sciendo

TATRA MOUNTAINS
Mathematical Publications
DOI: 10.2478/tmmp-2022-0019
Tatra Mt. Math. Publ. **82** (2022), 65–86

AUTOMATED TRANSCRIPTION OF HISTORICAL ENCRYPTED MANUSCRIPTS

Eugen Antal — Pavol Marák

Institute of Computer Science and Mathematics
Slovak University of Technology in Bratislava
SLOVAKIA

ABSTRACT. This paper deals with historical encrypted manuscripts and introduces an automated method for the detection and transcription of ciphertext symbols for subsequent cryptanalysis. Our database contains documents used in the past by aristocratic families living in the territory of Slovakia. They are encrypted using a nomenclator which is a specific type of substitution cipher. In our case, the nomenclator uses digits as ciphertext symbols. We have proposed a method for the detection, classification, and transcription of handwritten digits from the original documents. Our method is based on Mask R-CNN which is a deep convolutional neural network for instance segmentation. Mask R-CNN was trained on a manually collected database of digit annotations. We employ a specific strategy where the input image is first divided into small blocks. The image blocks are then passed to Mask R-CNN to obtain detections. This way we avoid problems related to the detection of a large number of small dense objects in a high-resolution image. Experiments have shown promising detection performance for all digit types with minimum false detections.

1. Introduction

An automated transcription of historical manuscripts is an open research question in general. Manuscripts may vary based on the time period, used language, writing style, etc. Moreover, transcribing a historical ciphertext (or a cipher key) can be an even more challenging task, because these systems may consist of a large number of various symbols (glyphs), numbers, and letters.

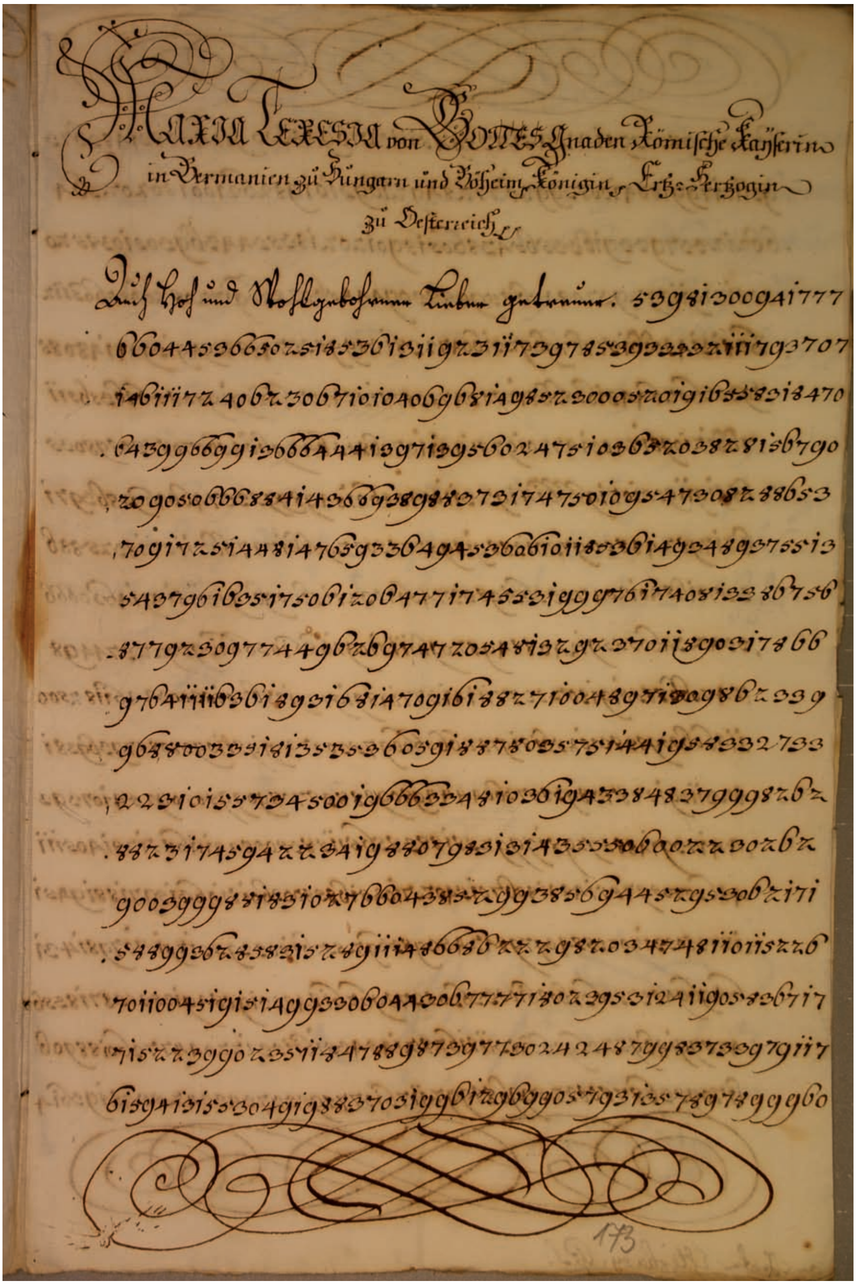
2020 Mathematics Subject Classification: 68T07, 94A60, 01A50.
Keywords: historical ciphers, nomenclator, manuscript, transcription, machine learning, deep convolutional neural networks, Mask R-CNN.
Supported by the Grant VEGA 2/0072/20.

In this work, we are focusing on the digitization and processing of historical ciphers used in the past by aristocratic families living in the territory of today's Slovakia. The archival documents which are the subject of our research are deposited in several preserved fonds of these aristocratic families in the Slovak National Archive in Bratislava. The encryption system used in these documents is called *nomenclator* [7, 15], which is a complex encryption system consisting of several simpler encryption subsystems linked together during the encryption. These subsystems are mostly based on different types of substitution. The main characteristics of a nomenclator are:

> *"A nomenclator mostly contains a substitution of letters (monoalphabetic or homophonic substitution) in a combination with substitution of $n-grams$ (bigram and/or trigram substitution), codes, and nulls. It is not widespread, but some nomenclators contain a polyalphabetic substitution, too. The sub-encryption systems (encryption rules) are described by a cipher key, which is very characteristic: the cipher key is mostly drawn on a large paper sheet; the individual sub-encryption systems are mostly graphically separated; the cipher text alphabet is often represented by (combinations of) letters, numbers, and special symbols/glyphs."* [3]

A typical ciphertext from our collection is shown in Figure 1 and cipher key in Figure 2. The used cipher symbol set from our collection consists of digits only (including special number modifications). Luckily, the writing style of the encrypted text is clean and the used digits are separated by relatively large spaces. The symbols are therefore easier to read. The aforementioned text readability may be attributed to an effort to minimize the possibility of error occurrence. Writing such clear and easy-to-read encrypted parts requires a lot of skill and patience.

In order to analyze and solve the manuscripts, we first need to perform a transcription of the cipher text represented by image to editable text. One may do it manually, which is a very time-consuming and error-prone process. Another possibility is to use a modern automated method, such as deep convolutional neural networks. Our method is based on Mask R-CNN which is a popular supervised object detector. Once the detector learns digit representation from a sufficient number of examples, it can be used to detect digits from new unseen documents. All the detected digits are finally read in the correct direction to form an editable text document which can be used for cryptanalysis purposes.


