The Impact of Optical Character Recognition with Putative Analysis for Legal Document Notarization in MATLAB for Forensic Analysis

Oleka Chioma Violet
Computer Engineering, Enugu State University of Science and Technology, Enugu, Nigeria

Abstract: **This paper presents the impact of optical character recognition (OCR) for legal document notarization to enhance forensic studies. The notarized process will be done employing image processing techniques such as background filtration, binarization, morphological dilation, putative, matching among others for accuracy. The main aim of this research is to provide a novel and reliable bibliography to enhance criminal investigation. Some existing literatures will be reviewed, analyzed and a new system will be proposed. Use case diagram and other universal modelling languages will be used for the system modelling while matlab will be used for the system implementation.**

Keywords: **Optical Character Recognition, Notarization, Recognition, Forensic.**

I. INTRODUCTION

Every process or act that involves human enrichment, money, asset or wealth has strong potential for fraud. Various agencies and techniques have been established and employed to combat this criminal act. Traditionally, there are two major means of recognition; one is the natural process using the sense organs (eyes, ears, tongue, nose and skin) while the other involves digital means, employing machine learning, pattern recognition, computer vision and optical character recognition (OCR) processes. Such digital techniques employ biometric traits such as face, iris, voice, handwriting and signature for authentication or recognition. A lot of research has been actively developed and some still on going in the general image and pattern recognition domain, notwithstanding efforts toward the dematerialization of documents, the need for fast and accurate paper-based document authentication is still growing in our society [1]. This work focus on legal documents recognition such as dead persons will, cheques, affidavits and lots more. This items required biometric authentication for public acceptance but unfortunately, various artificial means have been employed to replicate these documents with the aim of impersonation to make illegal gain of material resources (money, wealth etc), taking advantage of the setbacks (inaccuracy, delay and cost) in the conventional method of authentication (see section 2). Investigating cases of this type has become a major challenge to all security personnel, as various means have been applied yet to the best of our knowledge [9], no reliable solution in country like Nigeria for instance has been presented. This research provides a fast, reliable and easy to use novel solution for legal document notarization (forensic) using image processing techniques to extract characters from authentic and query document and comparing the label features for recognition using putative matching. The proposed system processes are discussed and mathlab will be used as an implementation tool.

Objectives of the Research

i. To develop a system that compares two documents based on handwriting characters using putative analysis
ii. To develop a system that decides if two documents are the same or not using optical character recognition process.
iii. To improve on the accuracy of existing system
iv. To develop a digital means of document notarization

II. LITERATURE REVIEW

In 2014, Chandan Kumar [4] proposed a novel technique for recognizing English language characters in documents using Artificial Neural Network. The persistency in recognition of characters by the network (AN) was found to be more than 90% accurate. Amit et al (2013) used binarization features alongside neural network classifier, employing back-propagation algorithm and delivering outstanding recognition accuracy of 85.62 % [3]. In 2006, Gatos et al. [5] used K-NN classifier to recognize 3799 words from IAM database and reported 81% accuracy [3]. In (2012) Manal et al., [2] researched using fixed segment and revealed that resizing the segment width for unrecognized characters to adapt with different character width fulfill better recognition percentage of 81%.

III. APPROACH

To solve the required optical character recognition problem we employed the MATHLAB computation software with image processing techniques which are:

Image Acquisition: this process involves capturing the real and query document using a scanner, HD camera or any image acquisition device.

Image pre-processing: this step is a preliminary image processing step to get a better image quality using histogram equalization. This technique reduces background noise from the scanned optical character document.

Image Binarization: this process converts the captured image into a bi-level (black and white color) in preparation for processing.

Digital segmentation: this is an image processing technique for finding the boundary of an object within image. This was used to locate characters in both documents to be recognized using both line and character segmentation techniques.

Morphological dilation and erosion: this involves processing of characters based on shape, applying structural elements to OCR images and creating resultant output image of similar shape.

Image Normalization: this procedure not only removes noise from the image but also bring the image to a range of intensity value that is normal for feature extraction process.

