



CLASSIFICATION AND DIGITAL RESTORATION OF DAMAGED ARABIC MANUSCRIPTS

Fawzy I. Elrefai^{1*}, Ali A. Halwa¹, Al Amira A. Hassan², Hany Gadelrab¹

¹ System & Computers Eng., Faculty of Engineering, Al-Azhar University, Cairo, Egypt.

² Conservator at Historic Cairo Administration, Ministry of Tourism and Antiquities, Cairo, Egypt.

*Corresponding Author E-mail: al_ame_ra@yahoo.com

Received: 21 June. 2021 Accepted: 29 Jan. 2022

ABSTRACT

This paper proposes a novel and effective approach to classify the damage in Ancient Arabic manuscripts through Damage Manuscripts Classification (DMC) Model; into two types, each type has spatial model. The first type is Fading Text Color (FTC) Model, and the second is Missing Part of Text (MPT) Model. For the first type, which is the Fading text color, it is done through (FTC) Model, where segmentation, contour strength and contour size algorithms were applied. As for the second type, we applied some segmentation algorithms to separate image of damaged manuscripts into foreground and background. Segmentation by thresholding applied on background to detect if there are missing in text to complete it based on database which we had prepared for this purpose including the handwritten Arabic fonts and the forms of letters in different styles such as Naskh , Reqaa,..etc. detection of style algorithm is also used to determine the style of missing text according to the same database , then digital restoration applied to the image. Applying Pre-processes on the used data yields good classifications' results. The contribution of this work is the introduction of synthetic features that enhance the classification performance.

For testing purposes, two famous books from the Islamic literature are used: 1) pages of the *Ottoman Quran* and 2) Some *Quran verses in Naskh script*.

KEYWORDS: Damaged Arabic Manuscripts, Classification, Foreground Separation, Machine Learning, Threshold Algorithm, and Digital Restoration, Missing Text, Color Fading, and Classify Damage in Manuscripts.

تصنيف نوع التلف بالمخطوطات العربية وكيفية عمل الترميم الإلكتروني لها

فوزي ابراهيم الرفاعي^١ ، علي عبد الرؤوف حلاوة^١ ، الأميرة أحمد حسان^٢ هاني جاد الرب السيد^١

^١ قسم النظم والحاسبات , كلية الهندسة , جامعة الأزهر , القاهرة , مصر

^٢ الإدارة العامة للقاهرة التاريخية , وزارة السياحة والآثار , القاهرة , مصر

* البريد الإلكتروني للمؤلف الرئيسي: al_ame_ra@yahoo.com

الملخص

تهدف هذه الدراسة إلى محتوى المخطوط العربي وكيفية الحفاظ عليه من خلال صيانتها عن العوامل التي يتعرض لها وقد تؤدي إلى إتلاف جزء منه أو الحيلولة دون الإستفاده القصوي من محتوى تلك المخطوطات وما تشتمل عليه من كنوز علمية وتاريخية ودينية لا تقدر بثمن. ومن هنا وجب علينا الربط بين العلوم الحديثة كمعالجة الصور وتعلم الآلة (Machine learning) في ترميم ومعالجة مظاهر التلف في المخطوط والتي تنتج عن التقادم الزمني أو الاستخدام الغير صحيح وأحيانا نتيجة للتداول بين المستخدمين. وقدمت هذه الدراسة نموذجا لتصنيف نوع التلف بالمخطوط العربي أوتوماتيكيا ودون الحاجة للعامل البشري ومن ثم فقد شملت تلك الدراسة أيضا علي كيفية جديدة للترميم الإلكتروني لأكثر أنواع التلف انتشارا في المخطوطات العربية.

الكلمات المفتاحية : المخطوطات العربية القديمة, نوع التلف في المخطوط, معالجة وتحسين الصور, فصل طبقات الصورة, الترميم الإلكتروني, بهتان لون النص بالمخطوطات العربية, فقد جزء من النص.

1. INTRODUCTION

Arabic Manuscripts are valuable sources of information, great evidence of our existence and our culture. Manuscripts have several different kinds of value. Some of them are valued as artifacts or objects of art other manuscripts are valued because of their association with a famous person. Some types of manuscripts naturally yield information more than other yields, although some of the division's manuscripts have art factual and associational value, most are collected for their informational or evidentiary value. Arabic manuscripts considered primary sources, often-unique ones, upon which the writing of history may be based on. They provide evidence of human activity. Often they suffer from degradation problems. Here the role of restoration process comes up to deal with degradation problems. Experts and specialists in the restoration and maintenance of manuscripts monitor the manifestations and factors of damage in the manuscript, and according to the type of damage [1], It been treated by the methods and means of careful restoration according to the original. **Fig. 1** shows some signs of damage in manuscript and the manual process of restoration.

