AN OPTICAL APPROACH FOR DECODING THE MYSTERIOUS VOYNICH MANUSCRIPT

Costin-Anton BOIANGIU [1*]
Ana-Karina NAZARE [2]
Andreea Dorina RACOVIȚĂ [3]
Iulia-Cristina STĂNICĂ [4]

ABSTRACT

The current paper presents several optical approaches used for investigating the decoding of the Voynich Manuscript. Over the past century, there have been numerous decipherment claims, but none of them can be accepted as the true solution. Most of them focus on the syntax analysis of the text, trying to correlate the Voynich text with well-known languages. In the current paper, we present a short history and description of the manuscript, as well as an innovative technique used for attempting its decoding. Our paper presents the use of several optical approaches, such as various distortions, words and character scrambling, in the attempt of correctly identifying voynichese words with the help of OCR.

KEYWORDS: *Voynich Manuscript; optical devices; OCR; decipherment.*

1. INTRODUCTION

The Voynich manuscript is a mysterious document, allegedly dated from the 15th century (based on radiocarbon dating) [1]. The book is written in an unknown language and contains many illustrations which can be used to divide the manuscript into six sections: herbal, astronomical, biological, cosmological, pharmaceutical and recipes [2]. There are approximately 240 pages left of the manuscript, as some others might have been lost over time. The author, date and origin of the book are not known; the name is based on that of the Polish book dealer Wilfrid Voynich, who got in the possession of the manuscript in 1912. Over time, before being bought by Voynich, the manuscript had numerous owners, such as Czech scientists, Emperor Rudolf II or members of the Jesuit order. Currently, it is kept in the Beinecke Rare Book & Manuscript Library of Yale University [3].

In order to study the text of the Voynich Manuscript, a transcription was made using the EVA (Extensible Voynich Alphabet), further influencing the numerical analysis, ‘word’

[1*] corresponding author, Professor PhD Eng., “Politehnica” University of Bucharest, Bucharest, Romania, costin.boiangiu@cs.pub.ro
[2] Engineer, “Politehnica” University of Bucharest, Bucharest, Romania, nazare.ana.karina@gmail.com
[3] Engineer, “Politehnica” University of Bucharest, Bucharest, Romania, andreea.d.racovita@gmail.com
[4] Engineer, “Politehnica” University of Bucharest, 060042 Bucharest, Romania, iulia.stanica@gmail.com

length, distribution and statistics. The text was analyzed and several observations were made along the years with regard to characters, words, sentences, paragraphs and sections [3]. Regarding the characters, it was observed that some of them appear generally at the end of lines, while others appear almost always at the beginning of paragraphs or on the first lines. Also, some characters seem to have separating or conjunctive function and are often found together. It is believed that some characters have the role of final letters, with varying frequencies [4]. Some characters appear very rarely and only on some pages of the manuscript, unusual fact for most languages. In terms of words, the same ones can be repeated two, three or more times in a phrase [5]. Also, many words differ by only one character and are placed in proximity to each other. Throughout the text, there are only a few words of one single character.

Although it is not certain from the manuscript what constitutes a 'word' in the grammatical sense or if the spaces are separators between words, analysts found 8114 different words and a total of 37919 tokens in the whole manuscript. Due to these uncertainties, the word statistics are the least reliable of all other statistics. However, although tokens appear to have a normal frequency distribution, they may represent syllables, as well as single characters, instead of real words. Additionally, there is a high number of word that appear only once in the whole manuscript [6].

Our paper consists of four sections: the first one is the introduction, in the second one we analyze the current attempts of decoding the manuscript, the third one presents our proposed approaches and the final one draws the conclusions of the research.

2. DECIPHERMENT CLAIMS

Over the years, a relatively high number of decipherment claims have appeared. Unfortunately, none of them could be labeled as being valid - some of them manage to identify just a few words, while others pretend to hold the key for the decoding of the entire manuscript. We will present further a few examples of deciphering claims, based on various criteria.

2.1. Old theories

There have been many theories about the meaning of the Voynich Manuscript since its discovery over a century ago. One of the earliest theories for deciphering the manuscript comes from the philosopher William Romaine Newbold. His claim is based on micrography, sustaining that the letters have no real meaning, as they are actually composed of many small signs, which can be seen only under a magnifying glass. This theory was disclaimed as being too speculative [7].

Joseph Martin Feely pretends it is a book written by Roger Bacon. By using substitution and starting from some words related to specific illustrations, he claims that it is written in a medieval, extremely abbreviated latin [5]. Leonell C. Strong suggests that after using some sort of double arithmetical progression, he discovered that the manuscript is written in a medieval form of English [8]. Both theories were considered as being extremely subjective.

