Document Image Converter to XML File, Case of Tifinagh

Mehdi Boutaounte
Faculty of science and techniques
Beni-mellal,
Morocco

Youssef Ouadid
Faculty of science and techniques
Beni-mellal,
Morocco

ABSTRACT
The document image converter is a necessity in our life for many reasons such as digitization of data in paper to secure them and gain time by automating this task, on the other hand preserving the environment by maintaining trees the first source of paper, for that this works based in a bi-cubic method for physical structure analyze and a graphs models representation for the characters recognition to generate at the end a standard XML file that can be used to create file is realised (doc, pdf ,html…).

General Terms
Pattern Recognition, Image processing..

Keywords
Document converter; physical structure; bi-cubic interpolation recognition; Key points.

1. INTRODUCTION
To create a converter from image of document to a digital document a combination is needed between first the physical structure recognition methods in which we found many works like the famous XY cut algorithm [1] that relies also L.Cinque [2] have proposed a method where they are using multi-resolution approach that give the essential information for structure analyses, but this method are destined to document with simple structure, so for the document with a complex layout structure we found the work of Jic Xi[3] that have proposed a method using the horizontal project -profile of document image to estimate the size of text area in order to determine the threshold used in RLSA[3], the problem that RLSA algorithm need to run through the image pixel by pixels also this method need extra treatment to combine the components resultant , so in order to resolve the problem a suggestion of a method based in bi-cubic interpolation explain in this paper. Second the optical character recognition methods as example Tifinagh characters because of the scarcity of works in it as the work of R.EL Ayachi [4] using neural networks, O.Bencharef [5] compare directly the forms of characters using a Riemannian metrics descriptors, A.Haidar [6] use the hidden Markov model in hybridization with the fuzzy logic. Another works proposed by B. El-Kessab [7] that combines between the neural networks and Markov model.

In this work a method of physical structure recognition applied in documents with a complex structure and optical character recognition using key points is used to have the results explain in this paper as follow: the first section describes the method used to analyze the physical structure of the document using the bi-cubic interpolation [8] the second section a classification of the components (images, Text, etc). The text components will be undergoing into next processing, segmentation and recognition of characters using the graphs model. The last section is reserved for the creation of XML file.

2. PHYSICAL STRUCTURE RECOGNITION

2.1 Bi-cubic interpolation
Bi-cubic interpolation, take into account the 16 pixels closest neighbors to the point to interpolate the value of the new pixel however the bilinear or nearest neighbor method does not exceed 8 pixels. The idea is to fit a polynomial model of the 16 gray levels of the source image, and then deduce the level of the interpolated point value calculation taken by model represented in equation 1.

$$M_g(x,y) = \sum_{p=0}^{p=3} \sum_{q=0}^{q=3} \alpha_{pq} x^p y^q \qquad (1)$$

2.2 Proposed Method
The method is based on bottom-up analysis. In order to create homogeneous blocks, we apply a smoothing on the binary document image using the method of image resizing based on bi-cubic interpolation algorithm [9] which will generated for each pixel a value calculated from neighboring pixels so the mean using this method is to reduce the size of the image to removes white spaces between characters, words and even lines then reset the image to the original size. to keep the original image size the method are divided into two part, first the vertical smoothing of the document image so we reduce the height of the image by a factor N (figure 1).

Fig 1: Result of reducing the height

The final phase is to restoring the original size of the image, the new image contain value between 0 and 1 so it need a binarisation to complete the vertical smoothing stage.

Fig 2: Restore the image to the original dimension

The same process is applied to the original image but this time horizontally to create a horizontal smoothing by reducing the weight ,the result in figure 3.