With the huge progress in the science of image processing and the development of new methods for dealing with images accurately, it was necessary to turn to digital restoration. The conservation and restoration of books, manuscripts, and documents is an activity dedicated to the preservation and protection of items of historical and personal value made primarily from paper, parchment, and leather. When applied to cultural heritage, conservation activities been undertaken by a conservator. The primary goal of conservation is to preserve the lifespan of the object as well as maintaining its integrity by keeping all additions reversible. Conservation of books and paper involves techniques of bookbinding, restoration, paper chemistry, and other material technologies including preservation and archival techniques [2]. Digital manuscripts restoration is a process that allow manuscripts viewed with their original quality without further compromising the condition of the physical material. the user in general will able to deal with the manuscript's image and perform all restoration operations for it and return it to the original again, and then the researcher will be able to deal with it and make maximum use of it without compromising the original .and until manuscripts is preserved from the damaged handling factors. The process of restoring a manuscript digitally includes a range of variables dependent on the condition of the manuscript and the desired final image quality. Looking at the significance of the Arabic manuscript and what it represents for us in terms of Arab culture and religious identity, it becomes clear to us the importance of research in the field of Arabic manuscript science and methods of preservation and linking these sciences to the field of image processing. In addition, the importance of this research stems from the temporal depth of the Arabic manuscripts, and the spatial dimension caused by the link between the Arabic language and the holy Qur'an accompanies it. Therefore, it spread among all the peoples that embraced Islam, another dimension is added to the temporal and spatial dimension, which is the civilizational dimension, and these three dimensions made the Arabic manuscript heritage longer

in life, larger in number, more diverse, stronger in spread and more authentic than the written heritage of any other nation. **Fig. 2** shows a copy of handwritten Qur'an before and after manual restoration.

2. RELATED WORK

A great body of research has discussed the damage and digital restoration of ancient manuscripts. Siddharth et. al [3] has focused Image processing techniques and applied them for digital restoration of Indian ancient manuscript. Ventzas et. al [4] worked on de noising and Binarization to introduce an innovative sequential procedure for digital image acquisition of historical documents including image preparation, image type classification according to their condition and their spatial structure, global and local features or both, including document image data mining. Image segmentation for object detection in indoor and outdoor environment had produced by Suganya and Latha[5] where these techniques is used to separate foreground image from background especially in dynamic environment. Keith Knox and William Christens-Barry in [6] use Modern imaging techniques have been applied to ancient manuscripts to recover writings that are not visible to the naked eye. Many algorithms were presented in [7] in the field of separation of foreground and background are based on the classical Shannon definition of entropy and a generalization defined as Tallies Entropy, It achieved good results by analyzing precision, recall, and accuracy.

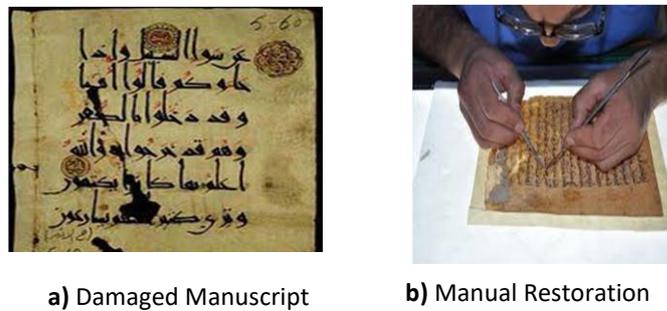


Fig. 1 Damage Manuscript and, Restoration Process

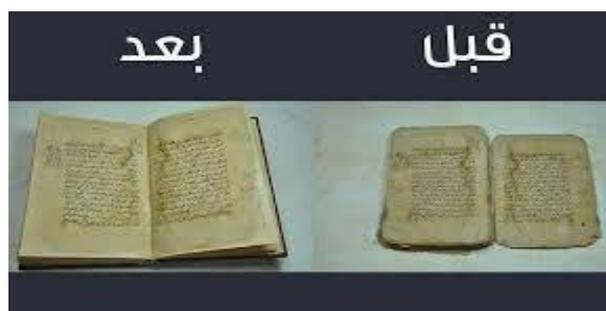


Fig. 2 : (**Right**) Damaged Manuscript, (**Left**) Restored Manuscript.

3. PROPOSED METHOD FOR CLASSIFICATION AND DIGITAL RESTORATION OF DAMAGED ARABIC MANUSCRIPTS

By studying in both of restoration and computer science we find it is amazing idea if we merge both of them to serve our heritage so we develop our completely new model to classify damaged manuscripts automatically to be ready to the next stage of digital restoration. We mentioned the importance of image

processing and machine learning in the process of digital restoration in a highly accurate manner, This, inspires us to introduce a novel model; **Fig. 3** focuses on the most common, present manifestations of damage in the Arabic manuscript. manifestations of damage may be presented in fading text color, which is produced because of light affection, all light, whether natural or artificial, causes damage, and its effects are cumulative, Ultraviolet radiation causes bleaching, discoloration and breakdown of organic materials, and missing part of text in ancient Arabic manuscripts is resulted from bad storage and handling [8].

At first, we make classification of damaged manuscripts according to the type of damage, and it had done through the first model: Damaged Manuscripts Classification (DMC) model. There are two classes of damage in manuscripts, each one has calcification model. The first one is Fading Text Color (FTC) model, and the second one is Missing Part of Text (MPT) model.

