinciam eft fa&a.
NNunc fub confulibus agit. IN umidia ab a£
cis regibus tenebatur. Sed iugurta obncca
tos ad berbalem.&hyem Pfalem micipfz re
Sei pionem Cart hago puinciam eft facta.
Nunc fub confulibus agit. Numidia ab ãi
cis regibus tenebatur. Sed iugurtæ obneca
tos ad herbalem.& hyem pfalem micipfæ re

Figure 3: Word confidence heatmap for Modern Latin model (top) and Historic font model (bottom)

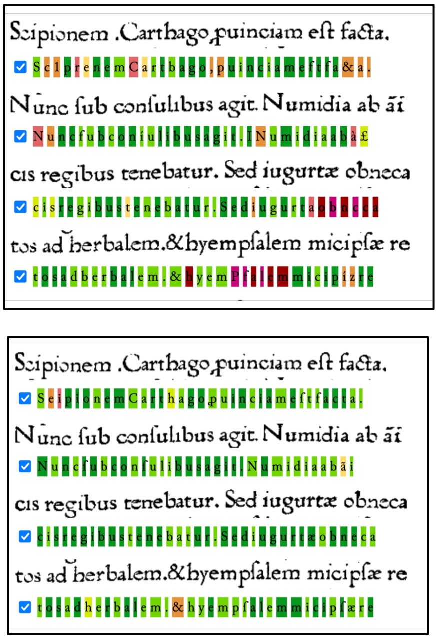


Figure 4: Character confidence heatmaps for Modern Latin model (top) and Historic font model (bottom)

confidence band – and neither one scores consistently higher than the other. Where we might have expected the Latin model to fare better than the English model, the confidence graphs do not hint at a significant nor even a consistent difference. Compared to the output of these two modern models, a Tesseract model trained on historic fonts from books printed in Latin, French and English, shows consistently higher confidence scores, no matter the lack of language focus.[4] This suggests that in this case, the historic font as a parameter in the training material has more impact on the OCR model's word confidence scoring than the focus on language.

In order to test this assumption in more detail, the word confidence scores for the modern Latin model and the historic font model were visualized next to the corresponding text lines using confidence heatmaps: in percentage blocks of 10, descending from 100%, word confidence scores were highlighted in colours shifting from green to red, the former signalling higher and the latter lower scoring (see Figure 3)

Comparing the two heatmaps, the historic font OCR model fares a lot better than the modern Latin OCR model. Yet the isolated problem areas in the historic OCR model map also put the low scores of the modern Latin model into context. Both models seem to struggle with unusual spacing typical for historic printed material, although in different ways. The modern Latin OCR model, moreover, assigns a low confidence score to a lot more words – most, though not all, are incorrectly transcribed or truncated.

In the next step, heatmaps were created to visualize character confidences (see Figure 4). As explained above, the character confidence output from Tesseract's LSTM model is normalized and moves within a smaller band than word confidence. The colour gradation changed per percentage point, from 99 downwards. When looking at the output from the two models on character confidence heatmaps, we can see more clearly where whole words are scoring low on account of single characters, and where whole groupings of letters are affected – and conversely, where unusual spacing rather than low-confidence characters lead to a low word confidence score. The character confidence heatmap for the Historic font model, for example, shows that all letters in "Carthago" are transcribed at relatively high confidence; the low

[4] The last page where the modern Latin model scores highest seems to contradict this; however, the actual page image does not contain any text and can be ignored in this case.

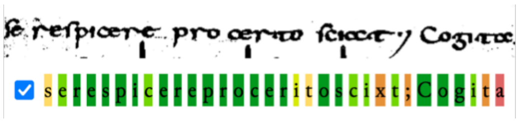


Figure 5: Confidence heatmap for Tesseract model trained on Carolingian minuscule run over Insular Minuscule manuscript

(correct transcription: "se respicere pro certo sciat; Cogita)

score of the split word presumably stems from the irregular spacing, which divides one legitimate word into two non-words. The character heatmap for the Modern Latin model shows different patterns: in some words, single characters are to blame for a whole word's score, elsewhere whole sequences of letters are affected, such as in the last two lines. In these latter cases, presumably, not only the single letters but also their surrounding characters affect the single character confidence scoring.

