In Codice Ratio: Scalable Transcription of Historical Handwritten Documents (Extended Abstract)

Serena Ammirati[1], Donatella Firmani[1], Marco Maiorino[2], Paolo Merialdo[1], Elena Nieddu[1], and Andrea Rossi[1]

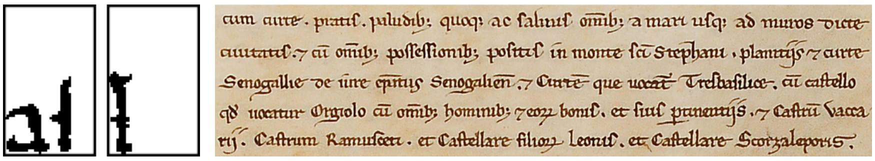
[1] Roma Tre University
serena.ammirati,donatella.firmani@uniroma3.it,merialdo@dia.uniroma3.it, ema.nieddu,and.rossi.516@gmail.com
[2] Vatican Secret Archives (Archivum Secretum Apostolicum Vaticanum)
m.maiorino@asv.va

Abstract. Huge amounts of handwritten historical documents are being published by digital libraries world wide. However, for these raw digital images to be really useful, they need to be annotated with informative content. State-of-the-art Handwritten Text Recognition (HTR) approaches require an impressive training effort by expert paleographers. Our contribution is a scalable, end-to-end transcription work-flow – that we call *In Codice Ratio* – based on fine-grain segmentation of text elements into characters and symbols, with limited training effort. We provide a preliminary evaluation of *In Codice Ratio* over a corpus of letters by pope Honorii III, stored in the Vatican Secret Archive.

1 Introduction

Large document collections are sources of important correlations between entities such as people, events, places, and organizations. Previous studies [7] have shown that it is possible to detect macroscopic patterns of cultural change over periods of centuries by analyzing large textual time series. Such automatic methods promise to empower scholars with a quantitative and data-driven tool to study culture and society, but their power has been limited by the amount of digitally transcribed sources. Indeed, the World Wide Web only contains a small part of the traditional archives. (It is evocative to think that it may only contain a few millimeters out of the 85km of linear shelves in the Vatican Secret Archives.) Recently, many historical archives have begun to digitize their assets, sharing high-resolution images of the original documents. Notable examples include the Bibliothque Nationale de France[3], the Virtual Manuscript Library of Switzerland[4], and the Vatican Apostolic Library[5]. In this scenario, expert paleographers

[3] http://gallica.bnf.fr
[4] http://www.e-codices.unifr.ch/en
[5] http://www.digitavaticana.org



(a) (b)

Fig. 1: (a) Fragments. (b) Sample text from the manuscript *Liber septimus regestorum domini Honorii pope III*, in the Vatican Registrers.

can largely benefit from computer-assisted transcription technologies. This includes not only full transcriptions, but also partial transcription and other kinds of automatically produced meta-data, useful for indexing and searching.

Popular automatic tools for transcribing the text content of digital images include Optical Character Recognition (OCR) systems, which work great for typewritten text but are not suitable for handwritten text recognition (HTR). Since most digitized documents by historical archives are manuscripts, HTR has recently gained more and more attention by researchers worldwide. Handwritten text is more challenging to recognize than typewritten one because characters have less regular shapes, and are often combined into single units known as a *ligatures*. Therefore, while OCR systems are trained to recognize individual typewritten glyphs, most state-of-the-art HTR systems use holistic approaches: all text elements (sentences, words, and characters) of a single text line are recognized as a whole, without any prior segmentation of the line into these elements. So-called *segmentation free* models [13] can be automatically obtained using well known training techniques, but they require the whole transcripts of a number of these unsegmented images. In order to use these technologies, users with experience in handwritten documents transcription are required to transcribe manually significant portions of the original documents. Currently available HTR technologies are still far from offering scalable automated solutions.