2.2. Recent theories

Recent theories have appeared, made by various linguists and researchers all over the world. Dr Arthur Tucker came in 2014 with a new theory: he identified a series of proper nouns, such as plants and constellations, based on their representative drawings. The linguist suggests that the mysterious language of the document is called Nahuatl (Aztec language), but he wasn't able to identify anything else except for those specific nouns [9]. Stephen Bax also sustained in 2014 that he succeeded in identifying several nouns (constellations or plants), by using a technique similar to the one used to decipher the hieroglyphs [10].

One of the most recent claims in deciphering the manuscript was made by Agnieszka Kałużna and Jacek Syguła. Their research, published in 2017 and entitled "The Key to The Voynich Manuscript", is based on correlating the symbols from the Voynich Manuscript with prefixes, suffixes or abbreviations from the Latin alphabet. They give some examples of translations by passing through numerous languages, such as Latin, Greek, French or Italian [11].

Another claim of decipherment comes from Canadian researchers, who used artificial intelligence to identify the language of the manuscript. They pretend that it is written in ancient Hebrew and that the words are actual alphagrams (words which must have their letters arranged alphabetically). Their results seem to be remarkable, with 80% of words making sense in Hebrew and the first sentence of the manuscript being identified as "She made recommendations to the priest, man of the house and me and people" [12]. After performing a quick OCR on some Voynich pages and setting the language as Hebrew, some words were identified, with various meanings: "ancestors", "forgive", "take a look", but these can be purely coincidences, as they are not necessarily correlated with the images or the theme of their corresponding chapter.

2.3. Hoax theories

Giving the fact that none of the theories proposed was accepted as being a valid decipherment of the Voynich Manuscript, people started to believe that it could be a hoax. Even if the parchment is allegedly dated from the 15th century, based on radiocarbon dating, it could have been used and written on centuries after being prepared. Some researchers say that Voynich himself created the manuscript and the now missing pages were removed later in order to make Roger Bacon as the intended author [13].

Some famous debate appeared between two scientists: Montemurro and Rugg. Gordon Rugg, a computing expert, claims that by using a card with randomly cut holes and moving it across various syllables, you can obtain a language that follows the statistical rules of true languages. On the other hand, physicist Marcelo Montemurro disagrees with the previously mentioned theory, stating that the manuscript itself is too complex and contains too many statistical similarities between its sections (regarding both words and images) [13] [14].

3. PROPOSED THEORY AND POSSIBLE ALGORITHM

Taking into account the fact that most decipherment techniques focused on searching for the meaning of the voynichese characters and words by correlating them with other languages, we considered that a totally different approach was needed. It is possible that several optical instruments (mirrors, lenses) were used in order to create the Voynich manuscript. We propose the simulation of different types of image transformations by using computer software in the attempt of correctly identifying the meaning of the Voynich manuscript.

3.1. Optical instruments for distortions

In the beginning, the best solution was to implement parameterized transformations which could be applied as filters to images containing selected chunks of texts. Prior to implementing the transformations as equations, they have been tested using specialized tools from dedicated programs (such as the Kaleidoscope tool from KrazyDad [15] or Adobe Photoshop filters).

The kaleidoscope effect (figure 1), while it is quite beautiful and interesting, doesn’t promise successful results in actually distorting the content, and the process of applying this effect on single words in order to encode such rich text is time consuming and highly unlikely to have been used.