The comparison of historic and modern printed font may not yield many surprises but serves as an example for what kind of issues the confidence visualizations can help to surface. Since the introduction of neural networks and LSTM to OCR and HTR technology also opened the door to processing more challenging materials, such as medieval manuscripts, it is conceivable that the same technique can be used for examining differences between different medieval script types. A bespoke model trained on one script type might reliably fail to transcribe unfamiliar ligatures or characters in a test manuscript, flagging up such phenomena with low confidence values (see Figure 5 – the model here struggles with the unfamiliar letter shapes for "a" and "t", untypical for Carolingian minuscule).

What these heatmaps also communicate, however, is that confidence scoring must be taken with a grain of salt: a low confidence score does not always mean that its cause is meaningful for the discussion, nor can we expect to securely identify the cause behind each scoring. Most poignantly, to understand whether an irregularity spotted is a systemic or an anecdotal occurrence, it is necessary to corroborate these findings with more data and evaluate them in the context of the entire document or sample. The next section is presenting a concrete example to showcase how to systematize such an approach with the help of statistical evidence.

2.3 Confidence heatmaps for scribal hands

Palaeographic analysis of medieval and Renaissance manuscripts deals with utterly human material – and to date relies on utterly human expertise. Often, a judgment call is made on account of intuition more than objective grounds. This is not only due to the blurry boundaries between script types and hands but perhaps also owed to the fact that differences between scripts are often difficult to describe or classify objectively. In the analysis of script types or manuscripts hands, visible differences between specific letter forms taken from different exponents can provide means for an objective comparison – except it is not always obvious where to start the search or how to weigh such discoveries.

The above experiment aimed to show how variegation in confidence levels can highlight pages, words, or letters outside the comfort zone of an OCR model. The same mechanism is applied in the following scenario by running a model trained on one manuscript over other exponents. A word-confidence graph maps the general affinity between the chosen manuscripts. Character confidence heatmaps were then used to identify low-scoring letters. Two metrics are used to then evaluate the relevance of these findings: the average confidence measured for all transcriptions of this letter (including scores for correct and incorrect transcriptions); and the overall error rate. These metrics are compared across all manuscripts under review. The average confidence and error rates from the manuscript the model is trained for serve as a baseline against which the results from the other exponents are compared. Said baseline helps to understand whether the reactions of the model cohere with its "comfort zone" or whether they signal a divergence from the baseline.

The process previously described showcased of models trained on a large number of text lines belonging to the same language or script group; the experiment here starts with training a bespoke model for just a single manuscript to then run it over test pages from the same and other manuscripts. Usually, Tesseract models are trained on hundreds of thousands of lines. Granted the small scope (and specific aim) of this experiment, the training of a bespoke model for a manuscript hand was performed with comparatively little material – hundreds, not thousands of lines.

