In Codice Ratio: Machine Transcription of Medieval Manuscripts

Serena Ammirati[1], Donatella Firmani[2](✉), Marco Maiorino[3], Paolo Merialdo[2], and Elena Nieddu[2]

[1] Department of Humanities, Roma Tre University, Rome, Italy
serena.ammirati@uniroma3.it
[2] Department of Computer Science, Roma Tre University, Rome, Italy
{donatella.firmani,paolo.merialdo,elena.nieddu}@uniroma3.it
[3] Vatican Secret Archives, Vatican City, Italy
m.maiorino@asv.va

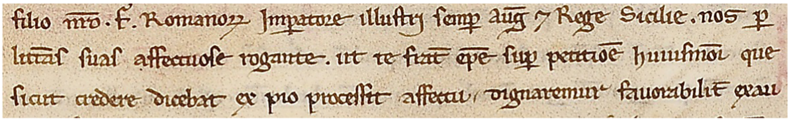
Abstract. Our project, *In Codice Ratio*, is an interdisciplinary research initiative for analyzing content of historical documents conserved in the Vatican Secret Archives (VSA). As most of such documents are digitized as images, Machine Transcription is both an enabler to the application of Knowledge Discovery techniques, as well as a useful tool to the paleographer for speeding up the transcription process. Our approach involves a convolutional neural network to recognize characters, statistical language models to compose and rank word transcriptions, and crowdsourcing for scalable training data collection. We have conducted experiments on pages from the medieval manuscript collection known as the Vatican Registers. Our results show that almost all the considered words can be transcribed without significant spelling errors.

1 Introduction

The research project *In Codice Ratio* has the goal of supporting humanities scholars in the content analysis and knowledge discovery activities on large collections of historical documents. Thanks to novel methods and tools that we aim at developing, paleographers, philologists and historians will be able to conduct data-driven studies at a large scale by quantitatively analyzing trends and evolution of writings and languages across time and countries, and by examining and discovering facts and correlations among information spread in vast corpora of documents. Our project concentrates on the collections preserved in the Vatican Secret Archives (VSA), one of the largest and most important historical archives in the world. In an extension of 85 km of shelving, it maintains more than 600 archival collections of historical sources on the Vatican activities – such as, official correspondence of the Vatican, account books, correspondence of the popes – starting from the end of the eighth century. We are currently working on the collection of the Vatican Registers, which record the inbound and outbound

This work is an extended abstract of [4].

P. Manghi et al. (Eds.): IRCDL 2019, CCIS 988, pp. 185–192, 2019.
https://doi.org/10.1007/978-3-030-11226-4_15



filio nr̄o. F. Romanoꝝ Imꝑatore illuſtri ſemꝑ aug 7 Rege Sicilie. nos ꝑ
litteras ſuas affectuoſe rogante. ut te frat̄ epe ſup petitioē huiuſmoi que
ſicut credere dicebat ex pio proceſſit affectu, dignaremur fauorabilit exau

Fig. 1. Sample text of the “Liber septimus regestorum domini Honorii Pope III”.

correspondence of the popes. A small sample is shown in Fig. 1. These registers have been continuously and systematically preserved since the middle age, hence most of these documents are *manuscripts*. The VSA has begun to acquire digital images of these documents but, unfortunately, complete transcriptions for the earliest registers do not exist. Therefore, a first fundamental step to develop any form of data-driven content analysis is to perform a transcription of the manuscripts. The problem is challenging: on the one hand, a manual transcription is unfeasible (at least in a reasonable amount of time), due to the volume (hundreds of thousands of pages) of the collection. On the other hand, although these manuscripts are written with a uniform style (a derivation of the *Caroline* style), traditional OCR does not apply here, because of irregularities of hand-writing, ligatures and abbreviations.

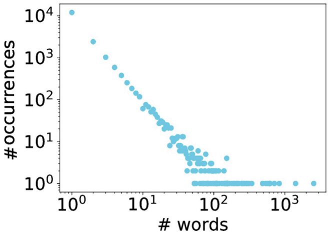


Fig. 2. Word count.

