A Real-time DSP-Based Optical Character Recognition System for Isolated Arabic characters using the TI TMS320C6416T

Haidar Almohri
University of Hartford
almohrih@hotmail.com

John S. Gray
University of Hartford
gray@mwd.hartford.edu

Hisham Alnajjar, PhD
University of Hartford
alnajjar@hartford.edu

Abstract

Optical Character Recognition (OCR) is an area of research that has attracted the interest of researchers for the past forty years. Although the subject has been the center topic for many researchers for years, it remains one of the most challenging and exciting areas in pattern recognition. Since Arabic is one of the most widely used languages in the world, the demand for a robust OCR for this language could be commercially valuable. There are varieties of software based solutions available for Arabic OCR. However, there is little work done in the area of hardware implementation of Arabic OCR where speed is a factor. In this research, a robust DSP-based OCR is designed for recognition of Arabic characters. Since the scope of this research is focused on hardware implementation, the system is designed for recognition of isolated Arabic characters. An efficient recognition algorithm based on feature extraction and using a Fuzzy ART Neural Network as well as the hardware implementation is also proposed in this research. A recognition rate of 95% is reported.

Introduction

Optical Character Recognition, usually referred to as OCR, is the process of converting the image obtained by scanning a text or a document into machine-editable format. OCR is one of the most important fields of pattern recognition and has been the center of attention for researchers in the last forty decades [1].
The goal is to process data that normally is processed only by humans with computers. One of the apparent advantages of computer processing is dealing with huge amounts of information at high speed [2]. Some other advantages of OCR are: reading postal address off envelopes, reading customer filled forms, archiving and retrieving text, digitizing libraries … etc. Using OCR, the handwritten and typewritten text could be stored into computers to generate databases of existing texts without using the keyboard.
The modern version of OCR appeared in the middle of the 1940's with the development of the digital computers [3]. Since then several character recognition systems for English,

Chinese and Japanese characters have been proposed [4, 5, 6]. However, developing OCR systems for other languages such as Arabic didn't receive the same amount of attention. Arabic is the official language of all countries in North Africa and most of the countries in the Middle East and is spoken by 234 million people [7, 8]. It is the sixth most commonly used language in the world. When spoken Arabic varies across regions, but written Arabic, sometimes called "Modern Standard Arabic" (MSA), is a standardized version used for official communication across the Arab world [9]. The characters of Arabic script and similar characters are used by a greater percentage of the world's population to write languages such as Arabic, Farsi (Persian), and Urdu [10]. Therefore, an efficient way to automate the process of digitizing the Arabic documents such as books, articles, etc. would be highly beneficial and commercially valuable. A 2006 survey cites that the first modern Arabic OCR approach took place in 1981 where Parhami and Taraghi presented their algorithm which achieved a character recognition rate of 85 percent. Since then, many attempts have been taken place and there have been numbers of commercially OCR products available in the market.

However, there is little effort done in implementing a hardware-based Arabic OCR device that has a small foot print and could be easily transported.
This research aims to design and implement an efficient hardware-based OCR using image processing and DSP techniques. The advantages of this OCR system include, but not limited to the followings:

- Small footprint
- Light and easy to carry
- Low power consumption
- High speed performance

Characteristics of Arabic Script

One of the reasons for slow advancements in Arabic OCRs is the characteristics of this script that makes it more challenging than other languages. Some of these characteristics are listed below:

- The Arabic script is cursive
- Characters can have different shapes in different positions of a word
- Most letters have one, two, or three dots
- A word is composed of sub-word (s)

In addition to the above characteristics, the Arabic font is written-read from right to left. These characteristics have made the progress of Arabic OCR more complex and difficult than other languages.