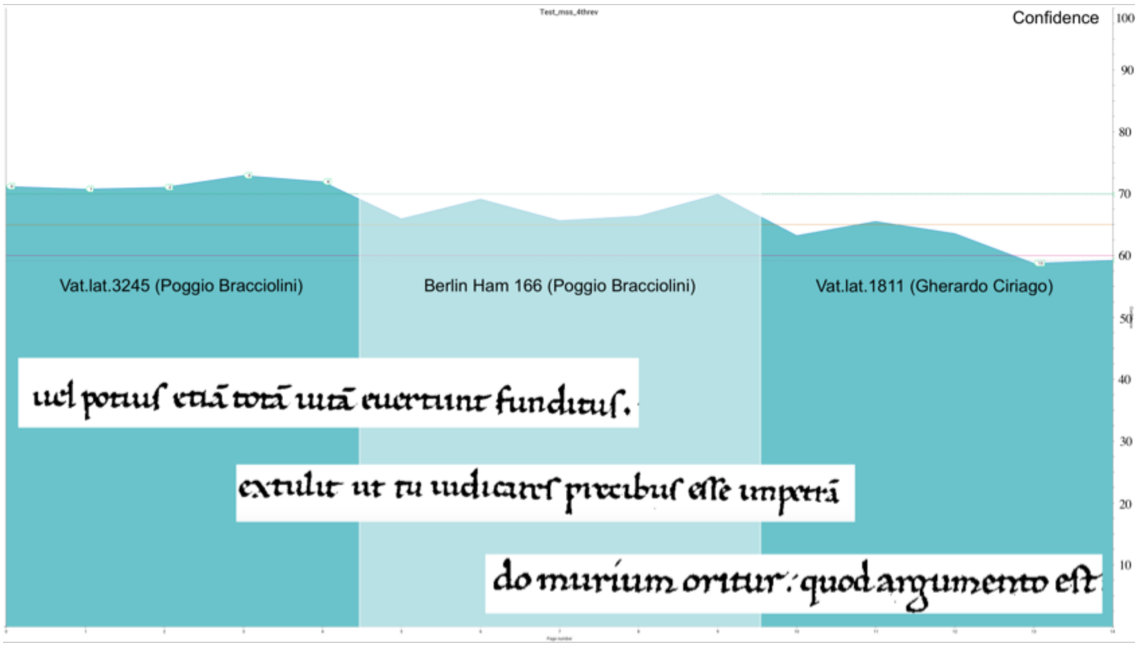


Figure 6: Word confidence graph for Vat.lat.3245, Berlin Ham 166 and Vat.lat.1811 processed with Tesseract model trained on Vat.lat.3235

For this experiment, the hand of Poggio Bracciolini (1380-1459), eminent Humanist and famous scribe, was compared to that of a follower – as also his own. Bracciolini, a trained notary and successful scribe, was the instrumental driver behind the development of what came to be called Humanist Minuscule (de la Mare 1977). Bracciolini is not only a rare example for the deliberate development of an idiosyncratic hand; it is equally uncommon that the evolution of a single hand can be traced with a number of surviving exponents (De Robertis 2017).

Whilst developing his style, Bracciolini also trained (and inspired) imitators. A limited number of manuscripts signed or authenticated by documentary evidence can be securely ascribed to Bracciolini. Based on visual comparison with these manuscripts, numerous others have been identified – and disputed in return (de la Mare 1973, Caldelli 2006). When trying to authenticate manuscripts putatively ascribed to Bracciolini, the palaeographic challenge is a war on two fronts: the first challenge is to identify differences or affinities between two manuscripts; in the second instance, the palaeographer must decide whether these findings signal the same hand or the penmanship of another. The OCR-based methodology cannot provide a secure answer – but as is to be shown, it can be used to identify samples for discussion.

The experiment used a model trained on lines taken from a manuscript identified as an autograph by Bracciolini, Vat.lat.3245 (785 lines for training, 87 lines for evaluation). The resulting OCR model was run over 5 pages each from Vat.lat.3245, Ms. Vat.lat.1811, a manuscript written by his close follower, Gherardo del Ciriago (1412-1472), and another manuscript ascribed to Bracciolini (Berlin Ms Hamilton 166).

The word confidence graph – unsurprisingly, perhaps – suggest that the model generally processed the other Bracciolini autograph at greater confidence levels than the hand of Gherardo del Ciriago (see Figure 6). The gap between the confidence scores for Vat.lat.3245 and Ms Ham 166, however, suggest that the hands, even though belonging to the same person, do differ somehow – perhaps a consequence of them being copied at different stages in Bracciolini's life: Ms Ham 166 was authored in 1408; Ms Vat.lat.3245 is dated to 1410-1415 (de la Mare 1973).

