Segmentation of Handwritten and Typewritten Tifinaghe Texts

Yassine Chajri
Sultan Moulay Slimane University
Beni Mellal
Morocco

Belaid Bouikhalene
Sultan Moulay Slimane University
Beni Mellal
Morocco

ABSTRACT

The recognition of Tifinaghe documents became a very important field of research. This importance is mainly due to the attention that has been given by the official institutions to the Amazigh language and culture. This paper represents our contribution in this process of Amazigh language's revitalization. In this paper we describe all the details concerning the necessary steps (pre-processing, text lines' segmentation, extracted lines' segmentation) of our approach for Tifinaghe text's segmentation (handwritten and typewritten). This approach is based on the Radon transform in terms of segmenting texts into lines and on the connected components algorithm in order to segment the extracted lines into Tifinaghe characters.

General Terms

Tifinaghe documents, Text segmentation, Text recognition.

Keywords

Tifinaghe- Amazigh language- Handwritten-Typewritten-Recognition- Segmentation- Radon Transform

1. INTRODUCTION

In this era of globalization, the language field has become an integrated market on a planetary scale. In this market, minority languages have become increasingly threatened by extinction. This reduces the diversity of the world's languages (450 languages are endangered, including 161 in the Americas, 175 in the Pacific, 55 in Asia, 37 in Africa and 7 in Europe and 1 language disappears on average every 15 days) [2].

Morocco has sensed this danger, leading it to take numerous measures aimed at revitalizing the Amazigh language (The Amazigh language is spoken currently by around thirty million throughout the whole world), namely:

- Foundation of the Royal Institute of Amazigh Culture (IRCAM).
- Article 5 of 2011 Moroccan Constitution introduced the Amazigh language as «An official language of the State, as a heritage common to all Moroccans without exception. ».
- Effective integration of the Amazigh language in public policies (Education, Media, etc.).

All of these measures herald a better tomorrow.

At the international level, more specifically, with regard to new information and communication technologies, the Amazigh language has experienced a process of standardization and integration:

- Character encoding specified by the extended ASCII.
- Integration in the Unicode Standard.
- Implementation of a standard Amazigh Keyboard layout.
- Building new Tifinaghe fonts.
- Development of converters that made Tifinaghe ANSI-Unicode transition and Arabic-Latin-Tifinaghe transliteration [1].

This paper's subject fits into the broader context of the Amazigh language revitalization. More specifically, in the context of the Tifinaghe text segmentation. We present the approach adopted for Tifinaghe text segmentation (handwritten and typewritten) which is based on Radon Transform. The basic idea who served us deeply for using this transform is its ability to extract lines even from noisy images. This transform represents image lines in form of peaks which facilitates their extraction.

2. RELATED WORKS

The text lines segmentation is a crucial step in any system of texts' recognition. The fact of properly segmenting the text into lines positively influences the recognition's results. In the literature, there are several methods that deal with the subject of texts lines segmentation. These methods (well described in [8]) can be classified in : Smearing methods, Grouping methods, Hough transform methods, Projection-based methods, stochastic methods and other methods [3][8].

Table 1: Related works of texts' segmentation methods

Smearing methods	For example Run Length Smoothing Algorithm [4][13] [8].
Grouping methods	The pixels or linked components, blocks or other features like salient points are aggregated in a bottom-up strategy. These units are linked to build alignments [12] [4] [8].
Hough transform	The algorithms transform the images in the Hough domain and assume that the local maxima correlate with text lines [15] [8].
Projection-based methods	The principle of these methods is to create a vector containing the sums of pixels of each line. In this vector, the local maxima represents the text line while the local minima represents the white area between lines [15] [12] [8].
Stochastic method	The principle is to extract non-linear paths between overlapping text lines by using Hidden Markov Modeling [12] [4].
Other methods	Processing of the overlapping and touching components method, Repulsive-Attractive network method, text line structure enhancing, etc [8].

3. APPROACH PROPOSED FOR TIFINAGHE TEXT'S SEGMENTATION (HANDWRITTEN AND TYPEWRITTEN)

This approach is based on Radon Transform and the connected components algorithm.