Our contribution. The approach of our project *In Codice Ratio* is in the middle of a spectrum, where on the one side there are OCR systems and on the other segmentation free HTR technologies, which recognize bigger handwritten elements. Rather than relying on well-segmented glyphs (as in typewriting) or training to recognize whole words, we focus on overlapping "fragments" of words composed of zero, one, or two (rarely three) characters, as in Figure 1a. For each word, we compute possible *cut-points* yielding different ways for segmenting the word into fragments. Then, we choose the best among different segmentation options using OCR and language models. Cut points are managed similarly to *on-line* HTR systems [6], where the text is written on a touch-screen and a sensor captures all the pen tip movements. Since we do not have access to pen movements, we need a number of labelled fragments[6] for training our model.

[6] A fragment can also have empty transcriptions when it does not contain any character.

Training with fragments has two advantages over training with words (as in segmentation free HTR):

- the number of fragments needed for training is much smaller, because it does not need to deal with the impressive variety of lexicon;
- fragments can be labeled by volunteers in large transcription projects, with little or no expertise, provided with adequate examples.

To this end, we set up a simple crowd-sourcing application.

Proof of Concept. Our hybrid work-flow takes the best of two worlds: we can handle challenging ligatures as in state-of-the-art HTR, and at the same time we require limited training effort like typical OCR systems. Our system is not mature enough for a thorough experimental evaluation, but for sake of demonstration, we take into account the "Vatican Registrers" corpus in the Vatican Secret Archives. The Vatican Registers is a huge collection of volumes (more than 18.000 pages) produced in the 13-th century, and containing official correspondence of the Roman Curia, such as political letters, opinions on legal questions, and documents addressed to various religious institutes throughout Europe. Such records are of unprecedented historical relevance, have not been transcribed yet, and have a regular writing style (see Figure 1b). For these reasons, we believe that they can motivate our work.

Our crowd-sourcing application enrolled 120 high-school students in the city of Rome, that did the labelling as a part of their work-related learning program. The program is jointly organized by the engineering and humanities departments of Roma Tre University, and includes frontal lessons in a variety of topics, such as paleography, history and machine learning, thus also serving as school guidance.

2 Related Works

The idea of exploiting large textual corpora to detect macroscopic cultural trends has been discussed for many years [11], promising to empower historians and other humanities scholars with a tool for the study of culture and society.

Data-driven history. Many studies have been published over the past few years [4, 10] about a quantitative and data-driven approach to the study of cultural change and continuity. A seminal study of 5 million English-language books published over the arc of 200 years [9] showed the potential of this approach, for example, measuring the time required by various technologies to become established or the duration of celebrity for various categories of people.

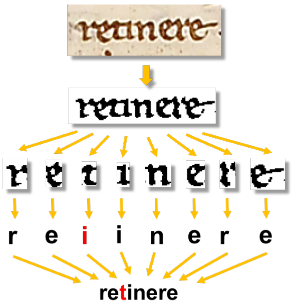


Fig. 2: Our workflow.

Handwriting text recognition. HTR can be defined as the ability to transform handwritten input represented as graphical marks into symbolic representation as ASCII text. According to the mode of data acquisition used, HTR can be classified into off-line and on-line. In off-line systems the handwriting is given as an image or scanned text, without time sequence information. In on-line systems the handwriting is given as a temporal sequence of coordinates that represents the pen tip trajectory. For high quality text images, current HTR state-of-the-art prototypes provide accuracy levels that range from 40 to 80% at the word level [12, 2]. On the other hand, on-line systems are more accurate [1, 5, 6], reaching 90% word-level accuracy in some cases.

Crowd-sourcing. Expected users of HTR technology belong mainly to two groups:

- individual researchers with experience in handwritten documents;
- volunteers which collaborate in large transcription projects.

Recent HTR project [14, 8] expose HTR tools through specialised crowd-sourcing web portals, supporting collaborative work.

Other works. Language modeling for on-line handwriting recognition bears many similarities with OCR and speech recognition, which often employ statistical n-gram models on the character or word level [15].

3 System Work-Flow

Our system first segments every word into small (possibly overlapping) fragments, then recognizes characters in fragments, and finally the entire word. The main steps are schematically shown in Figure 2.

