Unsupervised Alphabet Matching in Historical Encrypted Manuscript Images

Jialuo Chen, Mohamed Ali Souibgui, Alicia Fornés
Computer Vision Center
Computer Science Department
Universitat Autònoma de Barcelona
{jchen,msouibgui,afornes}@cvc.uab.es

Beáta Megyesi
Dept. of Linguistics and Philology
Uppsala University, Sweden
beata.megyesi@lingfil.uu.se

Abstract

Historical ciphers contain a wide range of symbols from various symbol sets. Identifying the cipher alphabet is a prerequisite before decryption can take place and is a time-consuming process. In this work we explore the use of image processing for identifying the underlying alphabet in cipher images, and to compare alphabets between ciphers. The experiments show that ciphers with similar alphabets can be successfully discovered through clustering.

1 Introduction

Historical ciphers contain many different symbols from various types of symbol sets. Although digits are the most popular types of symbols, we find alphabetical characters such as Latin or Greek letters, punctuation marks, diacritics, along with various types of graphic signs, such as Zodiac symbols or alchemical signs.

The first step in attacking a cipher is to digitize it and transcribe it by identifying each unique type of symbol that was used (namely, the 'cipher alphabet'). This is not easy if the cipher contains symbols from various symbol sets. The task is even more challenging when the symbols are touching or connected where individual symbols in the hand-writing are hard to segment.

Automatic methods using a kind of "AI-in-the-loop" strategy might help in the identification of symbol types, and assist the transcription process. This leads us to image processing, which has been shown its usefulness for handwritten recognition in historical manuscripts, including ciphers, see e.g. Fornés et al. (2017), Baró et al. (2019), Souibgui et al. (2020). However, as far as we know, there are no methods for searching and grouping ciphers with similar symbol sets. We believe that such a tool could help experts to identify the 'cipher alphabet' of any incoming new cipher, and also to retrieve similar ciphers that may help in the subsequent analysis and decryption stages. Thus, in this work, we explore the use of unsupervised clustering for the automatic identification and comparison of symbol types in ciphers. This process shall be done without the need of any transcribed, or annotated datasets.

2 Related Work

Encrypted manuscripts contain a wide range of symbols, especially those from Early Modern Times. An investigation of 700 historical cipher keys shows that the usage of digits, Latin characters, and graphic signs were evenly distributed in keys from the 15th and 16th centuries, as illustrated in Figure 1 (Megyesi et al., 2021). In fact, 30% of the symbols were graphic signs representing a large variety of symbols taken from symbol sets including not only the Zodiac or alchemical signs, but also various unknown, fancy symbols.

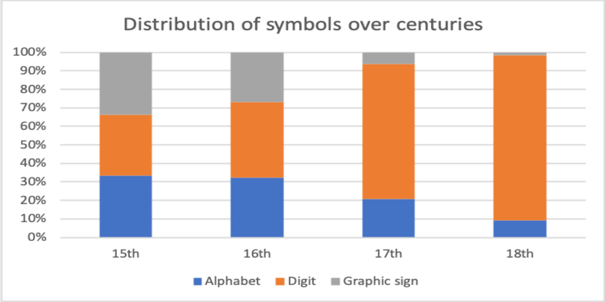


Figure 1: The usage of symbol types in cipher keys from the 15th to 18th centuries.

Image processing has proven to be useful for recognizing handwritten ciphers. Fornés et al. (2017) compared manual transcription versus automatic transcription with Recurrent Neural Networks with manual post-correction, showing that manual transcription was 15% slower if the model's accuracy was over 90%. Since then, more cipher transcription methods were proposed, using Siamese Neural Network and Gaussian Mixture Models (Yin et al., 2019), clustering (Baró et al., 2019; Chen et al., 2020), and few-shot learning (Souibgui et al., 2020).

As stated before, we are not aware of any existing image processing method for comparing and retrieving similar ciphers according to their symbol set. Thus, unsupervised clustering techniques (Gupta et al., 2019; Baró et al., 2019) are worth to explore since they can directly be applied to manuscript images without any transcription.

3 Methodology

The proposed method consists of three steps: a preprocessing stage consisting of binarization and segmentation into isolated symbols, a clustering phase where similar symbols are grouped together, and the analysis of the obtained clusters.