Since segmenting words in characters is tricky with handwritten texts, recent automatic transcription approaches typically aim at recognizing entire words. However, because of the variability and the size of the lexicon, they need a huge amount of training data, i.e., hundreds of fully transcribed pages. To illustrate this problem, consider Fig. 2: it reports the distribution of the occurrences of words in a corpus composed by a partial transcription of the Registers of Innocent III (in total, it is about 68,000 words). Observe that a few words (just 9) occur more than 100 times (the most occurring word is *“et”*, the Latin conjunction that corresponds to “and”), while the majority of words have less than 10 occurrences.

We follow a different approach, based on character segmentation. Our idea is to govern imprecise character segmentation by considering that correct segments are those that give rise to a sequence of characters that more likely compose a Latin word. We have therefore designed a principled solution that relies on a convolutional neural network classifier and on statistical language models. For every word, we perform a segmentation that can produce more segments than those actually formed by the characters in the word. Every segment is labeled by a classifier, which essentially recognizes the most likely character. We then organize the sequence of segments in a directed acyclic graph: the paths of such a graph represent candidate transcriptions for the word, and the most likely solution is selected based on language statistics.

Structure of this Paper. Section 2 contains an overview of our approach. Detailed description of our algorithms can be found in [4]. Main experimental results of [4] are reported in Sect. 3. Finally, Sects. 4 and 5 contain related work and concluding remarks. *In Codice Ratio* was introduced in [1]. All the code and data of the project is publicly online.[1]

2 Overview

We start from a set of high-quality scanned images of whole manuscript *pages*. Each page undergoes three standard preprocessing steps:

1. we transform the color image into a bi-chromatic one and crop white margins;
2. we correct *skew* and *slant*, i.e., page distortions due to acquisition process and calligraphy;
3. we crop lines and words according to horizontal and vertical white spaces, respectively.

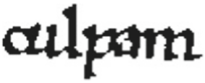


Fig. 3. A typical input image for our system.

Each word image is finally submitted to our transcription system. Figure 3 shows a pre-processed word image, of size 178 × 67 pixels. The correct transcription of the word in the figure is the Latin word "culpam" – the accusative singular of "culpa", that means "crime".

System Architecture. Our pipeline consists of four main components.

- **Training samples collector.** We implemented a custom crowd-sourcing platform, and employed 120 high-school students to label the dataset. To overcome the complexity of reading ancient fonts, we provided the students with positive and negative examples of each symbol. We trained a deep convolutional neural network character classifier on this dataset.
- **Character recognizer.** Recognizing characters within a handwritten word is challenging, due to *ligatures*. To this end, we first partition the input word into elementary text segments. Most segments contain actual characters, but there are also segments with spurious ink strokes. Then, we submit all the segments to the trained classifier. Computed labels are very accurate when the input segment contains an actual character, but can be wrong otherwise. We take into account minuscule characters of the Latin alphabet.
- **Transcription generator.** We reassemble noisy labels from the classifier into a set of candidate word transcriptions. Specifically, we select the best m candidate transcriptions for the input word image, using language models.
- **Word decoder.** We consider the m transcriptions from the previous step and revise character recognition decisions in a principled way, by solving a specific decoding problem on a high-order hidden Markov model. The most promising transcriptions are finally returned to the user.

[1] www.inf.uniroma3.it/db/icr/.

Discussion. It is worth noting that compared to a segmentation-free approach, training the classifier requires labeled examples for the limited set of character symbols, with a twofold advantage. First, the size of the training set is several order of magnitude smaller, as we need to provide examples only for the limited set of character symbols, and not for a rich lexicon of words. Second, producing the examples is much easier, as it does not require to transcribe whole words, an activity that can be carried on by expert paleographers. In our system, the production of the training set is accomplished by a crowdsourcing solution that consists of simple visual pattern matching tasks, similar to captchas.