1. **Pre-processing.** In this phase the color image of a page is cropped into lines and words. The image is also transformed into a bi-chromatic one.
2. **Cut-level operations.** For each word we guess cut-points for characters and build the so-called segmentation lattice data structure [6], such that each path in the lattice represent a way of **segmenting** the word.
3. **Fragment-level operations.** For each pair of cut-points we crop the corresponding text fragment and classify it, choosing among known characters.
4. **Word-level operations.** We finally return the best path in the labelled lattice, which represent a way for **transcribing** the word.

3.1 Transcription Algorithms

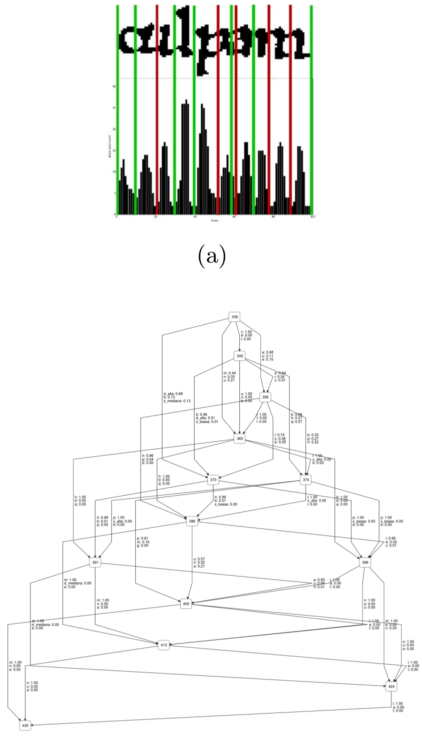
In this section we describe the inner phases of our system. Inner phases are designed for the Carolingian minuscule script, which is used in the manuscript for our proof of concept (see Figure 1b of Section 1).

Cut-level operations. The input for on-line HTR consists of pen up/pen down switching movements, therefore *cut-points* for character boundaries can be selected at certain positions of each stroke. Since we do not have access to stroke sequences, we design a simple heuristic based on black pixel distribution *local minima*, as shown in Figure 3a for a sample occurrence of the word “culpam”. Then, we consider all the possible segmentation options induced by the cut-points and let the further phases select the best one. We build a *segmentation lattice* where each start-end path represents a segmentation option.

- Each cut-point corresponds to a node. The leftmost cut-point is the *start* node, and the rightmost the *end* node.
- Each edge corresponds to the fragment of word bounded by its endpoints.

We observed experimentally that edges corresponding to fragments smaller than 8 pixels or larger than 34 pixels can be safely dropped our further consideration. The segmentation lattice for our "culpam" example is shown in Figure 3b. In the figure, edges have labels, as will be clarified in the next section.

Fragment-level operations. We call *fragment* any squared portion of a word that is bounded by two cut-points, including incomplete characters, combinations of incomplete characters, or multiple characters together. Each edge of the segmentation lattice corresponds to a different fragment, and each path corresponds to a different segmentation option of the word into fragments. Our next step is a OCR step, where each fragment/edge of the segmentation lattice is labeled with the result of a character classifier. There are many principled ways for the classification task at hand. We decide to use a *convolutional* Neural Network (NN), that is one of the most common and popular approaches [3]. Since the NN return a *score distribution*, rather than a unique answer, we add multiple edges between the same two nodes. Some of the fragments are submitted to a crowd-sourcing application, for training the NN. The application guides the workers, providing sample images of characters and highlighting variations in shape and style. Therefore, workers without experience in transcription can give answers based on their perceived similarity of fragments to sample images. We refer to the resulting data structure as *labeled lattice*.



(a)

(b)

Fig. 3: Cut-points for the word "culpam" and corresponding lattice. We use green for actual character boundaries, and red otherwise.

Word-level operations. Labeled paths represent candidate transcriptions for the current word. In order find the best transcription recognition result, we use language models (LM). A statistical LM is a representation of a certain language as a probability distribution over sequences of words or characters.

