Approach for Arabic handwritten image Processing: Case of Text Detection in Degraded Documents

Youssef Boulid
Department of Mathematics Faculty of Sciences
University Ibn Tofail, Kenitra, Morocco

Mohamed Youssfi Elkettani
Department of Mathematics Faculty of Sciences
University Ibn Tofail, Kenitra, Morocco

ABSTRACT
This study presents a new approach for processing of Arabic handwritten documents based on the extraction of characteristics and mechanisms involved in the process of human visual perception. The architecture which has been developed is based on the concept of multi-agent systems, allowing the integration of different stages of character recognition process in a cooperative way. This is illustrated using as example the prepossessing of binary noisy document. Therefore, a method was proposed, in order to distinguish between text and non-text components, using a new geometric primitives extracted from the analysis of the characteristics of Arabic script. Results show pixel-level precision and recall respectively of 98% and 93% for noise removal. This proves the effectiveness of the proposed approach in processing degraded documents and, consequently, improving the recognition performance.

General Terms
Arabic handwritten, Degraded binary documents, Text detection, Noise removal

Keywords
Stroke width, Intersections; Multi-agent systems, Distance transform

1. INTRODUCTION
Nowadays, Handwritten Character Recognition is of great importance; its offer the possibility to convert document images into text, allowing the indexation and search for information. There are several applications such as document digitization, storing, retrieving and indexing, automatic mail sorting, processing of bank checks and processing of forms. The importance of these applications has lead to intense research for several years.

The architecture of a handwritten character recognition engine often consists of five stages: pre-processing, segmentation, data representation, decision-making and post-processing. Most of the architectures are linear i.e. the phases are executed sequentially that may cause high amount of errors which accumulate from one stage to another.

The public datasets available for the learning/recognition stage often consist of clean images; therefore, building a system based on those datasets involves the preparation of the input images to have the same quality as those in the training set. For a large number of documents especially for historical documents, digitization has come too late and many documents suffer from physical degradation and are scanned in poor conditions. These degradations touch and overlap with text content resulting in changing the intensities, deleting parts of text or adding arbitrary patterns which strongly influence the recognition results. Thus, the preprocessing stage is the most critical since all the subsequent phases depend on it.

The degradations could be classified based on their nature:

- Dependent or Independent of location, size or other properties of text content.
- Regular or irregular in term of the consistency in properties like periodicity of occurrence in the document, its shape, position and grayscale.

For example complex background, salt-pepper and marginal noises are independent and irregular. By contrast, blur, ink bleed-through and ruled lines are dependent and regular [1].

Also the various types of noise could be classified according to their source:

Aging : The case of changing the color of the paper which introduces background noise and degrades the characters. Also, the fragility of the paper that introduces noise such as fractures, garbled characters, gaps in the lines.

Scanning : During this process, some types of noise could be introduced due to a low resolution, excessive compression or improper use of the scanner, such as skew, salt and pepper noise, shadowing, blurring, ink bleed-through, transformation of color, marginal noise.

Physical factors : Some of the physical factors that generate noise are : the carbon copy paper, the texture of the paper, the acidity of the ink, the humidity of the storage and the sunburn.

Factors related to document : The noise that is present in the content of the document such as signatures, underscores, figures, ruled-lines.

Haji et al [2] categorize types of noise in handwritten images firstly as a *Low-level noise* which is characterized by a random variation of intensity in document images that is produced by the hardware equipment during the scanning, transmission, storage or compression process, and secondly as a *High-level noise* that refers to anything other than text. They mentioned that the removal of high-level noise in handwritten images has been less studied compared with low-level noise, because of its application dependent nature.

As mentioned in the work of Agrawal and Doermann [3] there are two ways of cleaning document images: the first one is to detect and remove noise, and that is possible when the noise patterns have an independent set of features that could differentiate the noise from text content, and the second is to extract the text content leaving noise behind; therefore, the

contextual information and a priori knowledge about the text are required.

