Recognition of Tifinagh Handwritten Corona Virus COVID-19 Glossary

Yassine Chajri
Sultan Moulay Slimane University
Beni Mellal
Morocco

ABSTRACT

This paper aims to present a data set (images) of Tifinagh handwritten Corona virus COVID-19 glossary and a comprehensive approach that allows this glossary recognition. This approach is based on Radon transform in all steps of this process. In pre-processing phase, this transform intervenes for skew detection and correction. Regarding text segmentation, Radon transform allows transforming text lines into positioned peaks in order to easily extract all lines. Finally, this transform allows recognizing handwritten Tifinagh characters with a very high recognition rate.

General Terms

Documents recognition – Data set.

Keywords

COVID-19 – Tifinagh – Recognition – Segmentation - Handwritten – Radon Transform.

1. INTRODUCTION

The Corona virus pandemic (COVID-19) has exposed humanity to many risks threatening their lives. Because of this virus, suffering has spread and the global/world economy is in a major recession. Also, this pandemic has clearly demonstrated the weakness of health systems even for the largest countries in the world. This shows the great suffering of poor countries which lack the basic of living conditions (lack basic access to clean water, lack of security and stability, lack of hospitals and medical equipment, lack of balanced/ healthy diet, etc.). Today, it is very important for all countries of the world to unite their efforts in order to confront this great danger. Individuals and societies should act in solidarity with each other and contribute from their position to finding effective solutions that can alleviate their suffering. The dissemination of knowledge in general and those related to the Corona virus is one of the most important weapons which humanity must confront these dangers.

Currently, computing field witnesses an impressive dynamic in which the innovations and improvements affecting the hardware as well as the software, follow one another in a very accelerated way. This dynamic has changed our daily lives and offers us new perspectives by proposing new responses to different needs. These new technologies allow producing, transforming or exchanging information in large quantities in a very short time.

On our side, we propose in this paper a data set for Tifinagh handwritten Corona virus COVID-19 glossary. This data set includes a set of most frequently used terms regarding Corona virus topic writing in Tifinagh (alphabet used to write Amazigh language). Also, we present a recognition system able to recognize Tifinagh texts constructed in all of these terms.

2. RELATED WORKS

The present section focuses more on the field of characters recognition which knows considerable improvements (Latin characters, Arabs characters, etc.). However, some languages are still deficient in terms of works that are interested in knowing their characters. Regarding Amazigh language, the Tifinagh character is hardly treated in research. But, there are some approaches that are proposed in this context. These approaches can be classified into groups [1]:

Table 1. Related works.

Statistical approaches	(Oulamara, [2]), (Djematen et al., [3])
Neural network based approaches	(Ait Ouguengay, [4]), (El Yachi et al., [5]), (Es-Saady et al., [6])
Syntactic approach	(Es-Saady et al., [7])
Hidden Markov Model based approaches	(Amrouch et al., [8]), (Amrouch et al., [9])
Dynamic programming approach	(El Yachi et al., [10])

Concerning text lines segmentation, we have detailed in works [11] and [12] the most important techniques (Smearing methods, Grouping methods, Hough transform methods, Projection based methods, stochastic methods, etc.).

3. DATA SET

3.1 Value of data

- Given the importance of Health information technology in healthcare improvement, the recognition of Corona virus COVID-19 documents has become a very important area of scientific research.
- The Arabic language is spoken by more than 466 million people in the whole world.
- The Amazigh language is spoken currently by around 30 million throughout North Africa and the Sahel.
- The attention given by Moroccan official institutions to Amazigh language and culture: Amazigh language is recognized in Morocco's constitution, effective integration of Amazigh language in public policies, etc.

- This dataset contains Corona virus (COVID-19) terms, written in Arabic and Tifinagh, which represent the most frequently used terms.
- It is characterized by several styles of writing.
- It is very useful to implement a recognition system for handwritten documents related to Corona virus topic.
- It facilitates the research in this important area.