Maria Theresia von Gottes Gnaden Römische Kayserin
in Germanien zu Hungarn und Böheim Königin, Erz-Herzogin
zu Oesterreich etc.

Liebe hoch und Wohlgebohrene Lieben getreuen.

FIGURE 1. Encrypted message from 1756 (Slovak National Archives, fond Esterházi - čeklíska vetva, box n. 634).

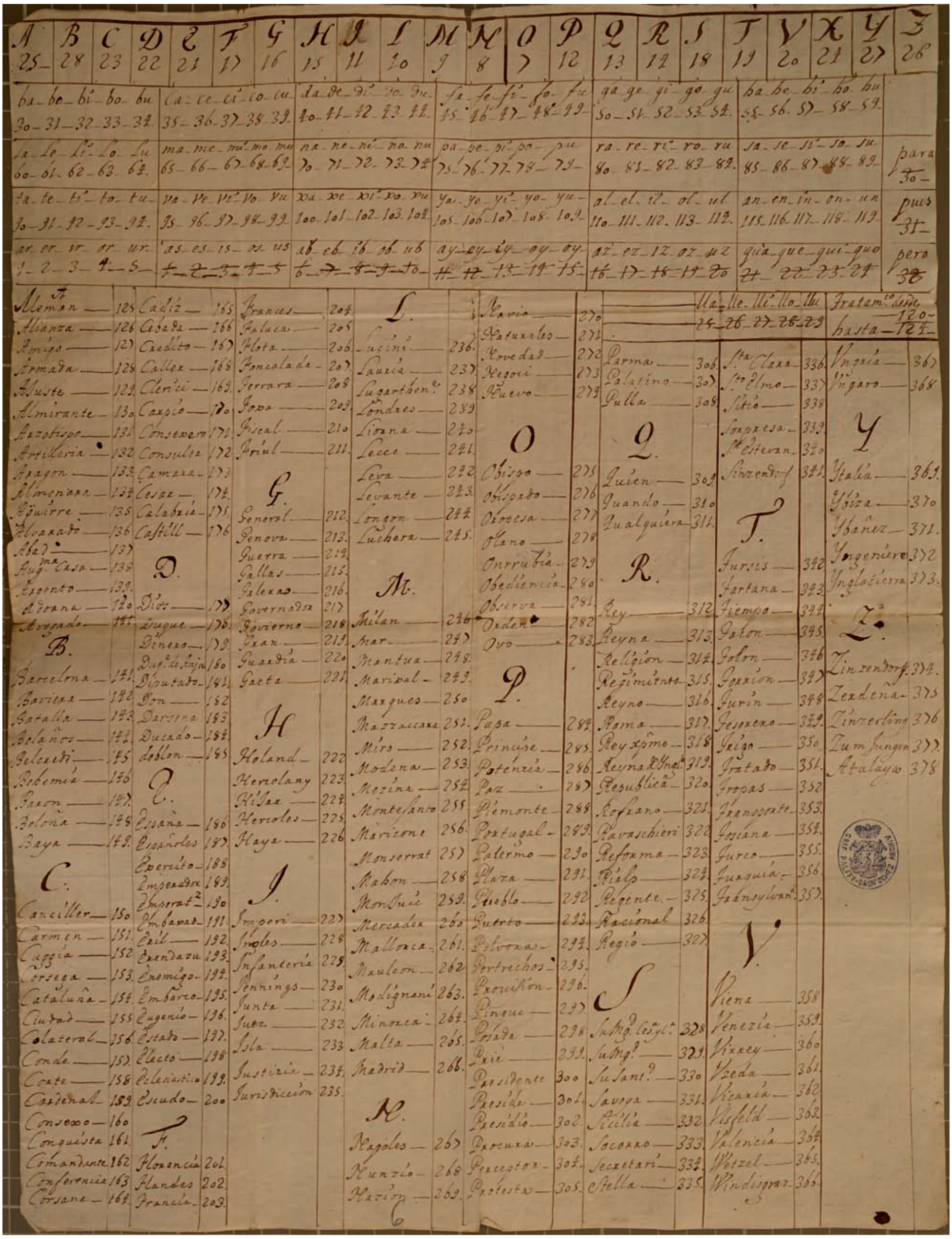


FIGURE 2. Cipher key example (Slovak National Archives, fond Pálffy--Daun, Klasse XXXIII – Wierich Daun, fasc. 22).

2. Related work

Historical ciphers (especially nomenclator systems) have been intensively researched in recent years. Many important publications on this subject are presented annually at the International Conference on Historical Cryptology (HistoCrypt). The design and structure of historical cipher keys were investigated in [3, 11, 14]. These publications are related to the two ongoing projects, namely:

DECRYPT (`https://de-crypt.org/`) [10] and

HCPortal (`https://hcportal.eu`) [4, 5].

These projects are mainly focusing on the digitization and processing of encrypted documents and cipher keys, and developing new methods to solve these ciphers. In some cases, large collections of documents from a particular time period or geographic location [6, 9] are studied which can also help to better understand some aspects of historical cryptography.

One of the fundamental stages of historical encrypted manuscript processing is its automated transcription which takes an image containing the original manuscript and produces an editable text corresponding to the manuscript. This is a rather challenging task that requires a robust method to address issues such as the recognition of complex patterns and handling poor image quality. Nowadays, these problems can be solved using the machine learning approach.

During the literature review, we came across several solutions intended for historical text recognition. There is a well-known web-based and offline solution called Transkribus capable of text recognition and transcription of documents written in any language [13]. Similarly, authors in [8] introduced a novel deep learning architecture named DIGITNET, and a large-scale handwritten digit dataset named DIDA, to detect and recognize handwritten digits in historical document images written in the 19th century. Their solution was based on a well-known YOLO detector. Another attempt was made by researchers in [12] who proposed a handwritten cipher text recognition based on few-shot object detection.

Our proposed solution is based on the modern and robust convolutional neural network Mask R-CNN [16], which belongs to the family of state-of-the-art supervised semantic segmentation methods. Mask R-CNN takes an image of a predefined size as input and performs detection producing a bounding box and polygonal mask for each detected object. Moreover, detected objects are assigned a class label.

Authors of this paper conduct their research at the Institute of Computer Science and Mathematics (Slovak University of Technology). There are several final theses dealing with the problem of historical encrypted document processing which were supervised at the institute. In [17], the comprehensive analysis and comparison of existing handwritten digit datasets are presented. In addition, two well-known object detectors, Mask R-CNN and YOLOv5, are examined and their detection accuracy is evaluated. Another research was conducted in [18], where image preprocessing and recognition of handwritten digits and their special modifications are presented. The recognition is performed using state-of-the-art convolutional neural networks (VGG, ResNet, ResNeXt, Inception) which form robust ensembles to boost classification performance. Moreover, the entire solution is developed as an interactive web application for highly customized handwritten document analysis. Finally, in [19] and [20], authors created a large collection of handwritten digit annotations and a method for digit detection and web-based document transcription.