Fig 3: Restore the image to the original dimension

The results of vertical and horizontal smoothing are fused to gives the following result in figure 4

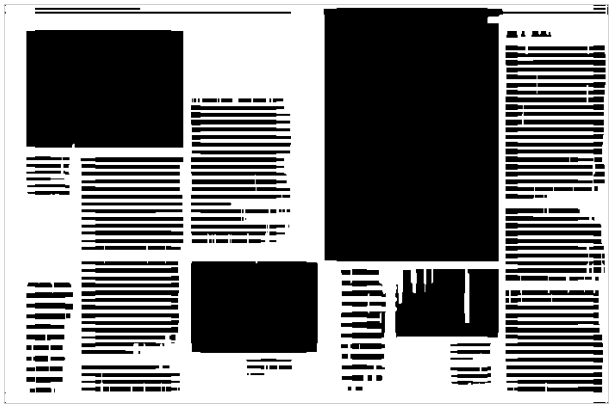

Fig 4: Result after detecting components

After carrying out smoothing on the picture, an algorithm is applied to correct the grouping of component. For this the distance between it is used [10] as the first parameter and the size of each one as a second parameter.

The Fuzzy k-Nearest [11] neighbor used to make decision of grouping areas or not based in the distance and size it give the fuzzy memberships of each component to it nearest and according to this value the decision is made using the following equation

$$u_i(x) = \frac{\sum_{j=1}^{k} u_{ij}\,(1/||x-x_j||^{2/(m-1)})}{\sum_{j=1}^{k}(1/||x-x_j||^{2/(m-1)})} \qquad (2)$$

where i = 1, 2, · · · , c. and j = 1, 2,· · · , k. with c number of classes and k number of nearest. m is used to determine how heavily the distance is weighted when calculating each neighbor's contribution to the membership value, and its value is usually chosen as m ∈ (1,+∞).

‖x − xj‖ is the Euclidean distance between x and its j th nearest neighbor xj . And uij is the membership degree of the pattern xj from the training set to the class i, among the k nearest neighbors of x.

$$u_{ij}(x_k) = \begin{cases} 0.51 + \left(\frac{n_j}{k}\right) * 0.49 & \text{if} \quad i = j \\ \left(\frac{n_j}{k}\right) * 0.49 & \text{if not} \end{cases} \qquad (3)$$

2.3 Classification of component

In this part we use two parameter to classify the images and the text areas, first the density of black pixels in the component is calculated

$$\text{Density} = \frac{\sum \text{black pixels}}{\sum \text{nombre of pixels}} \qquad (4)$$

Second, the gray levels detected in each component. A simple neural network is used in this step to classify each component in order to extract the text areas that need an extra treatment.

The histogram of projection is applied in text areas, first the vertical projection to segment text line figure 5.

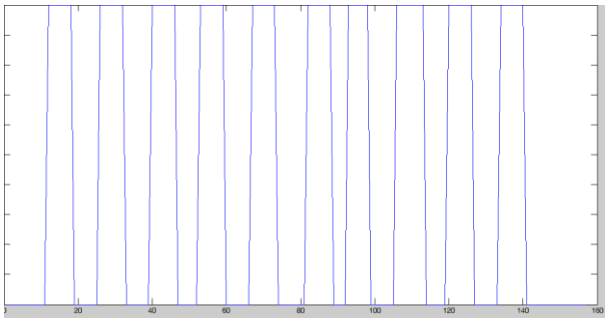

Fig 5: Vertical projection to segment text line

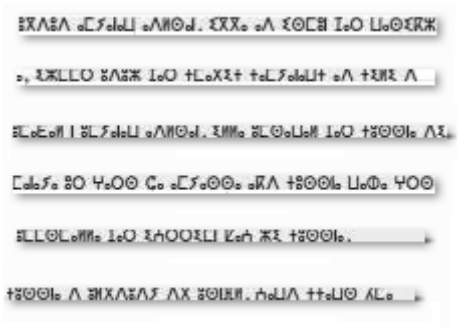

Fig 6: Example of text line segmented

The horizontal projection helps extracting each character individually (figure 7).