Table 2. Some translated terms of COVID-19

English language	Amazigh language	Arabic language
Virus	ⴰⴱⴰⵢⵔⵓⵙ	فيروس
Epidemic	ⴰⴱⴰⵢⵢⵓⵔ	وباء
Pandemic	ⵜⴰⵏⴱⴰⵣⵜ	جائحة
Corona	ⴽⵓⵔⵓⵏⴰ	كورونا
Sars	ⵙⴰⵕⵙ	سارس
Reanimation	ⴰⵙⴷⴼⵓ	الانعاش
Quarantine	ⴰⴱⵍⴰⴹ ⴰⴷⵓⵙⴰⵏ	الحجر الصحي
Isolation	ⴰⵙⵜⴰⵢ	العزل
Mask/Muzzle	ⵜⴰⴽⵎⴰⵎⵜ	الكمامة
Disinfection	ⴰⵙⵉⵣⴷⴳ	التعقيم
Alcohol	ⵍⴰⵍⴽⵓⵍ	الكحول
Sinopharm	ⵙⵉⵏⵓⴼⴰⵔⵎ	سينوفارم
Astrazeneca	ⴰⵙⵜⵔⴰⵣⵉⵏⵉⴽⴰ	أسترازينيكا
Moderna	ⵎⵓⴷⵉⵔⵏⴰ	موديرنا
Pfizer-BioNTech	ⴼⴰⵢⵣⵔ - ⴱⵢⵓⵏⵜⵉⴽ	فايزر- بيونتيك
Johnson & Johnson	ⵊⵓⵏⵙⵓⵏ	جونسون- جونسون
Symptoms	ⵜⵉⵎⵉⵜⴰⵔ	الأعراض
Fever	ⵜⴰⵡⵍⴰ	الحمى
Fatigue	ⴰⵔⵎⴰⵢ	الإرهاق
Headache	ⴰⴷⵉⴷⴰⵏ ⵉⵅⴼ	صداع الرأس
Diarrhea	ⵓⵛⴹⵉⵃ	الإسهال
Sense of Smell	ⵜⴰⴽⴰⵢⵜ ⵏ ⵜⵄⴹⵉ	حاسة الشم
Sense of taste	ⵜⴰⴽⴰⵢⵜ ⵏ ⵢⵉⵔⵎ	حاسة التذوق
Infection/Contagion	ⴰⵏⵉⵣⵖ/ ⵜⴰⵡⴰⵖⵉⵜ	العدوى
Rate of spread	ⴰⵙⵖⵍ ⵏ ⵓⵏⵖⴰⵍ	نسبة الإنتشار
Death rate	ⴰⵙⵖⵍ ⵏ ⵉⵎⵜⵉⵏ	نسبة الوفيات
Confirmed cases	ⴰⴷⴷⴰⴷⵏⵓⵙⵙⵉ ⴷⵏ	الحالات المؤكدة
Excluded/ Suspected cases	ⴰⴷⴷⴰⴷⵏⵉⵎⴰⴳ ⴳⵓⵅⵏ	الحالات المستبعدة
Critical cases	ⴰⴷⴷⴰⴷⵏⵉⵙⵏⵓⴳ ⵎⵓⵜⵏ	الحالات الحرجة
Recovery rate	ⴰⵙⵖⵍ ⵏⵡⵓⵊⵊⵓⵢ	نسبة التعافي
Cumulative incidence	ⴰⵏⵎⵎⴰⵍⵏ ⵜⵎⵙⵙⵡⵉⵜ ⴰⵙⴳⴷⵓⴷⴰⵏ	مؤشر الإصابة التراكمي
Epicenter	ⵜⴰⴽⴰⵜ	بؤرة
Immunity	ⵜⵉⵣⵣⵓⴳⵔⴰ	المناعة
Variant	ⵉⵙⵎⵙⵉ	متحور
Mutation	ⵜⴰⵖⵓⴼⴰⵏⵜ	طفرة
Alpha	ⴰⵍⴼⴰ	ألفا
Beta	ⴱⵉⵜⵜⴰ	بيتا
Gamma	ⴳⴰⵎⵎⴰ	غاما
Delta	ⴷⵉⵍⴹⴰ	دلتا
Omicron	ⵓⵎⵉⴽⵔⵓⵏ	أوميكرون
Mu	ⵎⵓ	مو
Lambda	ⵍⴰⵎⴷⴰ	لامدا
PCR- Test	ⵉⵔⵉⵎ- ⴱⵙ ⵔ	PCR- اختبار

3.2 Data preparation

For the preparation of the dataset we;

- Targeted the students (male and female) of Adouz Middle School in Beni Mellal (Beni Mellal-Khenifra, Morocco).
- Asked them to write a list of most frequently used terms about Corona virus.
- Translated these terms into Amazigh language.
- Used “HP LJ M 127128” to scan pages.
- Used Radon transform for skew detection and correction [13].
- Used histogram equalization for images normalization [14].
- Median filtering for image noise reduction [15].
- Normalized the images with a size of 200 * 80.