3.1. Damage Manuscripts Classification (DMC) Model includes important stages :

3.1.1. Preprocessing images of damaged manuscripts

Preprocessing starts after Data capture that is carried out by optically scanning a manuscript. The resulting data is stored in an image file, those images undergo Image enhancement processes, Objective of enhancement is to process an image so that the result is more suitable than the original image for a specific application such as Highlighting interesting detail in images, removing noise from images and Making images more visually appealing [9]. Referring to old Arabic manuscripts we find different types of materials that were used for writing such as paper, cloth or leather, and other materials with different colors so, a lot of manuscripts have dark background , changing of dark image into Negative images to make the details that embedded in dark regions of an image more appeared. Enhancement techniques helped more in these cases. The histogram of an image that shows us the distribution of grey levels in the image and gives assign to know the best group of pre-processing steps that are suitable with each image is very affective. Almost the most important step is Binarization process that convert the image format from the grayscale to binary: values of background pixels as one (white) and values of foreground pixels as zero (black). This process is carried out by choosing an efficient thresholding method value; this process increases the processing speed [10]. Binarization process is either global or local; In a global approach, threshold selection leads to a single threshold value for the entire image often based on an estimation of the background level from the intensity histogram of the image. Local Binarization process use different values for each pixel according to the local area information [11] we found global Binarization is more efficient in the case of handwritten Arabic text. **Fig. 4** States previous preprocesses applied on real Arabic manuscripts.

The last step in the preprocessing stage is filtering and smoothing to remove noise may appear in the images after scanning so, it is necessary to remove the noise and smoothing the input text image to prepare the data for the further processing. Noise in our used data has special nature; it appears usually on form of isolated dots that refers to the tiny drops of ink that falling of the pen during handwritten operation. By applying more than one filter on data the best result with Median filter using 3x3 window **Fig. 5** Median filter computes the median of all the pixels under the kernel window and the central pixel is replaced with this median value [12]. By applying the Damaged Manuscripts Classification (DMC) model on the used data without applying the previous steps of pre-processing, we find the results go to be unsatisfied as shown in **Table 1**.

3.1.2. Separation techniques of foreground and background

After pre-processing of the images Separation be applied to separate entered image to two images of foreground and background. There are many tools and algorithms allow you to separate foreground and background from the image. Before starting the separation process, we extracted the text from the manuscript by using divide image into blocks method [13] where The Gaussian filter is used and the image is divide into blocks, which are then binarized using an adaptive threshold. Then, the blocks are sequential to get the path of text lines, and finally lines of text are extracted, the process of extracting the text is carried out in many ways, and the most successful with our proposed model in order from the highest results to the lowest are:

- Divide image into blocks: using The Gaussian filtre to divide the image of manuscript into blocks, the blocks are sequential to get the path of text lines, and finally text lines are extracted by thinning the background of the path image.
- Projection Profile: The contour is used to extract the base line of the letter, which allows location the cut point between different adjacent lines.
- Projection profile and K-means: The image of manuscript is split vertically into several strips, and text is detected based on the histogram of the Projection Profile, Text blocks are clustered using K-means.

Saving Extracted lines of text as image (**img1**), that undergoes FTC model to apply morphological operation such as erosion or dilation, then treat any damage in some parts of the text; in the same time detect the style of the text and save in the Database of handwritten Arabic styles. Saving Background as image (**img2**). object detection algorithm is applied to detect the cuts' place (window) and save its dimensions to determine where the missing text must be locate, then image be classified as damaged image through next two stages.

3.1.3. Negative Image

A negative of an image is an image where its lightest areas appear as darkest and the darkest areas appear as lightest. it changes the black pixels of damage signs into white pixels with intensity greater than zero to make it easy to classify.

3.1.4. Image thresholding

Image thresholding is simple form of image segmentation that create a binary image from a grayscale. This is typically done in order to separate object "cut/damage place" or foreground pixels from background pixels to aid in image processing. if image has not any damage, its pixels go to be black pixels with intensity equal to zero .but in case of existence of damage white pixels will appears in the negative image and intensity is greater than zero, then this image is classified as damaged one. After detection the damage, images undergo MPT model to be restore and saved in (**img2**). At the end addition (**img1**), and (**img2**) to get the desired restored image of ancient Arabic manuscript **Fig. 6**.

