*ICIC Express Letters*
*Part B: Applications*
Volume **10**, Number **9**, September **2019**

 ISSN 2185-2766
pp. 763–771

# CLASSIFICATION APPROACHES' BEHAVIOR IN OPTICAL CHARACTER RECOGNITION SYSTEMS FOR DIACRITICAL LANGUAGES: CASE OF AMAZIGH LANGUAGE

Khadija El Gajoui[1,*], Fadoua Ataa Allah[2] and Mohammed Oumsis[3]

[1]LRIT – CNRST URAC 29, Rabat IT Center, Faculty of Sciences
[3]IT Department, School of Technology-Sale
Mohammed V University
Agdal 10080, Rabat, Morocco
[*]Corresponding author: Khadija.gajoui@gmail.com; oumsis@yahoo.com

[2]Center of Computer Science Studies, Information Systems and Communication (CEISIC)
The Royal Institute of Amazigh Culture
Madinat Al Irfane 2055, Rabat, Morocco
ataaallah@ircam.ma

Received January 2019; accepted April 2019

Abstract. *Optical Character Recognition (OCR) is based on converting a textual image to an editable text. To this end, OCR systems are composed of a set of modules including image preprocessing, segmentation, feature extraction, and classification. In this paper, we introduce our proposed OCR system for Amazigh language transcribed in Latin. In order to improve the performance of this system, we have, first, studied the influence of different preprocessing treatments on the recognition, typically for historical documents. Then, we have applied two approaches in the classification phase, which are neural network and adaptive classifier. To undertake this study, we have prepared a corpus including a set of documents extracted from books, written in Amazigh transcribed in Latin. The result of this study shows that neural network classifier gives best recognition rate for diacritic languages, particularly in the case of Amazigh language transcribed in Latin.*

**Keywords:** OCR, Amazigh, Neural network, Adaptive classifier, Diacritics, Transcription

1. **Introduction.** The Optical Character Recognition (OCR) [1-3] is a technology that allows converting various types of documents such as scanned paper documents, PDF files and digital images into editable and searchable format. OCR technology can be applied to both on-line and off-line writing aspect [4], and can be integrated in many application domains such as automatic processing, archiving, and indexing [5]. An OCR system [8] consists of pre-processing phase, segmentation phase [2,6], feature extraction phase [7,8] and classification phase [9].

With the officialization of the Amazigh language, several studies were made on this language. Existing studies on Amazigh OCR systems have focused on Amazigh writing in the Tifinaghe alphabet. However, this alphabet has been generalized and automated recently with the creation of the Royal Institute of Amazigh Culture in 2001. We notice the existence of large number of documents written in Amazigh language transcribed in Arabic and Latin alphabet. In this paper, we chose to treat Amazigh documents written in Latin alphabet.

In the remaining of this paper, we introduce the Amazigh language writing characteristics in Section 2. In Section 3, we present our proposed system. Then, we show, in

DOI: 10.24507/icicelb.10.09.763

Section 4, the evaluation of the proposed system tested on a set of documents extracted from different books. Finally, in Section 5, we draw conclusions and suggest further related research.

2. **Characteristics of Amazigh Language.** The Amazigh language, or Tamazight, is one of the oldest humankind languages [10]. There are no official data on the number of speakers, but the number of users is estimated to around thirty to fourty million. Three writing systems are used [11] to transcribe Amazigh language in Morocco: Tifinagh [12], the Arabic alphabet and the Latin alphabet which we focus on in this work [13]. The system used for the Amazigh transcription is not unique for all documents. There are standard transcription systems developed by known authors. Certain authors base their writings on referenced transcription tables, while others use their own charsets. Knowing that the charset used for transcription is not unique, and there is no standard system elaborated for this purpose, we have tried to collect the charsets used by different authors from a set of Amazigh documents transcribed in Latin [10] and form a list containing all characters used [14]. We can notice that the charsets used in the transcription of Amazigh in Latin is composed of the Latin alphabet with diacritics below, above and after the characters. In addition to the Latin alphabet, other special characters are used in the transcription. Thus, we have gotten a set of 23 consonants, 6 vowels and 10 diacritics [10]. We consider the Amazigh language transcribed in Latin as a diacritical language due to the presence of diacritics in a set of characters.

The difficulty facing the Amazigh language transcribed in Latin processing arises in two essential points.