As with the example running OCR on historic printed text, heatmaps were used to understand the confidence scores in more detail. Concretely, in this case, the heatmap provided a first point of contact to help identify letters of interest. In the second step, a closer analysis of the confidence scores for a particular letter was analysed across the whole sample.

A single character transcribed at low confidence would not yield credible data to support an analysis but looking at all transcriptions of the same letter across the test set, i.e. from a variety of contexts, can give a more balanced perspective on whether one or several low confidence ratings signal anecdotal or systematic failure. In a next step, therefore, the confidence values were analysed for all instances of "suspicious" letters across all manuscripts to understand whether we are looking at a meaningful difference or not.

The analytical framework builds on the assumption that the confidence values from Vat.lat.3245 provide the baseline to compare the ratings from the other manuscripts to. In addition to the overall confidence ratings (which included scores for correct and incorrect transcriptions) the error rate is calculated.

Using the heatmaps, following letters were singled out for analysis: "ct", for low scores in Vat.lat.1811; "h", for low scorings in Vat.lat.1811; and "ae" for low scorings in Ham 166. The confidence values and error rates were then collected from all pages in the test set.

In the case of "ct", for Vat.lat.3245 and Ham 166, average confidence and error rate were almost on par. The numbers for Vat.lat.1811, however, differed drastically, particularly the error rate.[5] When comparing samples from the ct ligature across all manuscripts, in fact, the difference is not only consistent but easily visible: the "c" is touching the middle stroke of the "t" (See Table 1).

The case of the letter "h" is less straightforward (see Table 2). Whilst the average confidence level for Ham 166 is not too far from Vat.lat.3245, the error rate is significantly higher; it is also puzzling that two thirds of the errors were transcribed to "b". Overall, the statistics are not definitive enough to support a divergence between Ham 166 and Vat.lat. 3245, nor did the samples surface a regular, visible difference. For Vat.lat.1811, however, the error rate was over 55%. Looking at the manuscript itself, the letter shape regularly differs from the samples found in the other two manuscripts: the belly tends to be wider, and the initiating stroke more horizontal than for the other manuscripts; the final stroke regularly reaches below the baseline.

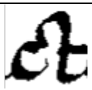
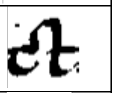
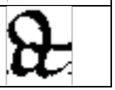
ct ligature	Avg conf	error rate	sample
Vat.lat.3245 (10 total)	98	20%	
Ham 166 (22 total)	97.7	22%	
Vat.lat.1811 (27 total)	95.0	85%	

Table 1: Statistics for "ct" ligature

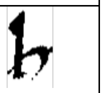
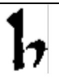
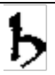
h	Avg conf	error rate	sample
Vat.lat.3245	97.8	7.8% (b: 3.9%)	
Ham 166	97.3	29.2% (b: 18.9%)	
Vat.lat.1811	96.2	55.2% (b: 11.9%)	

Table 2: Statistics for "h"

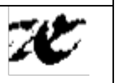
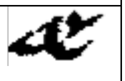
ae	Avg conf	error rate	sample
Vat.lat.3245 (1 total)	97.3	100%	
Ham 166 (24 total)	94.4	100%	
Vat.lat.1811	--	--	--

Table 3: Statistics for "ae" ligature

The "ae" ligature, meanwhile, presented a quite different situation (see Table 3). Initially chosen for low confidence and bad transcription in Ham 166, it turns out that the ligature appears not at all in Vat.lat.1811, and only once in Vat.lat.3245 – and is badly transcribed here, too. Granted "ae" is

[5] The ct ligature is usually transcribed by two letters, so the average confidence rating for both was used in the calculation.

included in the registered set of characters permitted in the Tesseract training, it would not be categorically excluded from recognition. The consistent failure to correctly recognize the glyph thus suggests it is scarce if not absent in the training material to start with – hence the low scores in its recognition. As it is, the use of "ae" or *e caudata* was not obligatory in either manuscript – these were orthographic novelties introduced by Humanist circles in Florence in the 14th and 15th century that aimed to replace the simple "e" common from Medieval times. The relative frequency of this glyph in Ms Ham 166 suggests that Bracciolini chose to deliberately employ it in Ms Ham 166, but mostly reverts to simple "e" in Vat.lat.3245.

