A novel comprehensive database for Arabic and English off-line handwritten digits recognition

Ahmed Subhi Abdalkafor¹, Waleed Kareem Awad², Khattab M. Ali Alheeti³
¹Career Development Center, University of Anbar, Iraq
²³Department of Computer Networking Systems, University of Anbar, Iraq

ABSTRACT
The recognition of Arabic handwritten is received at the same interest as other Latin languages. In Optical Character Recognition (OCR), handwriting Arabic recognition is considered as one of the critical and difficult tasks in the various scientific area. The main issues of this matter were due to the lack of public Arabic handwriting databases and the cursive nature of Arabic writing. In this paper, a new benchmark database is built for the Arabic and English off-line handwritten digits Recognition. The original form is divided into three groups: Arabic digits, English digits, and word Arabic digits which written five times by 100 different academic staff and students of university writers. Our database contains 14500 images; divided into two subsets of training and testing to help researchers through evaluating and comparing results obtained from their systems.

Keywords: Arabic, Cropping, Database, Handwritten, Word arabic digits

Corresponding Author:
Khattab M. Ali Alheeti,
Department of Computer Networking Systems,
College of Computer Science and Information Technology,
University of Anbar, Iraq.
Email: co.khattab.alheeti@uoanbar.edu.iq

1. INTRODUCTION

Pattern recognition has become an attractive filed, due to the wide uses for many applications that humans use frequently in their daily lives. Handwriting recognition is one of the methods of computerization to imitate or understand human behaviors in recognizing ways to convert handwritten or printed text into automatically encrypted numbers [1-3].

Many researches have focused on the recognition of Chinese, Japanese and other languages. In contrast, A little researches focusing on Arabic texts due to several challenges in terms of grammar, writing style, lack of access to a database in terms of Arabic as opposed to and other factors that make it difficult to identify this language [4-7]. Recently, Arabic Digit handwritten has a very high level of interest, it entire in wide applications in a different area, such as verification of bank-check verification, entry data applications criminal evidence and all types of vehicles plates to register the violations The recognition of digits is one of the issues in pattern recognition so it is adopted as a measure to compare different classification algorithms [8-9]. The researchers focused on neural network algorithms for wide use in recognition, prediction, and others [10]. In the recognition filed they have been implemented by using different databases to evaluate the proposed works. [11-13].

There are few numbers of databases have been proposed especially in Arabic digits to evaluate the classification algorithms. This proposed work focuses on some previous studies such as Alayba and et al. have been explained construct models of Word2Vec from large Arabic corpus [14]. However, these models are obtained from ten newspaper in various Arabic area. Grother in [15] proposed the EMNIST database that merges the lowercase and uppercase into a single 26-class task of digits handwritten. On the other hand, LeCun [16] proposed the database named MNIST for English handwritten digits recognition all images are
gray-scale and the size 28 by 28 pixels the overall of these images are 60000 images. R. Vijaya Kumar Reddy et al. [17] proposed different techniques of neural network approach to identify the Handwritten Hindi characters. They assessed the execution utilizing Convolutional Neural Network (CNNs) with improvement strategies and Deep Feed Forward Neural Network (DFFNN). The study identifies the set of database images that were collected from different users then train and test these techniques or strategies on it [18]. H. Alamri et al [19] proposed a new inclusive database for identifying Arabic handwriting. The submitted database looks at the first database that hiding place a different set of the Arabic language. It includes dates, isolated letters, strings of numerical, isolated Indian numbers, a set of (70) significant words and a set of distinct symbols. In 2011, I. Ahmad et al. [20] proposed a new comprehensive database of Arabic handwritten text (AHTD). This developed database consists of text written by (1000) writers from different countries. The submitted database includes images of the written text with different accuracy.

The basic database contains metadata describing the text written on the page, paragraph, and line levels [20-22]. This paper is structured as follows: The next section presents the characteristic of Arabic and English Digits. Then, the proposed system that involves four phases is presented in Section 3. Finally, Section 4 presents the conclusion of the study.