Algorithm 1 classify entered images of both signs of damage in Arabic manuscripts therefore, the output of the classification phase is either Missing part of text (MPT) or Fading Text color (FTC) **Fig. 7** addressed all of previous process applied on damaged Arabic manuscript the classification of damage in it.

```

Algorithm 1: Classification Algorithm
INPUT: <images> where images ∈ {'MPT', 'FTC'}.
OUTPUT: A specific image with MPT or FTC.
For  $i = 1$  to  $n$  # $\forall$  image  $i \in$  used data
1 |   Resize  $i$ 
2 |   Binaries  $i$ 
3 |   Apply noise remove filter
4 |   Apply separation algorithm.
5 |   IF threshold of negative = ZERO
6 |       Save image as FTC
7 |   Else
8 |       Save image as MPT
9 |   END IF
END For
    
```

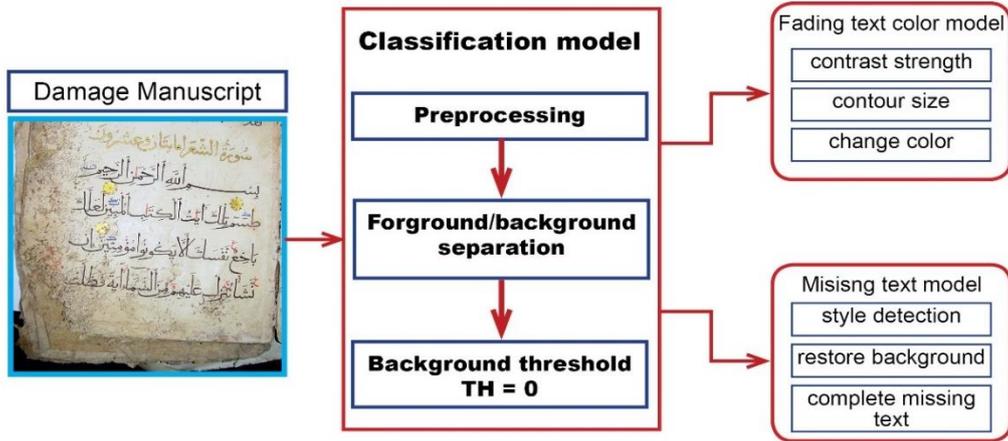


Fig. 3 Proposed Method for Classification

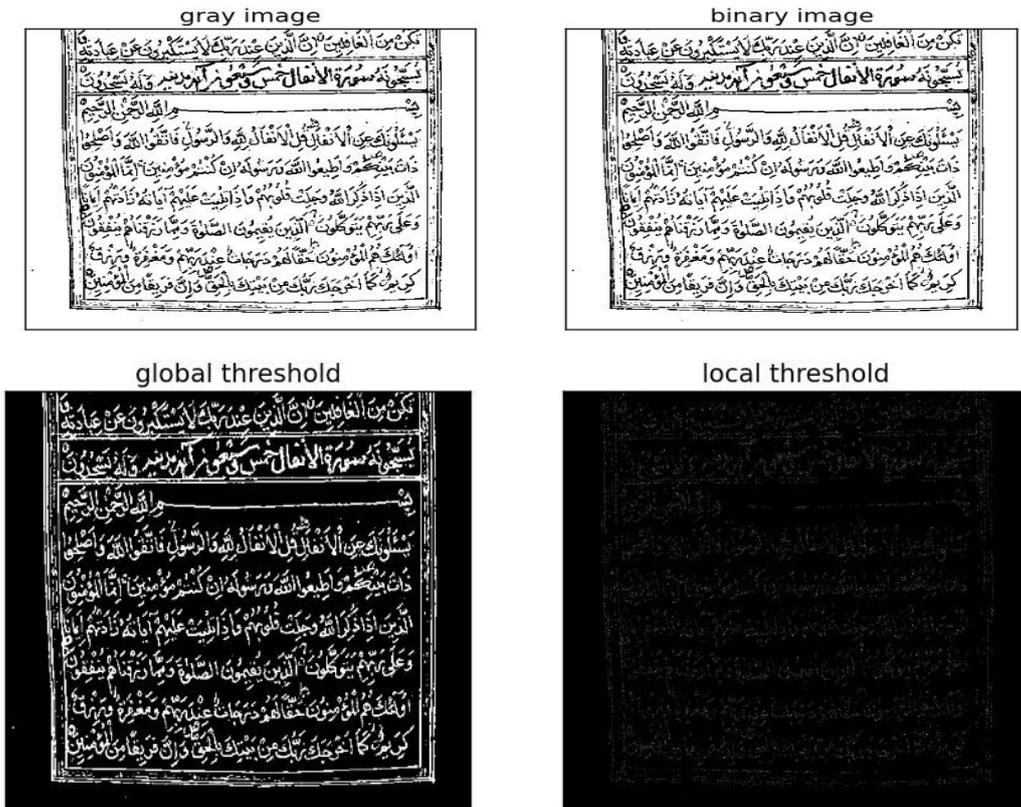


Fig. 4 Results of Pre-processes

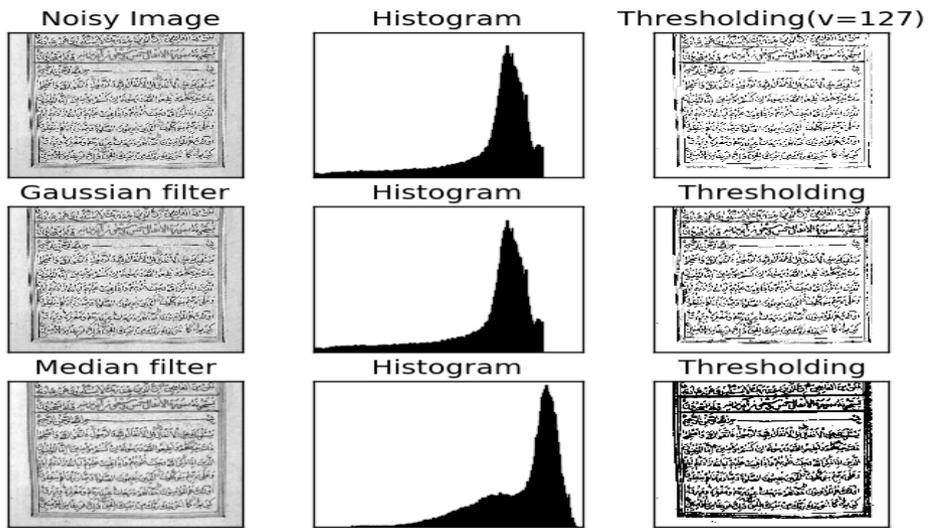


Fig. 5 Results of different filters

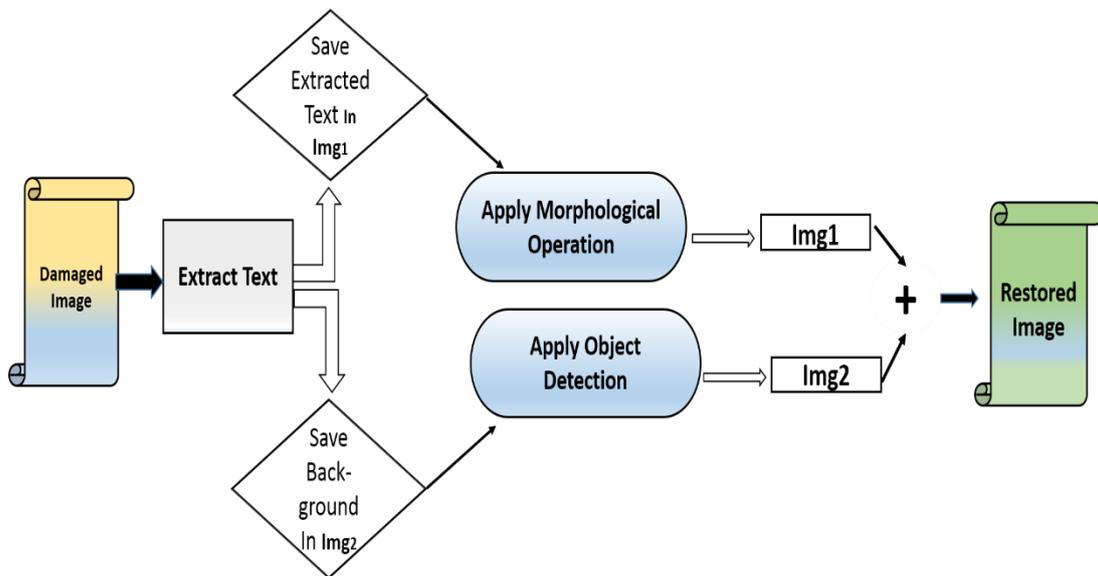


Fig. 6 Addition the Output of MPT Model & and FTC Model.

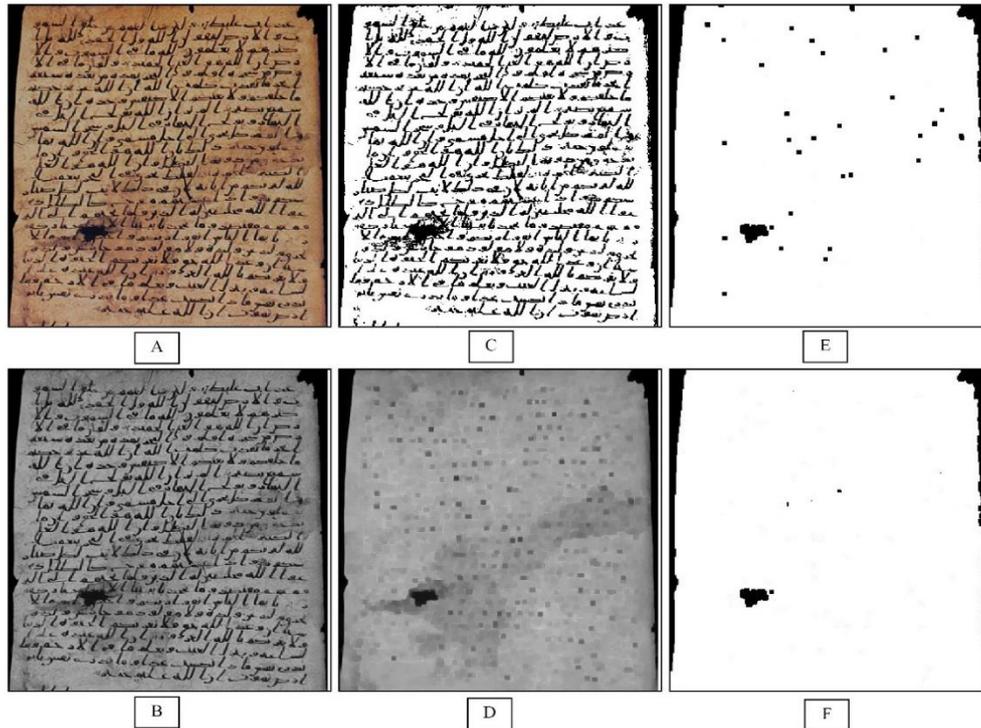


Fig. 7 Proposed Method For Classification (A) Damaged Image, (B) Gray Level, (C) Separated Foreground Layer, (D) Separated Background Layer, (E) Threshold Image, (F) Filtered Image With Threshold > Zero.