Above examples are not exhaustive but give an idea of the kind of material one might identify with the help of OCR confidence scores. Neither graphs nor heatmaps deliver very clean nor comprehensive evidence; they require human scrutiny and interpretation to yield up useful information. In that sense, the examples above do not intend to provide a clear-cut interpretation of the relationship between the three manuscripts. Rather, the intent is to showcase how different modes of visualizing confidence scores from OCR processing can aid the quest for material to feed into palaeographic analysis. How this evidence is ultimately to be weighed is left to the expert; however, the hope is that OCR confidence scores can serve as a heuristic tool to speed up the task in the first place.

3 Conclusion

In summary, this article presented an approach to re-purposing OCR technology for identifying peculiarities in historic scripts or differences in scribal hands. The argument aims to show that even though many standard OCR models, in this case Tesseract, are overtly trained to focus on the recognition of specific languages in print, the sensitivity of OCR models to differences in fonts can be exploited to highlight differences in script or scribal hand. This is done with the help of the so-called confidence score, which signals the certainty with which the OCR engine assumes the output to be correct. The argument above is outlining several methods of visualizing the confidence score and how this can aid palaeographic analysis. The method is emphatically not intended to classify scripts or to authenticate hands. The experiments merely test confidence scores for their heuristic potential in identifying differences between script types and hands.

There are some downsides to this approach; firstly, the confidence score is a blurry metric that can be influenced by many factors, not all of which are relevant to a palaeographic discussion (for example ink bleed-through, speckles or skewed pages). Which factor is chiefly to blame for a low confidence score is not necessarily clear. Gathering scores from every single exponent, by default from a variety of contexts, however, can help to gain a balanced perspective in that regard. Expectations should also be tempered with a view to comprehensiveness – OCR confidence scores cannot be expected to highlight *all* differences in a font or hand. In that sense, the method as sketched presents a means to break the ice.

From a practical perspective, creating ground truth to train bespoke models is time-intensive and low-volume models such as the one used here are not necessarily reliable. Within the small scope of this investigation, such a model might have been sufficient for proving a concept, but a better model trained on more ground truth or endorsed by more advanced technology might be needed for a more thorough analysis. With the steady advance of OCR technology and more and more sophisticated attempts to create models for low-volume scripts, it stands to hope that there will be new solutions to the latter issue before too long.

Lastly, one might argue that these experiments merely help to surface phenomena that are visible to the naked eye anyway. Without providing analytical value in itself, the method leaves the ultimate interpretation of these discoveries to the palaeographer's expertise. It should not be forgotten, however, that palaeographic analysis is a painstaking process, and the identification of such differences is extremely time-consuming when done by hand and from scratch. The real value of this method, therefore, is to direct said naked eye to the phenomena in the first place, that is, to speed up the discovery process – and at the best of times help palaeographers discover elements they did not know they were looking for.

References

Breuel, Thomas M. 2008. 'The OCRopus Open Source OCR System'. Proceedings Volume 6815, Document Recognition and Retrieval XV (68150F). https://doi.org/10.1117/12.783598.

Caldelli, Elisabetta. 2006. Copisti a Roma Nel Quattrocento. Roma: Viella.

Camps, Jean-Baptiste, and Florian Cafiero. 2014. “Genealogical Variant Locations and Simplified Stemma: A Test Case.” In Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches, edited by Tara Andrews and Caroline Macé, 69–94. Turnhout: Brepols.

Castro Correa, Ainoa. 2014. “Palaeography, Computer-Aided Palaeography and Digital Palaeography.” In Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches, edited by Tara Andrews and Caroline Macé, 69–94. Turnhout: Brepols.