3.3 Data organization

The dataset is divided into two parts:

- The first concerns images of terms written in Arabic.
- The second part is devoted to images of terms written in Tifinagh.

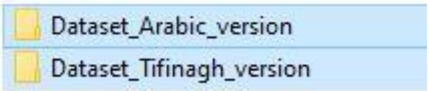
Dataset_Arabic_version
Dataset_Tifinagh_version

Fig 1: Dataset directory

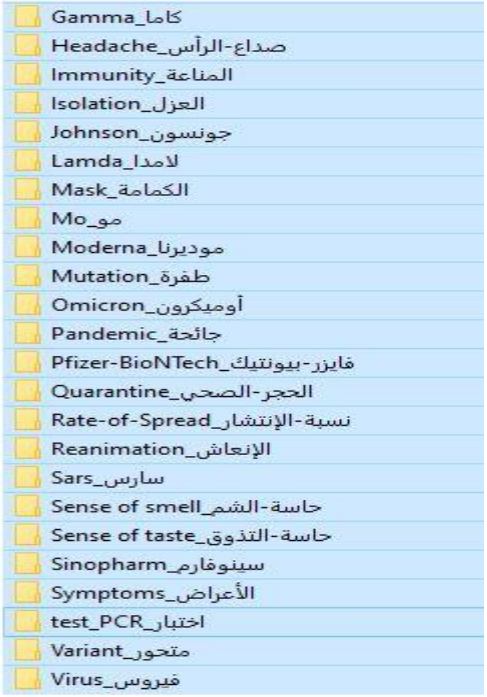
Gamma_گاما
Headache_صداع-الرأس
Immunity_المناعة
Isolation_العزل
Johnson_جونسون
Lamda_لامدا
Mask_الكمامة
Mo_مو
Moderna_مودرنا
Mutation_طفرة
Omicron_أوميكرون
Pandemic_جائحة
Pfizer-BioNTech_فايزر-بيونتيك
Quarantine_الحجر-الصحي
Rate-of-Spread_نسبة-الإنتشار
Reanimation_الإنعاش
Sars_سارس
Sense of smell_حاسة-الشم
Sense of taste_حاسة-التذوق
Sinopharm_سينوفارم
Symptoms_الأعراض
test_PCR_اختبار
Variant_متحور
Virus_فيروس

Fig 2: Arabic version of the dataset

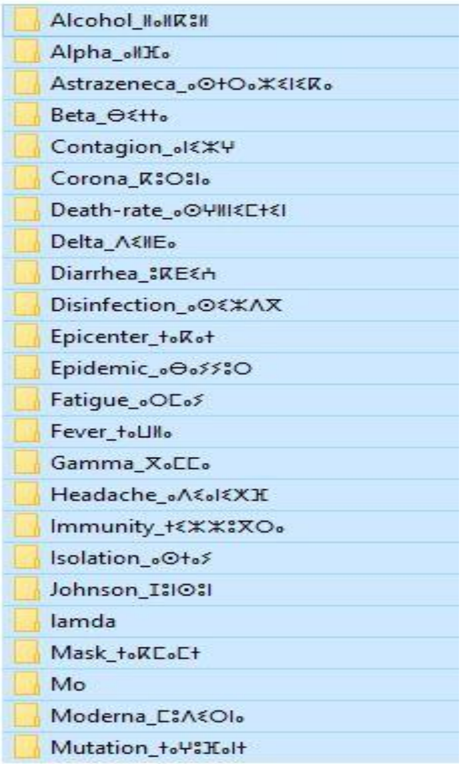
Alcohol_ⵍⴰⵍⴽⵓⵍ
Alpha_ⴰⵍⴼⴰ
Astrazeneca_ⴰⵙⵜⵔⴰⵣⵉⵏⵉⴽⴰ
Beta_ⴱⵉⵜⵜⴰ
Contagion_ⴰⵏⵉⵣⵖ
Corona_ⴽⵓⵔⵓⵏⴰ
Death-rate_ⴰⵙⵖⵍⵏⵉⵎⵜⵉⵏ
Delta_ⴷⵉⵍⵟⴰ
Diarrhea_ⵓⴽⵟⵉⵃ
Disinfection_ⴰⵙⵉⵣⴷⴳ
Epicenter_ⵜⴰⴽⴰⵜ
Epidemic_ⴰⴱⴰⵢⵢⵓⵔ
Fatigue_ⴰⵔⵎⴰⵢ
Fever_ⵜⴰⵡⵍⴰ
Gamma_ⴳⴰⵎⵎⴰ
Headache_ⴰⴷⵉⴰⵏⵉⵅⴼ
Immunity_ⵜⵉⵣⵣⵓⴳⵔⴰ
Isolation_ⴰⵙⵜⴰⵢ
Johnson_ⵊⵓⵏⵙⵓⵏ
lamda
Mask_ⵜⴰⴽⵎⴰⵎⵜ
Mo
Moderna_ⵎⵓⴷⵉⵔⵏⴰ
Mutation_ⵜⴰⵖⵓⴼⴰⵏⵜ