Preview of Existing Work

Typewritten vs. Handwritten Recognition

The problem of character recognition can be divided into two major categories: typewritten and handwritten. As their names describe their natures,typewritten recognition recognizes a document that has been previously typed and scanned prior to recognition progress. Such a system would be used as a way to digitize books, documents and papers in libraries, government, or held by companies. In handwritten recognition, the system attempts to recognize a text that has been written by a human (not a machine). This is usually more difficult as there is no standard way of writing and the handwriting of each person is different than the other. As a result, the recognition rate achieved for handwritten recognition systems is less than the typewritten.

Offline vs. Online Text Recognition

Character recognition systems may be further categorized to offline and online recognition systems. In offline recognition, the image of the type or handwritten text is acquired through scanning using an optical scanner. The image then is read by the system and is analyzed for recognition. In online recognition systems, input is an image of a hand-printed text which is usually acquired from a tablet computer or pen-based devices such as cell phone and sign pad. Online recognition is a fast growing technique for convenient human computer interface and it has a lot of advantages. For example, it can be used to help people such as computer novices, elderly people and house wives to conveniently use a computer. Additionaly, it makes a small size portable computer (PDA, handheld PC, palm PC, etc.) possible because there is no need for keyboard or keypad.

In this research, the developed system is designed for typewritten, offline character recognition; therefore the discussion will be focused on this area.

Basic OCR System's Architecture

Any offline OCR system contains of all or part of the following steps:

- Image Acquisition
- Preprocessing
- Line Segmentation
- Word Segmentation
- Character Segmentation
- Recognition
- Post Processing

Proposed Algorithm

Figure 1 shows the block diagram of the proposed algorithm.

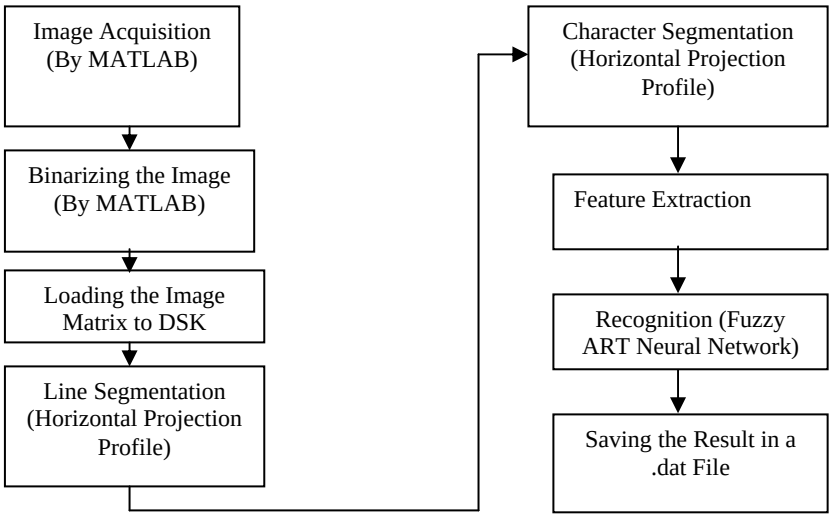

Figure 1: The block diagram of the proposed algorithm

Image Acquisition

The process starts by acquiring the image. Text is scanned using a 300 dpi scanner and the image of the text is saved as a .bmp file in a computer running MATLAB. MATLAB is used to read the image and convert it to black and white format. Using MATLAB, the pixel values of the binary image (represented as 0 or 1) are saved in a text file. The pixel values are then used to create a header file to represent the image in the main program.

As shown in figure 2, the scanned image always contains noise that usually appears as an extra pixel (black or white) in the character image. If the noise is not taken into consideration, it could subvert the process and produce an incorrect result.

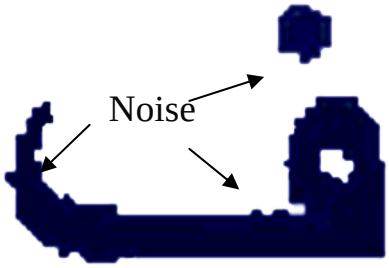

Figure 2: Letter “faa” corrupted by noise

Line Segmentation

When the image matrix is ready to be processed, the first step is to isolate each line of the text from the whole document. A horizontal projection profile technique is used for this purpose. A computer program scans the image horizontally to find the first and last black pixels in a line. Once these pixels are found, the area in between these pixels represents the line that may contain one or more character. Using the same technique, the whole document is scanned and each line is detected and saved in a temporary array for further processing.