The paper is organized as follows : The second section invokes some related works. In section 3 and 4 we outline respectively the proposed method and the architecture for text detection and noise removal. This is followed by evaluation in section 5, and finally, a conclusion and discussion about the future work.

2. RELATED WORKS

Several methods have been proposed in the literature for noise suppression in the case of binary handwritten documents :

Haji et al [2] have proposed an unsupervised learning approach for noise removal from handwritten images. They formulate noise removal and recognition as a single optimization problem involving latent variables, where the optimization criterion is the recognition score for the input image after noise removal. The values of the latent variable are estimated using an Expectation Maximization Algorithm. Improving the initial guesses for latent variables lead to speed up the convergence time. So given the prior knowledge about noise is available, a method based on Fuzzy Logic is used to incorporate prior knowledge into the optimization process. For this purpose, two classes of noise patterns (the impulsive noise and background lines) were chosen due to their nature that can be easily described by linguistic rules. Then, a set of geometric proprieties are incorporated in the Fuzzy inference system.

The work of Agrawal and Doermann [3] treats of binary images with type of noises that are similar in the size to the diacritics (secondary component) and near the text component under the name of "stroke like pattern noise (SPN)". Based on the characteristics of the text, the components of the image are classified as Prominent Text Components (PTC) if they can be identified without using neighborhood context, and non Prominent Text Components (non-PTC) that require neighborhood context. After labeling the components into PTC and other (included non-PTC and noise), a set of textual features are extracted in order to be used in a supervised classification algorithm using SVM with Radial Basis Function (RBF) kernel. After classification, the results are sent to a second stage to filter out the non-PTC components, for that, the width of the strokes and the property of cohesion (using distance transform on the background) are computed for each component and sent to K-means clustering method, and, finally, a verification step is performed to liberate the text components that are misclassified.

An algorithm was introduced in [4] for detection and suppression of the clutter noise attached to text from binary document images. Based on the assumption that the clutter noise is bigger than twice text maximum stroke width, distance transform is applied both for detection and suppression of noise in an iterative way. In the first stage all the foreground pixels are thinned to half maximum distance, resulting in an image with only the clutter components. Then, a set of features is extracted from this image in order to use them with a Support Vector Machine (SVM) classifier to decide whether the image contains clutter components or not. The second phase is to remove the attached clutter noise while keeping the text components. The main idea is to regenerate only the clutter component from the half-residual core obtained in the previous stage. For this reason, the distance transform from clutter boundary is computed for each connected component with clutter noise. The number of steps required (pixels with the same distance) for the regeneration change rapidly at the edge of the component. This distance (number of steps) is used to remove all pixels under it, and thus results in removing only clutter pixels.

The paper in [5] presents a set of binary image enhancement algorithms for clutter, salt-and-pepper and rule-lines noise. The clutter noise is removed using a multi-resolution approach. The first step is to erase the text and leave the clutter noise; the second step is to set the pixels in the original image corresponding to the marked noise (clutter), which are assigned to white. To detect salt and pepper noise, an estimation of the local connected component density is done. Then, a mathematical morphological operator and size based filter are used to remove this noise. A refined adaptive vertical runlength search (thickness of the text) is designed for removing the rule-line pixels without affecting text pixels. If a vertical run is entirely covered within a re-constructed rule line, the run is removed from the original document image. If a vertical run is longer than the width of a rule-line, the run is retained.

The contribution in this paper is composed of two parts: the first concern is the development of a preprocessing method for handwritten text detection in degraded documents based on extraction of specific characteristics of Arabic script, and the second is the integration of the proposed method in a new global architecture for Arabic handwriting recognition in order to simulate the process of human reading.

3. PROPOSED METHOD

We assume the input images to be a binary images of full Arabic text handwritten pages.