Fig 3: Tifinagh version of the dataset

In each of these datasets, the images obtained for each term are grouped in a file named in three parts:

- The first part concerns Corona virus term in Arabic/Tifinagh.
- The second part is devoted to Corona virus term in English.
- The third part concerns the number of units.

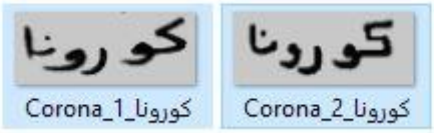
Corona_1_كورونا
Corona_2_كورونا

Fig 4: Corona term directory in Arabic dataset

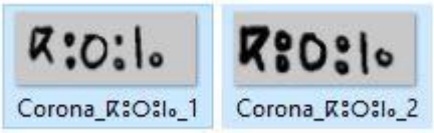
Corona_ⴽⵓⵔⵓⵏⴰ_1
Corona_ⴽⵓⵔⵓⵏⴰ_2

Fig 5: Corona term directory in Tifinagh dataset

4. RECOGNITION OF TIFINAGH HANDWRITTEN CORONAVIRUS COVID-19 GLOSSARY

The approach proposed in this work is based on Radon transform in all steps of Coronavirus Tifinagh text recognition.

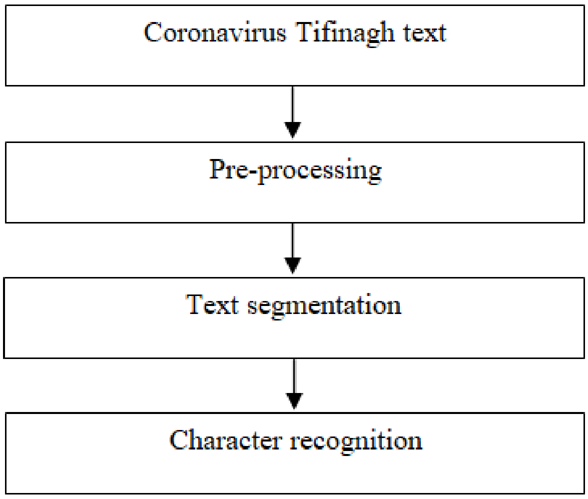

Fig 6: System architecture

4.1 Pre-processing

In image acquisition step, several factors negatively influence images quality such as: paper quality, fonts used in the text, scanner quality, scan resolution, etc. In order to resolve all of these problems, we applied the process presented in the part of data preparation.

4.2 Text segmentation

Text segmentation is a mandatory step in this system. It allows us to obtain isolated units (Tifinagh characters) which will be the objective of the recognition stage. In the paper [12], we have detailed the Tifinagh text segmentation process using Radon Transform.

- ***Radon Transform***

Radon transform is a mathematical technique developed by the mathematician Johann Radon [16]. The application of this transform to an image f (x, y) for a given set of angles can be considered as the projection of the image along the given angles. A projection at a given angle θ is obtained as the linear integration of the function on all parallel lines [17].

The result is a new image R(ρ,θ) that can be written mathematically by [18]:

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

Where:

$\rho = x\cos\theta + y\ \sin\theta$

δ(): is the Dirac delta function.