3 Experiments

For our experiments we use annotations from 2 pages of Vatican Register 12. This results in approximately 15K characters. Characters with less than 1K examples were augmented to match the required quantity and balance the training set. The augmentation process involves slight random rotation, zooming, shearing and shifting, both vertical and horizontal. The final dataset comprises 23K examples evenly split between 23 classes. We test our system on four pages belonging to the same Vatican Register, but spanning different ages and writers, transcribed entirely by volunteer paleographers. After undergoing the pre-processing, each word is transcribed independently by the system. Our system currently considers only word images without *abbreviated forms*. This is further discussed in Sect. 5. Our language model is composed of 716 ancient Roman Latin texts, spanning different ages and subjects, for a total of over 14M words. It is worth observing that the Latin language used in the Vatican Registers exhibits some differences with the ancient Roman Latin, which is typically used in publicly available corpora. These differences introduce some drawbacks, that we are currently overcoming, as we discuss later in Sect. 5.

Set-Up. We define the m**-precision** as the fraction of word images in our test set, for which the correct transcription is (i) generated by our system, and (ii) ranked in the top m positions. Classical definition of precision is captured by 1-precision. For the top few transcriptions, we provide edit distance statistics (**ED**) with respect to the correct transcription. Specifically, we use the distance metric in [3]. Our experiments are summarized below.

Results. In the *character recognition* step, average precision and recall of our neural network among all classes are both 96%. Precision ranges from 86% to 99%, whereas recall ranges from 74% to 99%. As frame of comparison, we trained a logistic regression model on the same dataset. Such baseline scored 80% and 79% average precision and recall. More results on this are in [5].

In the *transcription generation* step, the fraction of words for which our system yields the correct transcription is ≈65% (decoding can recover the correct transcription of approximately 9% of the remaining 35%), compared to much lower 20% achieved by the baseline in [1]. When the correct transcription is available, *language models* can rank the correct transcription of almost all the word images in the top 5. For remaining 35%, 16% of first-ranked transcriptions is at edit distance 1 from the correct transcription, 15% at distance 2 and 28% at distance 3. Figure 4 shows a sample word image of the 15% group, for which the first-ranked transcription is at distance 2 from the correct transcription.

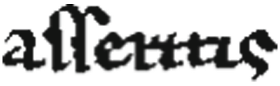
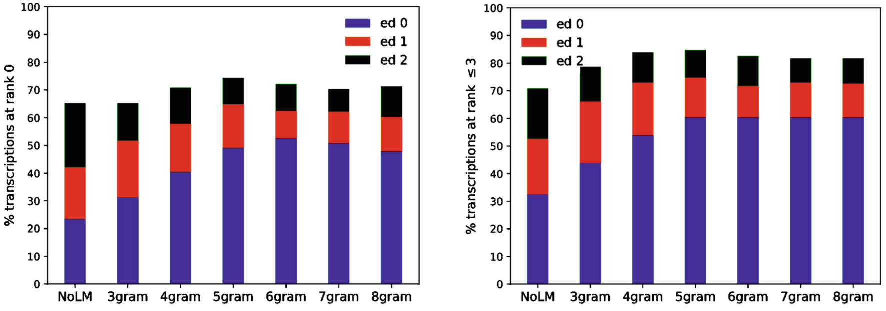


Fig. 4. The correct transcription of this word image is “asseritis”, while the first-ranked transcription is “afferitis”.

The purple bars in Fig. 5 show the 1-precision and 3-precision of our system for different q-gram sizes in the language model. The bar labeled as “NoLM” shows ranking results obtained without taking language models into account. The NoLM ranking was produced by multiplying network classification scores for each character. Figure 5b considers top 3 transcriptions of all the word images in the test set. We observe that, by using 6-g, almost 80% of our results are away from correct transcriptions by no more than 2 characters, and approximately 60% corresponds exactly to the underlying manuscript word. Figure 5a reports the corresponding results when considering only top 1 transcriptions. Another way for reading results in Fig. 5a is the following. Consider the 65% of the word images in our dataset for which we generate the correct transcription. Approximately 77% of the word images have the correct transcription ranked at position 1 when using 6-g (which is our default setting), but approximately 23% does not get the optimal ranking. Correct transcription, when generated, is in the top 5 for almost all the word images. Improving on the ranking produced requires a better model of the language used in the Vatican Register, included models of sentences, and is discussed in Sect. 5.

