# BPTI: Bilingual Printed Text Images Dataset for Recognition Purposes

Mohammad Yahia
Department of Computer Science
Al Hussein Technical University, Jordan
mohammad.yahia@htu.edu.jo

Husni Al-Muhtaseb
Information and Computer Science Department
King Fahd University of Petroleum and Minerals, Saudi Arabia
muhtaseb@kfupm.edu.sa

**Abstract:** *Datasets of text images are important for optical text recognition systems. Such datasets can be used to enhance performance and recognition rates. In this research work, we present a bilingual dataset consists of Arabic/English text images to address the lack of availability of bilingual text databases. The presented dataset consists of 97812 text images, which are categorized into two groups; Scanned page and digitized line images. Images of the two forms are written with 10 fonts and four sizes, and prepared/scanned with four dpi resolutions. The dataset preparation process includes text collection, text editing, image construction, and image processing. The dataset can be used in optical text recognition, optical font recognition, language identification, and segmentation. Different text recognition and language identification experiments have been conducted using images of the dataset and Hidden Markov Model (HMM) classifier. For the digitized images recognition experiments, the best-achieved recognition correctness is 99.01% and the best accuracy is 99.01%. The font that has the highest recognition rates was Tahoma. For the scanned images recognition experiments, Tahoma has also shown the highest performance with 97.86% for correctness and 97.73% for accuracy. For the language identification experiments, Tahoma has shown the performance with 99.98% for word-language identification rate.*

**Keywords:** *Optical character recognition, text images dataset, HMM.*

*Received February 9, 2022; accepted September 28, 2022*
*https://doi.org/10.34028/iajit/20/4/12*

## 1. Introduction

Multilingual text can appear in documents in different ways:

1. In the same document, where one of its components such as pages, columns, paragraphs, and/or lines could be written with a particular language and another component could be written with a second language.
2. Or the same line could have multilingual text. Some samples of these documents are shown in Figures 1 and 2.


Figure 1. Arabic/English signboard [39] (some lines are written with a particular language).

The need for developing multilingual recognition systems is increasing and becoming indispensable. This increasing demand is attributed to several reasons:

1. Increase communication between countries and between international organizations, which may lead to the appearance of several languages in one document.
2. Many countries such as India have several official languages and thus multiple languages may appear in the same document.
3. And despite the countries that have one official language, their organizations such as universities and banks may issue their documents written with two languages.

Multilingual printed text datasets are important for developing multilingual recognition systems, particularly for training and testing processes. There are numerous implementations and applications that can benefit from multilingual recognition systems and multilingual datasets. These implementations include recognizing multilingual documents such as university certificates, bank checks, and postal mails. Another important implementation is processing signboard banners.

Many multilingual text datasets that contain different languages other than Arabic are developed and used in recognition systems such as Chinse/English [23], English/Tamil [14], Farsi/English [21],

English/Gurmukhi [37], Thai/English [11], and Malayalam/English [35]. One main shortcoming that obstructs developing an Arabic/English recognition system is the lack of availability of bilingual (e.g., Arabic/English) text datasets. Several surveyed Arabic datasets were developed to recognize handwritten Arabic text [27, 30], Arabic printed characters [1], handwritten Arabic words [33], and Arabic fonts [26].

Unit (3)

Synchronic Versus Diachronic Approaches

Just as the term for science varies in implication from one culture to another, the abstractions used in science vary considerably, even in their most direct meaning, from one language to another. A famous Arabic word, ( الحكم ) is used to refer, as an immediate 'signified', to *judgment*, but has a multitude of other meanings in different contexts. The original or the most obvious and direct meaning is retained in such works as Kant's *Critique of Judgment* نقد الحكم where the term means 'understanding, good sense, discrimination' as is implied by the Quranic verse « اتيناه حكماً وعلماً » (Joseph, 22) (to him we gave judgment and knowledge). To this sense is allied the noun, another difficult abstraction, الحكمة — commonly rendered as 'wisdom', though in the Quran it can have the added meaning of 'profound discretion, good reasoning, or correct insight' as it is often 'collected' with (the Revelation or the Book) as in the verses « وما أنزل عليكم من الكتاب والحكمة » ( البقرة : الآية ٢٣١ ) and in ten other verses. The original meaning of discrimination is retained in another context, namely in the verse ( سورة ص – الآية ٢٠ ) « وآتيناه الحكمة وفصل الخطاب » where it is associated with the 'ability to decide, or to determine' and thus confirms that meaning of judgment. Similarly, the Bible adds another dimension to the more generalized meaning, which is justice or right (righteousness) or equity, *Isaiah*. (61/8) "For I the Lord love Judgment" — alternatively rendered in the Revised English Bible as *Justice*. How close the two meanings are can be seen from the other noun, again used significantly in the Quran with reference to God,

Figure 2. Arabic/English document (some lines are written with two languages).

## 1.1. Characteristics of Arabic Text

One characteristic of Arabic text is that it is written cursively. Cursive script adds challenges to Optical Character Recognition (OCR) systems, particularly to the segmentation process. In addition, Arabic language has 28 basic letters. Twenty-two of them can have four basic different shapes based on the letter position (i.e., start, middle, end, or isolated). These letters appear as standalone letters or connect from left, right or both sides. These letters are (Beh ب, Teh ت, Theh ث, Jeem ج , Hah ح, Khah خ, Seen س, Sheen ش, Sad ص, Dad ض, Tah ط, Zah ظ, Ain ع, Ghain غ, Feh ف, Qaf ق, Kaf ك, Lam ل, Meem م, Noon ن, Heh ه, and Yeh ي). The rest of them (Alef أ, Dal د, Thal ذ, Reh ر, Zain ز, and Waw و ) can have two basic shapes and they appear as isolated letters or they are only connect from right. The number and the position of dots that attached to Arabic letters play significant role in differentiating among letters. For example, the letter (Beh) has only one dot below the body of the letter, the letter (Teh) has two dots above the letter's body, and the letter (Theh) has three dots appear above the letter's body. All of these letters (i.e., Beh, Teh, and Theh) have the same body shape.

Table 1 shows different shapes for some basic Arabic letters based on their position within the word. In addition, overlapping and vertical stacking are resulting either from combining two letters or from prolonging for decorative purpose. Overlapping can cause difficulties in segmentation process since that the overlapped letters are difficult to be separated. Figure 3 shows some of these challenges:

a) Overlapping of two letters.
b) New shapes can be resulted from combinations of different letters, (e.g., vertical stacking and overlapping).
c) Different letters can have the same shape but dots distinguish them.
d) The same letter can have different shapes based on the position (e.g., start, middle and end) of the letter within a word.

Table 1. Basic shapes of some arabic letters.

| Isolated letter | English name | Initial shape | Medial shape | Terminal shape |
|---|---|---|---|---|
| ء | Hamza | ء | ء | ء |
| ا | Alef | ا | ـا | ـا |
| ب | Beh | بـ | ـبـ | ـب |
| ت | Teh | تـ | ـتـ | ـت |
| ث | Theh | ثـ | ـثـ | ـث |
| ز | Zain | ز | ـز | ـز |
| ش | Sheen | شـ | ـشـ | ـش |
| و | Waw | و | ـو | ـو |

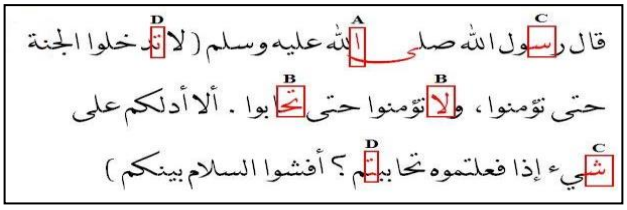


Figure 3. Some characteristics of Arabic text.

The organization of the paper is as follows: Section 2 introduces the literature review of developing and presenting bilingual datasets and bilingual recognition systems. The overview of the proposed dataset and the construction process are described in sections 3, and 4, respectively. Experiments that are conducted using some images of the dataset are reported in section 5. Section 6 presents the conclusion.

# 2. Literature Review

In this section, we present some research efforts pursued in developing bilingual/multilingual text datasets. We concentrate on Arabic text datasets. We also present some recognition systems of bilingual text.