Character Segmentation

Once each line of the text is stored in a separate array, using vertical projection profile, the program scans each array this time vertically to detect and isolate each character within each line. The first and last black pixels that are detected vertically are the borders of the character. It possible that when the characters are segmented, there is a white area above, below, or both above and below the character, except for the tallest character that its height is equal to the height of the line. Since the edges of each character box is needed for the recognition purpose, another horizontal scan is run to detect the top and bottom of the character and isolate the area that only contains the pixels of the character.

Feature Extraction and Recognition

At this point, the program has isolated each character in the document and the matrix representation of each character is ready to be processed for recognition purpose. In this research, several methods were examined to find the most suitable method for recognition. Several factors determine the efficiency of the recognition algorithm. The most important factors are the speed of the process and the accuracy of the result.

Feature Extraction: As discussed earlier, at the time of processing, a matrix of pixel values which contains the four boarders of each character image is extracted by the program and has been recognized in a manner similar to that shown in figure 3.

Figure 3: Extracted characters

Feature selection is one of the most critical issues in character recognition as the recognition rate is dependent on the choice of features. Every character has some features that distinguish it from the other characters. Some of the commonly used features for character recognition are loops, holes, strokes, vertical lines, cusps, etc. The majority of previous works focuses on these features, as they appeal to the human intuitive logic. Unfortunately, techniques using these features suffer a common drawback, namely, exhaustive processing time. The solution to this lies in the selection of an algorithm which effectively reduces image processing time while not compromising its accuracy.

An optimal selection of features, which categorically defines the details of the character and does not take a long processing time, is implemented to extract features from the character to be recognized prior to recognition.
This way, each character is distinguished by a set of features which are unique for the character. This information is used to train the Neural Network to learn and use these features to find the result rather than inputting all the pixel values for each character. The features extracted have the following properties:

- Easy to extract, which reduces the complexity of the program.
- Distinct, which eases the Neural Network's recognition process.
- Independent of font type and size, which is a big advantage since the system is capable of recognizing any font type with any size.

There are 14 features extracted from the character of which 4 of them are for the whole image as listed below:

1. Height / Width
2. number of black pixels / number of white pixels image
3. number of horizontal transitions
4. number of vertical transitions

The horizontal and vertical transition is a technique used to detect the curvature of each character and found to be effective for this purpose. The procedure runs a horizontal scanning through the character box and finds the number of times that the pixel value changes state from 0 to 1 or from 1 to 0 as shown in figure 8. The total number of times that the pixel status changes, is its horizontal transition value. Similar process is used to find the vertical transition value.

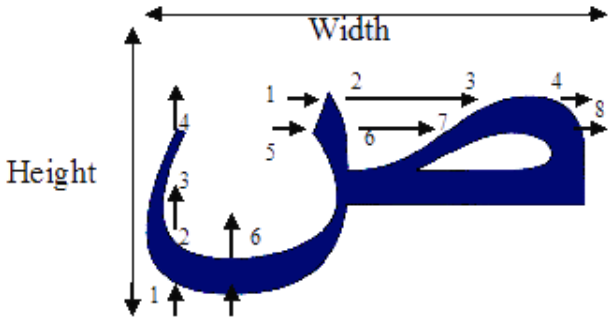

Figure 4: Horizontal and vertical transitions

In addition, the image is divided into four regions as shown in figure 9 and the following features are extracted from these regions:

1. Black Pixels in Region 1/White Pixels in Region 1
2. Black Pixels in Region 2/White Pixels in Region 2
3. Black Pixels in Region 3/White Pixels in Region 3
4. Black Pixels in Region 4/White Pixels in Region 4
5. Black Pixels in Region 1/Black Pixels in Region 2
6. Black Pixels in Region 3/Black Pixels in Region 4
7. Black Pixels in Region 1/Black Pixels in Region 3
8. Black Pixels in Region 2/Black Pixels in Region 4
9. Black Pixels in Region 1/Black Pixels in Region 4
10. Black Pixels in Region 2/Black Pixels in Region 3

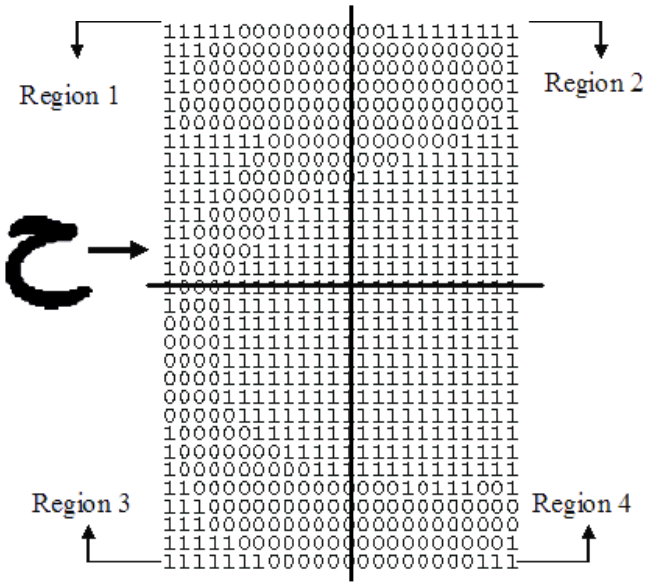

Figure 5: Dividing the image to 4 regions and extracting features

These features were found to be sufficient to distinguish between different characters. The extracted feature vector is to train the Neural Network.

ART Neural Network: A training set database has to be generated for the network to be trained. 700 sample characters chosen from the most popular Arabic fonts and sizes are used to generate the database. The 14 features described previousely are extracted from this set of characters using MATLAB and the results are saved in a text file which could be used by Professional II/PLUS as the training set.

The Fuzzy ART neural network has the following architecture: 14 Input, 1 Output, 60 F2 Layer, 0.0000 Vigilance. The network is trained for 20,000 times and tested using different samples to calculate the performance of the network. The test results showed that the network was able to predict about 95% of the input characters correctly. This accuracy range is an average as it varies depending on the resolution of the image, and the font type and size. If the input font is the same as the fonts available in the database (which are used to train the network) the accuracy goes up to 98% recognition, but if the font is unknown for the network, the error level increases yielding an accuracy of 92%. Since most of the popular Arabic fonts are defined in the training set, the network should be able to achieve a high accuracy rate in most cases.

After the network is fully trained and tested and a satisfactory result has been achieved, the C source code is generated using *flashcode* option which is a utility available in Professional II/PLUS.

Hardware Implementation and Results

This project is fully written in the C programming language using Code Composer Studio which is a fully integrated development environment (IDE) supporting Texas Instruments industry-leading DSP platforms. A C6416T DSK, which is a standalone development platform that enables users to evaluate and develop applications for the TI C64xx DSP family, is used to run the application. The Neural Network was designed in Professional II/PLUS software. The project was built and run on C6416T DSK and the results were saved on computer as a .dat file. The images were obtained by scanning different texts with a 300 dpi scanner and transferred to the system.

Conclusion and achievements

The aim of this work is to implement a hardware-based Optical Character Recognition system for printed Arabic characters. The goal was achieved using DSP techniques.
The following points summarize the conclusions of the work:

- Arabic character recognition is a research area involving many concepts and research points.
- Fuzzy ART Neural Network was implemented and tested.
- A noble and efficient recognition method with a high accuracy (plus 95%) was successfully developed and tested in this thesis.

- A complete hybrid hardware-based system for isolated Arabic character recognition was proposed and tested for performance.