Putative Match Analysis: The putative analysis (PA) was originally proposed for geographical aerial image search [6]. However, after careful review and experiment with this method, applying image processing techniques (see section 3) together with PA using Mathlab as shown in figure (9), comparing various OCR documents,the technique was able to differentiate between similar and similar fabricated documents based on the character content by comparing their angles (**P**) and the relative length (R) of its vectors (va and vb) as presented below in equation (1).

$$R = \frac{\min(\|va\|, \|vb\|)}{\max(\|va\|, \|vb\|)} \quad \text{.....equation (1)}$$

PA is constructed using what is called [6] a relative polar matrix (RPM) to create the vector with relative lengths (va) and (vb) from each pair of inliers (corresponding points) see figure (1a). Each RPM matrix column contains a cluster that can be visualized using both (R) and (**P**) as coordinates, employing the clustering technique as shown in figure (1a) which displays all points in the RPM plotted with the cluster set for each row and column related to each other.

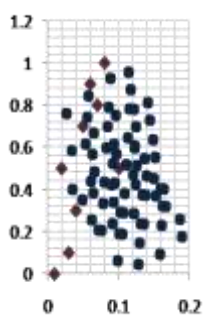

Figure (1a)

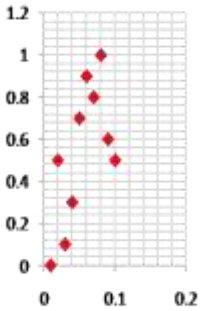

figure (1b)

The clustering technique is an unsupervised learning algorithm that solves the clustering problem, classifying each given set through a certain number of clusters. The process computes the average (k)for each row *i*of (R, **P**) giving a "centre" that is a bit perturbed by the outliers. Then the distance between each point on the row (R*i*, **P** *i*) and (k)is stored in a matrix, which is called [7] Relative Distance Matrix (RDM). In each iteration, the cluster set with an outlier furthest away from the average centre is located and that outlier is therefore removed and the RDM was updated accordingly. Figure (2a) shows the whole clustering point. Figure (2b) shows the set correspondence to the most extreme outlier. The outlier is easily found by locating the column and row with the largest value as also shown in figure (1b)

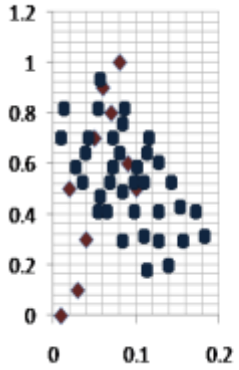

Figure 2a

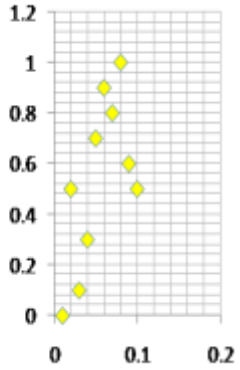

figure 2b

IV. MODELING AND SYSTEM ANALYSIS

In order to realize the model, Unified Modeling Language (UML) was employed. It uses diagrams to document an object-based decomposition of systems revealing the interactions between these objects and there dynamics. UML aims to provide a common vocabulary of object-based terms and diagramming techniques that is rich enough to model any system development project from analysis to design [9]. For this model, use case diagram was used, process flow chart, interface structure diagram and system flow diagram. Use case diagrams give a user point of view of a system [9], from figure 3: the use case diagram is used to model the operability of the system with the new user as the actor while in figure 4, the existing user is the actor, both requires authentication before the documents are uploaded for notarization. In figure 5: the interface structure diagram is the graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency in the system. Development of the system involves various tools from the graphical aspect of designing to the command aspect of writing codes.