## 2.1. Bilingual/Multilingual Text Datasets

Chtourou *et al*. [13] proposed an offline dataset for handwritten and printed text recognition purposes named Arabic-English Text Images Database (ALTID). The proposed dataset consists of 731 pages of English and Arabic printed documents. The scanned documents include 1845 Arabic text-block images and 2328 English text-block images. The handwritten dataset consists of 460 Arabic and 582 English text-blocks

images. All images were scanned at 300 dpi resolution. The ground truth of the dataset is provided. Pal and Chaudhuri [32] prepared a multilingual printed text line images dataset for script identification. The dataset consists of 700 document images containing around 25000 text line images written with five languages English, Chinese, Arabic, Devanagari and Bangla. In the Maurdor project [10], a dataset of handwritten and printed text images was prepared for improving automatic processing of digital documents. This dataset includes 10000 document images where French text images represent 50%, English text images represent 25%, and Arabic text images represent 25% of the documents. Chernyshova *et al*. [12] proposed a multilingual a dataset called MIDV-LAIT for identity documents recognition. The main feature of the dataset is the textual images in Perso-Arabic, Thai, and Indian scripts. Mobile Identity Document Video - Latin, Perso-Arabic, Indian, and Thai scripts (MIDV-LAIT) MIDV-LAIT dataset includes 180 unique documents and 3600 images. It contains identity cards, passports, and driving licenses from 14 countries. Al Maadeed *et al*. [4] presented an Arabic/English handwriting dataset called Qatar University Writer Identification dataset (QUWI). The dataset consists of 4086 scanned documents with 600 dpi resolution. These documents were written by 1017 writers. According to the authors, the dataset could be used in different research areas such as writer identification, gender identification, age identification, nationality identification, and handiness of a specific writer. Djeddi *et al*. [15] introduced an offline Arabic/French handwriting dataset LAMIS- Multi-Script offline Handwriting Database (LAMIS-MSHD) (LAMIS-MSHD) to be used in different researches such as signature verification, writer identification, and text segmentation and recognition. LAMIS-MSHD dataset consists of 600 Arabic and 600 French handwriting text samples, 1300 signatures and 21,000 digits. These components were written by 100 writers using 1300 forms. The documents were scanned using 300 dpi resolution.

Lin *et al*. [23] developed a dataset by collecting 1517 images from bilingual Chinese and English magazines and newspapers. These images included 29907 text lines and 548508 components (i.e., Chinese letters, English alphabets, and punctuations). In Hassan *et al*. [18], the dataset consists of 910 gray-scale images scanned at 300 dpi resolution. The 910 images contain English/Hindi script and English/Bangla scripts. These pages were collected from supplementary books such as guidebooks and language training books. The dataset of Dhanya *et al*. [14] consists of gray scale images scanned at 300 dpi resolution. The scanned documents contain English and Tamil scripts. Khoddami and Behrad [21] provided a dataset that contains gray scale images of 62 Farsi pages and 37 English pages. Each page consists of 28 text lines in average. The resolution of the page images is 300 dpi. The source of these images is the internet. The scripts of Farsi and English were written with three fonts, different sizes, and different font styles (e.g. natural, bold and italic). Rani *et al*. [37] prepared a dataset of gray scale images of 5212 Gurmukhi words and 6188 English words and numerals. The words were collected and scanned from books, magazines, and newspapers. Chanda *et al*. [11] prepared a dataset by scanning Thai/English text from newspaper and books at 300 dpi resolution. The dataset consists of 5000 gray scale images of Thai words and 5000 gray scale images of English words. The words were written using different fonts and different sizes. Philip and Samuel [35] prepared a dataset that consists of gray scale images of around 1000 Malayalam and English words. The words were collected from different textbooks and magazines and scanned at 300 dpi resolution. Mathew *et al*. [29] used a dataset of Hindi languages and English that consists of more than 55000 scanned images of printed documents. It includes around 1500 pages of English and the remaining pages represent 12 Hindi Languages. Contents of these images such as words were used for recognition purposes and script identification. Bartos *et al*. [9] introduced a public-domain dataset namely T-H-E Dataset for recognition purposes. T-H-E Dataset includes 156000 handwritten Turkish, Hungarian and English characters collected from 200 participants and scanned using 300 dpi resolution. The validity and applicability of the dataset were evaluated and confirmed by carrying out some recognition experiments using deep learning techniques. Hamdi *et al*. [17] proposed a multilingual dataset, namely NewEye, for developing and evaluating named entity recognition systems. The dataset is comprised historical newspaper images and contains four corpora German, Finnish, Finnish and Swedish. Each corpus is split into 80% for training process, 10% for validating process, and 10% for testing process. The dataset contains 30580 named entities such as person, location, organization, and human production.

## 2.2. Arabic Text Datasets

Al-Muhtaseb [6] proposed two datasets named PATS-A01 and PATS-A02 for Arabic text recognition. The first dataset consists of 2766 text line images and the second one consists of 318 text line images. The images of the datasets were created by the computer with 300 dpi resolution and binary colors (i.e., black background and white text). The text of the images was written with eight fonts (e.g., Arial, Tahoma, Akhbar, Thuluth, Naskh, Simplified Arabic, Andalus, and Traditional Arabic) and font size 18 points. Slimane *et al*. [40] proposed an Arabic printed word images dataset namely Arabic Printed Text Images (APTI). The proposed dataset included 45313600 Arabic word images with 72 dpi resolution with more than 250 million characters. These word images were resulted form 113284 various words written with 10 different fonts, 4 different styles,

and 10 different sizes (e.g., 6, 7, 8, 9, 10, 12, 14, 16, 18 and 24 points). The ground truth text of each word image was associated as an XML file. Luqman *et al*. [26] proposed a dataset named King Fahd University Arabic Font Database (KAFD). KAFD consisted of 430 page images and about 10000 line images. These images were organized into training, testing, and validation sets. The text of the KAFD dataset is freely available with 40 different fonts, different sizes (8, 9, 10, 11, 12, 14, 16, 18, 20, and 24 points), different styles (bold, bold-italic, normal, and italic), and different resolutions (100, 200, 300, 600 dpi resolutions). Amara *et al*. [7] developed an Arabic relational database for Arabic OCR systems namely ARABASE. The proposed database consists of digital images of documents, text phrases, word/subwords, isolated characters, digits, and signatures. Table 2 shows the survived datasets that Arabic is one of its languages.

Table 2. Surveyed printed text datasets.

| Authors | Dataset name | Language | Fonts | Font size (point) | Font styles | Images | Resolution (dpi) | Ground truth | Available |
|---|---|---|---|---|---|---|---|---|---|
| Pal and Chaudhuri [32] | - | Printed English, Chinese, Arabic, Devanagari and Bangla | Not disclosed | Not disclosed | Not disclosed | 25000 text line images | Not disclosed | Not disclosed | Not Available |
| Slimane *et al*. [40] | APTI | Printed Arabic | Multi fonts | 6, 7, 8, 9, 10, 12, 14, 16, 18, 24 | Normal, Bold, Italic, Bold and Italic | 45313600 Arabic word images | 72 | Attached | Available |
| Al-Muhtaseb [6] | PATS-A01/PATS-A02 | Printed Arabic | Multi fonts | 18 | Normal | 2766 text line images | 300 | Attached | Available |
| Luqman *et al*. [26] | KAFD | Printed Arabic | Multi fonts | 8, 9, 10, 11, 12, 14, 16, 18, 20, 24 | bold, bold–italic, normal, Italic | 10000 line images | 100, 200, 300, 600 | Attached | Available |
| Chtourou *et al*.[13] | ALTID | Handwritten and printed Arabic/English | Not disclosed | Not disclosed | Not disclosed | Printed: 1845 Arabic text-block images and 2328 English text-block Handwritten: 460 Arabic and 582 English text-block images | 300 | Attached | Available by contact authors |
| Brunessaux *et al*. [10] | Maurdor project | Handwritten and printed French, English, and Arabic | Not disclosed | Not disclosed | Not disclosed | 5000 French text images, 2500 English text images, and 2500 Arabic text images | 300 | Attached | available |
| Chernyshova *et al*. [12] | MIDV-LAIT | Printed Farsi, Arabic, Thai, and Indian | Multi fonts | Multi sizes | Normal, Bold, Italic, | 3600 images | Not disclosed | Not disclosed | Available |

## 2.3. Bilingual Recognition Systems

Tounsi *et al*. [43] proposed a recognizing system for Latin/Arabic text in natural scenes. They employed a standard Bag of Features (BoF) model using SIFT features. For the recognition purposes, they implemented the Hidden Markov Models (HMMs) computations using the HMM HTK toolkit [20]. Two datasets were used in the recognition step; ICDAR03 [25] and ARASTI [42]. The reported mean recognition accuracy was 79.2 % for Arabic and 91.7% for Latin script using hybrid features. Hegghammer [19] reported a benchmarking experiment comparing the performance of Tesseract, Amazon Textract, and Google Document AI on images of English and Arabic text. In their recognition experiments, two datasets were used; Old Books Dataset [8] and Yarmouk Arabic OCR Dataset [16]. Their results showed that Accuracy for English was considerably higher than for Arabic. Natarajan *et al*. [31] introduced a methodology for multilingual offline handwriting recognition using HMMs techniques. The system was built to process the scripts of different languages (i.e., English, Chinese, and Arabic) without the need of pre-segmentation or the need of word and character segmentation. Different datasets were used to evaluate the recognition performance. The IAM dataset [28] was used to evaluate English text, ETL9B corpus [38] was used to evaluate Chinese text, and IFN/ENIT corpus [33] was used to evaluate Arabic text. Lu *et al*. [24] presented a multilingual recognition system to recognize the text of three different languages: Arabic, English, and Chinese. The introduced system used the HMM classifier to build the character models. Eighty features such as intensities of black pixels were extracted using the technique of sliding window.