Most of algorithms for noise removal in the state of the art deals with one type of noise. They try to detect and process each kind of noise independently from the others in a sequential manner, which would be difficult and expensive if we want to build a complete system for noise processing.

Given the variability of noise types found in historical manuscripts, it is appropriate to deal with this kind of documents by extracting the text content while ignoring the noise and that which is done during the reading process. We studied some characteristics of Arabic script and found that the notion of intersections in strokes allows to distinguish text and non-text components.

In fact, the Arabic script is cursive (the letters are connected), letters may consist of lines, curves and loops, and can be connected horizontally. These connected points often represent intersections.

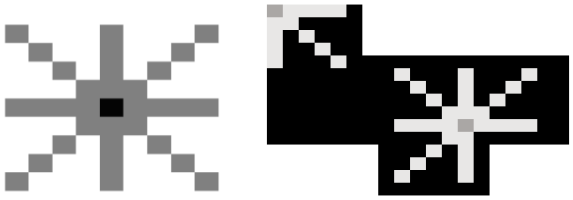
For instance, intersections in a written word are usually located on the baseline (a virtual line on which semi cursive or cursive text are aligned/joined [6]), and are dispersed in the form of group of pixels and represent a lower portion compared to the area of the word. By contrast, in arbitrary noisy patterns the amount of intersections is expected to be much higher.

In the following lines, the notion of intersections and how it is calculated, the proposed algorithms for text stroke extraction, text stroke width estimation, text height and width estimation are described.

3.1 Notion of intersection

For each black pixel in a Connected Component (CC), continuous elongations in the 8-directions from its neighbors are detected, i.e. window of fixed height and width (Fig 1-a), and the number of these elongations, if they exist, are

counted. The maximum number would be 8 if the pixel is located in the middle of a filled window.



(a) **(b)**

Fig 1 : Notion of intersection. (a) The 8-directions in the neighbourhood of the centred black pixel, represent the elongations, (b) for the pixel on the top left the number of elongations is equal to 3, while for the pixel in the middle, the number of elongations is equal to 7

One of the characteristics of the Arabic script is that the text strokes contain lower amount of intersections. So as to count the number of intersections in text, the height of the elongations is estimated as the average text stroke width. We do not know the actual width of the text stroke but rather we estimate it.

Therefore, there are three types of pixels:

- Pixels having zero elongations, which are called isolated pixels (correspond to diacritics or salt and pepper noise).
- Pixels having one or two elongations, which represent text pixels.
- Pixels having more than two elongations, which correspond to intersections pixels.

Finally counting the percentage of each type of pixels in the CC in order to decide whether this is a noisy component or not (Fig 2).

(a)

(b)

Fig 2 : Locations of intersections (gray dots). The percentage of intersections in text stroke (a) is less than that in the noisy pattern (b), 16% of intersections against 60% respectively for text and noise

3.2 Text extraction

To efficiently calculate the average text stroke width (thickness) and the average text height and width, the text components in the document have to be located. To achieve this, "K-means" method [7] is used and the diameter of a circle with the same area as the region of the CC is taken as criterion. The number of classes is chosen to be equal to three, which allows to group the components in three classes: class having diacritics, class containing text and class representing large components.

$$Diameter = \sqrt{\frac{4 * area}{\pi}} \qquad (1)$$

To locate the class having text component. For each class, the sum of the areas of all the CCs is counted and divide by the sum of their diameters. Several tests for different images have led to the conclusion that the ratio of the text class corresponds to the median value among the three values.

3.3 Average text stroke width

To estimate the stroke width of a CC, the Euclidean Distance Transform [8] is used (a number assigned to the foreground pixel, which is the distance between this pixel and its nearest nonzero pixel, see Fig 3-b).