3. Automated transcription

Over the course of our research, we have collected a large number of encrypted texts and nomenclator keys. In order to solve the ciphers, one needs to convert the image representation of the document to editable text or symbols for further cryptanalysis. Automated transcription of handwritten documents is the major contribution of this paper, however, it is just a single step in our research workflow which can be summarized as the sequence of the following steps:

(1) Research in archives which involves collection and digitization of cipher keys and encrypted documents.

(2) Automated transcription of the obtained documents based on machine learning:
   (a) Creating digit annotations.
   (b) Object detector training using the annotations.
   (c) Digit detection, classification, and transcription to obtain editable text for subsequent analysis.

(3) Analysis and solving the ciphers from the transcribed documents.

In Figure 3, we see the most important steps of the transcription procedure. For our needs, we manually created a new dataset of handwritten digit annotations using a Python graphical image annotation tool called LabelMe. We created 12 433 polygonal annotations of digits from several handwritten documents. Currently, we are not aware of any similar public dataset of such extent and precision so we consider this to be a substantial contribution in the field. The correctness of digit annotations was further verified by the experts from the Institute of History of the Slovak Academy of Sciences. Finally, the dataset was split into three subsets for training, validation, and testing. The next step was training our digit detector based on Mask R-CNN. The detector is responsible for locating digits within the document. The trained detector was evaluated on the test dataset. Digit bounding boxes and labels serve for line detection and final transcription.

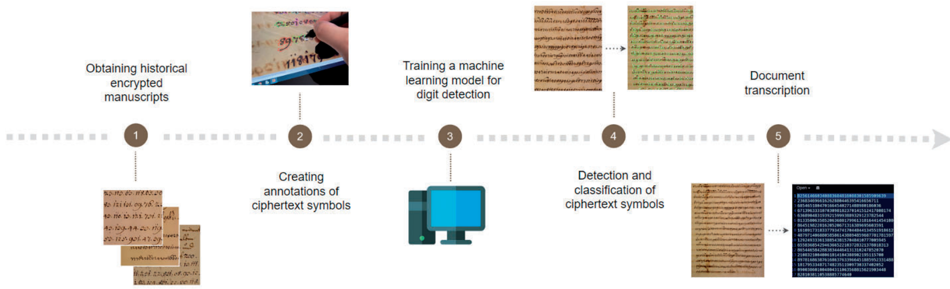


FIGURE 3. Automated handwritten document transcription workflow.

3.1. Database of manuscripts

Our collection of encrypted manuscripts and cipher keys consists of several hundred pages. These documents are deposited in the Slovak National Archive in three different fonds of aristocratic families:

- Esterházi,
- Pálffy-Daun,
- Amade-Üchtritz.

We made digital copies (photographs) of these documents in a high resolution (4160 $\times$ 6240 pixels). The used camera was mounted on a stand and we used additional light sources. The handwriting is clean and the cipher symbols are clearly separated and easy-to-read on most documents (see Figure 4). However, there are examples of lower quality (see Figures 5, 6, and 7).

The collected manuscripts can be separated into different types of encrypted documents:

- fully or partially encrypted messages (Figure 1),
- encrypted message where the plaintext is written above/below the lines of the ciphertext (Figure 8),
- encrypted parts in a diary (Figure 9),
- draft message containing encrypted passages,
- draft message containing encrypted passages where the plaintext is written above/below the lines of the ciphertext (Figure 6).

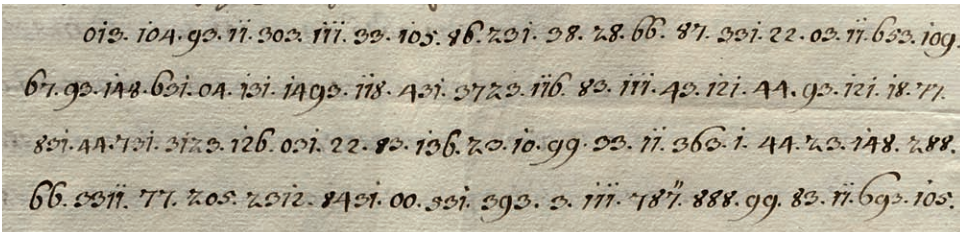


FIGURE 4. Clean and easy-to-read ciphertext example (Slovak National Archives, fond Esterházi - čeklíska vetva, box n. 634).

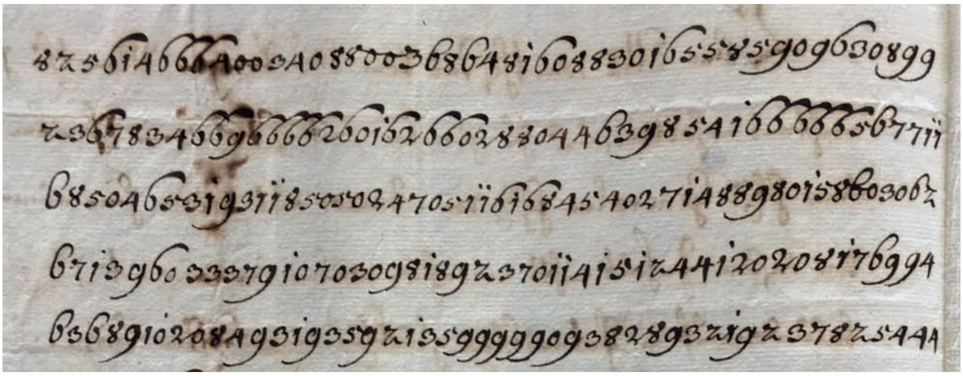


FIGURE 5. Lower ciphertext quality - weak background noise and smudged parts (Slovak National Archives, fond Esterházi - čeklíska vetva, box n. 634).

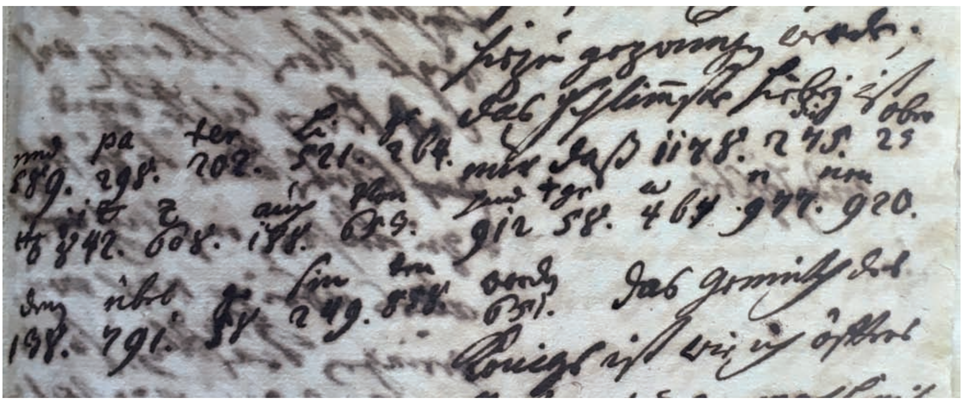


FIGURE 6. Lower ciphertext quality - strong background noise (Slovak National Archives, fond Esterházi - čeklíska vetva, box n. 635).