Bilingual OCR systems other than Arabic/English OCR system were also investigated in order to study their methodologies. Thomas and Venugopal [41] proposed an OCR system for recognizing Malayalam and English words without using script identification method. Two extraction approaches were used; Frequency capture and singular value decomposition. The recognition rates were 98.56% for Malayalam and 98.56% for English. Lehal [22] presented a bilingual

English/Gurmukhi OCR system using script identification technique. Statistical and structural features were used for script identification. A rule based bilingual engine along with language models such as trigram models were used for recognition purposes. The character recognition accuracy was 97.64%. Win *et al*. [44] proposed a bilingual recognition system for printed English/Myanmar text. The proposed system included five main processes: preprocessing, segmentation, feature extraction, classification, and post-processing. Support vector machine was used as a classifier. The reported overall recognition accuracy for the bilingual text was above 90%.

### 2.4. Limitations of Survived Datasets

To the best of our knowledge, there is no bilingual/multilingual OCR database available to the research community that provides bilingual text images at the page level as well as at the line level. Most of the available bilingual datasets are language-specific and provide their images in different sets based on the language of the text images. This behavior requires developing different recognition systems for each language or modifying a recognition system to work with different languages since that these systems are language/script-specific [34]. Another limitation is the size (i.e., number of images) of the datasets. The available bilingual datasets contain a small number of images, which can obstruct and affect the training and testing phases. Image resolution is an important feature in the image processing and text image recognition. Preparing images with several resolutions is challenging since that it required time, effort, and hardware equipment. Most images of the available bilingual dataset are provided with one resolution or low resolution. In addition, many of the available bilingual datasets are domain-specific, which may affect the representations of some characters, and lack statistics.

The lack of the availability of Arabic/English printed text datasets motivates us to prepare, build, and implement a dataset for Bilingual (Arabic/English) Printed Text Images (BPTI). BPTI addresses several obstacles and shortcomings of the available bilingual/multilingual datasets. BPTI has the following characteristics:

1. The elements of the datasets are Arabic, English, and bilingual printed text line images.
2. Two forms of images are included; scanned pages and digitized lines.
3. The Number of images are suitable.
4. The images are prepared with different fonts, multiple font sizes, and different resolutions
5. The ground truth text and statistics are available.
6. The texts of the Arabic images are written in Modern Standard Arabic (MSA).
7. The size of the data is manageable.
8. All possible Arabic letter basic-shapes and English letter cases are included.
9. Freely available.

## 3. BPTI Overview

BPTI is a multi-font, multi-size, and multi-resolution bilingual Arabic/English text images dataset. Text of BPTI is collected from several sources such as electronic books, electronic newspapers and magazines, and common news websites, and included different topics such as religion, history, geography, literature, sports, entertainment, technology, medicine, science, etc.

Two forms of images are included; digitized lines and scanned pages. Digitized images consist of 120 sets; each contains 777 binary images of Arabic, English, and bilingual text lines. Each set is written with one of 10 fonts (Adobe Arabic, Ae AlArabiya, Aljazeera, Calibri, courier New, MS Sans Serif, Simplified Arabic, Tahoma, and Times New Roman) and of four size (12, 14, 16, and 18 points), and prepared with one resolution (100, 200, and 300 dpi). All images are prepared without any font effects and styles such as bold, underline, or italic. The reason of selecting these fonts is that these fonts support Arabic and English text and are commonly used. The total number of the digitized line images in all sets is 93240. The scanned page images consist of 120 sets; each contains different number of page images with similar settings (e.g., font names, font sizes and resolution) to the digitized image sets. The total number of the scanned page images in all sets is 4572. The total number of the BPTI images is 97812. The structure of BPTI dataset is shown in Figure 4.

Table 3 shows statistics related to the two groups of the images of BPTI.

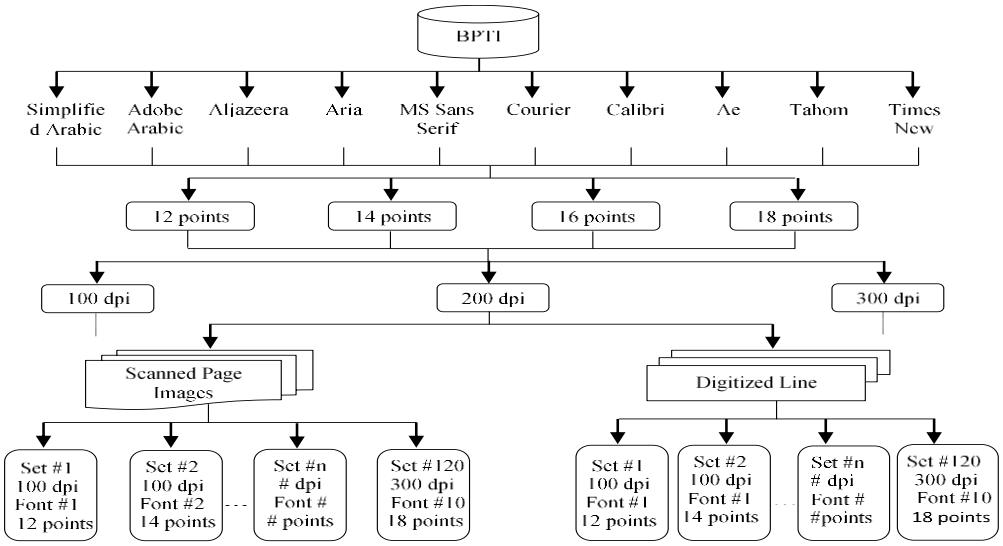


Figure 4. BPTI structure.

Table 3. Statistics of BPTI.

| Font | Scanned Page Images | | | | Digitized Line Images | | | |
|---|---|---|---|---|---|---|---|---|
| | Count | Resolution (dpi) | Font size (Point) | Total | Count | Resolution (dpi) | Font size (Point) | Total |
| **Adobe Arabic** | 34 | 100. 200, 300 | 12, 14, 16, 18 | 408 | 777 | 100. 200, 300 | 12, 14, 16, 18 | 9324 |
| **Ae AlArabiya** | 39 | 100. 200, 300 | 12, 14, 16, 18 | 468 | 777 | 100. 200, 300 | 12, 14, 16, 18 | 9324 |
| **Aljazeera** | 44 | 100. 200, 300 | 12, 14, 16, 18 | 528 | 777 | 100. 200, 300 | 12, 14, 16, 18 | 9324 |
| **Arial** | 36 | 100. 200, 300 | 12, 14, 16, 18 | 432 | 777 | 100. 200, 300 | 12, 14, 16, 18 | 9324 |
| **Calibri** | 37 | 100. 200, 300 | 12, 14, 16, 18 | 444 | 777 | 100. 200, 300 | 12, 14, 16, 18 | 9324 |
| **Courier New** | 36 | 100. 200, 300 | 12, 14, 16, 18 | 432 | 777 | 100. 200, 300 | 12, 14, 16, 18 | 9324 |
| **Microsoft Sans Serif** | 36 | 100. 200, 300 | 12, 14, 16, 18 | 432 | 777 | 100. 200, 300 | 12, 14, 16, 18 | 9324 |
| **Simplified Arabic** | 46 | 100. 200, 300 | 12, 14, 16, 18 | 552 | 777 | 100. 200, 300 | 12, 14, 16, 18 | 9324 |
| **Tahoma** | 37 | 100. 200, 300 | 12, 14, 16, 18 | 444 | 777 | 100. 200, 300 | 12, 14, 16, 18 | 9324 |
| **Times New Roman** | 36 | 100. 200, 300 | 12, 14, 16, 18 | 432 | 777 | 100. 200, 300 | 12, 14, 16, 18 | 9324 |

## 4. Construction of BPTI

As shown in Figure 5, the process of the construction starts with text collection and editing, which creates the ground truth text. For the scanned page images form, constructed pages, from the ground truth text with different fonts and font sizes, are printed and, then, scanned with different resolutions. Different tools and software are used in constructing the images of the digitized lines form.