The Radon transform has the ability to transform text lines into positioned peaks corresponding to lines parameters. As you can see in the figure (Figure 8) which represents Radon transform representation of Coronavirus Tifinagh text with 0° to 179° degrees of projection angle, there are seven colored spots which represent the seven lines in the text presented in figure (Figure 7).

This allows us to quickly extract text lines as we can see in the figure (Figure 9).

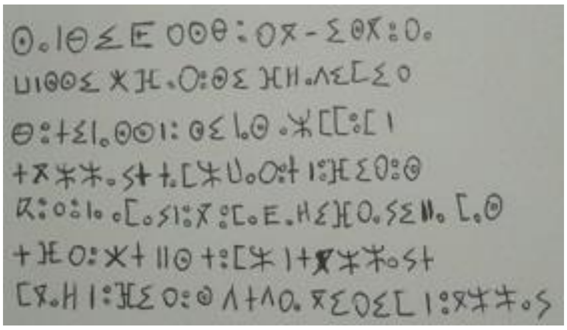

Fig 7: Handwritten Corona virus Tifinagh text

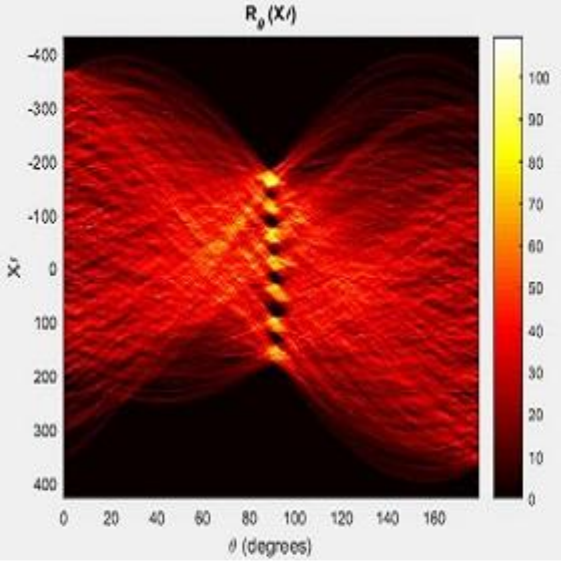

Fig 8: Radon Transform representation of handwritten Corona virus Tifinagh text

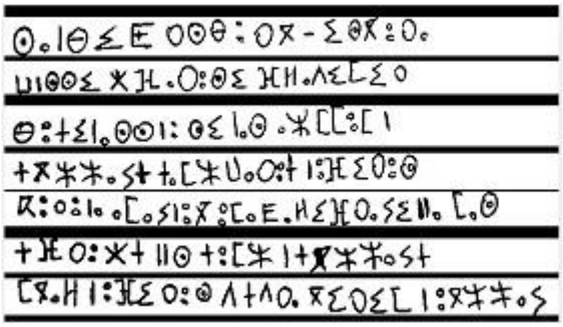

Fig 9: Segmentation of handwritten Corona virus Tifinagh text

After all, the algorithm of connected components intervenes to segment each line into Tifinagh characters [17].

4.3 Tifinagh characters recognition

We focus in this work on Tifinagh characters recognition. More precisely, we present an approach based on the Radon transform and we compare it with an approach based on HOG descriptor.

In the part which concerns text segmentation, we presented a very important characteristic of Radon transform which served us in this stage. Concerning Tifinagh characters recognition, we have exploited other properties characterizing this transform such as:

- ***Linearity:***

Let f and g be two functions and let c_1 and c_2 be two constants, then:

$$T_R(c_1 f + c_2 g) = c_1 T_R(f) + c_2 T_R(f)$$

- ***Translation:***

A displacement of the function $f(x,y)$by a distance (x_0, y_0) results in a change of its transform in the variable ρ by the distance $d = x_0 cos\theta + y_0 sin\theta$

$$T_R f(\rho,\theta) = T_R f(\rho\text{-}x_0\cos\theta\text{-}y_0\sin\theta, \theta)$$

- ***Rotation:***

The rotation of the function $f(x,y)$ by an angle θ_0 causes a phase shift.