3.1.5. Experimental Results

We have studied 73 samples of Qur’an Arabic manuscripts with signs of damage, and they been labeled by an expert in careful restoration and classified into two groups. Apply Classification on the images before pre-processes shown in **Table 1** the results are not good. After pre-processes, the performance of classification goes to be better than before, in case of Group (A) classifier yields accuracy that is greater than 93.5%. In case of Group (B) classifier yields, at least 90% as shown in **Table 2**. **Fig. 8** shows Performance of damaged manuscripts classifier before and after pre-processing..

Table1: proposed classifier without pre-processed images

Damage Signs	Samples No.	Right Classification	Wrong Classification	Results
Fading Text Color	32	21	11	65.6%
Missing Part Of Text	41	29	12	70.7%

Table 2: proposed classifier results

Damage Signs	Samples No.	Right Classification	Wrong Classification	Results
Fading Text Color	32	30	2	93.8%
Missing Part Of Text	41	37	4	90.2%

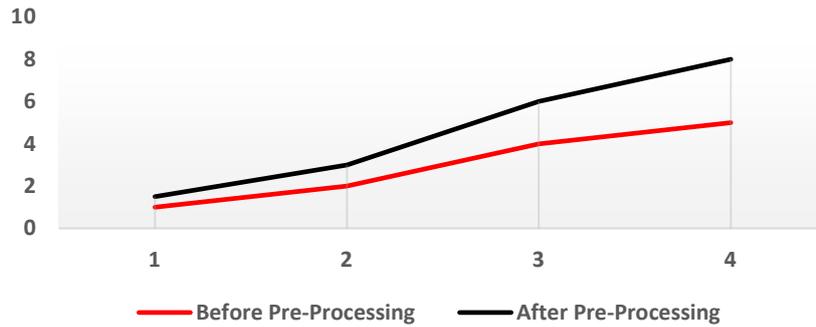


Fig. 8 Performance of damaged manuscripts classifier

Digital Restoration Models includes two main models: Fading Text Color (**FTC**) model studies type of damage can appear in the form of fading in the color of the text. Carving and erosion in the text structure, or text dilation because of heat factors that affect the natural colors used in writing the manuscripts are forms of FTC. Second model is Missing part of text (**MPT**) model is concerned with the other type of damage, which is damage in the body of the manuscript, which results as missing in the content of the text.

3.2. Fading Text Color (FTC) Model :

There are three sub-models inside this model each one addresses a specific aspect of fading text color to ensure a sound final output:

3.2.1. Contrast Strength Sub-Model

Is concerned with measuring the strength of the contrast in the text parts through the segmentation process, then finding the max value of the contrast and comparing, finally equality of all values with the max value which known by contrast uniformity as shown in **Fig. 9** simplicity of calculation equations.

3.2.2. Contour Size Sub-Model

Is concerned with the frame of the text (contour) in terms of size, and here we find that most types of damage in the text frame are erosion, which leads to the lack of continuity of the frame, then treated and completed.

3.2.3. Change Color Sub-Model

It completes The work of the previous sub-models to ensure accurate results, it is done by treating the color change, brightness strength and the intensity of clarity in each part, appearing the colors, apply thinning or dilation operations according to the state of damage. **Fig. 10** shows processes of FTC sub-model and the results by applying to Qur’an manuscript that been subjected to fading text color and how FTC sub-model restore its appearance according to the Original.

➤ Seg1, Seg2, Seg3, Segn	segmentation process.
➤ C (Seg1 > Seg2)	comparant equations.
➤ Put C Max in C Min	contraste uniformity.