2. CHARACTERISTIC OF ARABIC AND ENGLISH DIGITS

Arabic digits consist of 10 digits and written from right to left. Each digit has various shapes also in the word Arabic digits depending on the position of the character in the word and the writer himself. Conversely, the digits of the English language were written from left to right and not consists of complexity as the Arabic language [23].

3. METHODOLOGY

The evaluation process of any classification system had been heavily based on the database. As we know, we have a problem with type and number of databases whether Arabic or English that utilized to test identification systems in research area. In addition, there are few numbers of databases have been suggested especially in Arabic digits to test the classification algorithms. In this research, a new benchmark is generated for the Arabic and English off-line handwritten digits recognition. In more detail, the original form is separated into main three sets: Arabic digits, English digits, and word Arabic digits.

The methodology of the new benchmark is comprised of four phases which are acquisition images from the various volunteers, data extraction phase, pre-processing phase, and put the image into folders. All these phases are explained in the Figure 1. The lifecycle of the proposed system is shown in Figure 1 that explained the generation process from the collected image to the creation process of the new benchmark. Thus, all these phases are considered a very important part of this paper; all of them will explain in detail each follows the subsection of this paper.

![Figure 1. The block diagrams](image-url)
### 3.1. Data acquisition

The basic step to building any database is to find the appropriate data source [19]. The original form as shown in Figure 2 consists of three digits groups including (Arabic digits, English digits, and word Arabic digits) where each digit has been printed and the writers have been asked to write each digit in five empty boxes. The proposed database is collected from 100 writers, the average age range between 10 to 40 years from different levels of education that including the academic staff, employees, students of university and others. Little writers do not enter the schools of education, this status provides the proposed database with a diversity of shapes and patterns for each digit.

![Figure 2. Original form before scanning](image)

### 3.1.1. Form scanning

The original form was scanned using a high-resolution scanner. The scanner scans 300 pages using 400 dpi and the output is PNG format of the scanned process. Figure 3 illustrated the original form after the scanning method.

![Figure 3. Form after scanning](image)
3.2. **Data extraction phase**

In this study, we applied a software of Adobe Photoshop program to extract each image of digits from the five boxes, Figure 4 illustrated this stage. After this phase, our database has saved in three main an isolated folder and each folder contains 10 subfolders, each subfolder contains five images to produce 14500 digits of images.

![Figure 4. Extraction images](image)

3.3. **Pre-processing stage**

This process is considered a necessary method for any recognition system. One of the disadvantages resulting from the scanning process is the presence of spaces around the body of an image thus will generate a database of high-storage space and this will be affected the functioning of the system in terms of time. Therefore, a cropping manner was applied to delete all the increases and spaces around the body of the image [23-25]. This manner led to getting a low-storage database and reduce the time during the processing of the system. Thus, improving the performance of the recognition system. Figure 5 showed a universe of discourse manner.

![Figure 5. Cropping method](image)

After completely these methods, these images put in three isolated folders. Firstly, the folder contains five images of Arabic digits and secondly folder five English digits and Arabic word digits in the third folders all of them were written by 100 writers and the total is 15000 images. The statistics of the proposed database are illustrated in Table1.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Writers</td>
<td>100</td>
</tr>
<tr>
<td>First Folder (Arabic digits)</td>
<td>500</td>
</tr>
<tr>
<td>Second Folder (English digits)</td>
<td>500</td>
</tr>
<tr>
<td>Third Folder (word Arabic digits)</td>
<td>450</td>
</tr>
<tr>
<td>Bit Depth</td>
<td>24</td>
</tr>
<tr>
<td>Number of images</td>
<td>14500</td>
</tr>
<tr>
<td>Training Sets</td>
<td>10150</td>
</tr>
<tr>
<td>Testing Set</td>
<td>4350</td>
</tr>
<tr>
<td>Type of image</td>
<td>(PNG)</td>
</tr>
</tbody>
</table>

### 4. **CONCLUSION**

In this paper, a new database is presented for Arabic handwritten recognition and English digits for any recognition systems. It collected from 100 writers filled an original form that has five boxes. However,
more than 14500 images are extracted and implemented some suitable methods via cropping method for reducing some images of noise that may exist after the scanning process. A Novel Comprehensive Database is divided into 70% for training and 30% for testing.

REFERENCES