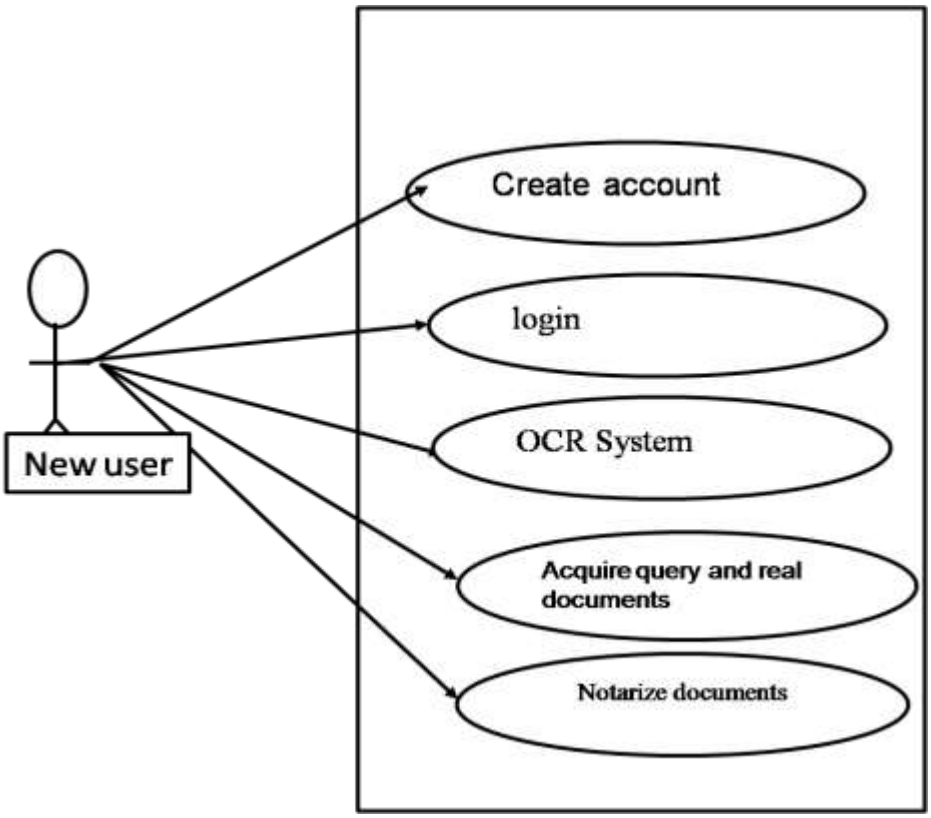

Figure 3: Use case diagram for new user

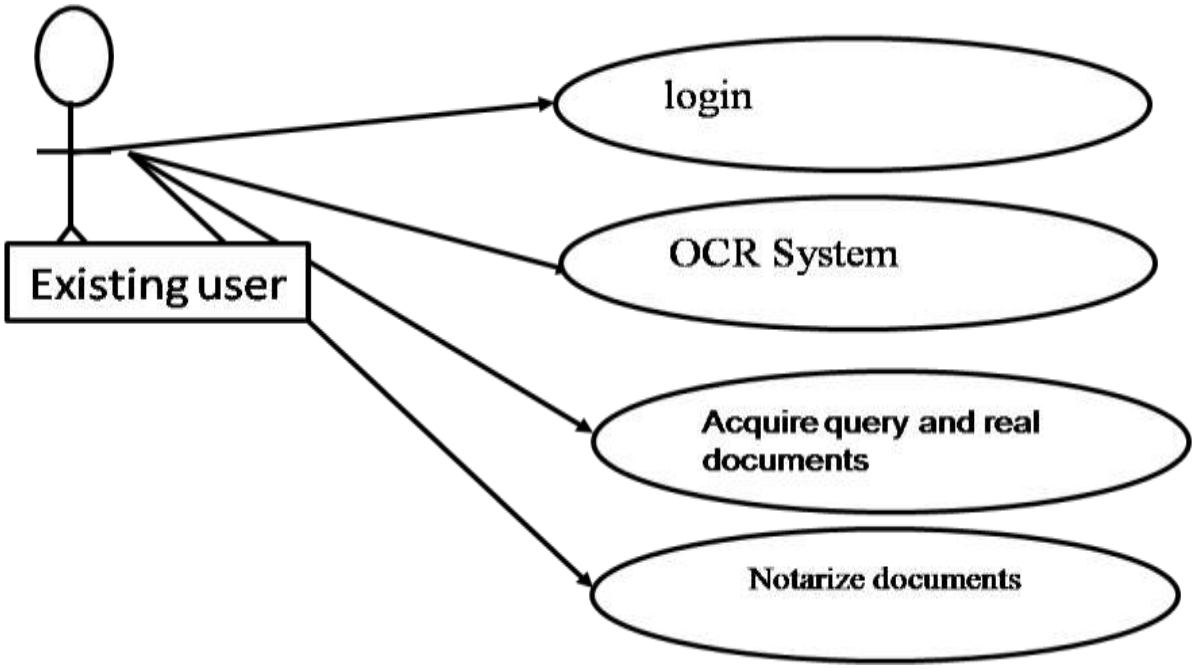

Figure 4: Use case diagram for existing user

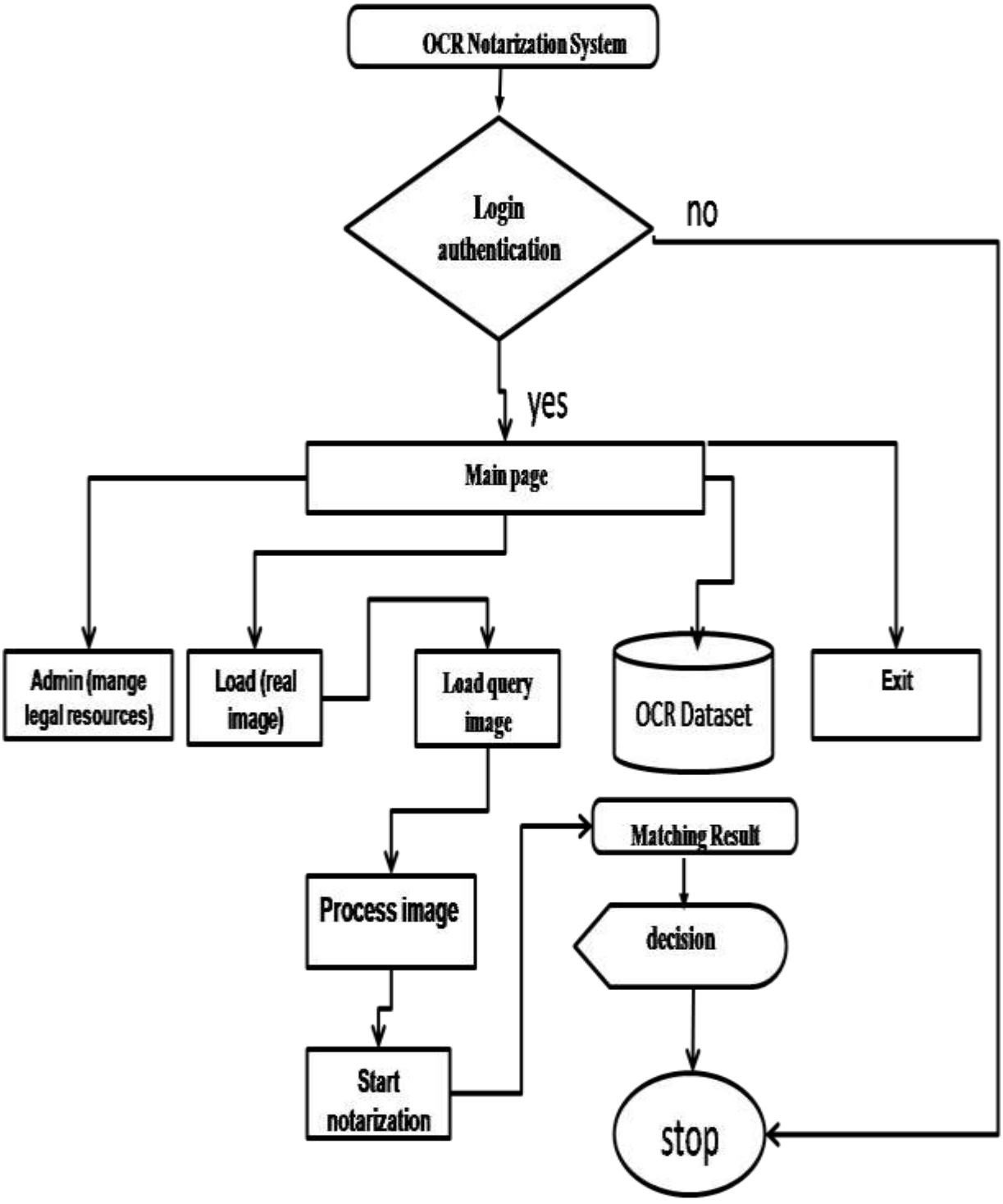

Figure 5: Interface Structure design