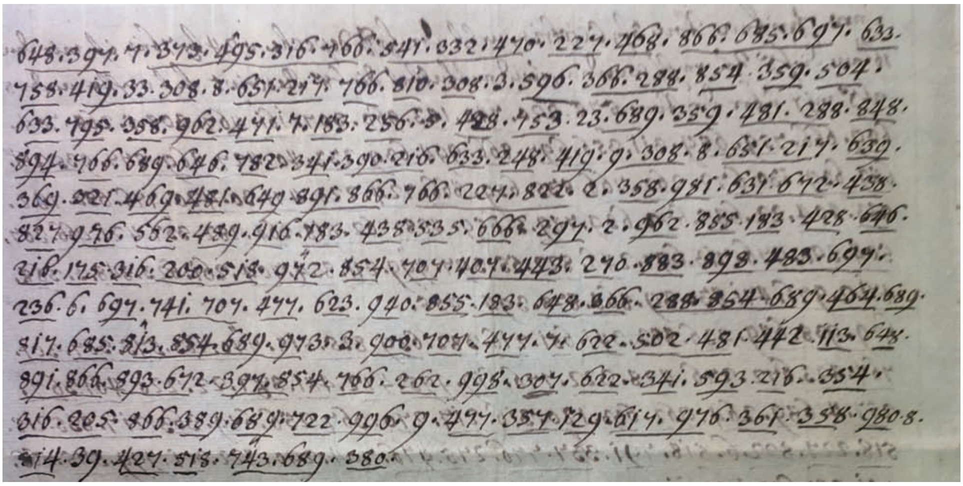


FIGURE 7. Lower ciphertext quality - underlined text and strong background noise (Slovak National Archives, fond Esterházi - čeklíska vetva, box n. 631).

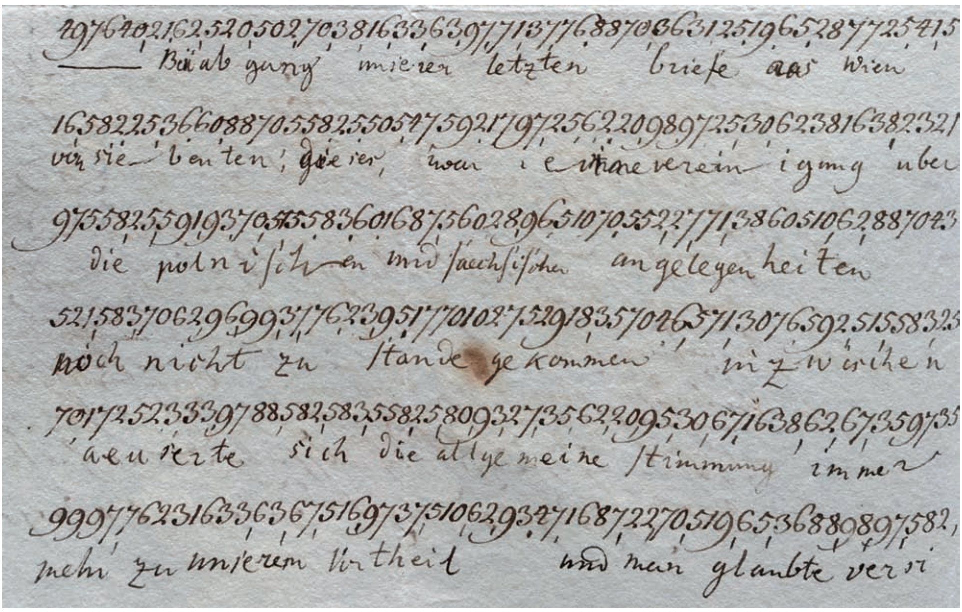


FIGURE 8. Plaintext written below the ciphertext (Slovak National Archives, fond Amade-Üchtritz, box n. 136).

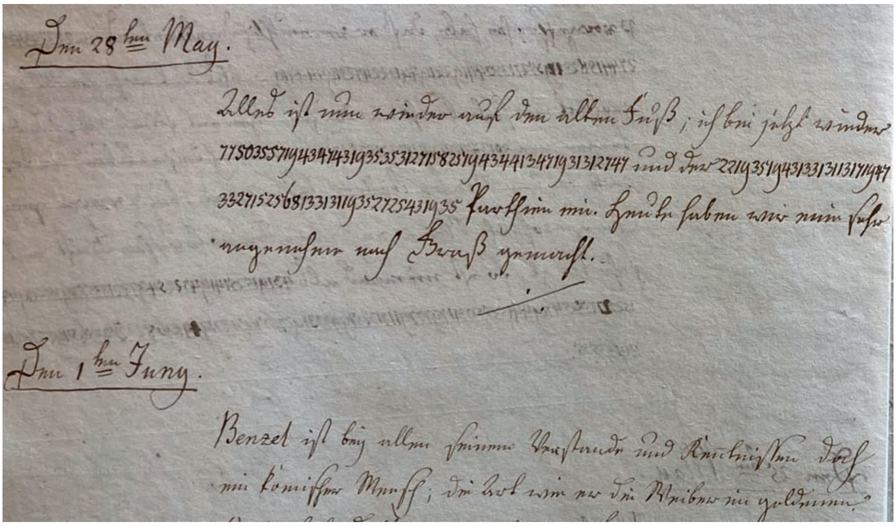


Figure 9. Encrypted diary parts (Slovak National Archives, fond Amade-Üchtritz, box n. 150).

The ciphertext symbol set examined so far consists of numbers only, including some markups (see Figure 10). Moreover, some of the ciphertexts consist of numbers separated with a dot (see Figures 4 and 7), other ciphertexts consist of numbers without separators (see Figures 1 and 5).

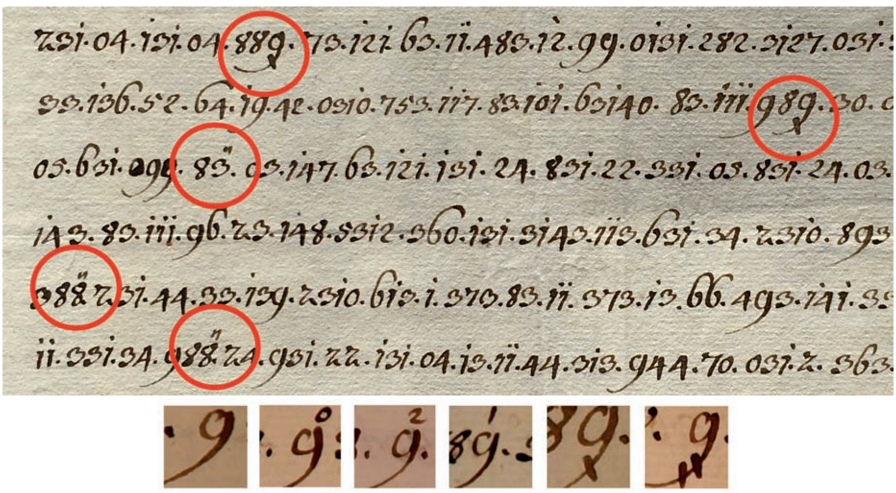


Figure 10. Special digit markups (Slovak National Archives, fond Esterházi - čeklíska vetva, box n. 634).