Future work

The implemented system suffers from some constraints and needs further work to become a reliable and commercially valuable product. The following list notes the limitations and suggested solutions for future research on this project:

- The image is obtained manually: as discussed earlier, the image is read by MATLAB outside the program and after preprocessing (converting to binary, etc.) its pixel values are transferred to the project as a header file. The possible solution for this limitation is to write C program code to read the image directly from the computer (or any other host device that the image is saved on).

- The system works only for isolated characters: since the scope of this research is focused on the recognition problem and the hardware implementation, the current system assumes that the characters are already isolated and the algorithm can only recognize the isolated characters. For this project to have a commercial value, the system should be able to isolate the connected characters from a word. Since there are techniques already developed for isolating Arabic characters, this system could be integrated with one of the existing character segmentation algorithms to overcome this limitation.

- The system is developed using a general-purpose DSP board (TMS320C6416T DSK). The algorithm could be developed on a single chip and a smaller board which is only designed for this purpose.
- A multi-language Optical Character Recognition hardware could be developed if several OCR applications for several languages are programmed on the same chip using the same board.
- The board could be integrated with a built-in scanner (like the scanners used to scan the bar code of the products) for image acquisition. This way, the system would work independently of any computer or host device. Such a board can scan a document and perform OCR without needing computer.

References

[1] Kavianifar Mandana, & Amin, Adnan (1999). Preprocessing and Structural Feature Extraction for a Multi-Fonts Arabic / Persian OCR. *Document Analysis and Recognition, 1999. ICDAR '99. Proceedings of the Fifth International Conference* .

[2] Vinciarelli, Alessandro (2003).Offline Cursive Handwritting: From Word to Text Recognition.

[3] Govindan V .K, & Shivaprasad, A.P (1990). Character Recognition – A Review. *Pattern Recognition*. *23*, 671-683

[4] Sekita, I., Toraichi, K., Mori, R., Yamamoto, K., & Yamada, H. (1988). Feature extraction of handprinted Japanese characters by spline function or relaxation matching. *Pattren Recognition*. *21*, 9-17.
[5] Xie X. L., & Suk, M. (1988). On machine recognition of handprinted Chinese characters by feature relaxation. *Pattern Recognition*. *21*, 1-7.
[6] Matsumura, H., Aoki, K., Iwahara,, T., Oohama, H., & Kogura, K. (1986). Desktop optical handwritten character reader. *Sanyo tech*. *18*, 3-12.
[7] Hashemi, M., Fatemi, O., & Safavi, R. (1995). Persian script recognition. *Proceedings of the third Int. Conference on document analysis and recognition*. *II*, 869-873.
[8] Allam, M. (1995).Segmentation versus Segmentation-free for Recognizing Arabic Text. *Document Recognition II, SPIE*. *2422*, 228-235.
[9] Ethnologue: Languages of the World, 14th ed. SIL Int'l, 2000.
[10] Lorigo, AuthorLiana M., & Govindaraju, Venu (2006). Offline Arabic Handwriting Recognition: A Survey. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*. *28*, 1.

Biography

HAIDAR ALMOHRI is currently an employed by Siemens Co. as a Communication Engineer in their branch in Kuwait. He completed his undergraduate and graduate studies in Electrical Engineering at the University of Hartford, Connecticut, USA.

John S. Gray is currently a Professor of Computer Science at the University of Hartford, West Hartford, CT. He is Chair of two degree programs – Computer Science and Multimedia Web Design and Development. His area of interest and expertise is UNIX system level programming with a focus on interprocess communications. As an educator, author and consultant, he has been involved with computers and software development for over 24 years.

Dr. Hisham Alnajjar is an Associate Professor of Electrical and Computer Engineering at the University of Hartford, Connecticut (USA), where he is also the Associate Dean of the College of Engineering, Technology, and Architecture (CETA). Ph.D. from Vanderbilt University, M.S. from Ohio University. His research interests include sensor array processing, digital signal processing, and power systems.