Consider now the 35% of word images for which our system does *not* generate the correct transcription is approximately.[2] Decoding can recover the correct transcription of approximately 9% of such word images. Other effects of the decoding phase is that top-ranked transcriptions become closer to correct transcriptions. For instance, the amount of word images having correct transcription



(a) 1-precision

(b) 3-precision

Fig. 5. Different values of q in the language model. “NoLM” represents ranking without language model, relying on character classification score only. (Color figure online)

[2] For such word images, most of first-ranked transcriptions have ≤3 spelling errors.

ranked as second increases by 30%. Overall, the amount of word images having correct transcription ranked as first does not change significantly.

4 Related Works

Handwritten Text Recognition (or HTR) is a research topic concerning the automatic transcription of handwritten text. Even though it extends to live-captured handwriting (online recognition), that is clearly not the case for historical documents. Offline recognition is generally regarded as harder than the online, due to the lack of temporal information: online handwriting recognition can leverage the order and timing of character strokes, while offline recognition cannot. Due to the many challenges involved in a fully automatic transcription system of historical handwritten documents, many researchers in the last years have focused on solving sub-problems, including word spotting [11], and text line segmentation [10]. Our goal is rather the creation of a full-fledged off-line HTR system: an effort shared by several ongoing projects, as more and more libraries and archives worldwide digitize their collections [7,13]. These systems generally work by a segmentation-free approach, where it is not necessary to individually segment each character. To deal effectively with ambiguity in segmentation and transcription, we map each word image to a lattice, whose source-to-sink paths represent alternative segmentations and corresponding transcriptions. Our approach is close to the technique in [8].

Crowdsourcing. Crowdsourcing solution for cultural heritage has been experienced in many projects. One of the pioneering initiative to crowdsource the transcription of manuscripts is the *Transcribe Bentham* project, a collaborative platform for crowdsourcing the transcription of the philosopher Jeremy Bentham's unpublished manuscripts [2]. Also the Transcriptorium project [12] exposes HTR tools through specialized crowd-sourcing web portals, supporting collaborative work. Our solution is more focused than the above ones: since it aims at producing training data, it relies on a much simpler solution based on visual pattern matching task that can be performed by unskilled workers.

Neural Networks. Our approach employs a convolutional neural network for character image classification. There has been an interest in applying recent results in recurrent and convolutional neural networks to achieve improved classification accuracy: [14] performs word spotting through a deep convolutional neural network, outperforming various word spotting benchmarks; while [6] adopts a bidirectional Long Short-Term Memory neural network to transcribe at word level, with high accuracy. In order to achieve complete transcriptions, these approaches would need thousands of word-level annotations, which is not a scalable task due to the expertise required. We will come back on this point when discussing future research directions for our project.

5 Conclusions and Future Work

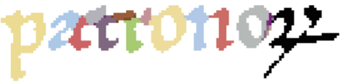


Fig. 6. A word (*patrono*rum) containing an abbreviation (in black).

Data science can deeply contribute to analyze and understand our historical and cultural heritage. Data acquisition and preparation from manuscript historical documents is done by means of a transcription process, whose scalability is limited, as it must be performed by expert paleographers. In this paper, we have presented a system, developed in the context of the *In Codice Ratio* project, to support the transcription of medieval manuscripts in order to improve the scalability of the process. We have followed an original approach that requires minimal training effort, and that is able to produce correct transcriptions for large portions of the manuscripts. Our approach, which relies on word segmentation, neural convolutional network, and language models, has been successfully experimented on the Vatican Registers.

We are currently working on the system in order to extend the set of symbols, and hence to improve the overall effectiveness of the process. In particular, we are adding the most frequent abbreviations, i.e., short-hands used by the scribes to save room or to speed up writing. In our process, the main issue with abbreviations is the lack of statistics on their occurrences, which prevents us to effectively apply the language models. Gathering statistics for the abbreviation is not trivial: the usage of these symbols depends both on the age and on the domain of the manuscripts. For instance, in the Vatican Registers, which have diplomatic and legal contents, some abbreviations are more frequent than in manuscripts with of literary works, even from the same age. Indeed, we have already collected training samples for the classifier also for many abbreviations: our crowdsourcing approach to collect labeled examples worked well also for these symbols, as it is based on simple visual pattern matching tasks. Figure 6 shows an example of a one of the most frequent abbreviations. The last symbol, in black, is a shorthand for the Latin desinence *"rum"*: notice that it is simple, given some sample images, to recognize it also without any paleography knowledge. Also the neural network performs well with the extended set of symbols, as abbreviations are typically well distinguishable from other symbols.