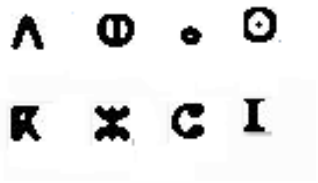

Fig 7: Example of characters segmented

The characters will be prepared for the recognition phase by normalizing its size to 20x20 pixels

3. CHARACTERS RECOGNITION

the algorithm of Zhang and Wang is adopted [12] due to its robustness and speed. This is a parallel algorithm in a Single iteration that produces perfectly skeletons 8-connected and which operates the collisions (figure .8). Then for each image obtained, we need to extract the simplest possible geometric elements of the graph which it is schematically represented by the primitive segments and key points.

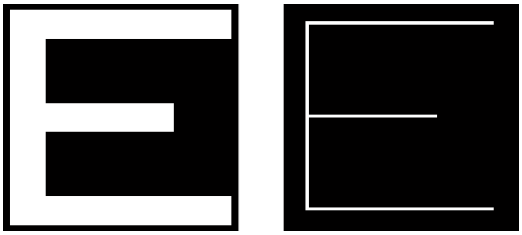

Fig 8: Skeletonized image character

For that we apply a translation, each pixel P of the character skeleton image will be translating in eight directions (figure 9).

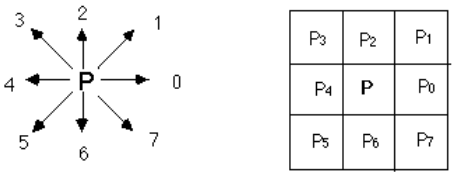

Fig 9: Eight direction of translation

The sum of the pixels of the images resulting from the transaction is placed in a matrix of zeros. The results are exploited to extract the key points , as follows:

- The pixels having a value of 1 represent the ends
- Values strictly upper to 2 represent the inflection and intersection points.
- While equal to 2 pixels represents the primitive segments

So the pixels with value equal to 1 or upper to 2 represents the key points that we need to represent the character (figure 10)

Fig 10: Extracting key points

A graph G is a pair G = (N, A), which N represents nodes or vertices and A represents the arcs or edges. The graphs of the most common structures are but not topologically complete mathematically, the model are a set of nodes in form of tree connected by distance between each pair of nodes and have a single parent nodes that is the origin of tree using this tree we can have the impact matrix [13] as follow:

An matrix M is a matrix of size n * m, such that

$$\begin{cases} n = |N| \; the \; number \; of \; nodes \\ m = |A| \; the \; number \; fo \; edges \end{cases} \quad (5)$$

An element $e_{i,j}$ (i= {1… n} and j= {1… m}) of the matrix M may have two values:

$$\begin{cases} e_{i,j} = 1 \; if \; arc \; j \; is \; ncident \; to \; node \; i \\ e_{i,j} = 0 \; if \; not \end{cases} \quad (6)$$

The recognition system consist on compared the impact matrix for each character.

Table 1. Recognition accuracy

	feature type	Key points
Adopted method	Training data	500
	Test data	1300
	Recognition rate	93,5
	Erreur rate	6,5

4. CHARACTERS RECOGNITION

In this part the final result of document structure recognition phase is used with the connected components [14], to extract more information about all zone detected in this phase which will be used to create the XML file.

For each component four values are extracted Imax, Imin, Jmax and Jmin (figure 11) that represent the position of each components in order to conserve the layout of document:

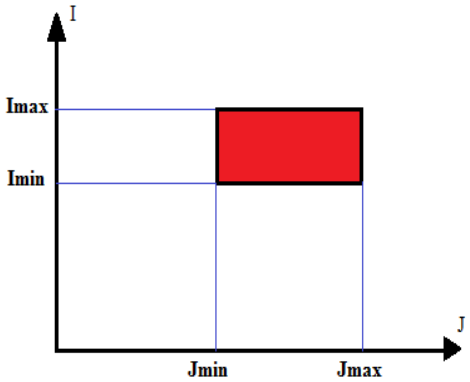

Fig 10: Determining the four values

Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable, with parameters that are extracted previously (Imin ,Jmin ,Imax and Jmax) as follows, for text the tag :

<Text>

<Imax></Imax>

<Jmax></Jmax>

<Imin></Imin>

<Jmin></Jmin>

</Text>

Non text blocks are presented using .