The Distance Transform assigns the shortest distance to every pixel of a set P from other set I as follows:

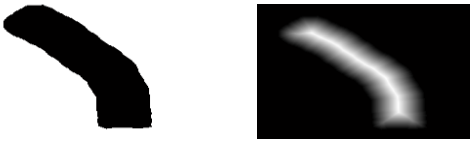
$$Dist_p(X) = \min_{Y \in I}(d(X,Y)) \qquad (2)$$

Where d is the Euclidean distance defined as :

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \qquad (3)$$

From the distance transform, the pixels having maximum intensities are located using regional maxima [9] (Fig 3-c). To explain more, pixels located are those found at a half distance from the edge of the stroke. Furthermore, a list with the values from these locations is constructed, and finally the average of these distances is multiplied by two.

$$S_w = \frac{\sum \max(Dist_{fg})}{n} * 2 \qquad (4)$$



(a) **(b)**

(c)

Fig 3 : Illustration for calculation of the stroke width, (a) the Connected Component, (b) the Distance Transform image, (c) the result of Regional Maxima. The average stroke width is equal to 54 pixels and the size of the image is 225x225

Finally the average text stroke width is equal to this of the stroke width of all the components in the text class.

3.4 Average text height and width

To efficiently decompose the document image into windows, the height and width of the window that may include maximum text components have to be taken into consideration. For this purpose, the average height and width of the bounding box (the smallest rectangle containing the region) from all the CCs in the text class are extracted. Finally, the height and the width of the window are estimated as three times the average height and width of the text.

4. PROPOSED ARCHITECTURE

In a system of handwriting recognition, there are several tools that interfere in different stages according to its objectives, such us: methods and tools for noise removal, skew and slant correction, binarization, feature extraction, learning, classification, lexical verification, segmentation into lines, words, characters and so forth.

Often such tools and methods are executed in a sequential manner. If a step is not properly performed, it may cause errors that accumulate from one stage to another, and therefore, influences the final recognition result.

The objective here is to simulate the process of human reading by extracting the characteristics and mechanisms involved in this process of reading. For this, an architecture that allows to structure and integrate methods in a scalable manner and ensure the cooperation between the different phases of the recognition process is needed.

Stanislas Dehaene [10] claims that in the case of reading, three sets of brain circuits are brought into play: invariant visual recognition networks involved to identify the writing followed by the conversion of the written characters to a phonological representation and in parallel access to the lexical and to the meaning of words and sentences.

As reported in [11] we have several strategies to address the document, such as the full reading (reading word after word), the inspection (the search in a specific region of the document) and, finally, the overview (preview the document).

The first observations and facts made about the nature of human visual exploration lead to reconsider the importance of the architecture of a character recognition system.

4.1 Fundamental Assumptions

Following the analysis on the behaviours of human beings in the search of information in the case of degraded document, some assumptions underlying the proposed approach are retained:

- To detect the handwritten script, the reader, unintentionally, detects the geometrical characteristics of the text objects, such as the homogeneity of the strokes in terms of thickness, lengths, shapes, dispersions and so forth.
- The reader does not look for a specific noise to remove. He ignores the type and the nature of the noise and look for the text in the image and within the noise.
- When observing a document for the first time, the reader does not read the text line after line sequentially. Consequentially he can skip paragraphs and analyze regions in any position of the document. By doing this, he learns the style of the writing.
- Facing a very distorted word, the reader is not able to fully distinguish this word until more information is collected during the reading of other parts of handwritten document [12].

4.2 Assumptions exploitation

Here are some hypotheses drawn from the above assumptions:

- Often the writer uses the same pen to write.
- Arabic text strokes tend to have a constant width.
- Dealing with noise is the same as dealing with all non-text components.
- When reading a document, readers can skip lines to see other areas while looking in every direction, which means that they have a complete visibility of the document. This process gives them information to distinguish between text and noise, and the possibility of extracting the text from the noise. Also, the previous process allow them to know if the document is written by more than one writer.
- Readers can learn from the document itself and use the context and the relationship between words in order to conclude the right meaning.