- The documents seniority: documents seniority is a major problem in the OCR. Several factors influence the recognition rate [13]. Among these factors, we can mention the quality and color of the paper. The quality of the documents may cause the transparency of the paper, where the characters on a page appear in another page in two-sided document case. On the other hand, the yellow color, characterizing such documents, turns into noise after scan. The old style of writing and breaks in the characters can also be an obstacle to good recognition.
  The figure below shows an example of these documents.


AD LECTOREM DE VSV

Figure 1. An example of an ancient document containing Amazigh text

- The presence of diacritics: the presence of diacritics represents, generally, an obstacle for the recognition. The diacritical marks can be considered as noise, and then eliminated in the pretreatment phase. Some characters have the same body with different diacritics so the chance to be confused in the classification phase is high.

## 3. The Proposed System.

3.1. **System architecture.** The architecture of the proposed system is displayed in Figure 2 [10].

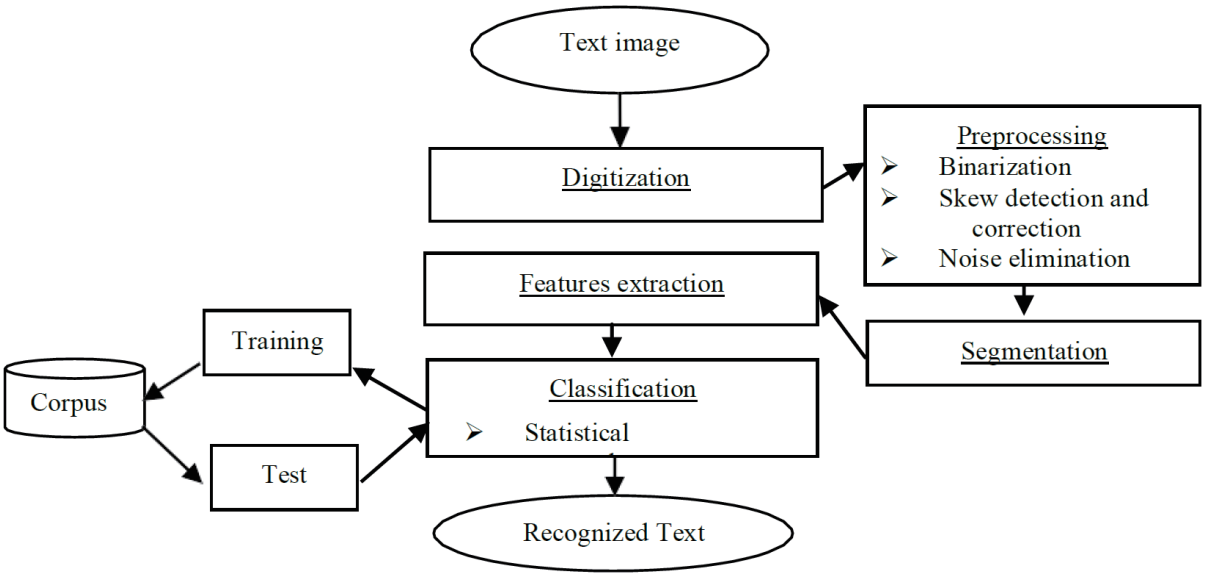


FIGURE 2. The architecture of our proposed system

3.2. **OCR system phases.** Our OCR system for the Amazigh language transcribed in Latin is composed as follows. The first phase is **the digitization**. After being scanned via an optical scanner, images are entered to the system, and then the systems steps are executed.

**Pre-processing phase**: In this phase, we choose to use tree treatments. The first one is binarization, where we compared tree binarization methods in order to select the most appropriate. The second treatment is skew detection and correction. In this treatment, we used Hough transformation. This transformation allows detecting writing lines and angle rotation of the document. It can be applied to any geometric shape that can be described by this equation:

$$\rho = x * \cos\theta + y * \sin\theta \tag{1}$$

where $(\rho, \theta)$ defines a vector from the origin to the nearest point on the line.

The last treatment is noise elimination. This treatment is very important for our system, since we deal with ancient documents that contain a special noise due to the paper quality. After several tests on different filters we chose to apply the median filter.

**Segmentation phase**: Since our language has a non-continuous cursive script, we used a horizontal histogram to extract lines and vertical histogram to extract characters.

**Features extraction phase**: At this level, we extracted different features such as gradients, and singular points of skeleton.