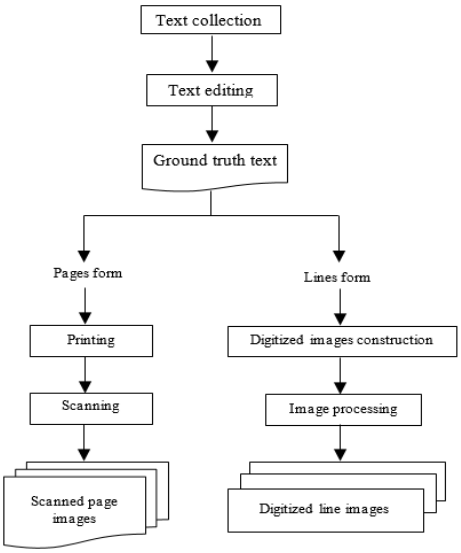


Figure 5. Construction processes of BPTI.

### 4.1. Text Collection

The first task of text collection stage in creating the dataset is collecting Arabic text, English text, and bilingual text from several sources such as electronic books as Express English [5] and common news websites such as Al Arabiya [2] and Al Ekhbaryia Saudi news channel [3].

### 4.2. Text Editing

The second task is to edit the extracted text and to remove diacritics. It includes trimming leading and trailing spaces, removing any two consecutive spaces, removing any non-Unicode characters, and correcting spelling mistakes. As a result of these processes (i.e., text collection and editing), we have prepared 777 bilingual text lines to represent the ground truth file.

#### 4.2.1. Ground Truth Text Description

The 777 text lines that resulted from the text collection stage represent the ground truth in the dataset. All Arabic and English letter shapes, all Arabic and English digits, and common punctuation marks are included in the dataset. Only one space separated each two adjacent words. In addition, no leading or trailing spaces are existing. The text lines include lines with only Arabic text, lines with only English text, and lines with

bilingual Arabic/English text. The statistics (e.g., lines, words, letters, punctuations, and numerals) of the components for each language are presented in Tables 4 and 5. The ground truth contains 410 bilingual text lines, which represents 52.7% of the total lines. Arabic text lines represent 24.97% of the dataset with 194 lines. English text lines represent 22.27% of the dataset with 173 lines.

Table 4. Statistics of arabic components.

| Arabic text | Count | Percentage |
|---|---|---|
| **Lines** | 194 | 24.97% |
| **Words** | 5304 | 59.58% |
| **Letters** | 22957 | 58.33% |
| **Digits** | 356 | 47.79% |
| **Punctuations** | 194 | 14.99% |

Table 5. Statistics of english components.

| English text | Count | Percentage |
|---|---|---|
| **Lines** | 173 | 22.27% |
| **Words** | 3598 | 40.42% |
| **Letters** | 16401 | 41.67% |
| **Digits** | 389 | 52.21% |
| **Punctuations** | 1100 | 85.01% |
| **Upper case letters** | 1673 | 4.25% |
| **Lower case letters** | 14728 | 37.42% |

Table 6. Statistics of Arabic characters (Letters, Digits, and Punctuations).

| Char | Count | % | Char | Count | % | Char | Count | % |
|---|---|---|---|---|---|---|---|---|
| ء | 83 | 0.36 | ـد | 428 | 1.86 | ـق | 151 | 0.66 |
| أ | 325 | 1.42 | ذ | 64 | 0.28 | ـقـ | 183 | 0.8 |
| ـأ | 221 | 0.96 | ـذ | 101 | 0.44 | ق | 40 | 0.17 |
| إ | 91 | 0.4 | ر | 261 | 1.14 | ك | 31 | 0.14 |
| ـإ | 93 | 0.41 | ـر | 622 | 2.71 | ـك | 272 | 1.18 |
| ؤ | 37 | 0.16 | ز | 45 | 0.2 | ـكـ | 238 | 1.04 |
| ـؤ | 35 | 0.15 | ـز | 74 | 0.32 | كـ | 70 | 0.3 |
| ئ | 31 | 0.14 | س | 34 | 0.15 | ل | 145 | 0.63 |
| ـئ | 72 | 0.31 | ـس | 244 | 1.06 | لـ | 1725 | 7.51 |
| ـئـ | 35 | 0.15 | ـسـ | 212 | 0.92 | ـلـ | 626 | 2.73 |
| ئـ | 30 | 0.13 | سـ | 44 | 0.19 | ـل | 244 | 1.06 |
| آ | 30 | 0.13 | ش | 30 | 0.13 | م | 130 | 0.57 |
| ـآ | 39 | 0.17 | ـش | 85 | 0.37 | ـمـ | 620 | 2.7 |
| ا | 1601 | 6.97 | ـشـ | 108 | 0.47 | مـ | 670 | 2.92 |
| ـا | 1465 | 6.38 | شـ | 30 | 0.13 | ـم | 160 | 0.7 |
| ى | 34 | 0.15 | ص | 35 | 0.15 | ن | 196 | 0.85 |
| ـى | 215 | 0.94 | ـص | 101 | 0.44 | ـنـ | 255 | 1.11 |
| ب | 55 | 0.24 | ـصـ | 132 | 0.57 | ـنـ | 396 | 1.72 |
| ـب | 368 | 1.6 | صـ | 30 | 0.13 | نـ | 294 | 1.28 |
| ـبـ | 281 | 1.22 | ض | 30 | 0.13 | ه | 60 | 0.26 |
| بـ | 97 | 0.42 | ـض | 72 | 0.31 | هـ | 222 | 0.97 |
| ت | 116 | 0.51 | ـضـ | 66 | 0.29 | ـهـ | 183 | 0.8 |
| ـت | 335 | 1.46 | ضـ | 31 | 0.14 | هـ | 135 | 0.59 |
| ـتـ | 512 | 2.23 | ط | 31 | 0.14 | و | 529 | 2.3 |
| تـ | 82 | 0.36 | ـط | 64 | 0.28 | ـو | 448 | 1.95 |
| ة | 101 | 0.44 | ـطـ | 125 | 0.54 | ي | 73 | 0.32 |
| ـة | 497 | 2.16 | ـط | 30 | 0.13 | ـي | 511 | 2.23 |
| ث | 35 | 0.15 | ظ | 30 | 0.13 | ـيـ | 603 | 2.63 |
| ـث | 35 | 0.15 | ـظ | 33 | 0.14 | يـ | 394 | 1.72 |
| ـثـ | 110 | 0.48 | ـظـ | 61 | 0.27 | ٪ | 22 | 11.34 |
| ثـ | 38 | 0.17 | ظـ | 30 | 0.13 | ، | 85 | 43.81 |
| ج | 31 | 0.14 | ع | 40 | 0.17 | ؛ | 25 | 12.89 |
| ـج | 177 | 0.77 | ـع | 325 | 1.42 | ؟ | 32 | 16.49 |
| ـجـ | 184 | 0.8 | ـعـ | 414 | 1.8 | ٠ | 38 | 10.67 |
| جـ | 32 | 0.14 | عـ | 86 | 0.37 | ١ | 30 | 15.46 |
| ح | 32 | 0.14 | غ | 30 | 0.13 | ٢ | 48 | 13.48 |
| ـح | 161 | 0.7 | ـغ | 50 | 0.22 | ٣ | 39 | 10.96 |
| ـحـ | 176 | 0.77 | ـغـ | 105 | 0.46 | ٤ | 33 | 9.27 |
| حـ | 31 | 0.14 | غـ | 30 | 0.13 | ٥ | 35 | 9.83 |
| خ | 32 | 0.14 | ف | 72 | 0.31 | ٦ | 34 | 9.55 |
| ـخ | 67 | 0.29 | ـف | 342 | 1.49 | ٧ | 31 | 8.71 |
| ـخـ | 118 | 0.51 | ـفـ | 164 | 0.71 | ٨ | 33 | 9.27 |
| خـ | 30 | 0.13 | فـ | 61 | 0.27 | ٩ | 31 | 8.71 |
| د | 151 | 0.66 | ق | 30 | 0.13 | ٪ | 34 | 9.55 |

Table 6 shows statistics related to all Arabic characters including letters, numerals, and punctuations. Regarding Arabic letters representation, the letter shape (ا) has the maximum occurrences of 1601 times, with 6.97%. The letter shapes (ـئ, آ, ـخ, ـش, ـص, ـض, ـط, ـظ, ـظ, ـغ, ق) have the minimum occurrences of 30 times, with 0.13% for each. Regarding Arabic numerals, the digit (١) has the maximum occurrences of 48 times, with 13.48%. The digits (٦ and ٨) have the minimum occurrences of 31 times, with 8.71% for each. Regarding Arabic punctuations, the mark (٠) has the maximum occurrences of 85 times with 43.81%. The mark (٪) has the minimum occurrences of 22 times with 11.34%. These statistics are calculated according to the same group. For example, the punctuation statistics are calculated based on the total number of punctuation marks.