To be able to simulate the process of human reading, an architecture is built on the concept of multi-agent system so as to easily implement algorithms based on the hypotheses listed above. Each agent is responsible for a region of the document. This allows to treat several parts of the document simultaneously and, therefore, learn from the same document since agents can communicate and exchange information with each other at different contextual levels (graphic, morphological, syntactic, lexical and semantic).

Information exchange between agents is done in a centralized manner through a central agent which has a total visibility of the document image (Fig 4).

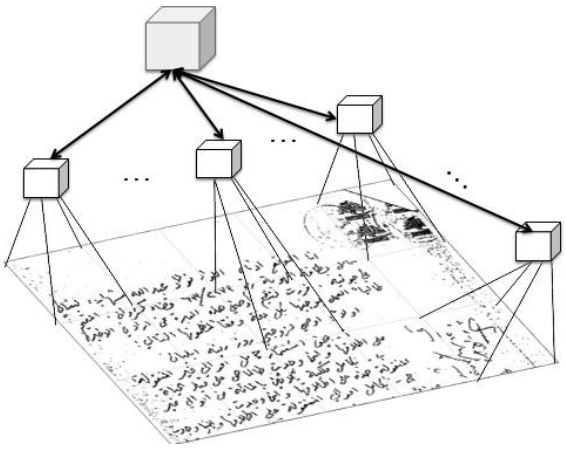


Fig 4 : Architecture of the system. Each agent (white cube) is responsible of a window thus having a local visibility, the communication between objects pass through the central agent (gray cube)

Treatments related to the extraction of information or execution of actions are delegated from the central agent to the simple agents in order to be executed simultaneously.

Each agent reacts in its context (a specific region of the image) and contains a set of functions with the possibility to do and undo some actions according to evaluation metrics. it can store different states of its region in order to be able to switch between them, i.e. switch from the binary image to a greyscale image in order to extract other information, and ensure maximum quality.

In the case of the presence of noise that degrades the quality of a word or character, agents can exchange information on the normal form of the characters. In order to look for the resemblance with degraded characters, these agents also can identify if there is more than one writer and ,also, objects that cannot be produced by a writer (do not respect the standards of the Arabic script). This process can be seen as a kind of supervised learning on the same document.

In the following, how to integrate the proposed method for text detection and noise removal in this architecture will be shown.

4.3 Text detection and noise removal

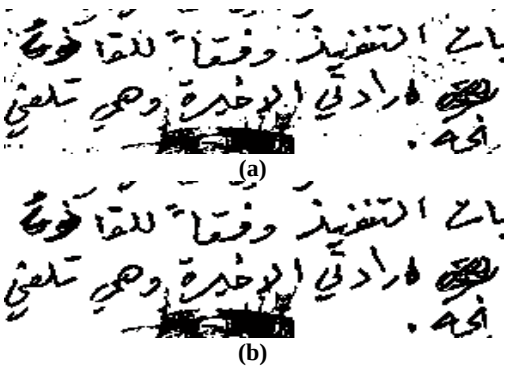
To imitate the overview of the document in order to look for text regions, the central agent concludes the average text height and width and the average text stroke width from the binary image and use them to decompose it into windows of fixed widths and heights and links agent to each region.

The first thing to do is to classify the content of the regions based on geometrical features, such as the percentage of intersections.

Three types of components based on the nature of intersections discussed above are distinguished:

- Text components having a percentage of text pixels exceed 50%.
- Diacritics components having a percentage of isolated pixels exceed 50%.
- Noisy components having a percentage of intersections pixels exceed 50%.

Before starting the classification of the components, the text stroke width information is used to remove the isolated components representing salt and pepper noise without touching the diacritics. Therefore, all the CC having areas less than three times of the text stroke width are suppressed. The idea behind this is that a perfect diacritic will be a filled circle with diameter equal to text stroke width, and thus having an area greater or equal to four time text stroke width (Fig 5).