Fig. 9 Contrast Uniformity Calculation

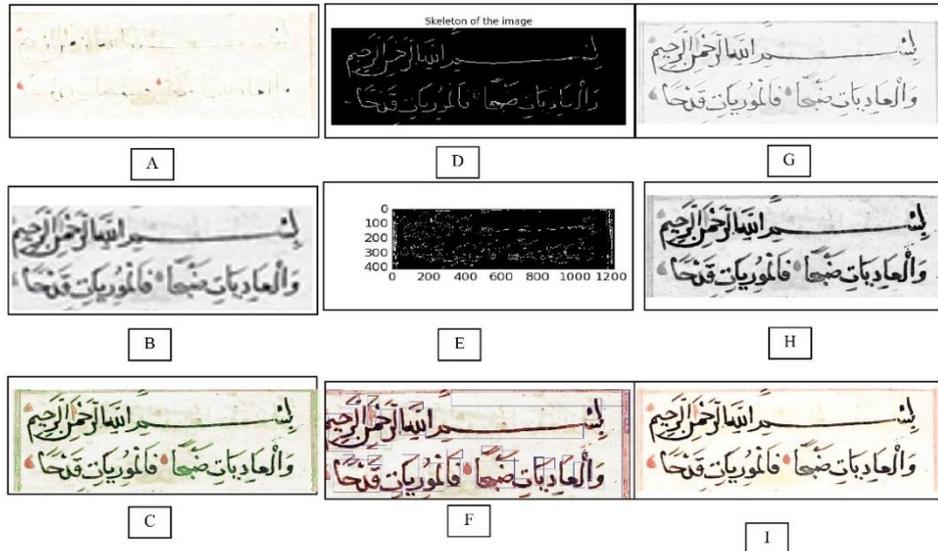


Fig. 10 Proposed Method of Fading Text Color Digital Restoration (A) Image with Faded Text Color, (B) Process of Tracing Contour, (C) Completed Contour Tracing , (D) Skeleton of Text, (E) Contrast Strength , (F) Segmented Text, (G) Color Uniformity, (H) Dilation Process, (I) Max Contrast and Color Appearance.

3.3. Missing Part of Text (MPT) Model

MTP model is concerned with the other type of damage in the body of the manuscript, which results as missing in the content of the text. At this stage, it is necessary to separate the content (foreground), treat the background, and then complete the text by specialists in the field of Arabic manuscripts, and according to established principles such as the Qur’an, hadith and some historical documents with known text.steps that were used in this model are :

3.3.1. Style Detection of Text

The need arises to a database of Arabic calligraphy style as shown in **Table 3** Samples of Arabic styles database, which we have prepared at priviouse time and include all forms of handwritten Arabic letters in various styles [14]. it is easy to complete the text with the same type of style found in the manuscript under restoration.

3.3.2. Background restoration

By defining the style, the next step is to restore the body of the manuscript, and it is called in the field of digital restoration, the restoration of the background, that is the treatment of cuts, protrusions and small holes and making them smooth.

3.3.3. Complete missing text

The last step is to complete the missing text based on the known text. The position of the text in the same damaged position is determined by specifying the location and dimensions of the cut in the x direction and y direction, This is known as a window (w) in which the text is completed with the same style, size, and characteristics as the text written around it using SIFT algorithm in extracting features. Through the database that was previously mentioned, the program replaces the entered text with a handwritten text with the same type of style and puts it in the specified place (w). **Fig. 11** shows processes of MPT sub-model and the results by applying to Qur’an manuscript that been subjected to fading text color and how MPT sub-model restore its appearance according to the Original.

Table 3: Samples of Arabic styles database

Letter's Arabic Name	Naskh Style	Reqaa Style	Letter's Printed Form
Saein			س
Haa'aa			ه
Noon			ن

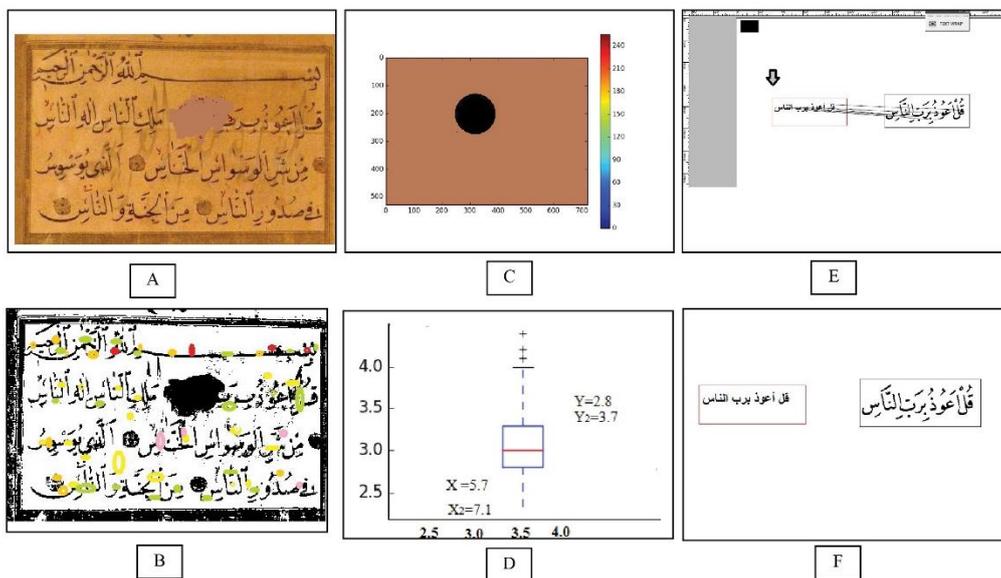


Fig.11 Proposed Method Of Missing Part Of Text (A) Image With missing text, (B) features extraction by SIFT to classify writing style, (C) determine damage window, (D) diminutions to locate text, (E) change printed text to calligraphy with same style , (F) output window with missing text.