3.2. Object detector

We used a machine learning approach to detect and classify digits in the handwritten encrypted documents. Specifically, we employ Mask R-CNN supervised instance segmentation algorithm. Mask R-CNN is a region-based deep convolutional neural network generating high-quality segmentation masks. Internally, Mask R-CNN takes an input image and extracts salient features using a predefined deep convolutional neural network, e.g., ResNet. Features are then passed to the subsequent layers responsible for a region proposal (RPN network) and prediction of a class and rectangular bounding box. Moreover, additional convolutional layers produce a high-resolution polygonal mask. Mask R-CNN scheme is depicted in Figure 11.

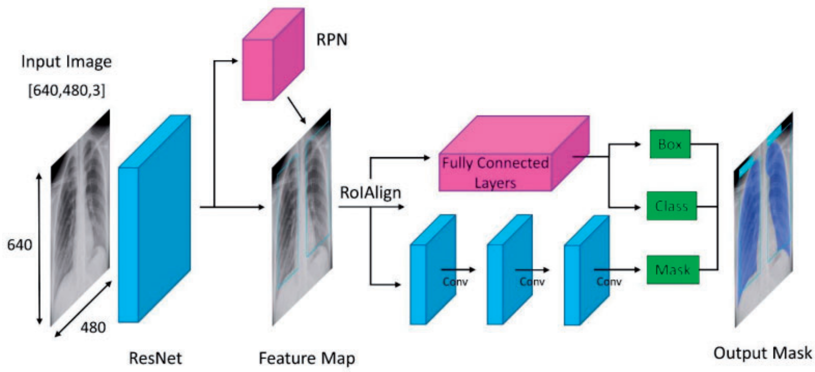


Figure 11. Mask R-CNN architecture [21] (this example shows an X-ray image object detection).

In our case, Mask R-CNN produces digit detections where each potential digit is represented by a bounding box, class label, pixel-level mask, and classification confidence as seen in Figure 12.

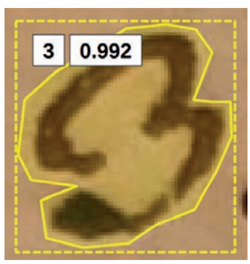


Figure 12. Sample digit detection performed by Mask R-CNN on our dataset (value 3 denotes a class label whereas value 0.992 represents a classification confidence from interval $\langle 0, 1 \rangle$).

3.3. Annotations

The training process of Mask R-CNN requires ground-truth annotations of objects of interest. An accurate digit detector must be trained on a rich training dataset covering intra-class and inter-class variability. Objects of interest, digits from 0 up to 9, were annotated using the LabelMe software tool (see Figure 13). LabelMe allows drawing geometric shapes to spatially delimit the object. We used polygonal annotations to create accurate masks and to avoid digit overlapping.

FIGURE 13. Polygonal digit annotations created by a touch pen using LabelMe software.

3.4. Training and testing

We have created 12 433 digit annotations from 18 document images using the LabelMe software (these images were manually analyzed and transcribed by experts so we could later verify the results of detection). This dataset was split into three subsets, namely training, validation, and testing subset.

The split ratio was 70 : 15 : 15, respectively. The digit distribution in the dataset is shown in the table in figure Figure 14. Digits 1 and 3 were more frequent than other digits which can be attributed to the characteristics of the encryption system.

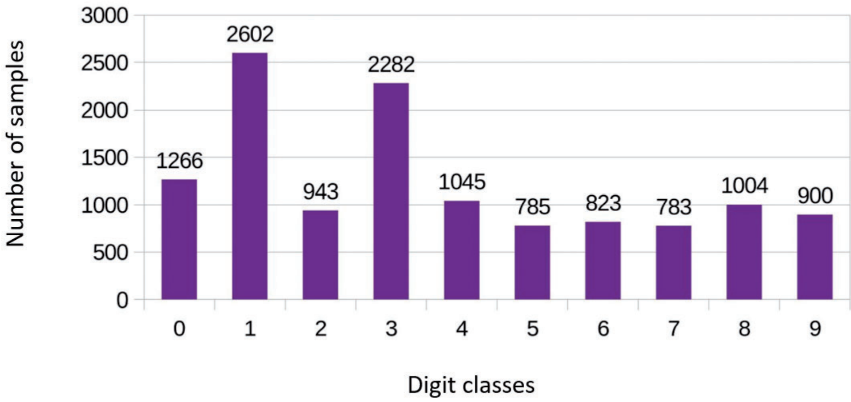


Figure 14. Distribution of digit annotations.

Detecting large number of digits in a high-resolution image is a difficult task since Mask R-CNN performance decreases when detecting small dense objects. To solve this issue, we divided the entire document image into smaller 128 × 128 pixel blocks (see Figure 15) and subsequently performed detection in the blocks.

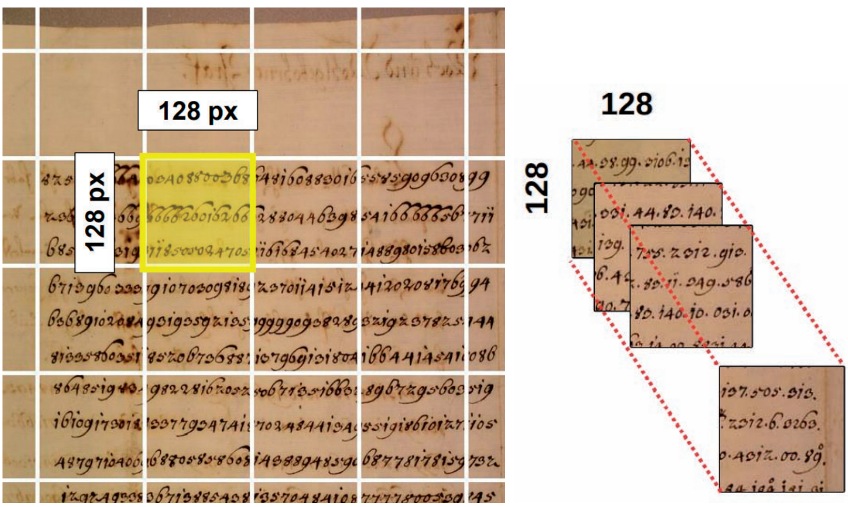


Figure 15. Division of the document image into smaller 128 × 128 blocks.

Figure 16 shows the original document divided into the blocks and the result of digit detections (green rectangles) inside each of the blocks.

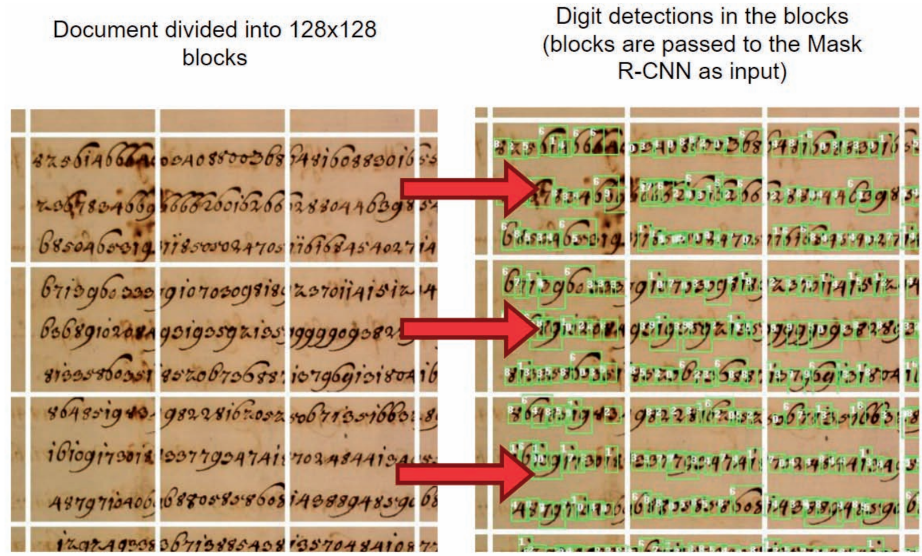


FIGURE 16. Block-level digit detections in the document.