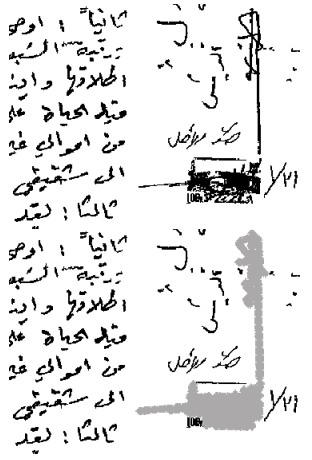
(a)

(b)

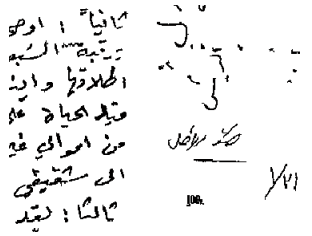
Fig 5 : Salt and pepper noise removal, (a) the original image, (b) the resulting image after noise removal

During the classification a case where some misclassified components are encountered as text, but in reality represent noise, and this is, mainly, due to the fact that they are attached to the text components. So when counting the number of intersections, the number of text pixels dominates the number of intersections pixels. To solve this problem another detail to distinguish this type of noise is added in order to calculate the number of intersections of all the components and extract their maximum, which represent the maximum number of intersections in a word.

Thus, the misclassified text components are filtered out when they have a number of intersections greater than the maximum number of intersections that can be found in a word (Fig 6).



(a) **(b)**



(c)

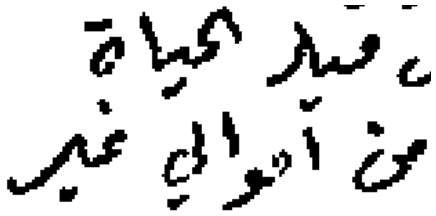
Fig 6 : Noisy components suppression, (a) the original image, (b) detection of noisy components (filled in gray), (c) the output image after noise removal

Another type of noise that may occur in the binary handwritten document is the stroke like pattern noise [3]. This latter looks like diacritics and dots (secondary components). This kind of noise can be suppressed using prior knowledge about their distances and positions.

In Arabic script diacritics are either located on above or below the baseline of the word. A smooth shift of their positions and distances to tolerate different handwriting styles. So, in order to distinguish between the real diacritics and the SPN, distances and positions from text components are used.

For each diacritic resulted from the classification done before (isolated pixels), the maximum between the height and the width is taken as the max-length. Then a vertical rectangle is drawn, which is centered on the diacritic and having width equal to twice the max-length, and height equal to six times the max-length. Then the content of that rectangle is verified to see if it contains portions of text components. If so, it is classified as diacritic belonging to the text. In the opposite case, it is removed (Fig 7).

(a)



(b)

Fig 7 : Extraction of diacritics, (a) the original image, the rectangles filled in gray represent real dots and diacritics, the empty rectangles represent the SPN to be removed, (b) the resulting image after noise removal

All these methods are incorporated in each agent. The central agent is the one which gives the orders to the other agents to execute each algorithm simultaneously.

5. EXPERIMENTAL RESULTS

Several tests, using different kinds of noisy handwritten Arabic binary documents are conducted. to illustrate this, here below are the images of the testament of the president Fouad Chehab[1] used for the demonstration and evaluation of the same image degraded manually.