The main steps of this approach are given below:

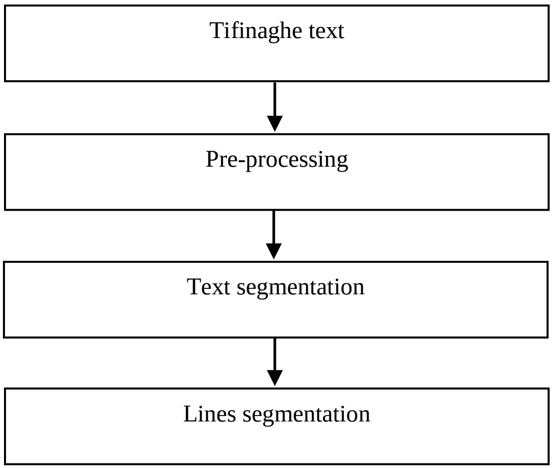

Fig 1: System architecture

3.1 Pre-processing

Image pre-processing is a mandatory step in our recognition system because of its ability to solve the problems associated with the images acquisition (scanner quality, scan resolution, type of printed documents, paper quality, fonts used in the text, etc.)[6].

In order to have an efficient system able to recognize the texts, we applied the following process:

- Convert RGB image to binary image.
- Apply median filter for noise removal [10].
- Image's skewing is one of the complicated problems generated by the texts' scan. This problem makes the text lines segmentation phase very difficult to realize. Therefore, the skew's detection and correction influence positively the results of the system and make it very efficient. For this, we have applied the Radon Transform [5][7].

3.2 Tifinaghe text's segmentation

The objective of this step is the Tifinaghe texts segmentation. This segmentation is very delicate because of the problems mentioned above. For this, we have applied the Radon transform. This choice is due to the ability of this transform to extract the lines even from noisy images. This is done by transforming the lines into positioned peaks corresponding to the line parameters.

Radon Transform

Radon Transform is a mathematical technique developed by the mathematician Johann Radon [16]. It is an integral transformation in multi-dimensional spaces. This transform allows transforming two dimensional images with lines into a domain of possible line parameters, where each line in the image will give a peak positioned at the corresponding line parameters [14]. In other words, this transform converts a function (image) in a series of projections for each angle θ ∈[0,π].

A projection at a given angle θ is obtained as the linear integration of the function on all parallel lines [8] [9].

The result is a new image R(ρ,θ) that can be written mathematically by [11]:

$$R(\rho,\theta) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,\delta(\rho - x\cos\theta - y\,\sin\theta)\mathrm{d}x\mathrm{d}y \qquad (1)$$

Where:

$$\rho = x\cos\theta + y\,\sin\theta \qquad (2)$$

δ(): is the Dirac delta function.

Since 1950, the Radon transform began to be applied in tomography, medicine, astronomy, optics, geophysics, etc.

3.3 Lines segmentation

After segmenting the text into lines, the process of lines segmentation begins in order to have the Tifinaghe's characters, which form each line, isolated. To do this, we have opted to apply the principle of the connected components algorithm. This principle is based on the fact of scanning the studied image (pixel-by-pixel from top to bottom and left to right) and grouping its pixels into components according to their connectivity relationship. In other words, all the pixels of a connected component share the same label in order to distinguish and extract the different disconnected structures.

The labeling is directly dependent on the connectivity considered (4-Connectivity and 8-Connectivity) [9].

Table 1. 8-Connectivity

1	**1**	**1**
1	**1**	**1**
1	**1**	**1**

Table 2. 4-Connectivity

0	**1**	**0**
1	**1**	**1**
0	**1**	**0**

4. RESULTS

In this paper, we have opted for the treatment of handwritten and typewritten Tifinaghe' texts. This choice is justified by the fact to determine the ability of our approach in terms of segmenting the different types of texts. The figures below (Figure 3, Figure 5 and Figure 7) show the Radon Transform representation for three Tifinaghe texts (handwritten and typewritten text) with 0° to 179° degrees of projection angle. In these representations, we can see some colored spots, which are the peaks of Radon Transform and which represent the lines in each text.