3.5. Transcription

Digit detection and classification are followed by an automated transcription. The overall procedure is as follows:

(1) Calculation of bounding box (B-box) centers.

(2) Calculation of histogram for vertical coordinates of B-boxes.

(3) Local extrema detection in the histogram which leads to line detection.

(4) Reading digits in the right direction.

(5) Exporting digits to the text file.

First, we need to compute a histogram of digit bounding box centers. The histogram reveals the distribution of centers on the vertical axis. We used 4-pixel wide bins when plotting the histogram as seen in Figure 17. Histogram peaks denote line positions. We used Python Scipy package to detect the peaks.

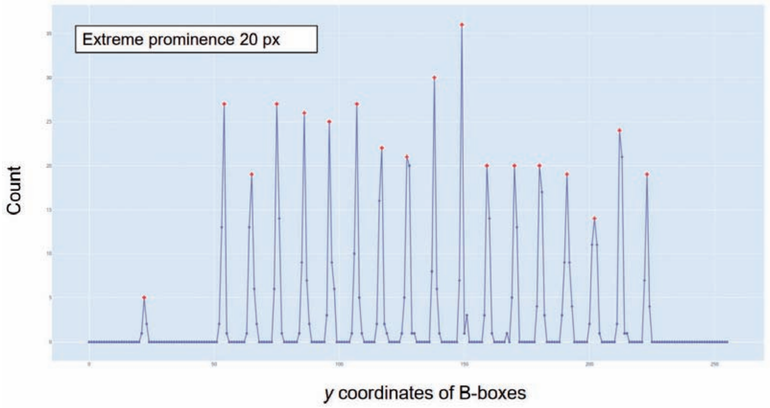


FIGURE 17. Histogram of bounding box centers revealing line positions.

Following histogram peak detection, we proceed to obtain line positions. Line positions correspond to the histogram peaks with slight vertical tolerance of $+/-$ 12 pixels. Line detections along with line height tolerance and detected digits assigned to the line are visualized in Figure 18.

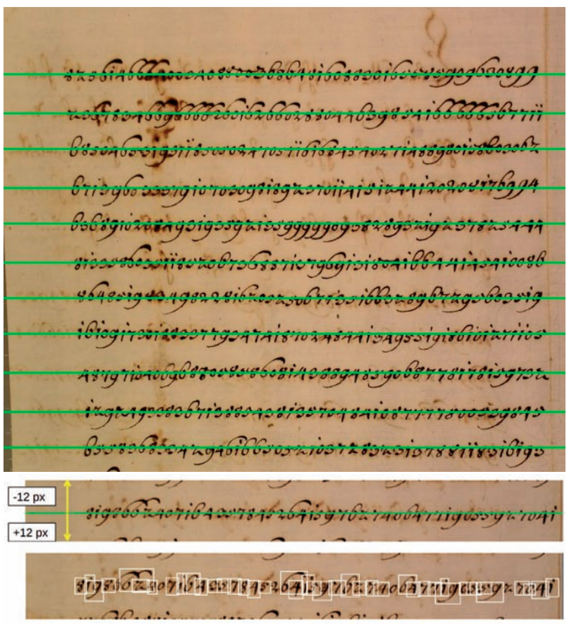


FIGURE 18. Line detection (green lines) and extraction of digits in the line (white boxes).

Detected digits are assigned to the line based on the distance of their B-box centers to the line's vertical position, taking the aforementioned tolerance into account. Reading digits assigned to the individual lines from left to right results in the transcription of the entire document. Digits are exported in the editable form to the text file (see Figure 19).

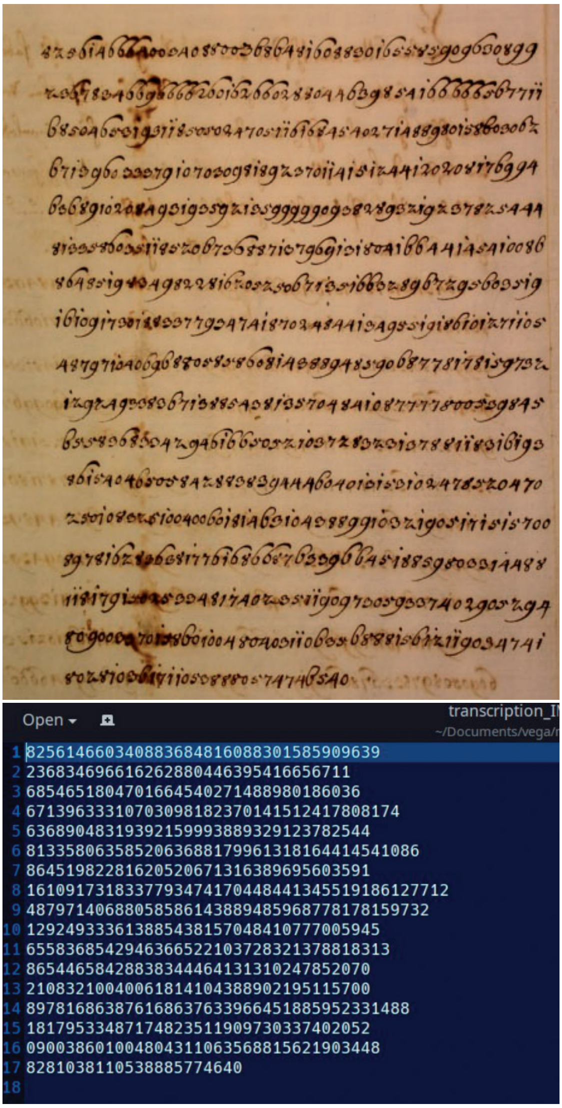


FIGURE 19. Encrypted document transcription (the original image is at the top, the transcribed digits are at the bottom).

4. Results and discussion

In this paper, we present our automated method for the detection, classification, and transcription of digits that are found in historical encrypted manuscripts. Our system based on Mask R-CNN achieves notable accuracy results on the test dataset (1 554 samples of digits) reaching overall digit classification accuracy as high as 99.5 %. These results were achieved after a relatively short period of training on GPU (100 epochs, using ResNet50 as the backbone feature extraction network). Accuracy results are summarized in Table 1.

TABLE 1. Digit classification accuracy achieved on the test set.

Digit	Number of incorrect classifications	Number of missed digits	Number of test samples	Classification accuracy (%)
0	0	2	186	98.92
1	1	0	223	99.55
2	1	0	111	99.09
3	0	0	159	100
4	1	0	146	99.31
5	1	0	129	99.22
6	0	0	157	100
7	0	0	124	100
8	1	0	182	99.45
9	0	0	137	100

Our system performs well on all digit classes and deals with image quality variations relatively well. Figure 20 shows digit detections across the entire encrypted document.