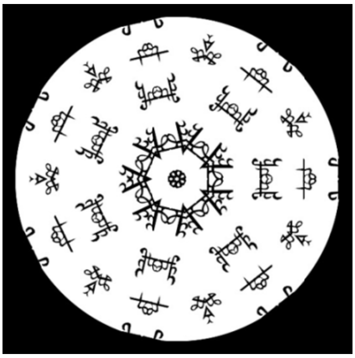

Figure 1. Kaleidoscope effect on Voynich words

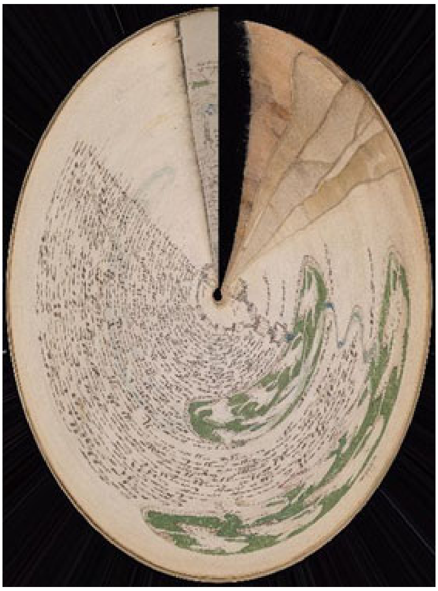

Figure 2. Polar coordinates distortion

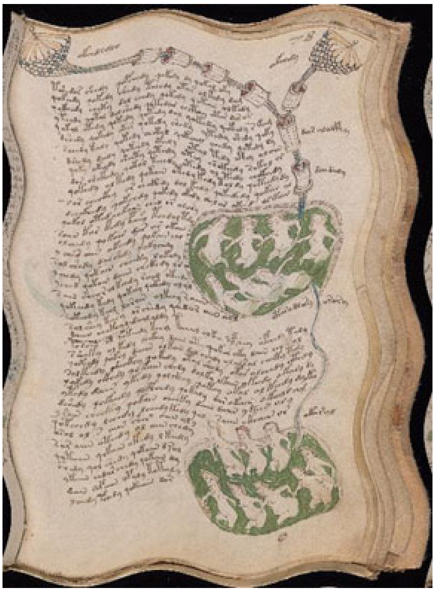

Figure 3. Wave (sine) distortion

We have tested several Photoshop effects, such as pinch, polar coordinates (rectangular to polar - figure 2), ripple, shear, spherize, twirl, wave (sine) - figure 3, zigzag, lens correction. Pinch, by example, could be simulated in reality using a conical mirror.

The transformation of polar coordinates from rectangular to polar corresponds to the use of a cylindrical mirror. Spherize, shear, ripple, wave can be simulated using lens of specific shapes.

Given the fact that the effects applied in Photoshop on the entire pages of the manuscript didn't give any relevant results, we decided to implement ourselves several geometric distortions which can be simulated using optical instruments available in the 15th century. They are: mirroring, rotation, twirl, fisheye, inverse conical deformation, inverse cylindrical deformation, skew, perspective transformation.

There are two approaches to this system. Either the text is considered ciphered and the deciphering method consists only in applying the inverse deformation, or the text has been encrypted in such a manner that it can be read only by applying a direct transformation (representing the physical object - the decypher key). Therefore we identify three classes of deformations to be applied:

- Direct optical deformations (Mirroring, Rotation, Twirl, Fisheye)
- Inverse optical deformations (Inverse cylindrical and conical deformations and, again, Mirroring and Rotation)
- General optical deformations (Skew, Perspective)

In order to test the implemented transformations, we selected four high resolution images from the Voynich Manuscript, which only contain text, as the drawings were irrelevant for our distortion techniques.

All complex distortions have customizable parameters, such as the twirl angle, the radius of the cone/cylinder base. For each transform and distortion technique, we can specify some width and height values and select a region of interest (ROI) from the original image. The distortion can, therefore, be applied also on that specific window, not just on the entire image.

We decided to create an interactive application, where the user can specify the wanted parameters and choose the center of the ROI using the mouse click (figure 4). The interface, as well as the loading and saving of the images were done using OpenCV as an external library.

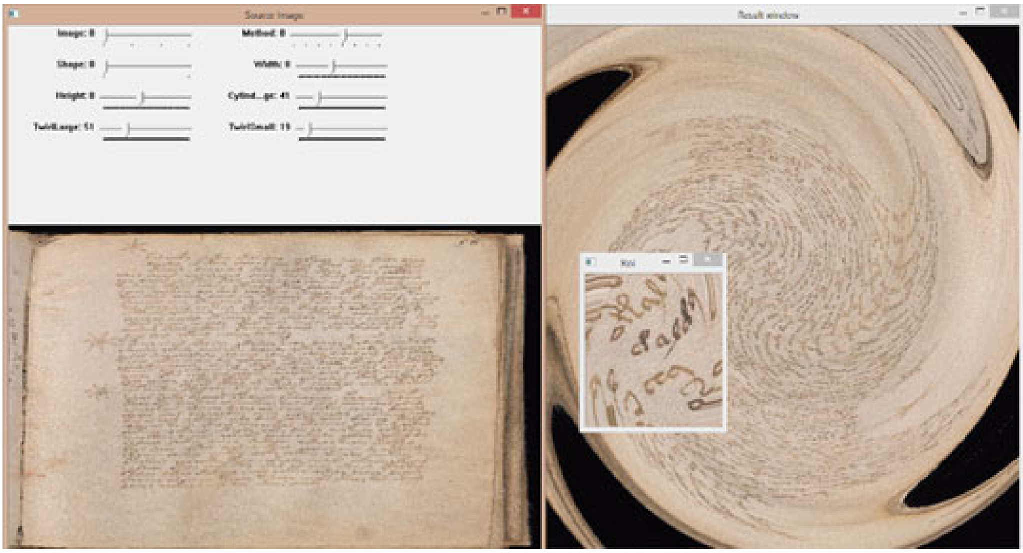

Image: 0
Method: 0
Shape: 0
Width: 0
Height: 0
Cylind...ge: 41
TwirlLarge: 51
TwirlSmall: 19