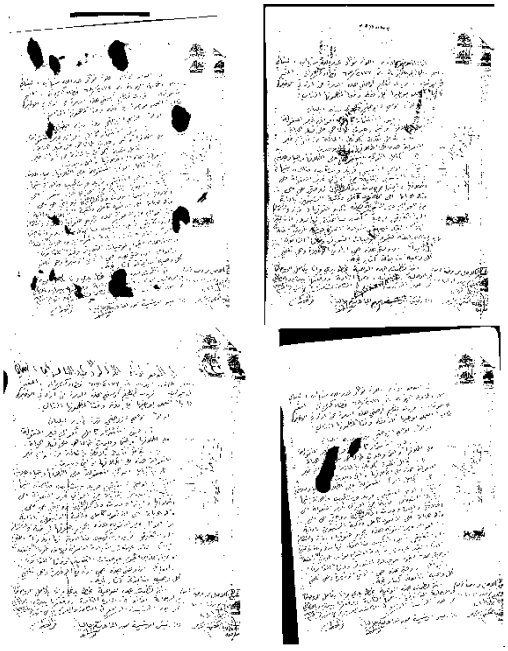


Fig 8 : Examples of noisy handwritten documents

[1] http://www.fouadchehab.com

The problem encountered is the lack of a common datasets to compare results between them. In each work found in the literature, the authors build their own data in order to evaluate the algorithms proposed. Such datasets containing noisy images of handwritten documents with their ground-truth are needed in order to further this field.

Here the ground-truth are done manually to remove the noisy pixels. Pixel-based evaluations are performed in order to assess the accuracy of the approach. Therefore, precision, recall, and deleted text are calculated using the following metrics :

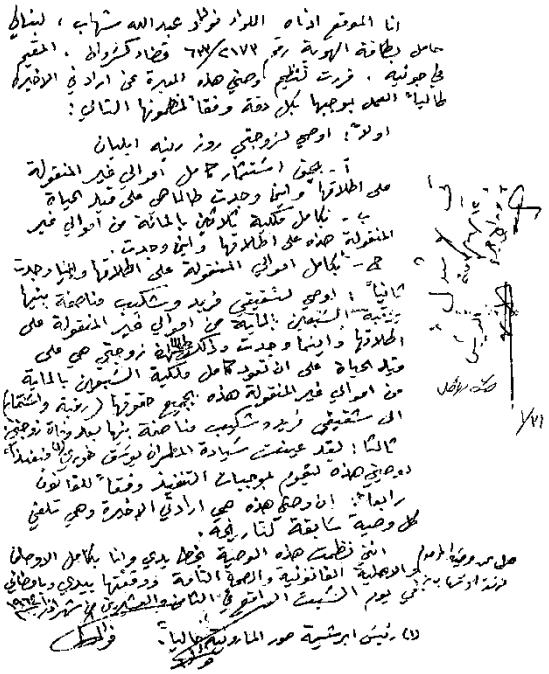
$$\text{Precision} = \frac{\text{Number of noise pixels removed}}{\text{Total pixels removed}}$$

$$\text{Recall} = \frac{\text{Number of noise pixels removed}}{\text{Total noise pixels}}$$

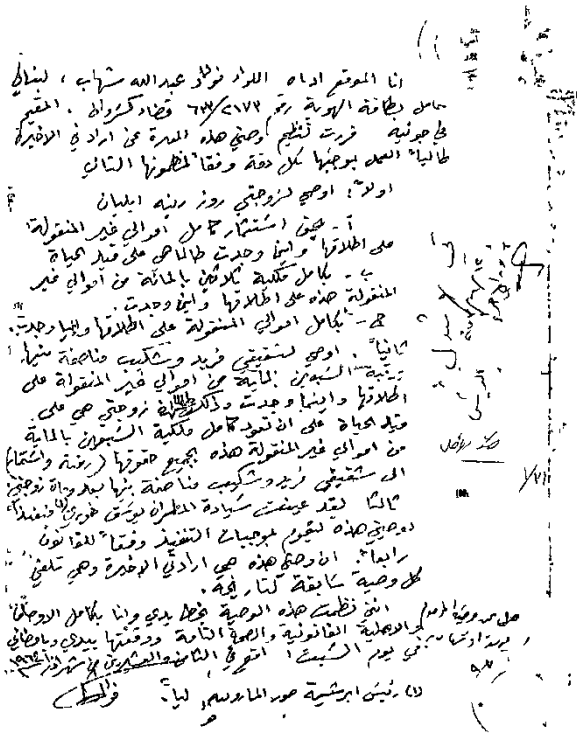
$$\text{Deleted text} = \frac{\text{1 - Precision}}{\text{Total text pixels}}$$

Precision, recall and deleted text accuracy are of 98%, 93% and 0.5% respectively. The result of the sample document is shown in (Fig 9). Comparative evaluation with existing methods of common dataset will be part of our future work.