<Picture>

<URL></URL>

<Imax></Imax>

<Jmax></Jmax>

<Imin></Imin>

<Jmin></Jmin>

</Picture>

The XML file is only an intermediate passage between the image of document and the document. The choice of XML

language is not arbitrary but due to the tree structure that modeling the majority of problems and the universality and portability of this type of file also is usable by any application equipped with a parser (software to analyze XML).

5. CONCLUSION

This work has for aim to solve two problems, first the recognition of document structure the physical structure without analyzing the image pixel by pixel as the majority of methods used in document structure analyze. Secondly to we presented a new technique for Tifinagh character recognition. The results obtained are exploited to create an image converter into an intermediate file (XML) that can be used by different programming language and to improve the efficiency of the system the integration of new methods of characters recognition is needed, also a system for language detection is under construction

6. ACKNOWLEDGMENTS

The authors are grateful to the Associate Editor and the referee for their valuable comments and suggestions.

7. REFERENCES

[1] Jean-Luc Meunier 'Optimized XY-Cut for Determining a Page Reading Order'Xerox Research Centre Europe 6, chemin de Maupertuis F-38240 Meylan.

[2] L. Cinquea, S. Levialdia, L. Lombardib, S. Tanimotoc "Segmentation of page images having artifacts of photocopying and scanning" aDipartimento di Scienze dell'Informazione, University of Rome "La Sapienza", Rome.

[3] Jie Xi, Jianming Hu, Lide Wu "Page segmentation of Chinese newspapers" Department of Computer Science, Fudan University, Shanghai, Fudan, People's Republic of China 2004.

[4] l Ayachi, R., Fakir, M., & Bouikhalene, B. . Recognition of TIFINAGHE characters using a multilayer neural networks. [IJIP]. International Journal of Image Processing, 5(2), 2011.

[5] M.Fakir, O.Bencharef, B.Bouikhalene, B. Minaoui " Tifinagh Character Recognition Using Riemannian Metric, SVM & Neural Networks " International Journal of Advances in Science and Technology, Vol. 2, No. 6, 2011.

[6] A.Haidar, M.Fakir, O.Bencharef "Hybridation des modèles de Markov cachés et de la logique floue pour la reconnaissance des caractères Tifinagh manuscrits " 5ème conférence internationale sur les TIC pour l'amazighe 2012.

[7] B. El Kessab, C. Daoui , B. Bouikhalene, S. Gounane "Using of networks neurons and Markov model for the recognition of Tinifagh characters " SITACAM Agadir pp -31-40 2011.

[8] P .Bonnet « Cours de Traitement d'Image USTL » Université des Sciences et Technologies de Lille 2008-2009.

[9] M.Boutaounte, D.Naji, M. Fakir, B. Bouikhalene "Tifinaghe Document Converter" International Journal of Computer Vision and Image Processing, vol 3 issue 3, 2013.

[10] M.Hanmandlu, A. A. Fuzzy Model Based Recognition of Handwritten Hindi Numerals using bacterial foraging. 6th IEEE/ACIS ICIS.

[11] Fu Chang, Shih-Yu Chu, Chi-Yen Chen "Chinese document layout analysis using an adaptive regrouping" Institute of Information Science, Academia Sinica, 128 Academia Road, Taipei, 115 Taiwan, 2004.

[12] Thi, N., Oanh & Tabbone, S. . Binarisation d'images de documents graphiques. Technical report Laboratoire Lorrain de Recherche en Informatique et ses Applications à Université de Nancy. Bellman 2004.

[13] D. Arrivault, «Apport des Graphes dans la ReconnaissanceNon-Contrainte de Caractères Manuscrits Anciens», pp 61, 2002.

[14] Luigi Di Stefano, Andrea Bulgarelli "A Simple and Efficient Connected Components Labeling Algorithm" DEIS, University of Bologna Via Risorgimento 2, 40136 Bologna,Italy