Figure 4. Output of the program - Twirl effect

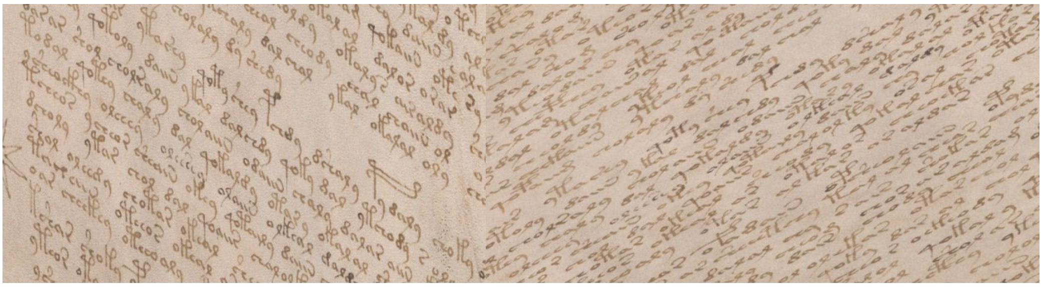

Figure 5. Skew results

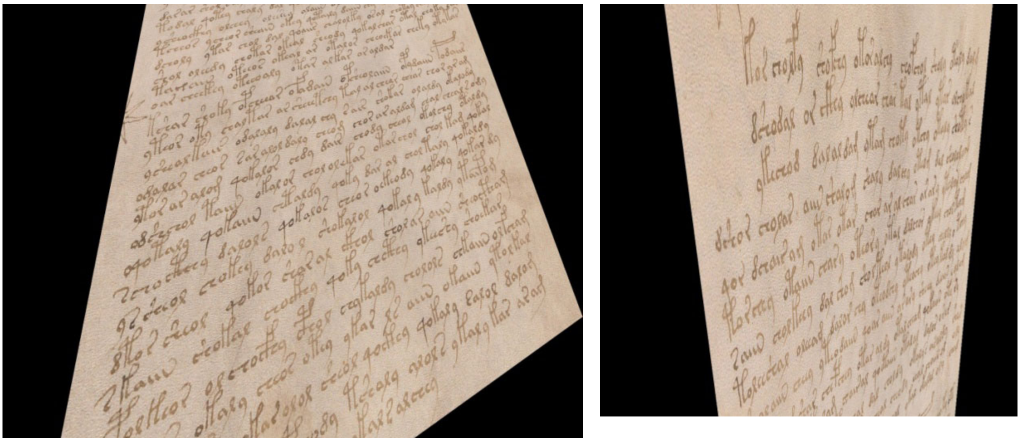

Figure 6. Perspective transformation results

3.2. Scrambling

Another approach taken into consideration is the scrambling of entire words or word parts situated in pseudo-random positions or around key positions marked by special characters. A certain number of ROIs are scrambled all around the page, in order to see if we can obtain some meaning to the text. The user can select in the interface the number of patches that will be scrambled, interchanging the location of each pair. The software saves the resulting images in a separate folder so that OCR can be applied to recognize words.

A preliminary step is providing an input set for the scrambling operations. This is represented by a set of words and word sections which must be detected in the source image and extracted. Word extraction was realized by identifying contours and extracting bounding rectangles in which words are identified. Scrambling is then applied on the detected words, with a fixed height and variable word length to select entire or partial words.

Word extraction

The implemented words extracting algorithm is a low-level image processing feature extraction algorithm [16]. As its initial set of data it receives a large resolution image of a whole, colored manuscript page and outputs a set of bounding rectangles for each individual word.

The first step in extracting words from the image is processing the image in order to easily identify regions of interest. Firstly, the image is converted to grayscale, as only the morphological information is needed, and then filtered using a bilateral filter, in order to reduce noise (such as the parchment texture artefacts). Next, the image is binarized and inverted (for further processing), using an adaptive threshold method offered by OpenCV. After thresholding the image, it is necessary to apply morphological operations of dilation and erosion respectively, in order to connect letters and disconnect words and lines, thus obtaining blobs for individual words (Figure 5. B).

The dilation is horizontal, while the erosion is vertical, because the letters in each word are disconnected, and the rows of the text are quite compact, and vertical neighboring words need to be separated. Lastly, by applying a custom function, the contours of the blobs are obtained. For each contour whose length is greater than a threshold calculated from the original dimensions of the source image, its straight bounding rectangle is determined (Figure 5. C) and then only this will be used for ulterior processing [17].

The algorithm also detects regions that do not correspond to words (such as drawings, sheet edges etc.), or even groups of connected words. Therefore, when selecting the correct bounding rectangles, several conditions must be taken into account. The following features are established through simple observation and direct testing, even though their limitations sometimes eliminate words correspondents in the final data set. Rectangles having an area which is too large or too small, or having a height more than three times larger than the average height of the rectangles (being assumed to be the average height of a word) are overlooked. Another condition is that words should have a greater length than height.

Finally, for each filtered rectangle, a mask from the binarized source image is selected. A ratio of non-zero pixels is calculated for each ROI, and if this ratio is greater than 0.45 (assuming words have a greater than 45% text surface), the rectangle is accepted.