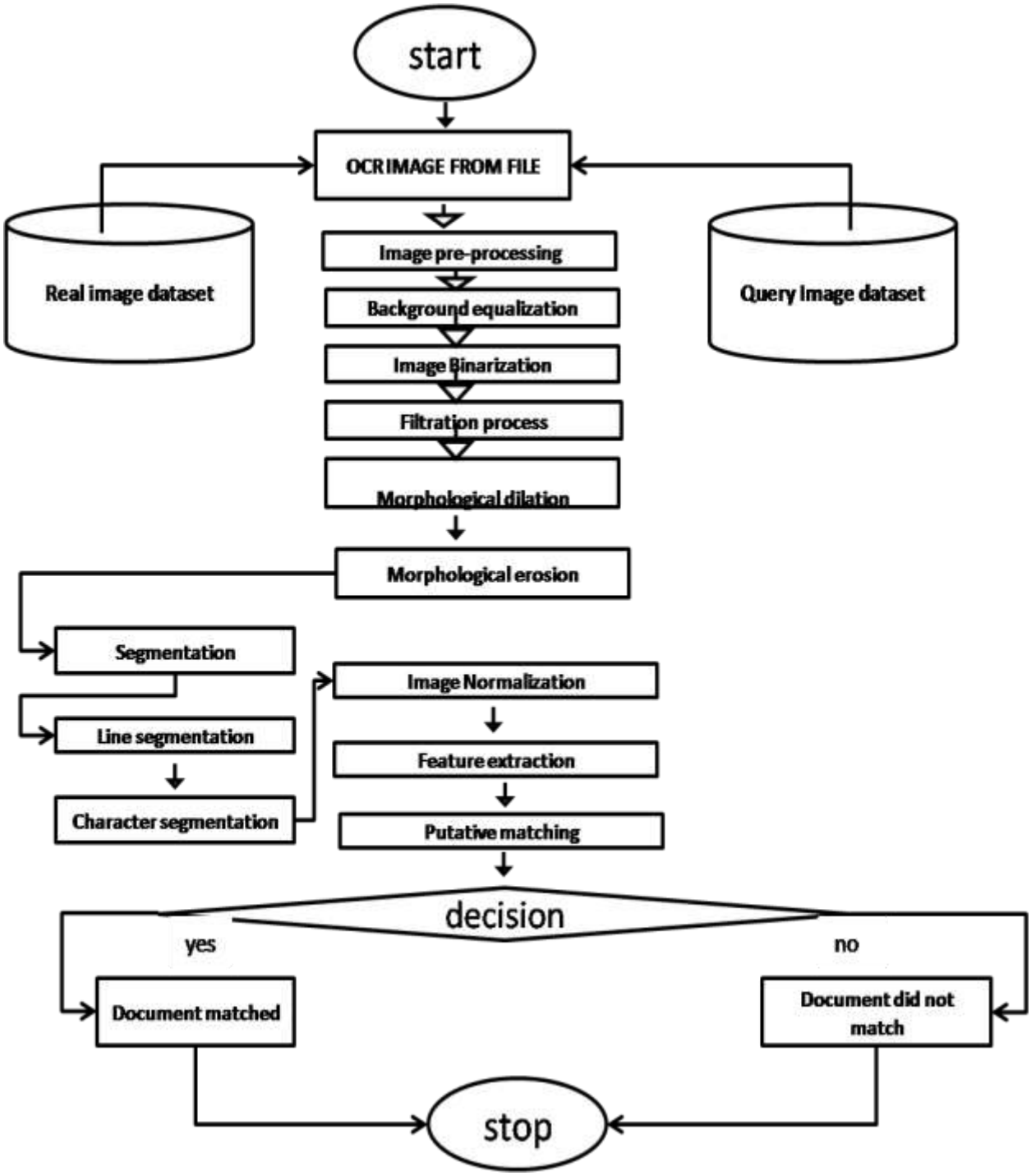

Figure 6: Process flow chart of the system

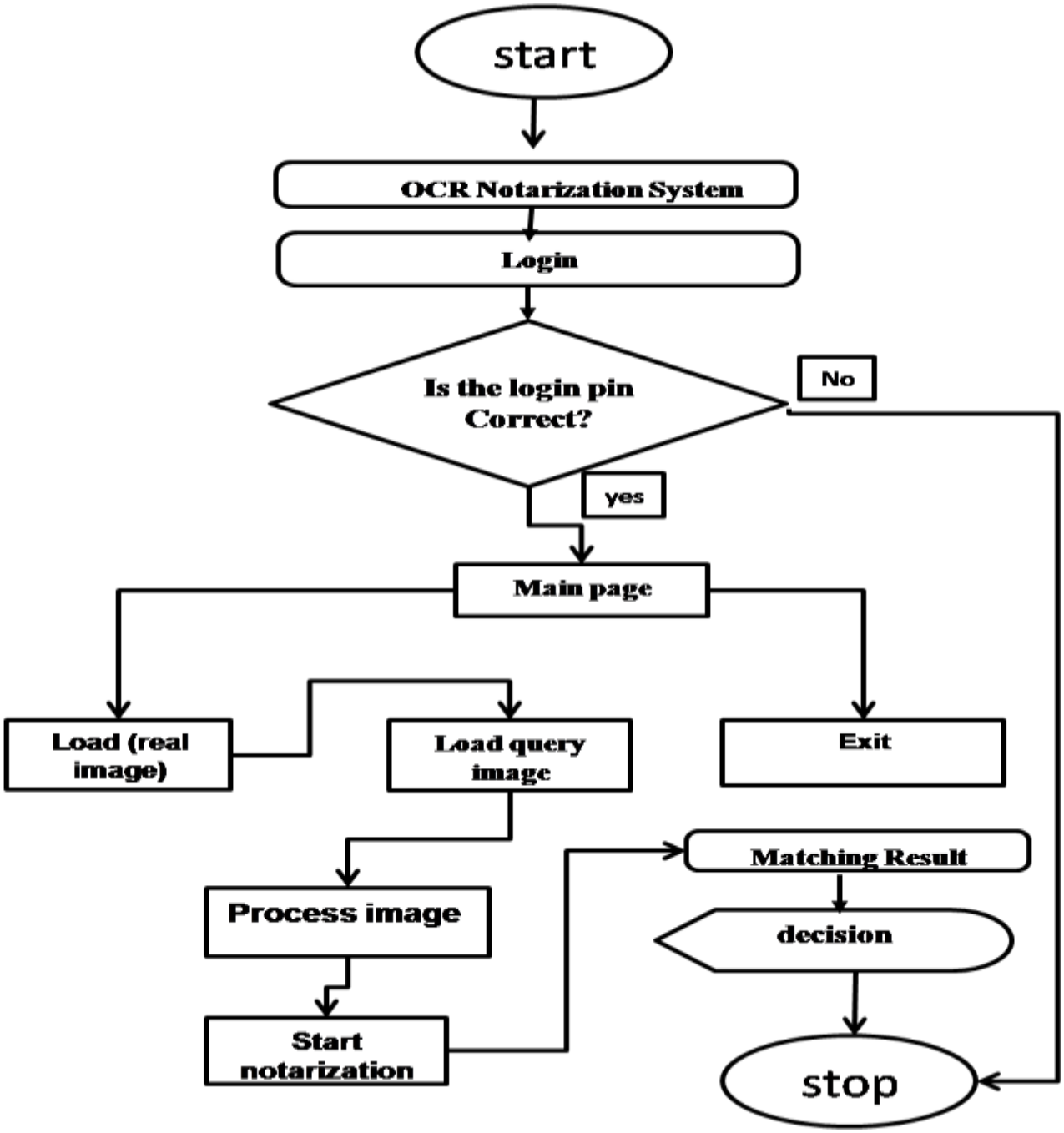

Figure 7: System Flow chart

4.1 Graphical User Interface (GUI)

The interface was designed to facilitate interactive system operation using matrix laboratory development tool. GUI can be used to setup the program, launch it, stop it and display results [8]. During setup stage, the operator is promoted to load real and query optical character images. The implementation result was accurate as shown, figure 6: demonstrates the image processing tools displaying the output processes in comparing two document as shown below, the result prompted "document matched" while in figure 9: after image processing techniques on the two documents, the result prompted "data not match".

