Transcription of Encoded Manuscripts with Image Processing Techniques

Alicia Fornés[1], Beáta Megyesi[2], Joan Mas[1]

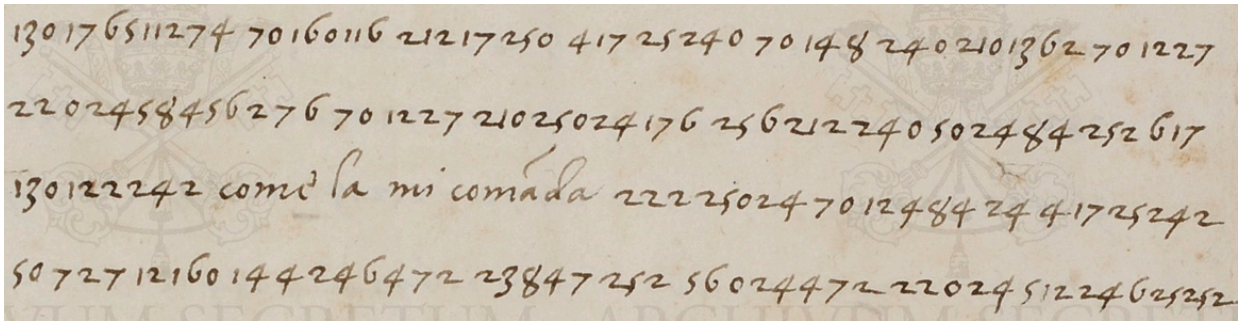
[1]. Computer Vision Center, Universitat Autònoma de Barcelona, Spain. afornes@cvc.uab.es
[2]. Dept. of Linguistics and Philology, Uppsala University, Sweden. beata.megyesi@lingfil.uu.se

Keywords: Image processing, Hand-written manuscripts, Automatic transcription, Historical cryptology.

Introduction

Historical hand-written manuscripts are an important source of information for our cultural heritage, and automatic processing of these can help exploring their content faster and easier.
A special type of hand-written manuscripts that are relatively common in archives and libraries are encrypted, secret documents, so called ciphers. Ciphers may contain and hide important information for the history of science, religion or diplomacy and therefore shall be decrypted and made accessible. The automatic decryption of historical hand-written ciphers is the main focus of the project *DECODE: Automatic decoding of historical manuscripts*[1]. In order to reveal the content of these secret messages, we collect and digitize ciphertexts and keys from Early Modern times, build a database, and develop software tools for (semi-)automatic decryption by cross-disciplinary research involving computer science, language technology, linguistics and philology.

Ciphers use a secret method of writing, often by transposition or substitution of characters, special symbols, already existing alphabets, digits, or a mixture of these. The encoded sequences are usually meticulously written and often segmented character by character to avoid any kind of ambiguity for the receiver to be able to decode the content, but continuous writing of some sequences where the symbols are connected also exists. In addition, the cipher sequences might be embedded in cleartext, i.e. texts in a known natural language, as illustrated in the picture below [ASV16:Segr.di.Stato/Portogallo/1A/16v@2016 Archivio Segreto Vaticano].



The first step for deciphering and making accessible the secret writing is their digitization and transcription. Transcription can be performed either by hand where a person types in the encrypted text symbol by symbol, or by (semi-)automatic means with a possible post editing by

[1] *DECODE* project: https://stp.lingfil.uu.se/~bea/decode/

manual validation. Manual transcription is time-consuming and expensive, and prone to errors. Automatic methods applied to ease the transcription process are preferable. However, image processing techniques developed so far for historical text manuscripts, such as the ones from the project *TransScriptorium*[2], are not fully adequate for dealing with encrypted documents for several reasons. First, the transcribing system cannot benefit from any lexicon or language model because the key is, a priori, unknown. Consequently, the use of an optical model alone is prone to errors, especially when there are ambiguities in the shape of digits/characters. Second, many ciphers contain a mixture of plaintext and encrypted text (ciphertext), which requires specifically adapted handwriting recognition methods. Third, the arcane nature of the symbols used calling for semiotic analysis, which requires the study of techniques closer to hand-drawn symbol recognition rather than handwriting recognition ones.