Table 7 shows statistics related to all English characters including letters, numerals, and punctuations. For English letters, the letter (e) has the maximum occurrences of 2067 with 12.6%. The letters (K, Q, and X) have the minimum occurrence of 30 times with 0.18% for each. Regarding English numerals, the digit (1) has the maximum occurrences of 57 times, with 14.6%. The digit (7) has the minimum occurrences of 33 times, with 8.48% for each. Regarding English punctuations, the mark (,) has the maximum occurrences of 278 times with 25.27%. The marks (@, #, and $) have the minimum occurrences of 20 times, with 1.82% for each. These statistics are calculated according to the same group. For example, the punctuation statistics are calculated based on the total number of punctuation marks.

Table 7. Statistics of english characters.

| Char | Count | % | Char | Count | % | Char | Count | % |
|---|---|---|---|---|---|---|---|---|
| **a** | 1192 | 7.27 | **D** | 42 | 0.26 | **:** | 96 | 8.73 |
| **b** | 206 | 1.26 | **E** | 93 | 0.57 | **?** | 22 | 2.00 |
| **c** | 440 | 2.68 | **F** | 37 | 0.23 | **,** | 278 | 25.27 |
| **d** | 471 | 2.87 | **G** | 43 | 0.26 | **.** | 180 | 16.36 |
| **e** | 2067 | 12.6 | **H** | 57 | 0.35 | **;** | 27 | 2.45 |
| **f** | 268 | 1.63 | **I** | 158 | 0.96 | **'** | 48 | 4.36 |
| **g** | 269 | 1.64 | **J** | 31 | 0.19 | **%** | 22 | 2.00 |
| **h** | 697 | 4.25 | **K** | 30 | 0.18 | **@** | 20 | 1.82 |
| **i** | 1020 | 6.22 | **L** | 58 | 0.35 | **#** | 20 | 1.82 |
| **j** | 39 | 0.24 | **M** | 52 | 0.32 | **$** | 20 | 1.82 |
| **k** | 123 | 0.75 | **N** | 79 | 0.48 | ***** | 24 | 2.18 |
| **l** | 667 | 4.07 | **O** | 82 | 0.50 | **{** | 26 | 2.36 |
| **m** | 391 | 2.38 | **P** | 71 | 0.43 | **}** | 26 | 2.36 |
| **n** | 1020 | 6.22 | **Q** | 30 | 0.18 | **[** | 24 | 2.18 |
| **o** | 1169 | 7.13 | **R** | 38 | 0.23 | **]** | 24 | 2.18 |
| **p** | 231 | 1.41 | **S** | 117 | 0.71 | **/** | 21 | 1.91 |
| **q** | 34 | 0.21 | **T** | 140 | 0.85 | **"** | 42 | 3.82 |
| **r** | 917 | 5.59 | **U** | 34 | 0.21 | **0** | 57 | 14.6 |
| **s** | 968 | 5.90 | **V** | 31 | 0.19 | **1** | 40 | 10.2 |
| **t** | 1237 | 7.54 | **W** | 71 | 0.43 | **2** | 38 | 9.77 |
| **u** | 465 | 2.84 | **X** | 30 | 0.18 | **3** | 38 | 9.77 |
| **v** | 177 | 1.08 | **Y** | 31 | 0.19 | **4** | 37 | 9.51 |
| **w** | 241 | 1.47 | **Z** | 34 | 0.21 | **5** | 34 | 8.74 |
| **x** | 55 | 0.34 | **!** | 21 | 1.91 | **6** | 38 | 9.77 |
| **y** | 327 | 1.99 | **&** | 22 | 2.00 | **7** | 33 | 8.48 |
| **z** | 37 | 0.23 | **+** | 28 | 2.55 | **8** | 35 | 9.00 |
| **A** | 145 | 0.88 | **-** | 27 | 2.45 | **9** | 39 | 10.0 |
| **B** | 62 | 0.38 | **(** | 41 | 3.73 | | | |
| **C** | 77 | 0.47 | **)** | 41 | 3.73 | | | |

## 4.3. Image Construction

The following two sub sections describe the process of creating the two forms of the images of the dataset; the scanned page and digitized line images.

### 4.3.1. Scanned Page Images Construction

To prepare scanned page images of BPTI dataset, we have followed the following procedure:

1. The text of the ground truth are saved as Microsoft Word documents.
2. All word documents are printed with ten fonts and four font sizes (i.e., each file is printed with one font and one font size).
3. The color of the text is black while the color of the background is white.
4. And the resulted printed pages are scanned using a scanner with 100, 200, and 300 dpi resolution and saved as grayscale Tagged Image File Format (TIFF) images. As a results, we have created 93240 scanned page images. Figure 6 shows an example of a scanned page image from the dataset BPTI.

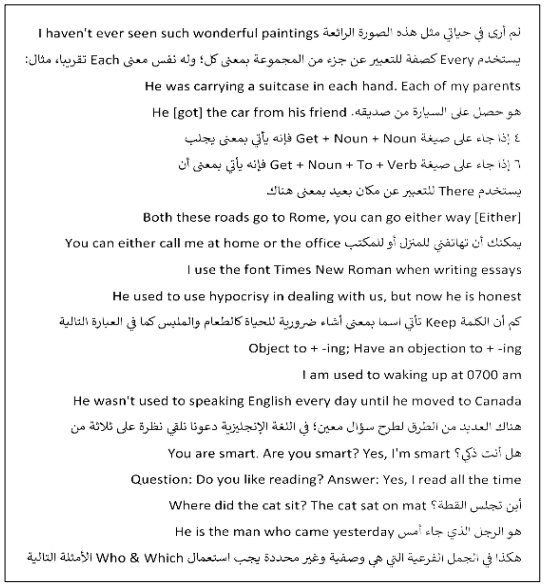


Figure 6. Scanned page image.

### 4.3.2. Digitized Line Images Construction

Line mages of the dataset were created using the ground truth as follows:

1. The 777 text lines were opened with Microsoft Word with ten fonts and four font sizes (i.e., each file is opened with one font and one font size).
2. Each text line is written in one page.
3. The color of the text is white while the color of the background is black.

Table 8 depicts a text sample written with the used 10 fonts. For each font size (12, 14, 16, and 18), we have 10 word files, which are saved as Portable Document Format (PDF) files in order to convert them into TIFF files. Each page, which represents a line text in the PDF files, is converted into separated TIFF images with the following settings:

1. Black and white for the color space.
2. 100, 200, and 300 dpi for the resolution. For each font with one font size and one resolution, 777 text line images are created and categorized as pone set. The resulted TIFF images from the previous stage need to be processed for recognition purposes. This process includes deleting the surrounded background white pixels and binarization. As a result, we have created 93240 binary line images. Figures 7, 8, and 9 depict some of the images that resulted from different stages.

Table 8. Bilingual text images written with 10 fonts.

| Font | Bilingual text |
|---|---|
| **Adobe Arabic** | What is Grammar? ما هو النحو؟ |
| **AeAlArabiya** | What is Grammar? **ما هو النحو؟** |
| **Aljazeera** | What is Grammar? **ما هو النحو؟** |
| **Arial** | What is Grammar? ما هو النحو؟ |
| **Calibri** | What is Grammar? ما هو النحو؟ |
| **Courier New** | What is Grammar? ما هو النحو؟ |
| **Microsoft Sans Serif** | What is Grammar? ما هو النحو؟ |
| **Simplified Arabic** | What is Grammar? ما هو النحو؟ |
| **Tahoma** | What is Grammar? ما هو النحو؟ |
| **Times New Roman** | What is Grammar? ما هو النحو؟ |




Figure 7. Bilingual text from ground truth.


Figure 8. Text line image surrounded by white area resulted from a PDF page.


Figure 9. Binary text line image after removing the surrounded white area.

Figure10 shows an Arabic text image written using Tahoma font. Figure 11 shows an English text image written using Times New Roman font. Figure 12 shows a bilingual text image written using Calibri font.


Figure 10. Arabic text image written with Tahoma font.


Figure 11. English text image written with Times New Roman font.