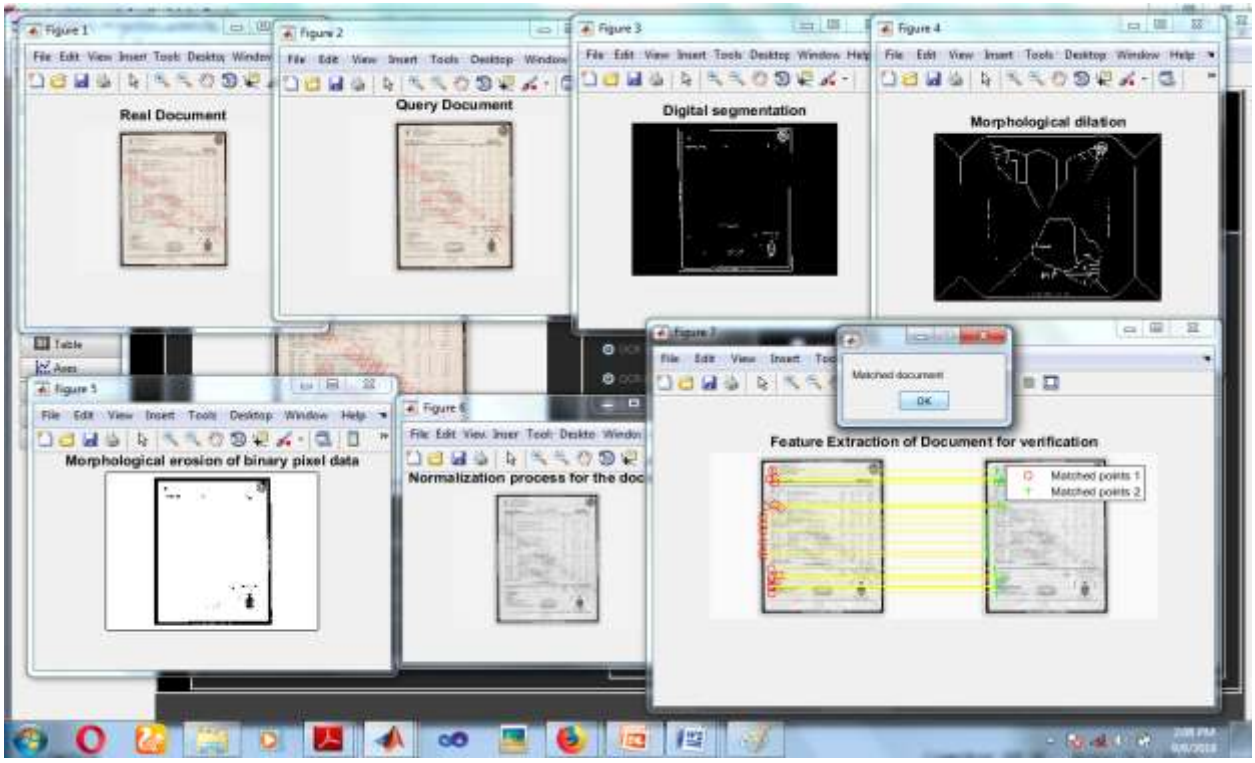

Figure 8: Framework for Matched document

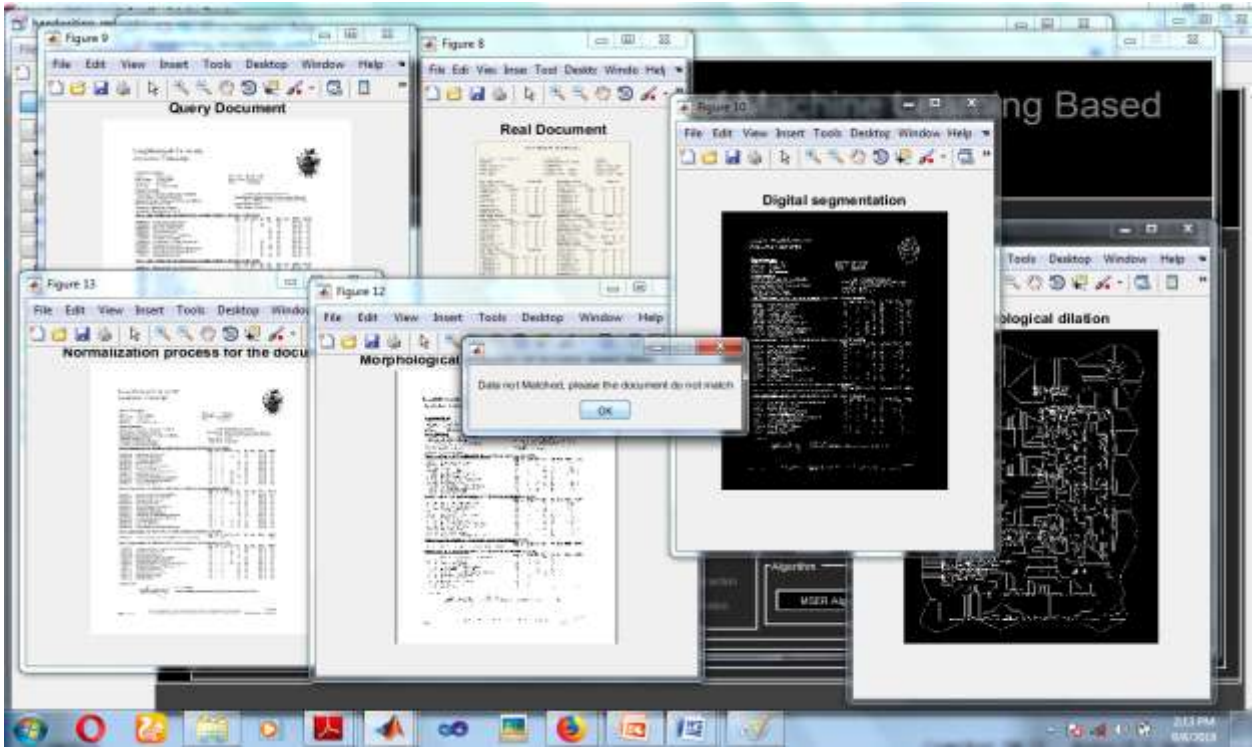

Figure 9: Framework for Rejected documents

V. CONCLUSION

This work is proposed for legal documents notarization processes using image processing techniques on optical character recognition. From the implementation result, it was shown that the proposed system was successfully designed using mathlab and the experimental screen shoots were shown for the recognition of documents accurately. Putative analysis was used as a good choice for a feature matches as it was proven [6] to find the inliers even for a rather high amount of noise in the scanned documents. The proposed system can also be modified as a robust tool for verifying signatures in the banks, offices and used for testing and developing computer vision applications.

REFERENCES

[1]. Ms. DeeptiJoon, Ms. ShalooKikon; An Offline Handwritten Signature Verification System - A Comprehensive Review; International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463; Vol. 4 Issue 6, June-2015, pp: (433-439)

[2]. Manal A. Abdullah, Lulwah M. Al-Harigy, and Hanadi H. Al-Fraidi; Off-Line Arabic Handwriting Character Recognition Using Word Segmentation; Journal Of Computing, Volume 4, Issue 3, March 2012, Issn 2151-9617.