**Classification phase**: In this phase, we choose to use statistical approaches for recognition. We compared two famous methods in order to visualize the behavior of each one against the script used in the transcription of Amazigh language that is characterized by the presence of diacritics. The two chosen methods are *the Recurrent Neural Networks, and the Adaptive Classifier.*

In the literature, it has been shown that the adaptive classifier gives good results for diacritical languages such as Greek, Urdu and Arabic. However, the neural network method is known by its robustness and capacity to be adapted to complicated cases. So, we will study the behavior of each of these two methods relatively to our studied language in order to choose the most appropriate. *Recurrent Neural Networks*: The RNNs are known by their capacity to learn and recognize complicated problems. They are used in many projects concerning OCR and they give important recognition rates [15,17]. The RNN chosen is **the Long Short-Term Memory (LSTM)**.

Proposed in the mid-90s, the LSTMs arrived to overcome the neural networks problem of forgetting information learned previously. They allow recurrent nets to continue to learn over many time steps. LSTM has proven to be very successful in machine learning and AI mostly in pattern recognition even in complex cases such as handwriting recognition [17]. The idea behind the LSTM is that each computational unit is related not only to a hidden state $h$ but also to a state $c$ of the cell that plays the role of memory. Each hidden layer in LSTM architecture is constituted of blocks. An example of block is shown in Figure 3.

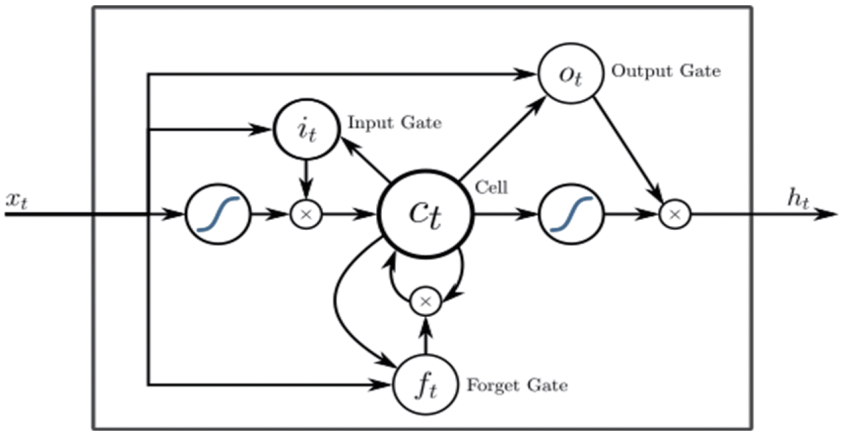


FIGURE 3. A block in LSTM

The LSTM block is composed of three non-linear gates namely, an input, an output and a forget gate. The gates intended role is to regulate the information flow into and out of the cell. The goal of this operation is to regulate long-range dependencies and in consequence, achieve successful RNN training. Each of the gates has its own parameters [18].

The equations for the LSTM memory blocks are given as follows:

$$i_t = \sigma\left(U_i h_{t-1} + W_i x_t + b_i\right) \tag{2}$$

$$f_t = \sigma\left(U_f h_{t-1} + W_f x_t + b_f\right) \tag{3}$$

$$o_t = \sigma\left(U_o h_{t-1} + W_o x_t + b_o\right) \tag{4}$$

$$c_t = f_t * c_{t-1} + i_t * \tanh\left(U_c h_{t-1} + W_c x_t + b_c\right) \tag{5}$$

$$h_t = o_t * \tanh(c_t) \tag{6}$$

where

$x$: input vector to the LSTM block
$i_t$: input gate's activation vector
$f_t$: forget gate's activation vector
$o_t$: output gate's activation vector
$c_t$: cell state vector
$h_t$: output vector of the LSTM block
$W$, $U$, $b$: weight matrices learned during training.

*Adaptive Classifier*: It has been suggested and demonstrated that OCR engines can benefit from the use of an adaptive classifier. The specificity of the Amazigh language transcribed into Latin characters is the presence of diacritic below and above a large number of characters. The experiments on this classifier for diacritical languages, such as ancient Greek [19] and Urdu [20], have shown that it is strong for this type of languages. The adaptive classifier in our case is based on polygonal approximation features. The polygonal approximation can be obtained by choosing the polygon vertices in such a way that the overall approximation error is minimized.

$$\text{Error measures:} \quad \begin{array}{l} \text{Mean square } E_2 = \sum_{i=2}^{N-1} |x_i - d_i|^2 \\ \text{Maximal } E_{\max} = \max_{2 \le i \le N-1} |x_i - d_i| \end{array}$$

4. **Experiments and Results.** The objective of our work is to elaborate an OCR system that can recognize the characters used in the Latin transcription of Amazigh language. To study the behavior of our proposed system against this language, we constructed a dedicated OCR corpus, and undertook a series of experiments focusing on preprocessing and classification phases.