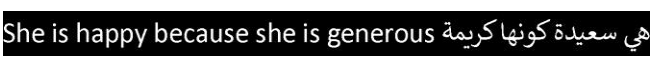


Figure 12. Bilingual text image written with Calibri font.

## 5. Experimental Evaluation

Our proposed BPTI dataset can be used for different research purposes and techniques other than text recognition, such as font recognition, language identification, and page and line segmentation. In order to evaluate the usefulness of the BPTI dataset with some of these techniques, a number of experiments are conducted and their results are reported. This Section presents the computations of the HMM and its parameters and settings, and the features that have been used in the classification experiments of the bilingual text using the scanned page and digitized line images.

### 5.1. HMMs

According to Plötz and Fink [36], HMM is a common classifier used for text recognition. Each shape of Arabic letter is represented by an HMM and the entire text line is represented by composed HMMs. Each letter image is represented by a two-dimension feature vector $O=o_1, o_2, o_3, ..., o_n$, where $o_f$ is a character feature vector observed at frame $f$. The problem of character recognition can be considered as finding a sequence of characters that maximizes the probability $(C_i|O)$ as:

$$\underset{i}{arg\,max} = \{P(C_i|O)\} \quad (1)$$

Where $C_i$ is the ith character and $O$ is observation vectors.

Using Bayes' Rule, the probability of Equation (1), can be computed as:

$$P(C_i|O) = \frac{P(O|C_i)P(C_i)}{P(O)} \quad (2)$$

Where $(O|C_i)$ is a likelihood, $(C_i)$ is the prior probabilities of a character and $(O)$ is the observations probability.

Thus the most probable character, for a given prior probabilities of a character $(C_i)$, depends only on the likelihood function $(O|C_i)$. The probability of generating the observation sequence $O$ by the model $M$ moving through the state sequence $S$ is calculated as the product of the transition probabilities $aij$ and the emission probabilities $bj(oi)$:

$$P(O,S|M) = a_{12}b_2(o_1)a_{22}b_2(o_2)a_{23}b_3(o_3)\cdots \quad (3)$$

Given that $S$ is unknown, the required likelihood is calculated by summing over all possible state sequences $S = s(1), s(2), s(3), ..., s(f)$.

$$P(O|M) = \sum_S a_{s(0)s(1)} \prod_{t=1}^{F} b_{s(f)}(o_f) a_{s(f)s(f+1)} \quad (4)$$

Where $(0)$ is the entry state and $(f+1)$ is the exit state.

The likelihood function can be approximated by only considering the most likely state sequence:

$$\hat{P}(O|M) = \max_S \left\{ a_{s(0)s(1)} \prod_{t=1}^{F} b_{s(f)}(o_f) a_{s(f)s(f+1)} \right\} \quad (5)$$

Where $\hat{P}(O|M)$ is an alternative to Equation in (4).

### 5.2. Codebook Sizes and Number of States

We have conducted different experiments with different combinations of codebook sizes and number of states. A codebook is a collection of nodes where each node represents a set of observation vectors. A state is an observations vector of a vertical segment of a given text. The range of the used codebook sizes was between 40 and 350 with step 10 (e.g., 40, 50,…, 340, and 350). The range of the used number of states was between 4 and 20.

### 5.3. Feature Extraction and Sliding Window

We use statistical information (e.g., pixels' density) as feature type with a sliding window as a feature extraction technique. A sliding window consists of a specified number of blocks of predefined width and height stacked vertically. The feature extraction process works by sliding a window, horizontally, from the beginning of the text line image to the end in order to extract features. These features include, for the current position of the sliding window over the text image, the summation of white pixels in each block, the summation of white pixels in each two consecutive blocks, the summation of white pixels in all blocks, and the summation of black pixels in all blocks. Figure 13 shows the concept of the sliding window; segments (A), (C), and (D) are examples of a sliding window at different positions on a line image. Segment (B) represents two-overlapped windows.

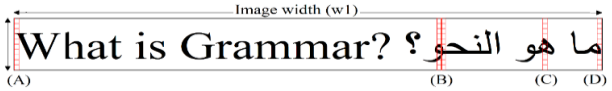


Figure 13. A sliding window of 10 vertical blocks.

### 5.4. Performance Metrics

We use two metrics in order to evaluate the recognition performance: Correctness and Accuracy. These two metrics are based on counting of total samples (*samples*), substitutions (*Sub*), insertions (*Ins*), and deletions (*Del*) of samples. The equations of these metrics are defined as follow:

$$Correctness = \frac{Samples - Sub - Del}{Samples} \times 100 \quad (6)$$

$$Accuracy = \frac{Samples - Sub - Ins - Del}{Samples} \times 100 \quad (7)$$

### 5.5. Experimental Results Using Scanned Page Images

In these experiments, we used all pages with 300 dpi resolution, 10 fonts, and 18 points for the font size. As a result, we have 381 page images. We have segmented these images by specifying the centroids of the blank

areas between the text lines. This method depends on, first, computing the horizontal projections of the inverted binary page image. Then, finding the zeros projections, which represent the blank regions between the text lines. Third, computing the centroids of the blank regions. Each centroid represents the end of a text line image and the start of the next line image. The regions located between these two centroids are segmented out of the image page as a text line image. As a result of the segmentation process, we have 777 text line images for each font. We have selected 155 text line images randomly to represent the testing group. The remaining unselected 622 text line images are used for training the HMM. The generated testing group was used in each single font classification experiments. Table 9 shows the highest recognition rates of each single-font text recognition experiments. The achieved highest correctness was 97.86% and the highest accuracy was 97.73% for Tahoma font. These rates were achieved using a codebook size of 340 and 17 HMM states.

Table 9. Correctness and accuracy of single font experiments in the scanned page images.

| Font | CB Size | States | Correctness % | Accuracy % |
|---|---|---|---|---|
| **Adobe Arabic** | 202 | 9 | 96.1 | 95.76 |
| **Ae Alarabiya** | 119 | 14 | 97.18 | 96.99 |
| **Aljazeera** | 170 | 11 | 96.45 | 96.12 |
| **Arial** | 140 | 15 | 96.86 | 96.3 |
| **Calibri** | 155 | 14 | 95.14 | 95 |
| **Courier New** | 170 | 18 | 97.31 | 97.01 |
| **Microsoft Sans Serif** | 310 | 14 | 97.82 | 97.7 |
| **Simplified Arabic** | 171 | 13 | 97.79 | 97.69 |
| **Tahoma** | 340 | 17 | 97.86 | 97.73 |
| **Times New Roman** | 114 | 16 | 97.06 | 96.95 |

## 5.6. Experimental Results Using Digitized Line Images

For the training and testing in the classification of the digitized line images, we have used 10 sets from the digitized line images form. Each set is prepared with 300 dpi resolution, written with one of 10 fonts, and 18 points for the font size. We have selected 155 text line images randomly to represent the testing group. The remaining unselected 622 text line images represent the training group and are used to train the HMM. The generated testing group was used in each single font classification experiments. The highest results for each font are summarized in Table 10. One main reason for the high achieved rates is that these digitized text images, which are created by a computer, are free from noise and also free from any degrees of skew (tilt). In addition, the used technique bypasses the need for segmentation of Arabic/English text on character or word levels, which is error-prone. Finally, feature extracting technique was designed and implemented to extract features that could represent, accurately, Arabic and English characters.

Table 10. Correctness and accuracy of single font experiments in the digitized line images.

| Font | CB Size | States | Correctness % | Accuracy % |
|---|---|---|---|---|
| **Adobe Arabic** | 110 | 7 | 98.43 | 98.39 |
| **Ae Alarabiya** | 330 | 11 | 98.89 | 98.89 |
| **Aljazeera** | 160 | 12 | 97.98 | 97.76 |
| **Arial** | 150 | 12 | 98.77 | 98.64 |
| **Calibri** | 170 | 8 | 98.36 | 98.21 |
| **Courier New** | 150 | 20 | 98.63 | 98.63 |
| **Microsoft Sans Serif** | 310 | 12 | 98.67 | 98.67 |
| **Simplified Arabic** | 100 | 8 | 97.64 | 97.33 |
| **Tahoma** | 320 | 15 | 99.01 | 99.01 |
| **Times New Roman** | 130 | 12 | 98.87 | 98.86 |

## 5.7. Language Identification

Language identification could be used in multilingual OCR systems to identify used languages to enhance preprocessing tasks, feature selection, and increase recognition rates. In the language identification experiments, we used the same training and testing sets that were used in the classification of the digitized line images. The HMM classifier is used to detect the language of text images at the word level. The ground truth text of our datasets was modified to represent the language of the character instead of the character itself. We used "A" for an Arabic character and "E" for an English character. Figure 15 shows the coded ground truth of the text line image shown in Figure 14 using the new coding method. Then, we train the system to identify the language of each character using the same modules of our recognition system. Finally, a number of language identification experiments were conducted using the images of the testing set.