In this paper, we study the feasibility of the current image processing techniques in order to digitize ciphers by recognizing and transferring the symbols into a computer-readable format. For this purpose, we present a semi-automatic transcription method based on Deep Neural Networks, followed by a manual validation. We compare the results with a complete manual transcription, and analyze the human time effort of the two scenarios.

Image Processing Methodology

The handwriting recognition system has the following steps: First, each document has been binarized, deskewed, and the text lines have been segmented using projection profiles. The images of the text lines are the input of the Multi-Dimensional Long Short-Term Memory Blocks Neural Networks (MD-LSTMs) [GS08, VDN16]. Contrary to previous techniques applied for recognition (e.g. HMMs), MD-LSTMs obtain good results without the need of computing feature vectors from the image.

For each text line, the output of the network is a sequence of digits. The system also detects when a digit has a dot above or below. Whenever the confidence of the system when transcribing a certain digit is low, the symbol "?" is used. This denotes that an expert user must check it.

In this work, the networks were trained using 15 cipher pages from six different ciphers (with six different handwritings) in order to learn the handwriting style variability. For validation set, we used 5 cipher pages from the same ciphers as in the training set but different pages. It must be noted that the amount of training pages is not enough for the transcription of cleartext, so all text words appearing in the document are denoted as "x". The transcription must be performed by an expert.

Finally, and with the aim of improving the visualization of the results and facilitating the posterior validation and correction task, we used force alignment between the result of the neural network and the input image.

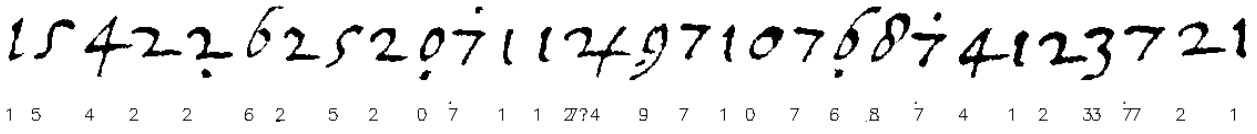
[2] *Transcriptorium* project: http://transcriptorium.eu

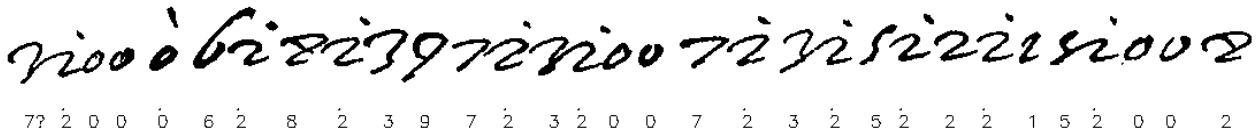
Manual vs. Automatic Transcription and Correction

For the tests, we chose 14 new unseen cipher pages, of which 12 pages were taken from four ciphers in the training set, and two pages came from two new, previously unseen ciphers, which means that these handwriting styles have not been learned during training. In this way, we can also analyze the generalization and scalability degree of the method for new handwriting styles.

To compare the speed of automatic versus manual transcription in a fair way, each manuscript was manually transcribed by one person, whereas the output from the automatic transcription was corrected and validated by a different person.

In the manual transcription, the transcriber opened the image of the cipher, and transcribed it symbol by symbol in a text file. Contrary, for validation, the transcriber opened the output from the automatic transcription as a picture where the cipher page was segmented line by line and the suggested transcription was reproduced below. As it can be observed in the figures below, the symbol "?" appears when the system is not confident on the transcribed digit. Also, if the system detects a dot above the digit, then the transcription also contains the dot.





The results obtained by manual transcription and validated transcription from automatic output were compared and shown in Table 1. In average, the automatic system transcribes the digits with an average accuracy of 88 %. However, one of the ciphers (Francia_18_3_233r), written by a writer whose handwriting was not represented in the training set, was more difficult to the system to automatically transcribe and accuracy decreased to 61%.