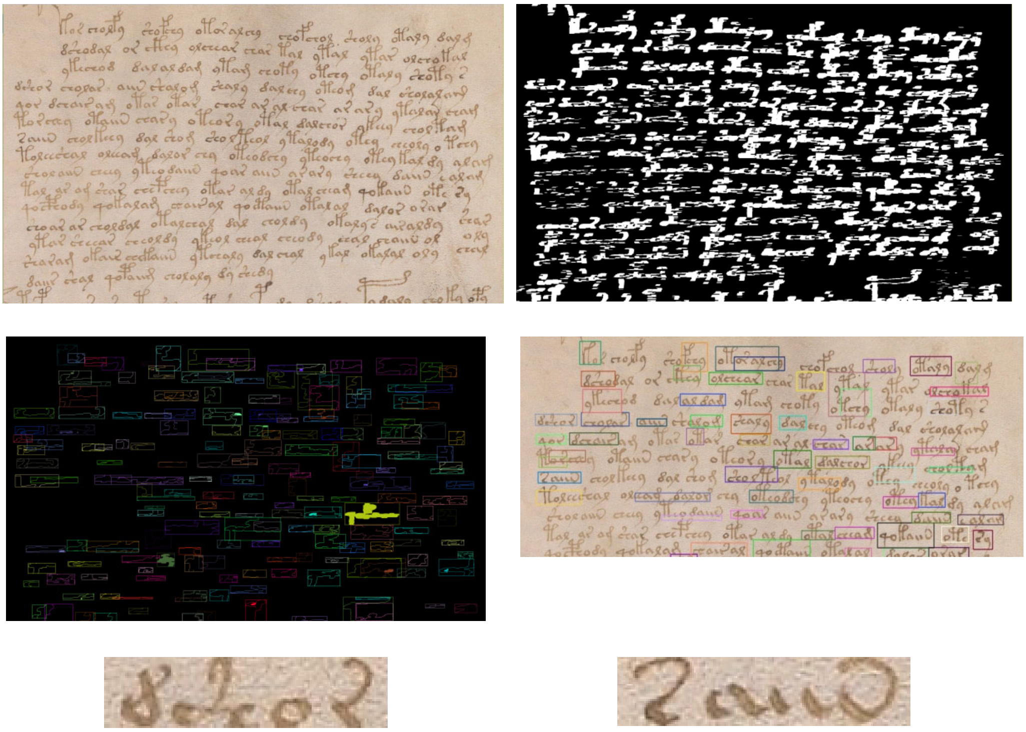

Figure 5. Image processing steps in detecting words. A. Part of source image B. Binarized and dilated image. C.1., C.2. Identified contours and bounding rectangles D. Two of the identified words

OCR

Tesseract OCR is probably the most famous optical character recognition system, developed initially by Hewlett Packard and currently sponsored by Google. We used the API on each image saved after performing the scrambling in order to see if any text can be extracted from it. There were no concluding results, so further investigations are needed.

Template matching

Template matching is a well-known technique used in image processing, which has the goal of finding a sub-image (template) in a bigger image (search area). The process implies the translation of the template over the search area and calculating the similarity between each translated window and the original template [18]. An essential step is represented by the way the similarity measure is chosen is order to quantify the "matching". Some of the most used similarity metrics include cross correlation and sum of absolute differences [19].

We used OpenCV in order to perform the template matching. By sliding the template window from left to right and up to down, a resulting matrix is created with the value obtained with the similarity matrix in that exact position (x, y). We used a trackbar with a certain threshold in order to accept a lower or higher degree of similarity between the wanted characters (template) and the specific Voynich page area. Template matching was essential in order to identify gallows characters (characters which raise above the other characters) (figure 6).

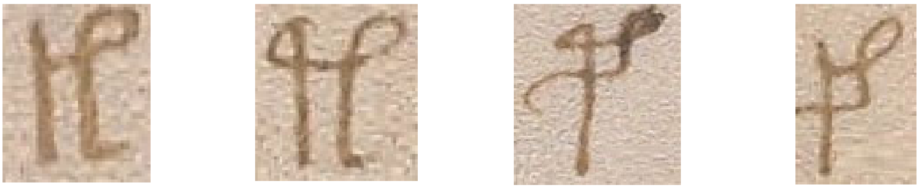

Figure 6. Voynich gallows characters

We used template matching on both pages of the Voynich Manuscript (figure 7) and various Medieval English manuscripts (figure 8). The interesting fact is that the template matching technique returned positive results also on some Medieval English pages, showing similarities between characters.