[3]. Amit Choudharya, Rahul Rishib, Savita Ahlawatc; Off-Line Handwritten Character Recognition using Features Extracted from Binarization Technique; ScienceDirect2013, AASRI Conference on Intelligent Systems and Control.

[4]. ChandanKumar; Hand-Written Character Recognition Using Artificial Neural Network. 2014.

[5]. Gatos, B., Pratikakis, I. &Perantonis, S. J., 2006. “Hybrid off-line cursive handwriting word recognition in proceedings of 18th international conference on pattern recognition (ICPR’06), 2, pp. 998–1002.

[6]. Anders Hast And Andrea Marchetti . Putative Match Analysis A Repeatable Alternative To Ransac For Matching Of Aerial Images;- International Conference On Computer Vision Theory And Applications. 2012.

[7]. Zuliani, M., 2009. RANSAC for dummies. pp. 42. http://vision.ece.ucsb.edu/~zuliani/Research/RANSAC/docs/RANSAC4Dummies.pdf

[8]. Habib, Mohammed and Hussien; Detection and Tracking System of Moving Objects Based on MATLAB; (IJERT), ISSN: 2278-0181, IJERTV3IS100721 Vol. 3 Issue 10, October- 2014.

[9]. Ambrose A. Azeta, Nicholas A. Omoregbe, AdewoleAdewumi, DolapoOguntade, Design of a Face Recognition System for Security Control. International Conference on African Development Issues (CU-ICADI) 2015: Information and Communication Technology Track. Department of Computer and Information Sciences, Covenant University, Ota, Ogun-State, Nigeria