4.1. **Corpus construction.** As none of the previous work has been done on the Amazigh language transcribed on Latin, there is no standard corpus for this language, which justifies the need to construct our own corpus. We have created 2 corpuses: reference corpus to train and test system models and validation corpus to validate the built model. The creation of reference corpus, associated to the Amazigh language transcribed in Latin, passed through tree steps.

- We collected different charsets used in the transcriptions from different books.
- We applied many fonts with varied sizes. The used fonts must be adequate to writing characterized by the presence of diacritics.
- We created a set of lines containing the extracted characters in different fonts and sizes. Each line is represented the corpus by an image and a text file including the transcription of this line.

The reference corpus is composed of training and test corpora. The two corpora contain respectively 7,000 and 3,000 images. The validation corpus is composed of raw document images containing a text written in Amazigh language transcribed in Latin. It is prepared based on scanned pages collected from different books [10]. Some books in this collection are ancient and others are recent. Therefore, the quality of the documents differs from one document to another depending on the state and the book seniority. There are 3 types of documents:

- Doc 1: recent document with good quality
- Doc 2: ancient document
- Doc 3: ancient document with transparent paper

4.2. **Preprocessing phase experiments.** As explained in Section 3, we chose to apply tree different treatments in the preprocessing phase.

**Binarization.** We compared tree famous binarization methods that are Sauvola's method, non-linear method and Otsu's method. Table 1 gives the recognition rates according to the three different binarization methods.

Table 1. Recognition rates for binarization methods

| **Binarization method** | Sauvola | Otsu | Nonlinear |
|---|---|---|---|
| **Recognition rates** | 66% | 82% | 96% |

The experiment is applied on scanned documents. Unfortunately, the scanner degrades usually the image quality and leaves some dark pixels and gray spots that may prevent good character recognition. According to the experimentations, the Otsu's method does not allow eliminating the image defects completely, which influences the recognition rate. On the other hand, Sauvola's method manages to remove the imperfections. However, it generates some breaks on the characters structure and causes the deletion of some text parts, which explains the low recognition rate noted for this method. However, the nonlinear method gives best result in binarization and character pattern conservation, which is illustrated by the high recognition rate in Table 1.

**Skew detection and correction.** The second treatment is the skew detection and correction. It aims to rotate the document to the right direction and correct the skew caused generally by the scanner. For these reasons, we used the Hough transformation.

The skew in an image makes the horizontal and vertical segmentation a very difficult task, since the lines are not straight, and **decrease greatly the recognition rate**.

To overcome this problem, the Hough transformation method is used. It is able to detect the angle rotation by defining the baseline of the writing, and to allow a good segmentation [10].

**Noise elimination.** The third treatment in our system preprocessing phase is the noise elimination. Noise in a document can be confused with characters, which influences the recognition performance, especially in the case of diacritical languages. To deal with this problem, we apply median filter. The noise elimination result illustrated in Table 2 shows that the median filter arrives to increase the recognition rate by 9%.

Table 2. Recognition rates for noise elimination method

| **Noise elimination** | Raw | Median filter |
|---|---|---|
| **Recognition rate** | 86% | 95% |

4.3. **Classification phase experiments.** To study the behavior of classification approaches against one of diacritical languages, which is the Amazigh language transcribed in Latin in our case, we chose two approaches: neural network and adaptive classifier. To conduct experimentation based on neural network, we used OCRopus System [14]. It is a free document analysis and optical character recognition tool. It is based on statistical approaches using Multi-Layer Perceptrons (MLPs). While, we used Tesseract [21] system to study the behavior of the adaptive classifier against the neural network model.

**Model training and test**

The training phase is a primary step in the two approaches. We used our created corpus to train both systems. To train the neural network system, we undertook the learning for more than 30,000 iterations. After each 1000 iterations, a model is created based on the previous models. We generated 20 different neural network models in total.

The test shows that the best model is **obtained after 20,000 iterations and gives a recognition rate of 98%**. The training of the Tesseract system passes through tree steps: the generation of boxes, the creation of the trained data file and the training [21]. **Test on the obtained system succeeded in 96%**.