Christlein, Vincent. 2018. “Handwriting Analysis with Focus on Writer Identification and Writer Retrieval.” PhD of Engineering, Erlangen-Nürnberg: Friedrich-Alexander-Universität.

Cilia, Nicole Dalia, Claudio de Stefano, Francesco Fontanella, Claudio Marocco, Mario Molinara, and Alessandra Scotto di Freca. 2019. “A Two-Step System Based on Deep Transfer Learning for Writer Identification in Medieval Books.” In CAIP 2019, LNCS 11679, edited by M. Vento and G. Percannella, 305–16. Springer Nature. https://doi.org/10.1007/978-3-030-29891-3_27.

“ICDAR2017 Competition on the Classification of Medieval Handwritings in Latin Script.” 2017. Classification of Medieval Handwritings in Latin Script (blog). 2017. https://clamm.irht.cnrs.fr/icdar-2017. [accessed 30 September 2023]

Kestemont, Mike, Vincent Christlein, and Dominique Stutzmann. 2017. “Artificial Palaeography: Computational Approaches to Identifying Script Types in Medieval Manuscripts.” *Speculum* 92 (1).

de la Mare, Albinia. 1973. The Handwriting of Italian Humanists. Oxford: Oxford University Press.

de la Mare, Albinia. 1977. ‘Humanistic Script: The First Ten Years’. In Das Verhältnis der Humanisten zum Buch, edited by Fritz Krafft and Dieter Wuttke, 89–110. Boppard: Harald Boldt.

Perveen, Shaheen. n.d. ‘TESSERACT’. HackMD (blog). https://hackmd.io/@rDplrV2BTM-mnyMNpr9TfQ/Hy4ccns2I. [accessed 30 September 2023]

de Robertis, Teresa. 2017. “Scritture Umanistiche Elementari (e Altro).” Scrineum Rivista 14. http://dx.doi.org/10.13128/Scrineum-21994.

Smith, Ray. 2007. ‘An Overview of the Tesseract OCR Engine’. In Document Analysis and Recognition, 2007. ICDAR 2007., 2:629–33. IEEE.

Smith, Ray. 2016. ‘Tesseract Blends Old and New OCR Technology’. Tutorial presented at the Document Analysis Systems, Santorini.

Tesseract, https://github.com/tesseract-ocr [accessed 30 September 2023]

‘Tesseract Documentation FAQ’. n.d. https://tesseract-ocr.github.io/tessdoc/tess3/FAQ-Old.html. [Accessed 16 November 2023.]

Ul-Hasan, Adnan, Faisal Shafait, and Thomas Breuel. 2013. ‘High-Performance OCR for Printed English and Fraktur Using LSTM Networks’. In Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. 10.1109/ICDAR.2013.140.

A Appendix A: OCR models used

rescribev9_fast.traineddata via Rescribe Desktop tool: https://rescribe.xyz/rescribe/ [accessed 30 September 2023]

lat.traineddata and eng.traineddata: https://github.com/tesseract-ocr/tessdata [accessed 30 September 2023]

carolinemsv1_fast.traineddata: https://rescribe.xyz/rescribe/trainings.html [accessed 16 November 2023]

B Appendix B: Test Books and manuscripts

Boccaccio, Giovanni. 1585. *Il Decameron*. Giunti: Venice.

Festus, Rufius. 1472. *Breviarum rerum gestarum populi* Romani: Venice. https://digitale-sammlungen.de/en/view/bsb00006378?page=,1 [accessed 30 September 2023]

Rome, Vatican Library, Ms Vat.lat.3245

Rome, Vatican Library, Ms Vat.lat.1811

Berlin, Ms Hamilton 166

Einsiedeln, Stiftsbibliothek, Codex 281(886): Ascetica; Glossa psalmorum; Poenitentiale (https://www.e-codices.unifr.ch/en/list/one/sbe/0281)