Our plan to collect statistics for the abbreviations is to use our current system to produce partial transcriptions for a number of pages, a few dozens, highlighting the words where the character classifier recognizes an abbreviation. Then, we will ask to the paleographers to transcribe these words. Based on these semi-automatic transcriptions, we will progressively update the language models. So far, we took a probabilistic approach on language modeling: we plan, however, to investigate character-level neural language modeling, similarly to [9].

Acknowledgments. We thank NVIDIA Corporation for the donation of a Quadro M5000 GPU, and Regione Lazio (Progetti di Gruppi di Ricerca) and Roma Tre University (Piano Straordinario di Sviluppo della Ricerca di Ateneo) for supporting our project "In Codice Ratio".

References

1. Ammirati, S., Firmani, D., Maiorino, M., Merialdo, P., Nieddu, E., Rossi, A.: In codice ratio: scalable transcription of historical handwritten documents. In: Proceedings of the 25th Italian Symposium on Advanced Database Systems, Squillace Lido (Catanzaro), Italy, 25–29 June 2017, p. 65 (2017)
2. Causer, T., Terras, M.: 'Many hands make light work. Many hands together make merry work': transcribe Bentham and crowdsourcing manuscript collections. Crowdsourcing Our Cultural Heritage, pp. 57–88 (2014)
3. Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks. In: Proceedings of the 2003 International Conference on Information Integration on the Web, IIWEB 2003, pp. 73–78 (2003)
4. Firmani, D., Maiorino, M., Merialdo, P., Nieddu, E.: Towards knowledge discovery from the Vatican secret archives. In codice ratio - episode 1: machine transcription of the manuscripts. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, 19–23 August 2018, pp. 263–272 (2018)
5. Firmani, D., Merialdo, P., Nieddu, E., Scardapane, S.: In codice ratio: OCR of handwritten Latin documents using deep convolutional networks. In: Proceedings of the 11th International Workshop on Artificial Intelligence for Cultural Heritage (2017)
6. Fischer, A., et al.: Automatic transcription of handwritten medieval documents. In: 15th International Conference on Virtual Systems and Multimedia. IEEE (2009)
7. Flaounas, I., et al.: Research methods in the age of digital journalism: massive-scale automated analysis of news-content-topics, style and gender. Digit. J. **1**(1), 102–116 (2013)
8. Keysers, D., Deselaers, T., Rowley, H.A., Wang, L.-L., Carbune, V.: Multi-language online handwriting recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1180–1194 (2017)
9. Kim, Y., Jernite, Y., Sontag, D., Rush, A.M.: Character-aware neural language models. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI 2016, pp. 2741–2749. AAAI Press (2016)
10. Likforman-Sulem, L., Zahour, A., Taconet, B.: Text line segmentation of historical documents: a survey. Int. J. Doc. Anal. Recogn. **9**(2), 123–138 (2007)
11. Puigcerver, J., Toselli, A.H., Vidal, E.: ICDAR 2015 competition on keyword spotting for handwritten documents. In: 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1176–1180. IEEE (2015)
12. Sánchez, J.A., et al.: tranScriptorium: a European project on handwritten text recognition. In: Proceedings of the 2013 ACM Symposium on Document Engineering, pp. 227–228. ACM (2013)
13. Sánchez, J.A., Romero, V., Toselli, A.H., Vidal, E.: ICFHR 2014 competition on handwritten text recognition on tranScriptorium datasets (HTRtS). In: 2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 785–790. IEEE (2014)
14. Sudholt, S., Fink, G.A.: PHOCNet: a deep convolutional neural network for word spotting in handwritten documents. In: 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 277–282. IEEE (2016)