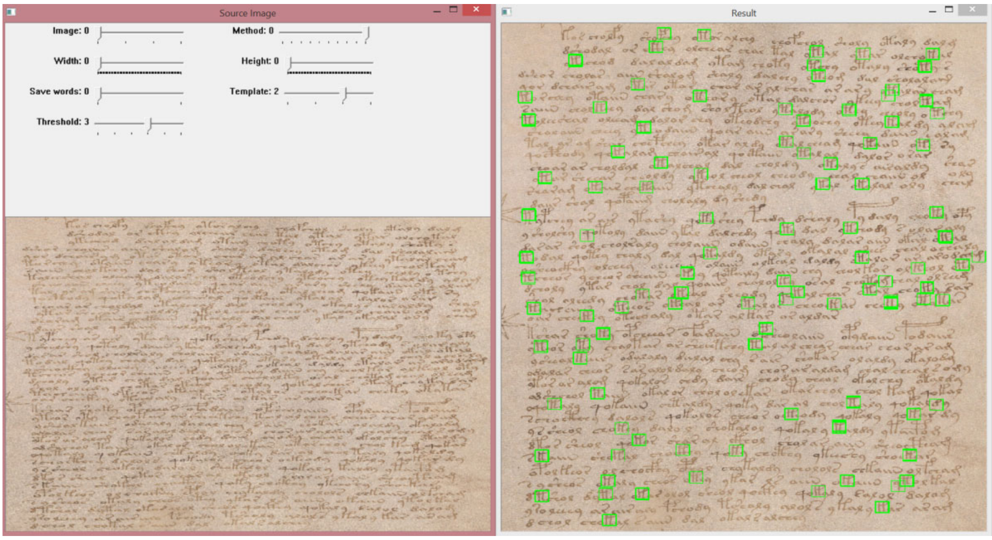

Image: 0
Method: 0
Width: 0
Height: 0
Save words: 0
Template: 2
Threshold: 3

Figure 7. Template matching on a Voynich page

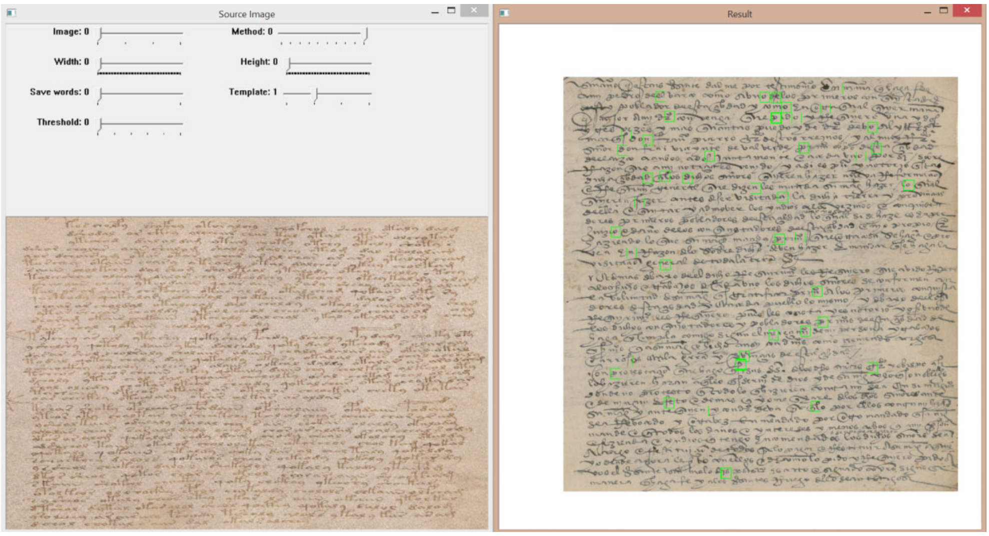

Image: 0
Method: 0
Width: 0
Height: 0
Save words: 0
Template: 1
Threshold: 0

Figure 8. Template matching on a Medieval English manuscript

3.3. Mirrors

Our third approach for deciphering the Voynich manuscript consists in using various sizes and shapes of mirrors (rectangular, oval) in order to transform the text (figure 9). We tried to place the mirrors in different ways: facing each other (figure 10), perpendicular (figure 11), at an acute or obtuse angle. When positioning the mirrors facing each other, a special phenomenon called "infinite reflection" is produced, but in the end, the original image is obtained. A single mirror produces one image, two mirrors perpendicular on each other produce three images, while two mirrors at an acute angle produce more than three images. The mirrors were also positioned in different spots on a page: in the upper part, at the middle or at the bottom. We also tried to fold the page and position the mirrors between words, between lines or next to gallows characters. An interesting result (figure 12a) was obtained on a specific Voynich page, which contained symmetrical text and images from the beginning (figure 12b).