In other words, a LM expresses the likelihood that certain sequences of words or characters appear in texts written in the language under analysis. To this end, we downloaded a large medieval Latin corpus ($\approx 1.5M$ words) and computed 3-grams frequencies. Then, we select the best path maximizing the corresponding **word** probability. For instance, for the path "culham" in Figure 3b we have[7]$p(\$culham^\wedge) = p(c|\$)p(u|\$c)p(l|cu)p(h|ul)p(a|lh)p(m|ha)p(|am)$, where \$ and $^\wedge$ are special symbols denoting the beginning and the end of a word. Every term in the product can be computed directly from language models.

4 Experiments

We describe in this section our preliminary experimental results, which serves us as a *proof of concept* for *In Codice Ratio*. Ideas and methods in the proof have been realized in collaboration with the Vatican Secret Archives, with the aim of demonstrating in principle the practical potential of our system.

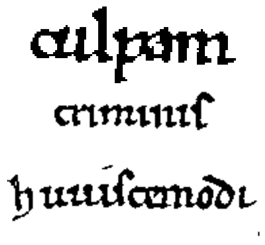


Fig. 4: Sample words, for our proof of concept.

Dataset. The "Vatican Registrers" corpus consists of 43 parchment registers, for a total of 18650 pages (i.e., writing facades). All the registers are written with the same script: the so-called *Cancelleresca.* Our dataset consists of 30 pages ($\approx 15K$ characters) of register 12 by Pope Honorii III, that is the only Pope with un-transcribed registers and therefore is of most interest for the VSR.

Results and future works. Our NN shows 95% accuracy and recall for our dataset, with 4.6% error. Typical errors of the NN are the following.

- Characters "f" and "s" are easily confused, due to their similar shapes. Specifically, $\approx$ 20% of "s" are labelled as "f" and 25% of "f" as "s".
- Characters "l" is often mis-classified as other "upper" characters, due to spurious ink in the fragment. Specifically, $\approx$ 72% of "l" are labelled as "b" and only $\approx$ 17% of "l" are labelled as such.

Other characters are labelled correctly more than 96% of the times.

We select 3 words, showing strengths and limits of our method, dubbed "culpam", "criminis" and "uiuscemod(i)"[8]. In Table 1 we show the top three paths in the labelled lattice, according to word probability. For "criminis", the right transcription is first in the ranking. For "uiuscemod", the right transcription is not contained in any path, but the maximum probability path "uiufemod" is correct except for the ligature "sc", which is labelled as "f" as mentioned earlier in this section. This specific problem is due to errors in the training set, where a number of "sc" fragments has been selected as "f" by workers. In the near

[7] Using 2-nd order Markov assumption.

[8] "uiuscemod" and "i" are processed as separate words.

future, we plan to repeat the training phase by removing such wrong items. Finally, “culpam” is overthrown by “cullum” because of more serious errors in the OCR process (“p” and “a” are labelled as “l” and “u’, respectively). While word probability does not help (3-gram “llu” is more frequent than “lpa” in our LM), we are confident to correct this kind of error in future works. Furture works include the following.

- More sophisticated path selection criteria, for instance, taking into account NN output score (second “l” has lower score than “p”).
- More advanced language model tools, such as Hidden Markov Models. This is currently being developed with promising result.
- Excluding labels that do not fit in current line margins, such as excluding an “l” where the lower margin contains black pixels, such as for “p”.

5 Conclusions

In Codice Ratio is an automatic transcription workflow with low training effort. Our proof of concept is done on a high-resolution digitized copy of the registers by pope Honorii III. Manuscript pages first undergo a series of transformations, that extract a clean version of the text image. Then, preprocessed pages are decomposed into fragments containing basic text elements, such as characters and symbols. Some fragments are labelled by unskilled crowd-sourcing workers, which are asked simply to select matching images to template symbols selected by paleographers. Labelled symbols are used to train a Neural Networks, that is in charge of automatically compute labels for all the un-labelled fragments. Automatically computed labels are aggregated at the word-level into a segmentation lattice, in which all the traversing paths represent candidate transcriptions for the word. Selection of the best path is done based on language models. Our proof of concept suggests that *In Codice Ratio* can be applied to the large collection of Vatican Registers in the Vatican Secret Archives, subject to specifics improving, that are matter of ongoing work.