(a)



(b)



(c)

Fig 9 : Result of the text detection and noise removal algorithm, (a) the original noisy image, (b) the ground-truth image, (c) the resulting image

6. CONCLUSION

In this paper, an architecture was proposed for binary handwritten document processing, especially, for text detection and noise removal by analyzing the characteristics of the Arabic script, intersections found in this work give us a set of good features to distinguish between text and noise.

The method suggested here differs from previous approaches in that it does not look for a type and nature of noise to remove, but detect and extract text regardless of its direction from degraded document. Also it does not require a training stage to estimate its parameters. It means that it extracts the features and learn their parameters directly from the document itself.

The advantage of this approach compared to traditional ones is that it could integrate all stages of the characters recognition process in a cooperative way to overcome the problems related to the loss of information in sequential approach. More explicitly, instead of executing the methods in a sequential manner on all the image, agents treat each area of the document and cooperate to improve the recognition rate.

To imitate the mechanisms involved in the reading process, the proposed architecture allows us to improve the simple agents making them to interact and cooperate in an autonomous way at different contextual levels (graphical, morphological, syntactical, lexical, and even semantically) to maximize the recognition accuracy.

In the future improvement of the system by adding more features, such as the analysis of dispersion of the intersections is expected as well as completion of the proposed architecture by allowing the agents to integrate other stages of characters recognition process, such as segmentation and recognition.

7. REFERENCES

[1] Farahmand, A., Sarrafzadeh, A., & Shanbehzadeh, J. (2013). Document Image Noises and Removal Methods. In Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 1)

[2] Haji, M., Bui, T. D., & Suen, C. Y. (2012). Removal of noise patterns in handwritten images using expectation maximization and fuzzy inference systems. Pattern Recognition, 45(12), 4237-4249

[3] Agrawal, M., & Doermann, D. (2011, September). Stroke-like pattern noise removal in binary document images. In Document Analysis and Recognition (ICDAR), 2011 International Conference on (pp. 17-21). IEEE

[4] Agrawal, M., & Doermann, D. (2013). Clutter noise removal in binary document images. International Journal on Document Analysis and Recognition (IJDAR),16(4), 351-369

[5] Shi, Z., Setlur, S., & Govindaraju, V. (2011, September). Image enhancement for degraded binary document images. In Document Analysis and Recognition (ICDAR), 2011 International Conference on (pp. 895-899). IEEE

[6] Bahaghighat, M. K., & Mohammadi, J. (2012). Novel Approach for Baseline Detection and Text Line Segmentation. International Journal of Computer Applications, 51(2)

[7] MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations.

In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297)

[8] G. Borgefors. Distance transformations in digital images. Comput. Vision Graph. Image Process., 34(3):344–371,1986

[9] Vincent, L. (1993). Morphological grayscale reconstruction in image analysis: applications and efficient algorithms. Image Processing, IEEE Transactions on, 2(2), 176-201

[10] Dehaene, S. (2007). Neurones de la lecture (Les): La nouvelle science de la lecture et de son apprentissage. Odile Jacob

[11] Eglin, V., Bres, S., & Emptoz, H. (1999). Structuration de documents par repérage de zones d'intérêt. TS. Traitement du signal, 16(3), 217-239.

[12] Heutte, L., Nosary, A., & Paquet, T. (2004). A multiple agent architecture for handwritten text recognition. Pattern Recognition, 37(4), 665-674