4. PERFORMANCE COMPARISON

Finally, we conduct a performance comparison between our proposed model and similar models. We found that our model is distinguished by its focus on the Arabic manuscript, as it was exposed to the algorithms of classification damaged manuscripts according to the type of damage. In [15] a restoration algorithm for the Pahlavi or middle-age Persian manuscript had been provided. The central idea was based on the morphological analysis and connected component concept which used the mathematical morphology and connected component concept to segment the line, word, and character overlapped Pahlavi documents and prepares those texts for OCR application. In [16] Proposed an adversarial network generator that focused on the damaged frescoes and did not expose the written manuscript, The network uses the image stitching input and the output is a 30×30 matrix. The model has been tried to repair severely worn frescoes without structures, and the proposed algorithm has a better restoration effect in terms of color restoration, texture similarity and structure continuity of the damaged frescoes. In [17] only one type of damage had been studied, which is MPT by identifying text block, applied on

one manuscript in the Hindi language by level success is 83, 07%. Thus, our comprehensive model provided results that exceeded 93.5% in classifying the type of damage, and great results in digital restoration of damaged Arabic manuscripts, which opens the way in this part of the study for future researches.

SUMMARY AND CONCLUSIONS

In this work, we propose a model for classifying the type of damage that founded in ancient Arabic manuscripts using novel models for classifications and digital restoration. To these models, we attribute the superior performance. First, the used data have been captured, preprocesses have been applied, the classification stage has employed Separation techniques of foreground and background, Threshold methods to classify entered images of both signs of damage in Arabic manuscripts either Missing part of text (MPT) or Fading Text color (FTC). Then, we present the empirical performance results of classifications on 73 samples of Qur'an Arabic manuscripts with signs of damage, and they were labeled into two groups: Group (A) classifier yields accuracy that is greater than 93.5%. In case of Group (B), classifier yields at least 90%. Second, the digital restoration, which includes two main models Fading Text Color (FTC) model, inside it there are three sub-models: contrast strength model, contour size model and change in color model, and Missing part of text (MPT) model.

REFERENCES

1. Bendix C. (2010). The Preservation Advisory Centre. The British Library Design Office.
2. Tamimi F., Hirayama H. (2019). Digital Restorative Dentistry.
3. Kota S.S., Massand R. and Singh P. (2018). Digital Enhancement Of Indian Manuscript, Yashodhar Charitra. Computer Science and Engineering Department, The LNM Institute of Information Technology, Jaipur, India.
4. Ventzas D., Ntogas N and Ventza M. (2012). Digital Restoration by Denoising and Binarization of Historical Manuscripts Images. Department of Computer Science and Telecommunications. Technological Educational Institute of Larissa.
5. shri S.P., Latha L. (2015). Volume 1 Issue: A Survey On Foreground Subtraction In A Dynamic Environment. Elk Asia Pacific Journal Of Computer Science And Information Systems .
6. Knox K. , Christens-Barry W. (2008). Image restoration of damaged or erased manuscripts 16th European Signal Processing Conference ,Lausanne, Switzerland.
7. Gonzales C. , Richard E. , Woods. (2002). Digital Image Processing. 2nd ed. Englewood Cliffs, NJ: rentice-Hall.
8. Gadelrab H. (2012). Experimental Study on Chemically Alternated Microbiological Deteriorated Leather Artifacts and Evaluation of Chosen Methods of Treatment with Application on Selected Examples.
9. Lowe D. G. (2004). Vis., vol. 60, no. 2, pp. 91–110: Distinctive image features from scale-invariant keypoints.
10. Plamondon R. , Srihari S. N. (2004). vol 22, no. 1. : Online and Offline Handwriting Recognition. Comprehensive Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence.
11. Gatos, Pratikakis I., Perantonis. (2006). Adaptive degraded document image binarization. Pattern Recognition .

12. Gonzales C., Richard E., Woods. (2002). *Digital Image Processing*. 2nd ed. Englewood Cliffs, NJ: Prentice-Hall.
13. Boulid Y., Souhar A., Elkettani Y. (2016). Detection of Text Lines of Handwritten Arabic Manuscripts using Markov Decision Processes. *International Journal of Interactive Multimedia and Artificial Intelligence*.
14. Ezz M., Al Amir M., Hassan A. (2019). Classification of Arabic writing styles in ancient Arabic manuscripts. *International Journal of Advanced Computer Science and Applications*.
15. Aghaeinia H. (2005). An efficient restoration algorithm for the historic middle-age Persian (Pahlavi) manuscripts. *Computer Science, IEEE International Conference on Systems, Man and Cybernetics*.
16. J., Wang H., Deng Z., Pan M., Chen H. (2021). volume 9: Restoration of non-structural damaged murals in Shenzhen Bao'an based on a generator–discriminator network. *Heritage Science*.
17. Moshino H., Koyano Y., Mouri S. (2018). Text block identification in restoration process of Javanese script damage”, *Journal of Physics: Conference Series*.