path	prob
culpam	
cullum	$1 \cdot 10^{-4}$
culpam	$8 \cdot 10^{-7}$
culluni	$3 \cdot 10^{-7}$
criminis	
criminis	$8 \cdot 10^{-7}$
crinunis	$6 \cdot 10^{-8}$
crinuius	$4 \cdot 10^{-8}$
uiuscemod(i)	
uiufemod	$2 \cdot 10^{-2}$
uuifemod	$2 \cdot 10^{-3}$
uiiifemod	$5 \cdot 10^{-4}$

Table 1: Path probabilities for the words in Figure 4.

Acknowledgments

We thank Gaetano Bonofiglio, Veronica Iovinella e Andrea Salvoni for their work on the OCR neural network. We thank Matteo Mariani for helping with all the pre-processing steps. We also thank Gianlorenzo Didonato for developing the front-end of the *In Codice Ratio* crowd-sourcing application. Finally, we are indebted to all the teacher and students of Liceo Keplero and Liceo Montale who joined the work-related learning program, and did all the labelling effort.

References

1. Character recognition experiments using unipen data. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, ICDAR '01, pages 481–, Washington, DC, USA, 2001. IEEE Computer Society.
2. R. Bertolami and H. Bunke. Hidden markov model-based ensemble methods for offline handwritten text line recognition. *Pattern Recognition*, 41(11):3452 – 3460, 2008.
3. D. Ciregan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3642–3649. IEEE, 2012.
4. I. Flaounas, O. Ali, T. Lansdall-Welfare, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini. Research methods in the age of digital journalism: Massive-scale automated analysis of news-contenttopics, style and gender. *Digital Journalism*, 1(1):102–116, 2013.
5. S. Jaeger, S. Manke, J. Reichert, and A. Waibel. Online handwriting recognition: the npen++ recognizer. *International Journal on Document Analysis and Recognition*, 3(3):169–180, 2001.
6. D. Keysers, T. Deselaers, H. A. Rowley, L.-L. Wang, and V. Carbune. Multi-language online handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.
7. T. Lansdall-Welfare, S. Sudhahar, J. Thompson, J. Lewis, F. N. Team, and N. Cristianini. Content analysis of 150 years of british periodicals. *Proceedings of the National Academy of Sciences*, 114(4):E457–E465, 2017.
8. A. Marcus, A. Parameswaran, et al. Crowdsourced data management: Industry and academic perspectives. *Foundations and Trends® in Databases*, 6(1-2):1–161, 2015.
9. J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, , J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.
10. F. Moretti. *Distant reading*. Verso Books, 2013.
11. R. Reddy and G. StClair. The million book digital library project. `http://www.rr.cs.cmu.edu/mbdl.htm`, 2016. Accessed December 19, 2016.
12. V. Romero, V. Alabau, and J. M. Benedí. Combination of n-grams and stochastic context-free grammars in an offline handwritten recognition system. In *Proceedings of the 3rd Iberian Conference on Pattern Recognition and Image Analysis, Part I*, IbPRIA '07, pages 467–474, Berlin, Heidelberg, 2007. Springer-Verlag.
13. V. Romero, J. A. Snchez, V. Bosch, K. Depuydt, and J. de Does. Influence of text line segmentation in handwritten text recognition. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 536–540, Aug 2015.
14. J. A. Sánchez, G. Mühlberger, B. Gatos, P. Schofield, K. Depuydt, R. M. Davis, E. Vidal, and J. De Does. transcriptorium: a european project on handwritten text recognition. In *Proceedings of the 2013 ACM symposium on Document engineering*, pages 227–228. ACM, 2013.
15. A. Stolcke et al. Srilm-an extensible language modeling toolkit. In *Interspeech*, volume 2002, page 2002, 2002.