Image Preprocessing: The preprocessing stage starts by binarizing (Sauvola et al., 1997) the document image to facilitate the succeeding segmentation. Then, symbols are segmented using two different approaches. In case symbols are easy to segment because they are mostly isolated (i.e. there are very few touching symbols), we opt for a connected component analysis to obtain the segmented symbols. Contrary, if symbols are frequently touching, the symbol segmentation becomes difficult. Therefore, we opt for a more sophisticated method based on deep learning and proposed in Axler and Wolf (2018). Although the method was designed for word segmentation, we have adapted it for symbol segmentation. For this purpose, we have re-trained the model on 7000 synthetically generated document pages, which have been created by concatenating Omniglot symbols (Lake et al., 2015) and adding some random transformations to make them look similar to real ciphers. An example of a training page is shown in Fig. 2-A, and a segmentation example of a real cipher using the trained model is shown in Fig. 2-B.

Clustering: Once symbols are segmented, we compute the SIFT descriptor for each symbol and we apply the k-means clustering algorithm. Clustering consists in grouping those visually similar symbols in sets, named clusters. Since we are interested in comparing the different 'cipher alphabets', it is important to avoid unbalanced data. Thus, we take the same amount of symbols from each encrypted document to balance the data for a fair comparison in the clustering analysis stage.

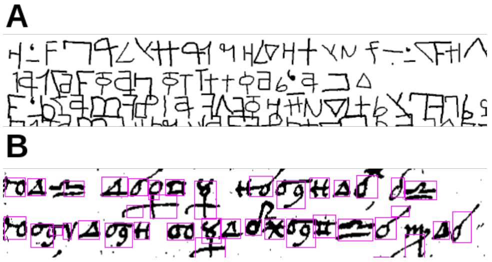


Figure 2: A: An example of a synthetic page created from Omniglot symbols. B: The segmentation output on the Borg cipher.

Clusters Analysis: Once we obtain the clusters from the two ciphers to compare, namely Cipher A and Cipher B, we analyze the similarity of their symbol elements. The goal is to analyze each cluster and verify the origin of its elements, whether they belong to Cipher A or B, or both. A cluster can have different levels of mixing, as shown in Figure 3. Depending on the frequency of each type of cluster, two ciphers will be considered more or less similar:

- If the 'cipher alphabets' are different, most clusters will contain symbols belonging to the same cipher (many clusters of type 1, 2 or 3, see Fig. 3).
- It the 'cipher alphabets' are similar, most clusters will contain symbols belonging to both ciphers (e.g. many clusters of type 4, see Fig. 3).

Being C_{mix} the number of clusters with mixed symbols (belonging to both Ciphers A and B) and C_{total} the total amount of clusters, the alphabet similarity is computed as follows:

$$Similarty(Cipher_A, Cipher_B) = \frac{C_{mix} \times 100}{C_{total}} \quad (1)$$

In this similarity computation we omit those clusters with very few elements (probably they are infrequent symbols). It is worth to observe that this analysis is sensitive to the symbol segmentation and the handwriting styles. For example, ciphers with different alphabets but similar handwriting styles could produce mixed clusters.

4 Experimental Results

We have evaluated our approach on encrypted manuscripts, most of them from the Decode