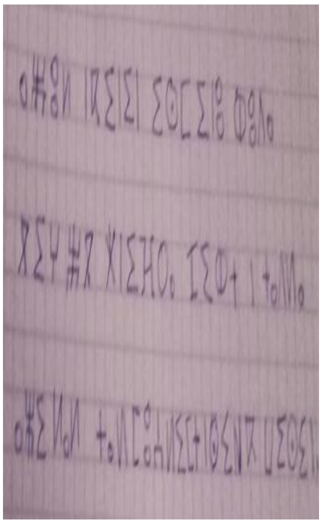

Fig 2: Text number 1

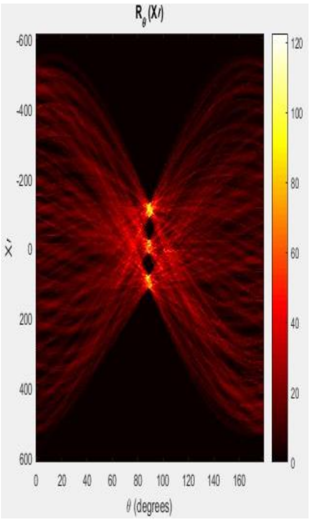

Fig 3: Radon transform representation of text number 1

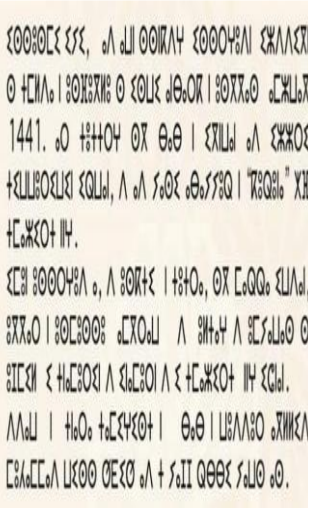

Fig 4: Text number 2

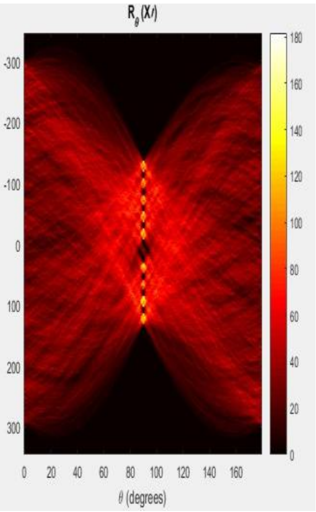

Fig 5: Radon transform representation of text number 2

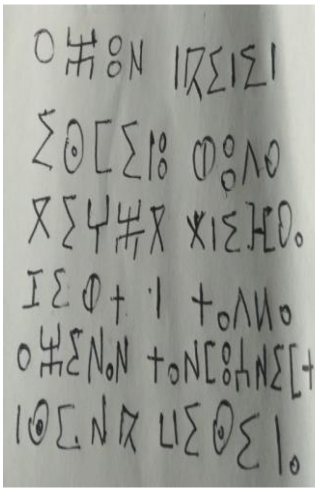

Fig 6: Text number 3

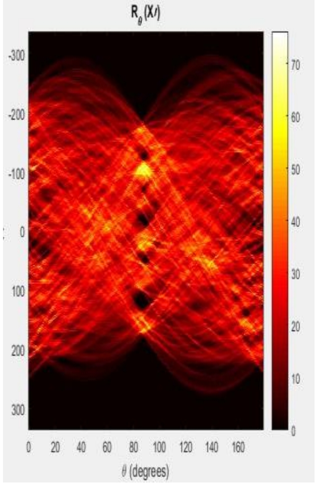

Fig 7: Radon transform representation of text number 3

Finally, all the extracted lines can be segmented in Tifinaghe's characters with a very precise way (As we can see in the figure (Figure 9)).

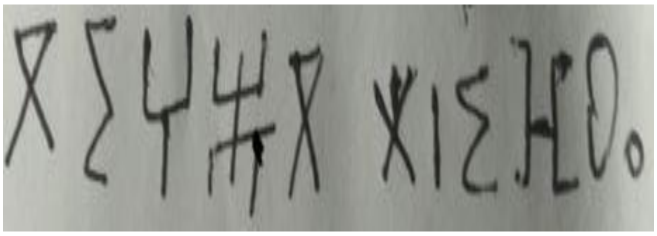

Fig 8: Line of text number 3

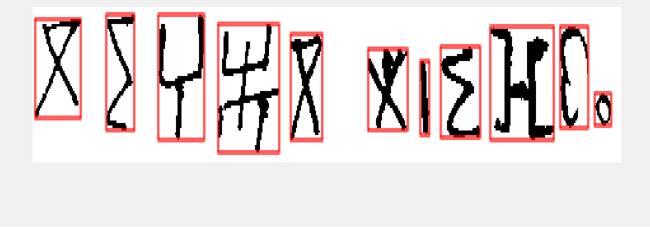

Fig 9: Line segmentation into Tifinaghe's characters

Our approach has been applied on a dataset-image of Tifinaghe documents. This dataset contains 95 images which are composed by handwritten and typewritten Tifinaghe's texts. To evaluate the performance of our approach, we have determined two performance indices: Texts Segmentation Rate (TSR) and Lines Segmentation Rate (LSR).

$$\text{TSR} = \frac{\text{Number of texts correctly segmented}}{\text{Number of texts}} \quad (3)$$

$$\text{LSR} = \frac{\text{Number of lines correctly segmented}}{\text{Number of lines}} \quad (4)$$

The table below (Table 4) presents the results obtained by using the approach presented in this paper. These results clearly show that this approach allows segmenting the Tifinaghe's texts into lines and these extracted lines into Tifinaghe's characters with a very precise way.