$$T_R\ Rot^{\theta_0}\ f(\rho,\theta) = T_R f(\rho, \theta + \theta_0)$$

- ***Scale:***

A scale change of the function $f(x,y)$ by $\alpha \neq 0$ implies a change of scale on its transform $\frac{1}{|\alpha|} T_R f(\alpha\rho, \theta)$

Concerning the feature vector, we calculated the mean according to the following formula:

$$Vector\ (\theta_i) = \frac{\frac{1}{n}\sum_{j=1}^{n} R_{ij}}{R_i}$$

Where $R_i = MAX\,(R_{i1}, R_{i2}, \ldots\ldots, R_{in})$; *i=1,……, 180; n=77.*

The three figures below respectively show the graphical representation of Radon transform of three Tifinagh characters ⵛ, ⴹ and ⵟ.

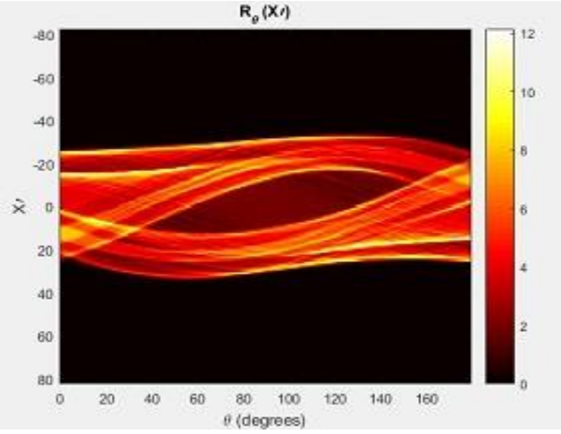

Fig 10: Radon Transform representation of ⵛcharacter

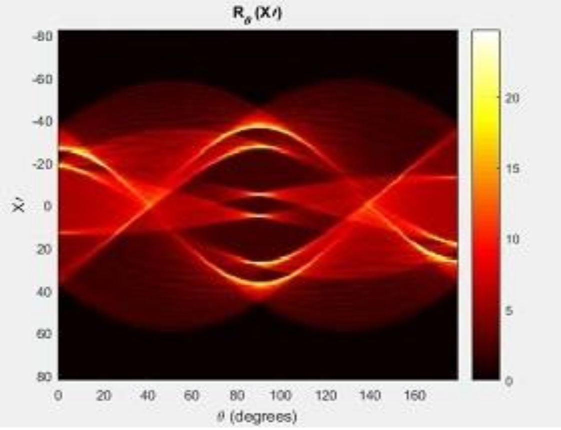

Fig 11: Radon Transform representation of ⴹcharacter

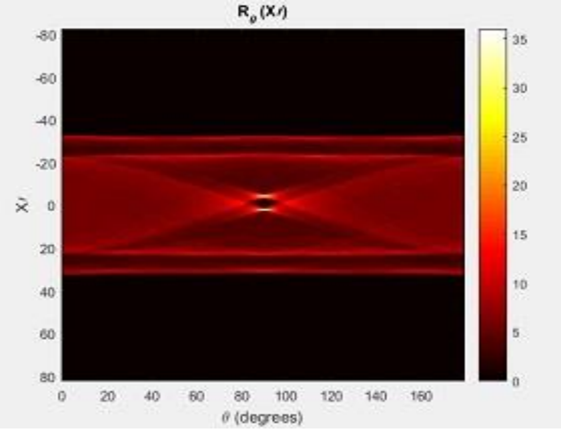

Fig 12: Radon Transform representation of ⵟ character

5. RESULTS

This part will be divided into two sections: the first is devoted to results obtained in texts segmentation stage while the second concerns the results obtained in characters recognition phase.

5.1 Segmentation

The table (Table 3) presents the results obtained by the proposed approach for handwritten Tifinagh texts segmentation.

Table 3. Segmentation rate of handwritten Coron avirus Tifinagh texts

Texts Segmentation Rate (%)	Lines Segmentation Rate (%)
91	97

5.2 Tifinagh characters recognition

The combination of Radon Transform with artificial neural networks (ANN) has given a very high recognition rate which reaches 98.3%. Regarding the approach based on HOG descriptor, its combination with ANN allowed to obtain 95% of recognition rate.