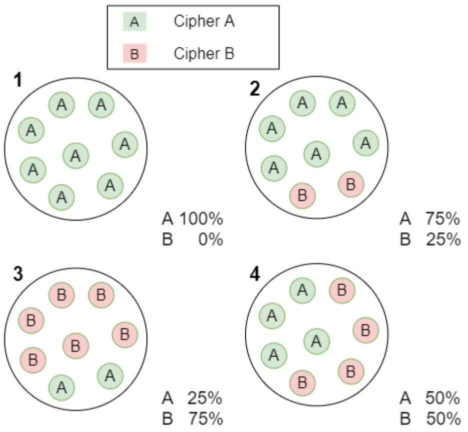
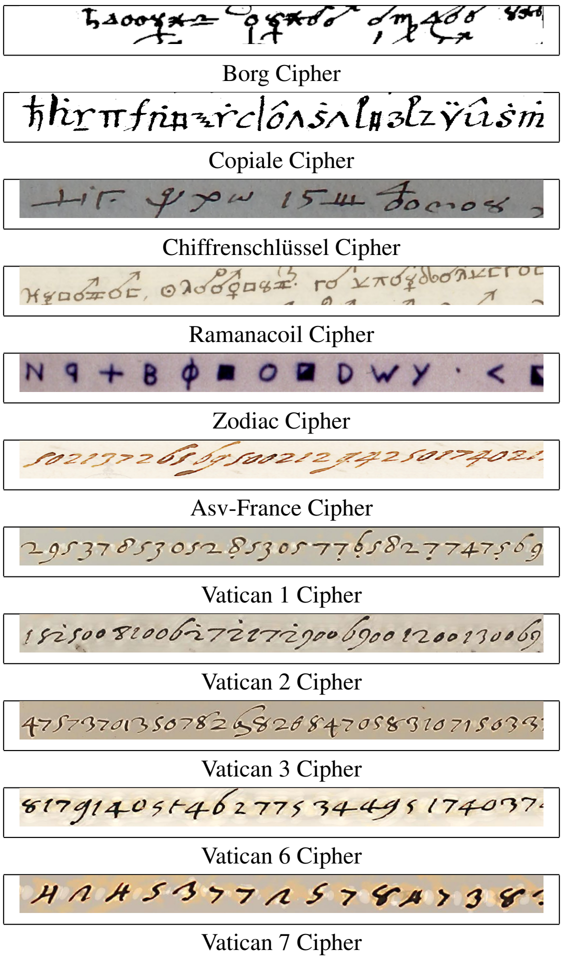


Figure 3: Cluster analysis. Cluster 1: All elements are from Cipher A. Cluster 2: There are more elements from Cipher A than from B. Cluster 3: There are more elements from Cipher B than from cipher A. Cluster 4: There is the same amount of elements from cipher A and B.


Borg Cipher

Copiale Cipher

Chiffrenschlüssel Cipher

Ramanacoil Cipher

Zodiac Cipher

Asv-France Cipher

Vatican 1 Cipher

Vatican 2 Cipher

Vatican 3 Cipher

Vatican 6 Cipher

Vatican 7 Cipher

Figure 4: Samples from the evaluated ciphers.

database (Megyesi et al., 2019). Figure 4 shows some examples. As it can be seen, some documents contain similar symbols, especially for the Vatican ciphers, with Arabic digits. However, these have different handwriting styles. During experiments, we took 5 pages from each cipher.

The obtained results are presented in Table 4. As can be seen, the similarity percentages range between 2.77% and 62.91%. Note that we are not reaching a higher similarity score probably because all the compared ciphers are different from each other in hand-writing style. The first observation is that ciphers with similar alphabets, such as the Vatican ones, are getting the highest similarity scores, compared to the rest of the ciphers. However, as we said before, the alphabet similarity can be easily affected by the writing styles. This is indeed the case: We obtain the highest score (62.91 %) when the writer style is similar, such as in the case of Vatican 3 and Vatican 6 with similar writing style of the digits ”2”, ”4” and ”7”, as shown in Figure 4). In the case of different writing styles, like Vatican 1 and Vatican 7, or between Asv-France and all the Vatican ciphers, we obtain a low similarity (20.94 %) though they all share the same cipher alphabet, digits.

We also observe a low similarity between the Zodiac and the rest of ciphers because Zodiac's cipher alphabet does not share overlapping symbols with the other cipher's alphabets. The other ciphers mainly use well-known graphic signs and digits and their similarity is medium to the rest of ciphers, without being too high or too low, indicating that these alphabets contain more or less overlapping symbols (e.g. digits) and are similar to each other.

Figure 5 illustrates some obtained clusters where symbols from different ciphers are grouped together if their shape appearance is similar.

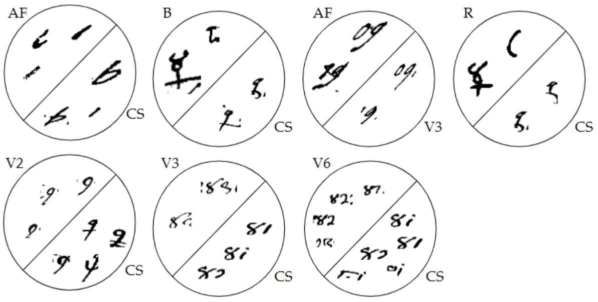


Figure 5: Results. Examples of mixed clusters.