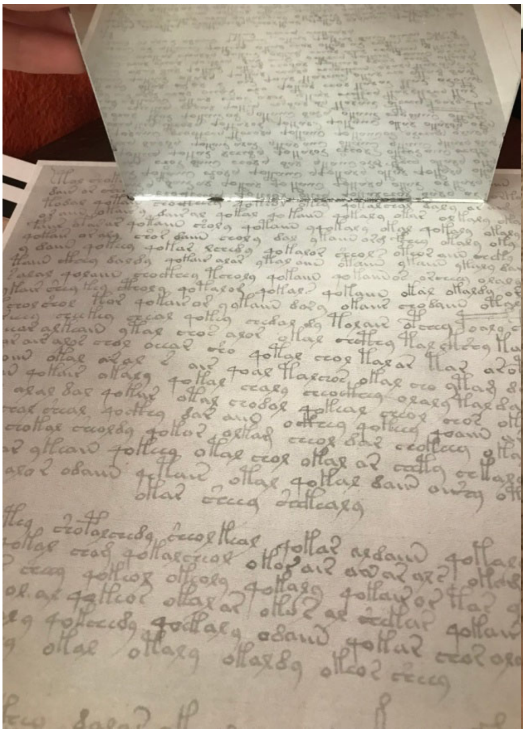

Figure 9. Single mirror placed at the top

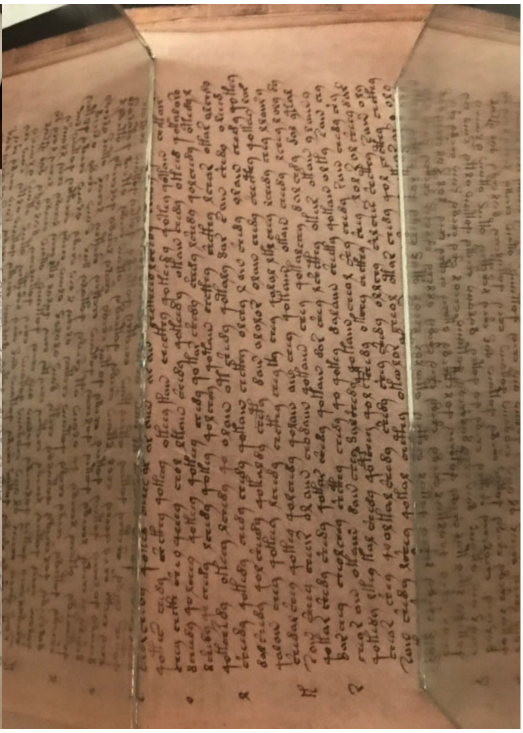

Figure 10. Two mirrors facing each other

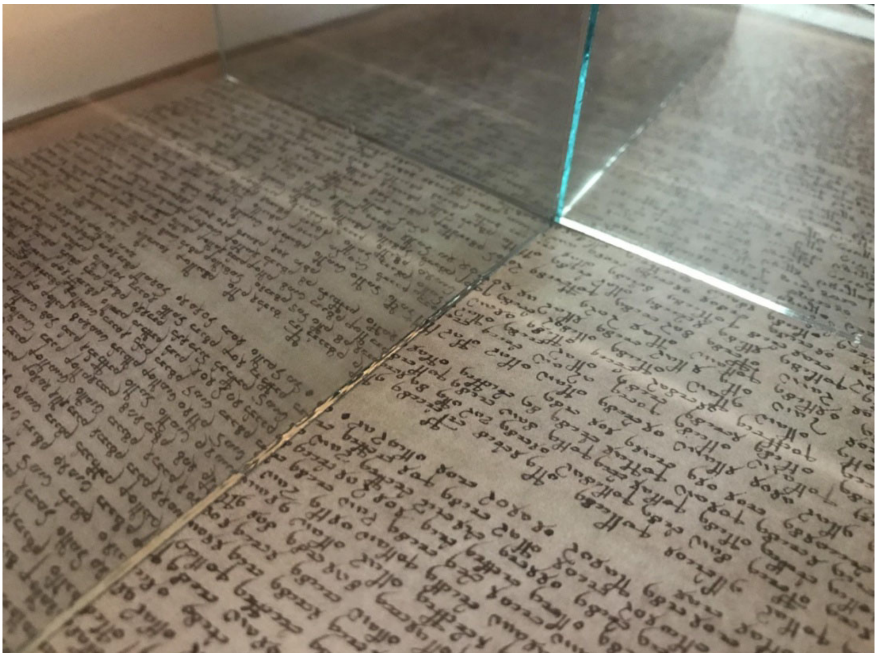

Figure 11. Two perpendicular mirrors

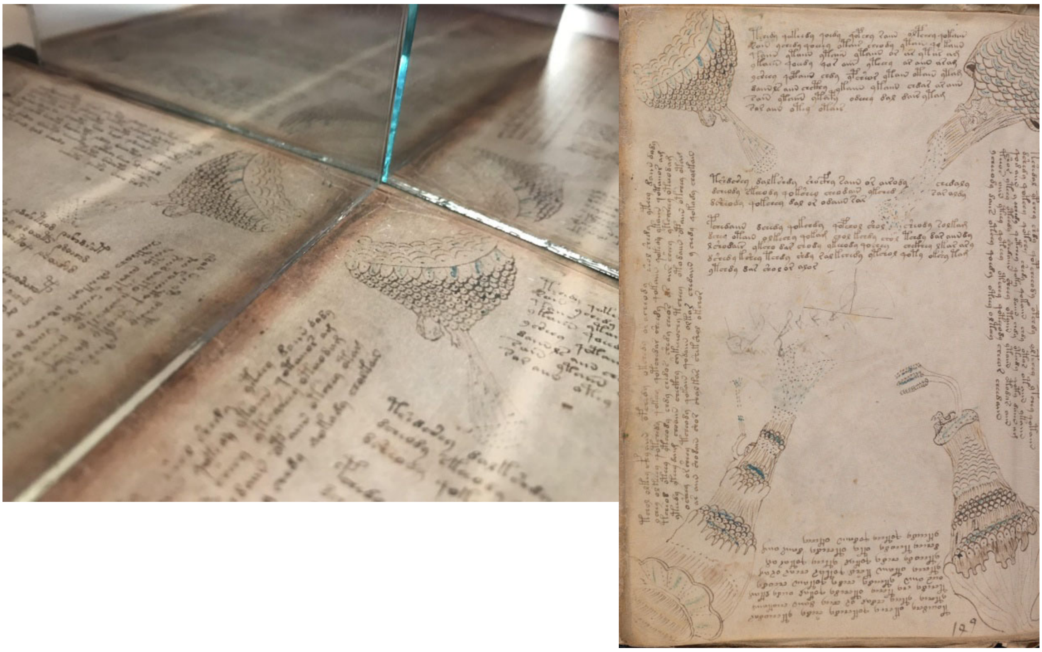

Figure 12. Symmetrical images. Left: Mirrored drawings; Right: Original Voynich page.

Our tests with the mirrors did not lead however to any conclusive results. Even if the text might be mirrored (as we also simulated through computer software) or written from right to left (similar to Hebrew writing), the images obtained did not bring any known meaning to the mysterious writing.

4. CONCLUSIONS

After experimenting with a wide variety of simulated distortion devices, we can conclude that it is unlikely that the Voynich manuscript was written using one such optical instrument (mirrors or lenses). Other techniques, such as word scrambling based on keywords or gallows characters, didn't show concluding results either. Combining the power of artificial intelligence, OCR and trying to identify meaningful words and sentences seem to represent the best option to finally decode the mysterious manuscript.

REFERENCES

[1] Steindl, Klaus; Sulzer, Andreas (2011). "The Voynich Code - The World's Mysterious Manuscript"

[2] Schmeh, Klaus (2011). "The Voynich Manuscript: The Book Nobody Can Read". Skeptical Inquirer.

[3] Zandbergen, René. "Voynich MS - Long tour: Known history of the manuscript". Voynich.nu

[4] Tiltman, John (1967). "The Voynich Manuscript, The Most Mysterious Manuscript in the World". NSA Technical Journal 12, pp.41-85.

[5] D'Imperio, M. E. (1978). "The Voynich Manuscript: An Elegant Enigma". National Security Agency. pp. 1–152.