During our experiments, we also investigated situations where the manuscript contains mixed encrypted/unencrypted parts. This is where we wanted our algorithm to detect numbers in the encrypted parts only. There is a related ongoing research focused on the detection of encrypted regions which may decrease false digit detections. As depicted in Figure 21, we see that the detector produces only a very limited number of false detections outside the encrypted regions.

We also address the issue of missing digit detections which may occur when digits are located on the interface of the two consecutive image blocks. In [19] and [20], authors contribute to our research by introducing an algorithm that divides the input image into blocks using a differently shifted grid. This way we perform detection in the blocks with various offsets and combine the detection results so that the duplicate detections are removed and the results do not suffer from missing digit detections on the block interfaces.

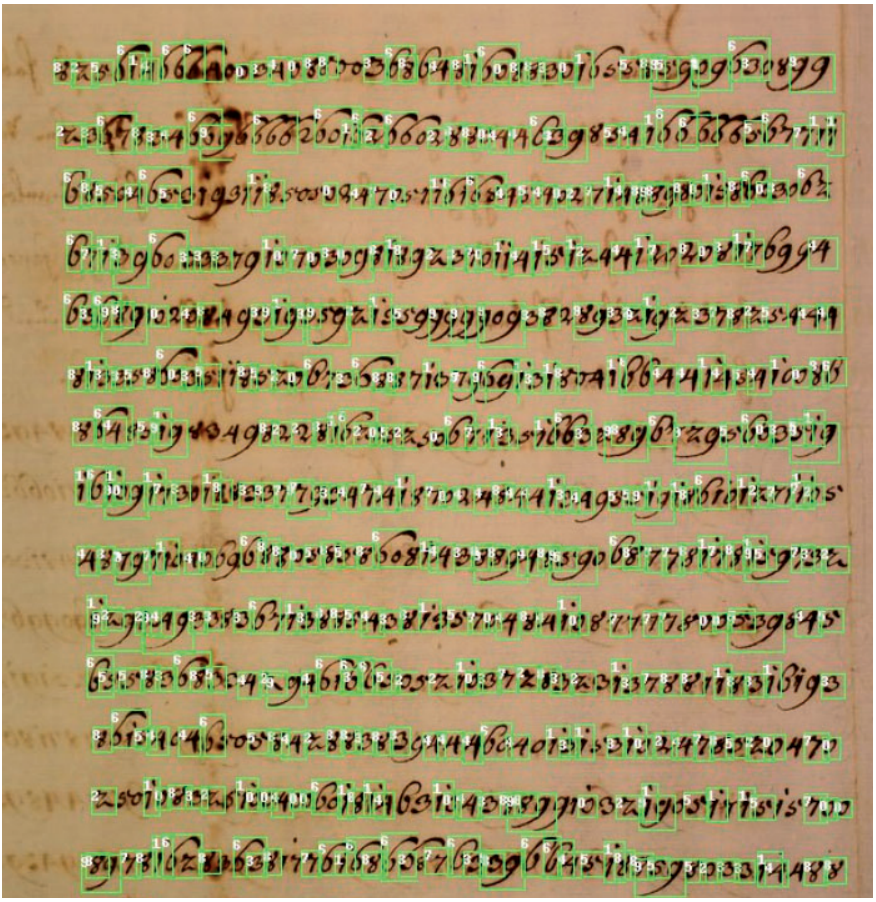


FIGURE 20. Example of digit detection in the entire encrypted document.

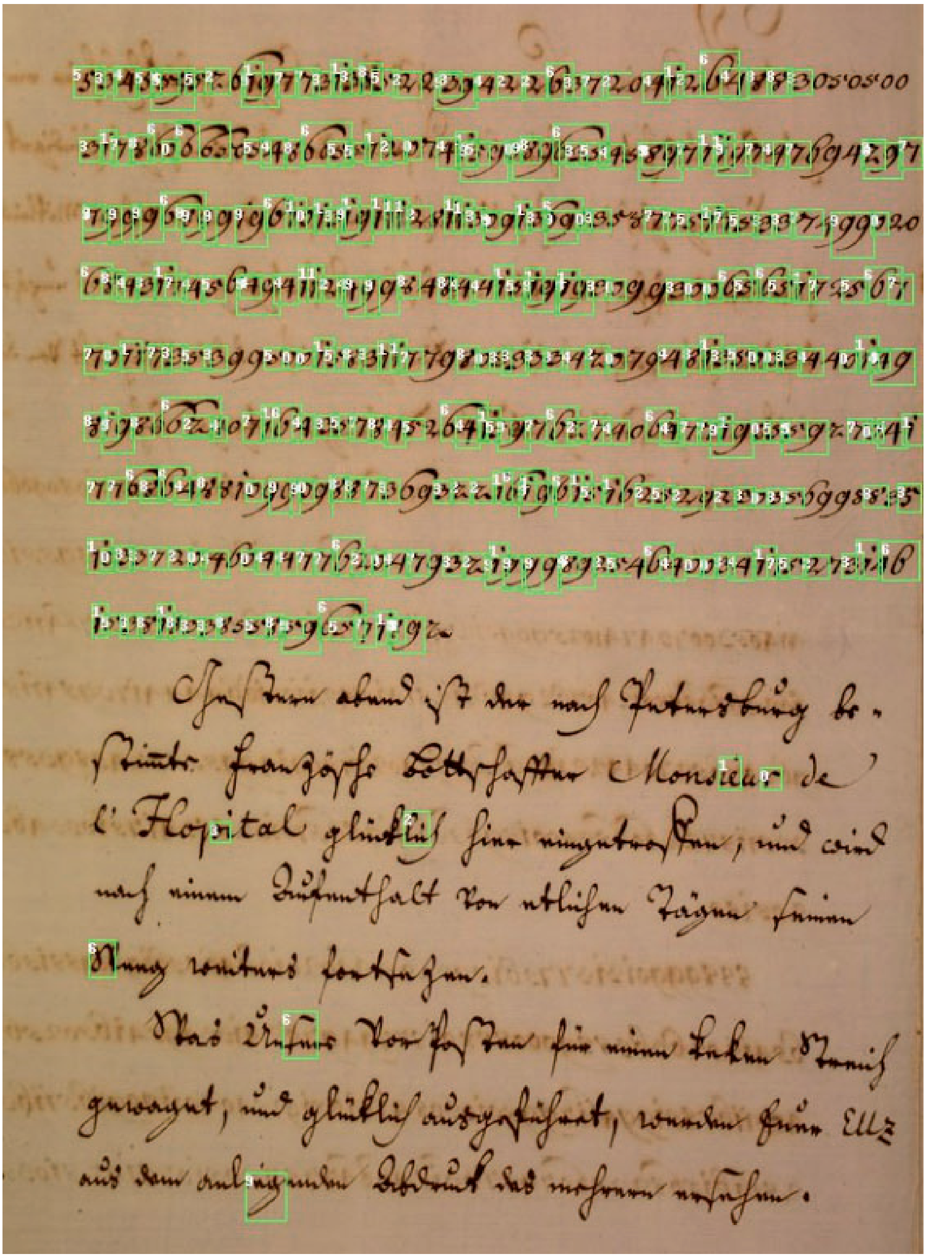


FIGURE 21. Example of digit detection in the document with mixed encrypted (top region) and unencrypted (bottom region) content.