The table (Table 4) summarizes these results:

Table 4. Recognition rate of handwritten Tifinagh characters

Approach	Recognition rate (%)
Radon transform with ANN	98.3
HOG descriptor with ANN	95

For this comparison be more comprehensive, the table (Table 5) shows the number of characteristics extracted by both approaches:

Table 5. Number of characteristics extracted by Radon transform and HOG descriptor approaches

Radon Transform	HOG descriptor
180	45

6. CONCLUSIONS

The world today is living under the brunt of Corona pandemic, which has cast a shadow over all fields of life. This pandemic has turned our daily lives upside down and has made the simplest things the most difficult things to achieve. To face this danger and to return to normal life, all efforts must be combined and knowledge spread to different regions in all world languages. For example, the importance of scientific research appears in its great contribution to overcoming any crisis. This is what we notice clearly in this crisis represented by COVID-19 pandemic, where scientific researches is accelerating in all fields in order to limit its impact. This paper fits into this broader context and provides a data set of Tifinagh handwritten Corona virus COVID-19 glossary. Also, it proposes a comprehensive approach allowing this glossary recognition. This approach is based on Radon transform in all steps of this process (pre-processing, text segmentation, handwritten Tifinagh characters recognition). This approach has shown its effectiveness by allowing handwritten Tifinagh texts segmentation with a success rate of 91%, lines segmentation with a success rate of 97% and handwritten Tifinagh characters recognition with a success rate of 98.3%.

7. REFERENCES

[1] A Rachidi. Reconnaissance automatique de caractères et de textes amazighes : état des lieux et perspectives », Asinag. 9, 2014,119-132

[2] A Oulamara, and J. Duvernoy. An application of the Hough transform to automatic recognition of Berber characters. Signal Processing, vol.14, 1988, 79-90.

[3] A Djematen , B Taconet. and A Zahour. A Geometrical Method for Printing and Handwritten Berber Character Recognition. ICDAR'97, 1997,564.

[4] Y Ait Ouguengay and M Taalabi. Elaboration d'un réseau de neurones artificiels pour la reconnaissance optique de la graphie amazighe: Phase d'apprentissage. Systèmes intelligents-Théories et applications, Paris : Europia, cop. (impr. au Maroc), ISBN-102909285553,2009.

[5] R El Ayachi, K Moro, M Fakir and B. Bouikhalene. Recognition of Tifinaghe Characters Using a Multilayer Neural Network. International Journal Of Image Processing (IJIP), vol. 5, Issue 2, 2011.

[6] Y Es Saady, A Rachidi, M El Yassa and D. Mammass. AMHCD: A Database for Amazigh Handwritten Character Recognition Research. International Journal of Computer Applications , Vol.27, N°.4, 2011, 44-48.

[7] Y Es Saady, A Rachidi, M El Yassa, D Mammass. Printed Amazigh Character Recognition by a Syntactic Approach using Finite Automata. ICGSTGVIP Journal, vol.10, Issue 2, 2010,1-8.

[8] M Amrouch, A Rachidi, M El Yassa and D Mammass. Handwritten Amazigh Character Recognition Based On Hidden Markov Models. ICGST-GVIP Journal, vol.10, Issue 5, 2010,11-18.

[9] M Amrouch, Y Es Saady, A Rachidi, M El Yassa and D Mammass. Handwritten Amazigh Character Recognition System Based on Continuous HMMs and Directional Features. IJMER journal, Vol.2, Issue 2, 2012, 436-441.

[10] R El Ayachi, K Moro, M Fakir and B Bouikhalene. On the Recognition of Tifinaghe Scripts. Journal of Theoretical and Applied Information Technology, vol.20, 2, 2010, 61-66.

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

[12] Y Chajri and B Bouikhalene. Segmentation of Handwritten and Typewritten Tifinaghe Texts. International Journal of Computer Applications, vol 183, 27, 2021, 49-52.

[13] M Hasegawa and S Tabbone. Histogram of radon transform with angle correlation matrix for distortion invariant shape descriptor. NeuroComputing.

[14] S Parker and J Kemi. Ladeji-Osias Implementing a Histogram Equalization Algorithm in Reconfigurable Hardware.

[15] K ManglemSingh. Fuzzy rule based median filter for gray-scale images. J.Inf.Hiding Multimed.SignalProcess. vol 2, 2, 2011.

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

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

[18] C Hoilund.The Radon Transform. Aalborg University, VGIS, 07gr721, 2007.