**Model evaluation**

To evaluate the two approaches, we used the validation corpus defined in the corpus constructions section. In order to analyze the system behavior towards pre-processing, we ran the system on documents in two steps. In the first step, we used raw documents, without any preprocessing. In the second step, documents have undergone pretreatments, which are binarization skew detection and noise elimination, previously discussed in the preprocessing phase. The recognition rates are shown in Table 3.

Table 3. Recognition rates

| Recognition rates | | Approaches | |
|---|---|---|---|
| | | **NN** | **AC** |
| **Document quality variation** | **Raw Doc 1** | **95%** | 94% |
| | **Preprocessed Doc 1** | **97%** | 95% |
| | **Raw Doc 2** | 60% | **70%** |
| | **Preprocessed Doc 2** | **96%** | 90% |
| | **Raw Doc 3** | 54% | **60%** |
| | **Preprocessed Doc 3** | **91%** | 86% |

In order to observe the impact of each approach on the recognition of character with/ without diacritic, we have calculated the classification rate of these characters for both approaches. For that, we used documents of type Doc 2 with preprocessing. The results are displayed on Table 4.

Table 4. Character with/without diacritic recognition rate for each approach

| | Approaches | |
|---|---|---|
| | **NN** | **AC** |
| **Character with diacritic** | 80% | 71% |
| **Character without diacritic** | 98% | 94% |

4.4. **Results.** According to Table 3, we remark that the adaptive classifier gives better result while dealing with raw ancient documents. Furthermore, we notice that the recognition rate decreases as the document quality deteriorates from Doc 1 to Doc 3. While, the preprocessing importance increases for both approaches, even it is more efficient for the neural network approach. Thus, the neural network approach gives better results compared to the adaptive classifier that reach **respectively 96% and 90% for preprocessed ancient documents**. In processed documents, the recognition rate is remarkably low compared to documents with processing in the two approaches. Several recognition errors appear in both cases. Those errors are usually due to noise, or to the characters breaks caused by some treatments. However, there are some misclassifications errors, for example,

- The capital letters and lowercases are sometimes confused;
- Problem in detecting the absence or the presence of diacritics for some characters like "G" is confused with "Ḡ", and "Ů" with "U";
- "w" is generally not recognized;
- Spaces are sometimes missed.

Comparison of recognition rates of characters with and without diacritics, for both approaches, shows that the classification errors are made mainly in the characters with diacritics. The difference is remarkable for both approaches but the NN is more suitable for the recognition of those characters. The errors remarked on diacritical characters are such as:

- Characters that are recognized as two characters, for example, "ů" is recognized as "ii" or "u" as "rr";
- Confusion between characters: "d̲" and "t̲" with "l̤", "g" with "ġ", ....

There are also some errors in characters with no diacritics such as confusion between "e" and "c", "a" and "u", "nn" and "m", "h" and "lr".

Discrimination between diacritics differs from an approach to another. Some of these diacritics are distinguishable but others are confused. Table 4 shows that the rate of

diacritical marks recognition is high for the two approaches. However, the fusion of diacritics with the body of the character prevents the recognition of retching 100%. Fusion problems can happen in the acquisition phase, when we scan the document, or in the preprocessing phase. Some diacritics are similar, so the chance to be confused increases. Other diacritics are in the form of superscript character (written under, above, to the right or to the left of the character). Those diacritical marks are confused with the original character. The confusion of character with their exponent is due to problems of positioning the character relatively to the baseline. These experiments show that preprocessing phase is an important phase for the OCR system. However, the treatments chosen must be adequate to documents category. Insufficient treatment can produce damages in the characters structure that influence recognition performance. Both approaches gave an interesting percentage of recognition, which shows that the learning based on our constructed corpus is quite successful. The results of these experiments prove that the neural network approach is the best approach for the classification of diacritical language, which is, in our case, the Amazigh language transcribed in Latin. The adaptive classifier gives also good recognition rate but the error on diacritical characters is much more important than the neural network approach. We note that there is no reference system to compare with, as a result of the lack of research developed for this language.

5. **Conclusion.** In this paper, we have introduced the studied language, which is the Amazigh language transcribed in Latin. Then, we have described our proposed system composed of a set of modules. With the aim to study the behavior of two statistical classification approaches, which are neural network and adaptive classifier, we created a corpus and we trained both classifiers. On the other hand, we compared some preprocessing treatments, in order to study the importance of this phase in the recognition performance. Experiments show that results, with and without pre-processing phase, are dissimilar. Recognition rates increase remarkably when applying the treatments on the image. Consequently, we can conclude that pre-processing phase is a primary step in OCR systems. On the other hand, comparison of the two classifiers, on the constructed corpus, gives a recognition rate of 95% for the adaptive classifier vs. 97% for the NN which proves that the neural network approach is better than the adaptive classifier. In perspective of this work, we will develop the corpus and combine the classifiers to give a better recognition. On the other hand, we will work on a second transcription of the Amazigh language which is the Arabic transcription.