Figure 14. Example of a Bilingual text line image.

EEESEEEEEESEEEESEESEEEESEESEEEESAAAASAAAAAASAASAAAAAAAASAAAAA

Figure 15. Updated ground truth of the example of Figure 14.

Table 11 shows the average rates of identifying word-language per used font. The average rate is computed by dividing the count of correct identified words by the total words. Word-language identification average is computed by calculating the average of all testing group rates for a font. The highest rate was 99.98% for Tahoma font.

Table 11. Word-Language identification rates.

| Font | Word-Language Identification Rate % | Font | Word-Language Identification Rate % |
|---|---|---|---|
| **Adobe Arabic** | 99.88 | **Courier New** | 99.84 |
| **AE Alarabiya** | 99.92 | **Microsoft Sans Serif** | 99.87 |
| **Aljazeera** | 99.94 | **Simplified Arabic** | 99.79 |
| **Arial** | 99.75 | **Tahoma** | 99.98 |
| **Calibri** | 99.60 | **Times New roman** | 99.81 |

## 5.8. Tools and Programming Code

All tools and programming code used in this research work can be reached through the following link:

https://drive.google.com/drive/folders/1iFt12jrkbZDR0xtRoQYp-akmuZpjm5G0

## 6. Conclusions

In this paper, we presented a bilingual dataset namely BPTI. The dataset BPTI consists of two forms of images; Scanned page and digitized line images. The scanned page images consist of 120 sets and a total of 4572 images. The digitized images consist of 120 sets; each contains 777 binary images of Arabic, English, and bilingual text lines with a total of 93240 images. For both forms, each set is written with one of 10 fonts and one of four sizes (12, 14, 16, and 18 points), and prepared/scanned with one resolution (100, 200, and 300 dpi). Statistics, which are related to the ground truth and images, were presented. In addition, the ground truth text, which represents the text of the images, is available. BPTI is available freely throughout communicating with the 1st author of this research work. To assure the usefulness of the dataset, a number of recognition, segmentation, and language identification experiments were conducted. For the digitized text recognition experiments, Tahoma has shown the highest performance with 99.01% for correctness and 99.01% for accuracy. For the scanned text recognition experiments, Tahoma has also shown the highest performance with 97.86% for correctness and 97.73% for accuracy. For the language identification experiments, Tahoma has shown the performance with 99.98% for word-language identification rate.

## Acknowledgement

The authors wish to thank Al Hussein Technical University (HTU) and King Fahd University of Petroleum and Minerals (KFUPM).

## References

[1] AbdelRaouf A., Higgins C., and Khalil M., "A database for Arabic Printed Character Recognition," *in Proceedings of the International Conference on Image Analysis and Recognition*, Povoa de Varzim, pp. 567-578, 2008. Doi: 10.1007/978-3-540-69812-8_56

[2] Al Arabiya Middle East Broadcasting Center MBC, Available: https://www.alarabiya.net, Last Visited, 2022.

[3] Al Ekhbariya Saudi News Channel, Available: http://www.alekhbariya.net, Last Visited, 2022.

[4] Al Maadeed S., Ayouby W., Hassaine A., and Aljaam J., "QUWI: An Arabic and English Handwriting Dataset for Offline Writer Identification," *in Proceedings of the International Conference on Frontiers in Handwriting Recognition*, Bari, pp. 746-751, 2012. Doi: 10.1109/ICFHR.2012.256

[5] AL-Hourani O., Express English, Available: http://www.expenglish.com, Last Visited, 2022.

[6] Al-Muhtaseb H., Arabic Text Recognition of Printed Manuscripts, PhD Theses University of Bradford, 2010.https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.529712

[7] Amara N., Mazhoud O., Bouzrara N., and Ellouze N., "ARABASE: A relational Database for Arabic OCR Systems," *The International Arab Journal of Information Technology*, vol. 2, no. 4, pp. 259-266, 2005.

[8] Barcha P., PedroBarcha/old-books-dataset, UNICAMP, University of Campinas, Brazil, Available: https://github.com/PedroBarcha/old-books-dataset. Last Visited, 2022.

[9] Bartos G., Hoşcan Y., Kauer A., and Hajnal É., "A multilingual Handwritten Character Dataset: THE Dataset," *Acta Polytechnica Hungarica*, vol. 17, no. 9, pp. 141-160, 2020. DOI: 10.12700/APH.17.9.2020.9.8

[10] Brunessaux S., Giroux P., Grilheres B., Manta M., Bodin M., Choukri K., Galibert O., and Kahn J., "The Maurdor Project: Improving Automatic Processing of Digital Documents," *in Proceedings of the 11th IAPR International Workshop on Document Analysis Systems*, Tours, pp. 349-354, 2014. DOI: 10.1109/DAS.2014.58

[11] Chanda S., Pal U., and Terrades O., "Word-wise Thai and Roman Script Identification," *ACM Transactions on Asian Language Information Processing*, vol. 8, no. 3, pp. 1-21, 2009. DOI: 10.1145/1568292.1568294

[12] Chernyshova Y., Emelianova E., Sheshkus A., and Arlazarov V., "MIDV-LAIT: A challenging Dataset for Recognition of IDs with Perso-Arabic, Thai, and Indian Scripts," *in Proceedings of the 16th International Conference on Document Analysis and Recognition*, Lausanne, pp. 258-272, 2021. https://doi.org/10.1007/978-3-030-86331-9_17

[13] Chtourou I., Rouhou A., Jaiem F., and Kanoun S., "ALTID: Arabic/Latin Text Images Database for Recognition Research," *in Proceedings of the 13th International Conference on Document Analysis and Recognition*, Tunis, pp. 836-840, 2015. DOI: 10.1109/ICDAR.2015.7333879.

[14] Dhanya D., Ramakrishnan A., and Pati P., "Script Identification in Printed Bilingual Documents," *Sadhana*, vol. 27, no. 1, pp. 73-82, 2002. DOI: 10.1007/BF02703313

[15] Djeddi C., Gattal A., Souici-Meslati L., Siddiqi I., Chibani Y., and El Abed H., "LAMIS-MSHD: A multi-script Offline Handwriting Database," *in Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition*, Hersonissos, pp. 93-97, 2014. DOI: 10.1109/ICFHR.2014.23

[16] Doush I., AIKhateeb F., and Gharibeh A., "Yarmouk Arabic OCR Dataset," *in Proceedings of the 8th International Conference on Computer Science and Information Technology*, Amman, pp. 150-154, 2018. DOI: 10.1109/CSIT.2018.8486162

[17] Hamdi A., Pontes E., Boros E., Nguyen T., Hackl G., Moreno J., and Doucet A., "A multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers," *in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Virtual Event, pp. 2328-2334, 2021. https://doi.org/10.1145/3404835.3463255

[18] Hassan E., Garg R., Chaudhury S., and Gopal M., "Script Based Text Identification: A multi-level Architecture," *in Proceedings of the Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data*, Beijing, pp. 1-8, 2011. https://doi.org/10.1145/2034617.2034630

[19] Hegghammer T., "OCR with Tesseract, Amazon Textract, and Google Document AI: A Benchmarking Experiment," *Journal of Computational Social Science*, vol. 5, no. 1, pp. 861-882, 2022. DOI:10.31235/osf.io/6zfvs

[20] HTK3, Cambridge University Engineering Department, Available: http://htk.eng.cam.ac.uk, Last Visited, 2022.

[21] Khoddami M. and Behrad A., "Farsi and Latin Script Identification Using Curvature Scale Space Features," *in Proceedings of the 10th Symposium on Neural Network Applications in Electrical Engineering (NEUREL)*, Belgrade, pp. 213-217, 2010. Doi: 10.1109/NEUREL.2010.5644061

[22] Lehal G., "A Bilingual Gurmukhi-English OCR Based on Multiple Script Identifiers and Language Models," *in Proceedings of the 4th International Workshop on Multilingual OCR*, Washington, pp. 1-5, 2013. https://doi.org/10.1145/2505377.2505381

[23] Lin X., Guo C., and Chang F., "Classifying Textual Components of Bilingual Documents with Decision-Tree Support Vector Machines," *in Proceedings of International Conference on Document Analysis and Recognition*, Beijing, pp. 498-502, 2011. DOI: 10.1109/ICDAR.2011.106

[24] Lu Z., Bazzi I., Kornai A., Makhoul J., Natarajan P., and Schwartz R., "Robust Language-Independent OCR System," *in Proceedings of the 27th AIPR Workshop: Advances in Computer-Assisted Recognition*, Washington, pp. 96-104, 1999. Doi: 10.1117/12.339811

[25] Lucas S., Panaretos A., Sosa L., Tang A., Wong, S., and Young R., "ICDAR 2003 Robust Reading Competitions," *in Proceedings of the 7th International Conference on Document Analysis and Recognition*, Edinburgh, pp. 682-687, 2003. Doi: 10.1109/ICDAR.2003.1227749.