[6] Reddy, Sravana; Kevin, Knight (2011): "What we know about the Voynich Manuscript". Proceedings of the ACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.

[7] "Penn Biographies - William Romaine Newbold (1865–1926)". University of Pennsylvania.

[8] Winter, Jay (October 17, 2015). "The Complete Voynich Manuscript Digitally Enhanced Researchers Edition". Lulu Press. pp. 1–259.

[9] Flood, Alison (2014). "New clue to Voynich manuscript mystery". The Guardian, Science and Nature.

[10] Bax, Stephens (2014). "A proposed partial decoding of the Voynich script".

[11] Kaluzna, Agnieszka; Syguła, Jacek; Jaśkiewicz, Grzegorz (2017). „The Key to The Voynich Manuscript".

[12] Hauer, Bradley; Kondrak, Grzegorz, (2018). "Decoding Anagrammed Texts Written in an Unknown Language and Script". Transactions of the Association for Computational Linguistics

[13] Rugg, Gordon; Taylor, Gavin (2016): "Hoaxing statistical features of the Voynich Manuscript". Cryptologia Journal 41, pp. 247-268.

[14] Montemurro, Marcelo A.; Zanette, Damian H. (2013): "Keywords and Co-Occurrence Patterns in the Voynich Manuscript: An Information-Theoretic Analysis". Plos One Journal.

[15] Krazy Dad (2017). Make your own kaleidoscope. Available online: https://krazydad.com/kaleido/, Accessed at: April 27, 2018.

[16] Mark S. Nixon, Alberto S. Aguado (2012). "Feature Extraction & Image Processing for Computer Vision", Academic Press 2012

[17] OpenCV Official documentation, Image Thresholding, Finding contours in your image, Contour features

[18] Kim, Yongmin; Sun, Shijun; Park, HyunWook (2004). "Template Matching Using Correlative Autopredicative Search". University of Washington.

[19] Yang, Ruigang (2016). „Object recognition and template matching", Universitatea din Kentucky, SUA.