## REFERENCES

[1] M. D. Felse, R. A. Banks Jr., J. Benton et al., *Optical Character Recognition Pre-Verification System*, U.S. Patent Application, No. 14/834591, 2018.

[2] I. Chaker and R. Benslimane, New approach for the recognition of printed Arabic characters, *Mediterranean Telecommunications Review*, vol.1, no.2, 2014.

[3] G. Mathur and S. Rikhari, A review on recognition of Indian handwritten numerals, *International Journal of Advance Research, Ideas and Innovations in Technology*, vol.3, no.3, pp.545-548, 2017.

[4] A. Verma, S. Arora and P. Verma, OCR – Optical character recognition, *The 7th International Conferernce on Recent Innovations in Science, Engineering and Management*, 2016.

[5] P. Jan, P. Jiri and N. Milan, OCR systems in language specific environments, *Recent Researches in Telecommunications, Informatics, Electronics & Signal Processing*, 2011.

[6] J. R. Bruce, *Mathematical Expression Detection and Segmentation in Document Images*, Ph.D. Thesis, Virginia Tech, 2014.

[7] M. Askari, M. Asadi, A. Bidgoli and H. Ebrahimpour, Isolated Persian/Arabic handwriting characters: Derivative projection, profile features, implemented on GPUS, *Journal of AI and Data Mining*, vol.4, no.1, pp.9-17, 2016.

[8] A. Chaudhuri, K. Mandaviya, P. Badelia et al., Optical character recognition systems for Latin language, in *Optical Character Recognition Systems for Different Languages with Soft Computing*, Springer, Cham, 2017.

[9] P. Sharma and R. Singh, Survey and classification of character recognition system, *International Journal of Engineering Trends and Technology*, vol.4, no.3, 2013.
[10] K. El Gajoui and F. Ataa Allah, Optical character recognition for multilingual documents: Amazighe-French, *Proc. of the 2nd World Conference on Complex Systems (WCCS)*, pp.84-89, 2014.
[11] A. Skounti and A. Lemjidi, Tirra: The origins of writing in Morocco, *The Royal Institute of Amazigh Culture*, vol.1, 2003.
[12] R. El Ayachi, M. Fakir and B. Bouikhalene, *Recognition of Tifinaghe Characters Using Dynamic Programming & Neural Network*, InTech, 2011.
[13] F. Drira, Restoring historic documents degraded over time, *Proc. of the 2nd International Conference on Document Image Analysis for Libraries (DIAL'06)*, Washington, DC, USA, pp.350-357, 2016.
[14] K. El Gajoui, F. Ataa Allah and M. Oumsis, Diacritical language OCR based on neural network: Case of Amazigh language, *Procedia Computer Science*, vol.73, pp.298-305, 2015.
[15] T. Ganesh and N. Jyothish, Optical character recognition of tamil characters based on DCT features using SVM and neural network, *International Journal of Engineering & Future Technology*, vol.14, no.2, pp.33-40, 2017.
[16] J. Palka and J. Palka, OCR systems based on neural network, *Proc. of Annals of DAAAM and Proc. of the International DAAAM Symposium*, 2011.
[17] U. Springmann and A. Lüdeling, OCR of historical printings with an application to building diachronic corpora: A case study using the ridges herbal corpus, *arXiv*, vol.1, 2016.
[18] Y. Lu and F. M. Salem, Simplified gating in long short-term memory recurrent neural networks, *arXiv*, vol.61, 2017.
[19] W. Nick, *Training Tesseract for Ancient Greek OCR*, Google Inc. eutypon, 2012.
[20] A. A. Quratul, H. Sarmad, N. Aneeta, A. Umair and I. Faheem, Adapting Tesseract for complex scripts: An example for Urdu Nastalique, *Proc. of the 11th IAPR Workshop on Document Analysis Systems*, vol.1, 2014.
[21] K. El Gajoui, F. Ataa Allah and M. Oumsis, Recognition of Amazigh language transcribed into Latin based on polygonal approximation, *International Journal of Circuits, Systems and Signal Processing*, vol.10, no.1, 2016.