Cipher	No. of lines per page	Manual (mins)	Validation (mins)	Accuracy automatic	Manual mins/line	Validation mins/line
Francia_4_1_221r	3	5	4	92%	1.67	1.33
Francia_6_1_236r	31	50	47	92%	1.61	1.52
Francia_18_2_206v	24	45	41	81%	1.88	1.71
Francia_18_3_233r	20	45	30	61%	2,25	1.50
Francia_64_2_040v	24	25	30	92%	1.04	1.25
Francia_64_4_056v	26	20	52	87%	0.77	2.00
Francia_64_5_060v	25	20	26	94%	0.80	1.04
Francia_64_6_064v	16	10	13	94%	0.63	0.81
Spagna_423_2_297r	8	15	4	98%	1.88	0.50
Spagna_423_3_300v	2	3	3	74%	1.50	1.50
Spagna_423_4_374r	10	15	10	85%	1.50	1.00
Spagna_423_6_388v	21	35	15	95%	1.67	0.71
Spagna_423_7_391r	13	15	8	97%	1.15	0.62
Spagna_423_9_491v	21	25	20	93%	1.19	0.95
Average	**17.43**	**23.43**	**21.6**	**88%**	**1.39**	**1.17**

Table 1. Summary of results per line and cipher given the time in minutes for manual transcription, as well as the validation and correction of automatic transcription; the accuracy of automatic transcription, and the average rate of manual transcription and validation per line. Rows in red color denote those where the manual transcription is faster.

The results show that in most cases manual transcription is 15% slower on average compared to the automatic transcription with post-editing if the accuracy of the image processing is above 90%. When accuracy is lower, validation time usually increases because the more transcription errors we find, the more effort it takes to localize both the wrong symbol(s) in the picture and in the transcription file. For each error, the user usually starts to read the line from the beginning. Noteworthy also that we do not count the time it takes to prepare and train the automatic transcription models, including the preprocessing of the images (cut the margins, clean the bleed through, etc.) and the time for training the models. We also noted that the validators would have benefited from a user-friendly transcription tool where transcription suggested by the model was aligned with the original symbol in the picture. However, the automatic transcription clearly helped to differentiate between two different symbols written similarly, thereby helping the user to identify the symbol set represented in the cipher.

Conclusion

We have shown that image processing can be used as base for transcription followed by a post-processing step with user validation and correction. Even though image processing techniques need to be trained today on individual handwritings to reach high(er) accuracy, they might be of great help to identify the symbol set represented in the manuscript and to make clear distinctions between symbols, hence can be used as a support tool for the transcriber.

In this work, we focused on ciphers without any esoteric or other symbol sets, which might be more difficult for an automatic recognition system. Also, we have identified only cipher sequences; cleartext was only detected without any further transcription.

In the future, we would like to test to combine image processing and automatic decryption in one step to skip the time-consuming transcription step and create synergy effects as both image processing and automatic decryption tools rely on language models that could be used simultaneously. Another alternative is image processing for validation of the manual transcription, which might be an interesting alternative to investigate in the future.

Acknowledgement
This work has been partially supported by the Spanish project TIN2015-70924-C2-2-R, the Swedish Research Council DECODE project grant E0067801, and the *Ramon y Cajal* Fellowship RYC-2014-16831.

References

- [ASV16] Segr.di.Stato/Portogallo/1A/16v@2016 Archivio Segreto Vaticano. The picture has been reproduced by the kind permission of Archivio Segreto Vaticano, all rights reserved.

- [FB14] V. Frinken, H. Bunke. Continuous Handwritten Script Recognition. Book chapter, Handbook of Document Image Processing and Recognition. Springer-Verlag, 2014.

- [GS08] A. Graves and J. Schmidhuber. “Offline handwriting recognition with multidimensional recurrent neural networks”. Neural Information Processing Systems, 2008.

- [VDN16] P. Voigtlaender, P. Doetsch, H. Ney. Handwriting Recognition with Large Multidimensional Long Short-Term Memory Recurrent Neural Networks. International Conference on Frontiers in Handwriting Recognition, 2016.