From the different quantitative and qualitative results, we note that it is hard to assess the perfor-

Table 1: Results. Percentage of similarity between different pairs of ciphers. AF: Asv-France, B: Borg, CS: Chiffrenschlüssel, C: Copiale, R: Ramanacoil, Vn: Vatican n, Z: Zodiac.

%	B	CS	C	R	V1	V2	V3	V6	V7	Z
AF	11.00	20.94	05.95	07.73	07.41	11.01	18.59	09.95	07.33	04.55
B	—	21.46	19.11	13.27	14.15	20.18	25.91	23.81	08.66	05.20
CS	—	—	14.74	18.13	17.48	37.04	43.81	35.21	14.90	12.20
C	—	—	—	10.33	21.07	14.62	21.08	20.37	09.39	07.14
R	—	—	—	—	08.89	05.43	08.07	07.56	03.71	08.83
V1	—	—	—	—	—	32.21	39.61	39.00	20.94	06.12
V2	—	—	—	—	—	—	54.78	46.17	24.70	07.85
V3	—	—	—	—	—	—	—	62.91	25.00	07.66
V6	—	—	—	—	—	—	—	—	24.05	05.00
V7	—	—	—	—	—	—	—	—	—	02.77

mance of the proposed method without any access to the ground-truth. Thus, we opted for visually checking the manuscripts. A thorough evaluation would be necessary, preferably by an expert in paleography who could establish the ground truth to set the similarity degree between ciphers and unify symbol sets across different ciphers.

5 Conclusion

We have presented an unsupervised method for identifying the symbol set in cipher images, avoiding the need of manual transcription or human intervention. The experiments show that it can provide an intuition of the underlying symbol set, and group ciphers with similar cipher alphabets. The presented results are promising and encourage us to further explore image processing for automatic alphabet recovery and transcription of ciphers.

Acknowledgement

We would like to thank all master's students in language technology at Uppsala University who helped us producing the manual transcriptions. This work has been supported by the Swedish Research Council, grant 2018-06074, DECRYPT – Decryption of Historical Manuscripts, the Spanish project RTI2018-095645-B-C21 and the CERCA Program / Generalitat de Catalunya.

References

Gregory Axler and Lior Wolf. 2018. Toward a dataset-agnostic word segmentation method. In *ICIP*.

Arnau Baró, Jialuo Chen, Alicia Fornés, and Beáta Megyesi. 2019. Towards a generic unsupervised method for transcription of encoded manuscripts. In *DATECH*, pages 73–78.

Jialuo Chen, Mohamed Ali Souibgui, Alicia Fornés, and Beáta Megyesi. 2020. A web-based interactive transcription tool for encrypted manuscripts. In *HistoCrypt 2020*, pages 52–59.

Alicia Fornés, Beáta Megyesi, and Joan Mas. 2017. Transcription of encoded manuscripts with image processing techniques. In *Digital Humanities*.

Divam Gupta, Ramachandran Ramjee, Nipun Kwatra, and Muthian Sivathanu. 2019. Unsupervised clustering using pseudo-semi-supervised learning. In *ICLR*.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Beáta Megyesi, Nils Blomqvist, and Eva Pettersson. 2019. The decode database: Collection of historical ciphers and keys. In *HistoCrypt*, pages 69–78.

Beáta Megyesi, Crina Tudor, Benedek Láng, and Anna Lehofer. 2021. Key Design in the Early Modern Era in Europe. In *HistoCrypt*.

J. Sauvola, T. Seppanen, S. Haapakoski, and M. Pietikainen. 1997. Adaptive document binarization. In *ICDAR*, pages 147–152.

Mohamed Ali Souibgui, Alicia Fornés, Yousri Kessentini, and Crina Tudor. 2020. A few-shot learning approach for historical ciphered manuscript recognition. In *ICPR*.

Xusen Yin, Nada Aldarrab, Beáta Megyesi, and Kevin Knight. 2019. Decipherment of historical manuscript images. In *ICDAR*, pages 78–85.