5. Conclusion

This paper presents an ongoing research focusing on historical manuscripts encrypted using the nomenclator cipher system and their automated transcription. This cipher system uses digits for the representation of a ciphertext. We have a large collection of unsolved digitized encrypted documents and keys of varying structure and quality. We have developed an automated method for the detection, classification, and transcription of handwritten digits. Our system is based on the popular Mask R-CNN object detector. We created a large database of digit annotations and trained the detector. Testing and experiments indicate promising results when it comes to classification accuracy and capability to deal with images of varying quality. Furthermore, the adopted transcription technique turned out to be relatively accurate in detecting lines and reading symbols to form a final editable text document.

In addition to ciphertexts, we are also addressing the processing of cipher keys (nomenclators), which is a challenging task. These keys are mostly drawn on a paper sheet and the individual sub-encryption systems are visually separated. We are working on a (semi)automated computer vision method to identify, separate, and process the individual sub-encryption parts.

We plan to publish our developed transcription tools and our dataset (polygonal annotations) of historical handwritten digits and cipher symbols as open-source projects (available for other researchers). All projects will be documented and integrated into the *Portal of Historical Ciphers*[1] [4, 5] which is a special online project focusing on historical cryptology. We believe that our results can help other researchers avoid the need for time-consuming manual transcription of handwritten documents, not only in the field of historical cryptology.

REFERENCES

[1] ANTAL, E.: *Modern Cryptanalysis of Classical Ciphers*. PhD. Thesis, STU in Bratislava, 2017. (In Slovak).

[2] ANTAL, E.—ELIÁŠ, M.: *Evolutionary computation in cryptanalysis of classical ciphers*, Tatra Mt. Math. Publ. **70** (2017), 179–197.

[3] ANTAL, E.—MÍRKA, J.: *Wrong design of cipher keys: Analysis of historical cipher keys from the Hessisches Staatsarchiv Marburg used in the Thirty Years' War*, in: Proceedings of the 5th International Conference on Historical Cryptology, HistoCrypt 2022, Linköping University Electronic Press, pp. 1–11, DOI: https://doi.org/10.3384/ecp188387

[1] `https://hcportal.eu`

[4] ANTAL, E.—ZAJAC, P.: *HCPortal oderview*, in: Proceedings of the 3rd International Conference on Historical Cryptology, HistoCrypt 2020 no. 171, Linköping University Electronic Press, pp. 18–20.

[5] ANTAL, E.— ZAJAC, P.: *HCPortal modules for teaching and promoting cryptology*, in: Proceedings of the 4th International Conference on Historical Cryptology, HistoCrypt 2021, Linköping University Electronic Press, pp. 1–11. `https://doi.org/10.3384/ecp183151`

[6] ANTAL, E.—ZAJAC, P.—MÍRKA, J.: *Solving a mystery from the Thirty Years' War: Karel Rabenhaupt ze Suché's Encrypted letter to Landgravine Amalie Elisabeth*, in: Proceedings of the 4th International Conference on Historical Cryptology, HistoCrypt 2021, Linköping University Electronic Press, pp. 12–24. `https://doi.org/10.3384/ecp183152`

[7] KAHN, D.: *The Codebreakers: The Comprehensive History of Secret Communication from Ancient Times to the Internet*, Scribner, New York, 1996.

[8] KUSETOGULLARI, H. ET AL.: *DIGITNET: A deep handwritten digit detection and recognition method using a new historical handwritten digit dataset.* Big Data Research, **23** (2021), 100182, `https://doi.org/10.1016/j.bdr.2020.100182`

[9] LÁNG B.: *Was it a sudden shift in professionalization? Austrian cryptology and a description of the staatskanzlei key collection in the Haus-, Hof- und Staatsarchiv of Vienna*, in: *Proceedings of the 3rd International Conference on Historical Cryptology, HistoCrypt 2020*, Linköping University Electronic Press, pp. 87–95.

[10] MEGYESI, B.—ESSLINGER, B.—FORNÉS, A. —KOPAL, N.—LÁNG, B.—LASRY, G.—DE LEEUW, K. —PETTERSSON, E.—WACKER, A.—WALDISPÜHL, M.: *Decryption of historical manuscripts: the DECRYPT project.* Cryptologia, **44** (2020), no. 6, 545–559.

[11] MEGYESI, B.—TUDOR, C.—LÁNG, B.—LEHOFER, A.: *Key design in the early modern era in Europe*, in: *Proceedings of the 4th International Conference on Historical Cryptology, HistoCrypt 2021*, pages 121–130. Linköping University Electronic Press.

[12] SOUIBGUI, M. A.—FORNÉS, A.—KESSENTINI, Y.—TUDOR, C.: *A few-shot learning approach for historical ciphered manuscript recognition*, in: 25th International Conference on Pattern Recognition (ICPR 2020), IEEE (2021), 5413–5420.

[13] TRANSKRIBUS TEAM: Transkribus: `https://readcoop.eu/transkribus/`

[14] TUDOR C.—MEGYESI B.—LÁNG B.: *Automatic key structure extraction*, in: *Proceedings of the 3rd International Conference on Historical Cryptology, HistoCrypt 2020*, Linköping University Electronic Press, pp. 146–152.

[15] VON ZUR GATHEN, J.: *CryptoSchool.* Springer-Verlag, Berlin, 2015.

[16] HE, K.—GKIOXARI, G.—DOLLÁR, P.—GIRSHICK, R.: *Mask R-CNN*, in: *IEEE International Conference on Computer Vision (ICCV), 2017*, pp. 2980–2988, DOI: 10.1109/ICCV.2017.322

[17] MIKUŠ, F.: *Comparison of Artificial Intelligence Methods for Handwritten Digit Recognition.* Bachelor Thesis, FEI STU, Bratislava, Slovakia, 2022. (In Slovak)

[18] KIRSCHOVÁ, P.: *Handwritten Digit Recognition Based on Deep Learning Methods* Master Thesis, FEI STU, Bratislava, Slovakia, 2022. (In Slovak)

[19] TÓTHOVÁ, L. : *Segmentation and Recognition of Encrypted Handwritten Historical Documents*. Bachelor Thesis, FEI STU, Bratislava, Slovakia, 2022. (In Slovak)

[20] ŽÚDEL, P.: *Interactive System for Processing of Handwritten Encrypted Documents*, Bachelor Thesis, FEI STU, Bratislava, Slovakia, 2022. (In Slovak)

[21] PODDER, S. ET AL.: *An efficient method of detection of COVID-19 using mask R-CNN on chest X-Ray images*, AIMS Biophysics, **8** (2021), no. 3, 281–290, DOI: 10.3934/biophy.2021022.

Received October 2, 2022

Institute of Computer Science and Mathematics
Slovak University of Technology in Bratislava
Ilkovičova 3
812 19 Bratislava
SLOVAKIA

E-mail: eugen.antal@stuba.sk
pavol.marak@stuba.sk