[26] Luqman H., Mahmoud S., and Awaida S., "KAFD Arabic Font Database," *Pattern Recognition*, vol. 47, no. 6, pp. 2231-2240, 2014. https://doi.org/10.1016/j.patcog.2013.12.012

[27] Mahmoud S., Ahmad I., Alshayeb M., Al-Khatib W., Parves M., Fink G., Margner V., and Abed H., "Khatt: Arabic Offline Handwritten Text Database," *in Proceedings of International Conference on Frontiers in Handwriting Recognition*, Bari, pp. 449-454, 2012. DOI: 10.1109/ICFHR.2012.224

[28] Marti U. and Bunke H., "The IAM-Database: An English Sentence Database for Offline Handwriting Recognition," *International Journal on Document Analysis and Recognition*, vol. 5, no. 1, pp. 39-46, 2002. DOI:10.1007/s100320200071

[29] Mathew M., Singh A., and Jawahar C., "Multilingual OCR for Indic Scripts," *in Proceedings of the 12th IAPR Workshop on Document Analysis Systems*, Santorini, pp. 186-191, 2016. DOI: 10.1109/DAS.2016.68.

[30] Mezghani A., Kanoun S., Khemakhem M., and El Abed H., "A Database for Arabic Handwritten Text Image Recognition and Writer Identification," *in Proceedings of the International Conference on Frontiers in Handwriting Recognition*, Bari, pp. 399-402, 2012. DOI: 10.1109/ICFHR.2012.155

[31] Natarajan P., Saleem S., Prasad R., MacRostie E., and Subramanian K., "Multi-lingual Offline Handwriting Recognition Using Hidden Markov Models: A Script-Independent Approach," *in Proceedings of the Arabic and Chinese Handwriting Recognition*, College Park, pp. 231-250, 2006. https://doi.org/10.1007/978-3-540-78199-8_14

[32] Pal U. and Chaudhuri B., "Automatic Identification of English, Chinese, Arabic, Devnagari and Bangla Script Line," *in Proceedings of the 6th International Conference on Document Analysis and Recognition*, Seattle, pp. 790-794, 2001. DOI: 10.1109/ICDAR.2001.953896

[33] Pechwitz M., Maddouri S., Margner V., Ellouze N., and Amiri H., "IFN/ENIT-dtabase of Handwritten Arabic Words," *in Proceedings of the CIFED*, Hammamet, pp. 127-136, 2002. https://www.researchgate.net/publication/228904501_IFNENIT-database_of_handwritten_Arabic_words

[34] Peng X., Cao H., Setlur S., Govindaraju V., and Natarajan P., "Multilingual OCR Research and Applications: An Overview," *in Proceedings of the 4th International Workshop on Multilingual OCR*, Washington, pp. 1-8, 2013. https://doi.org/10.1145/2505377.2509977

[35] Philip B. and Samuel R., "A Novel Bilingual OCR for Printed Malayalam-English Text Based on Gabor Features and Dominant Singular Values," *in Proceedings of International Conference on Digital Image Processing*, Bangkok, pp. 361-365, 2009. DOI: 10.1109/ICDIP.2009.50

[36] Plötz T., and Fink G., "Markov Models for Offline Handwriting Recognition: A Survey," *International Journal on Document Analysis and Recognition*, vol. 12, no. 4, pp. 269-298, 2009. DOI: 10.1007/s10032-009-0098-4

[37] Rani R., Dhir R., and Lehal G., "Performance Analysis of Feature Extractors and Classifiers for Script Recognition of English and Gurmukhi Words," *in Proceedings of the workshop on Document Analysis and Recognition*, Mumbai, pp. 30-36, 2012. DOI: 10.1145/2432553.2432559

[38] Saito T., "On the Data Base ETK9B of Handprinted Characters in JIS Chinese Characters and its Analysis," *IEICE Trans*, vol. 68, no. 4, pp. 757-772, 1985.

[39] Saudi Press Agency, Available: https://www.spa.gov.sa/viewstory.php?lang=ar&newsid=941565, Last Visited, 2021.

[40] Slimane F., Ingold R., Kanoun S., Alimi A., and Hennebert J., "A New Arabic Printed Text Image Database and Evaluation Protocols," *in Proceedings of the 10th International Conference on Document Analysis and Recognition*, Barcelona, pp. 946-950, 2009. DOI: 10.1109/ICDAR.2009.155

[41] Thomas B., and Venugopal C., "Bilingual Malayalam English OCR System Using Singular Values and Frequency Capture Approach," *in Proceedings of International Conference of Advances in Computing, Communication and Control*, Mumbai, pp. 372-377, 2011. DOI: 10.1007/978-3-642-18440-6_47

[42] Tounsi M., Moalla I., and Alimi A., "ARASTI: A Database for Arabic Scene Text Recognition," *in Proceedings of the 1st International Workshop on Arabic Script Analysis and Recognition*, Nancy, pp. 140-144, 2017. DOI: 10.1109/ASAR.2017.8067776

[43] Tounsi M., Moalla I., Pal U., and Alimi A., "Arabic and Latin Scene Text Recognition by Combining Handcrafted and Deep-Learned Features," *Arabian Journal for Science and Engineering*, vol. 47, pp. 9727-9740, 2022. DOI: 10.1007/s13369-021-06311-1

[44] Win H., Khine P., and Tun K., "Bilingual OCR System for Myanmar and English Scripts with Simultaneous Recognition," *in Proceedings of the International Journal of Scientific and Engineering Research*, vol. 2, no. 10, 2011.

**Mohammad Yahia** received his B.Sc. degree in Computer Science from Isra University (IU), Amman, Jordan in 1996, the M.Sc. degree in Computer Science from King Fahd University of Petroleum & Minerals (KFUPM), Dhahran, Saudi Arabia, in 2012, and the Ph.D. degree in Computer Science and Engineering from KFUPM in 2018. From 2013 to 2021, Dr. Yahia was a lecturer at the Information and Computer Science Department in KFUPM. Dr. Yahia Joined the Deanship of Scientific Research at KFUPM as a Research Specialist from 2019 to 2020. Currently, he is an Assistant Professor at the School of Computing and Informatics in AL Hussein Technical University, Amman, Jordan. Dr. Yahia received the Excellence Teaching Award at the lecturer level in King Faisal University, Hofuf, Saudi Arabia, 2010. During his academic career, Dr. Mohammad Yahia participated in many academic committees in different universities and organizations concerned with developing and improving public education in Saudi Arabia. His research interest includes Arabic Computing, Natural Language Processing, Artificial Intelligence, and Machine Learning. Dr. Yahia is an active participant in different community services and charity works as a volunteer.

**Husni Al-Muhtaseb** received his Ph.D. degree from the Department of Electronic Imaging and Media Communications (EIMC) of the School of Informatics in the University of Bradford, UK in 2010. He received his M.S. degree in Computer Science and Engineering from King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia, in 1988 and the B.E. degree in Electrical Engineering, computer option, from Yarmouk University, Irbid, Jordan in 1984. He is currently an Assistant Professor of Information and Computer Science at KFUPM. Dr. Husni Al-Muhtaseb is a member of Association of Jordanian Engineers, Electrical Engineering Division and Saudi Computer Society. Dr. Al-Muhtaseb developed the first course in Arabization of Computers in the world. The course is now being taught in 10s of universities and colleges. Dr. Al-Muhtaseb has participated in several industrial projects with different institutes/organizations including KACST, STC, MOHE and Aramco. He also worked as a consultant for different entities including KFUPM schools and Ministry of Education. Dr. Husni Al-Muhtaseb has more than 50 research publications. He got the first excellence award in instructional Technologies at KFUPM for year 2007. His Research Interests include: Computer Arabization, Arabic Computing, English-Arabic Machine Translation, Arabic Text Vocalization, Arabic Text recognition, OCR, Arabic Speech Synthesis and Recognition, Evaluation of Arabic Software, Software Development, E-learning and Instructional Technologies.