Table 4. Results obtained for Tifinaghe texts' segmentation

TSR	LSR
91 %	97 %

Despite the many obstacles that make difficult the Tifinaghe handwritten texts' segmentation, this approach makes it possible with very high success rates.

5. CONCLUSION

Text segmentation is an indispensable step in all document recognition systems. For all documents types (handwritten and typewritten), good segmentation positively influences the recognition result. This step consists of decomposing, isolating and segmenting the text into classifiable units called characters. The tools allowing the acquisition of document images are still far from perfect. This has a direct effect on the images' quality. Thus, to improve their qualities and to increase the efficiency of our system, we applied three families of pre-processing techniques:

- Binarization
- Filtering
- Skew detection and correction

The main objective of this paper was to present the approach that we have adopted for the Tifinaghe text's segmentation (handwritten and typewritten).This approach is based on the Radon transform in terms of segmenting texts into lines and on the connected components algorithm to segment the extracted lines into characters.

The results obtained clearly show that this approach makes to segment Tifinaghe texts with very high success rates (TSR = 91% and LSR = 97%).

6. REFERENCES

[1] Ataa Allah, F. and Frain, J. 2013. Amazigh Converter based on WordprocessingML.

[2] Boukous, A. 2010. Revitalisation de la langue Amazighe : Défis, Enjeux et Stratégies.

[3] Brodic, D., Milivojevic Z. N. and Milivojevic D. R. 2012. Approach to the Improvement of the Text Line Segmentation by Oriented Anisotropic Gaussian Kernel. electronics and electrical engineering, no. 2(118).

[4] Brodic, D. 2015. text line segmentation with water flow algorithm based on power function. Journal of electrical engineering, vol. 66, no.3, pp. 132–141.

[5] Chajri, Y. , Bouikhalene, B. , and Maarir, A. Segmentation of Text/Graphic from Handwritten Mathematical Documents Using Gabor Filter. The International Arab Journal of Information Technology, Vol. 14, No. 4A, Special Issue 2017.

[6] Chajri, Y., Maarir, A. and Bouikhalene, B. 2016. A Comparative Study of Handwritten Mathematical Symbols Recognition. IEEE 13th International Conference Computer Graphics Imaging and Visualization, pp. 448-451.

[7] Chajri Y. and Bouikhalene B. 2016. Handwritten mathematical symbols dataset. Data in brief, vol. 7, pp. 432–436.

[8] Chajri, Y. and Bouikhalene, B.2016. Recognition of Handwritten Mathematical Text. International Journal of Future Generation Communication and Networking, vol. 9, no. 8, pp. 307-316.

[9] Chajri, Y. and Bouikhalene, B. 2016. Handwritten Mathematical Expressions Recognition. International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 9, no. 5, pp. 69-76.

[10] Erkan, U., Gökrem, L. and Enginoğlu, S.,2019. Adaptive Right Median Filter for Salt-and-Pepper Noise Removal". 11. 542-550. 10.29137/umagd.495904.

[11] Hoilund, C. 2007. The Radon Transform. Aalborg University, VGIS, 07gr721 November 12.

[12] Likforman-Sulem, L., Zahour, A. and Taconet, B. Text Line Segmentation of Historical Documents: a Survey".

[13] Louloudis, G., Gatos, B., Pratikakis, I. and Halatsis, C. 2009. Text line and word segmentation of handwritten documents", Pattern Recognition 42, pp. 3169 – 3183.

[14] Miciak, M. Character Recognition Using Radon Transformation and Principal Component Analysis in Postal Applications. Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 495 – 500.

[15] Nicolaou, A. and Gatos, B. 2009. Handwritten Text Line Segmentation by Shredding Text into its Lines. 10th International Conference on Document Analysis and Recognition.

[16] Radon, J.1986. On the determination of functions from their integral values along certain manifolds. IEEE Transactions on Medical Imaging, vol.5, no.4, pp. 170-176.