International Journal of Advanced Research in Engineering and Applied Sciences
ISSN: 2278-6252

# INTRODUCTION TO OPTICAL CHARACTER RECOGNITION

**Zohreh Mousavinasab***

**Sara Bahadori***

**Abstract**: *The recognition of handwritten notes when the information is entered into computer this way is essential. The first line of research in this field has been done by the Chinese and Japanese. Extensive research conducted in the field of Latin texts has been done and is still ongoing. For better recognition, a special set of characters has been introduced which may have a fundamental difference with the main characters in some situations. There are commercial systems that use these kinds of alphabets. Using these alphabets tend to be very good and it has been reported that they can be recognized precisely more than 99%. However, in some situations for instance speed writing these systems (e.g., taking notes), cannot be applicable.*

*Department of Computer Engineering, Omidiyeh Branch, Islamic Azad University, Omidiyeh, Iran

## 1- STATEMENT OF THE PROBLEM

Usually a keyboard for entering information into a computer is used. Entering data via handwriting or speech in some situations is preferred for the following reasons:

- ✓ Writing with a pen or speaking much faster and easier.
- ✓ There are some places, such as in a classroom where students can't type texts but can write them.
- ✓ Pocket computers cannot use a full keyboard, and only some of them have keyboards.
- ✓ Some natural languages have many symbols, such as in the language of Kanji which has 4000 letters, it is too difficult to enter data via the keyboard [1].

Other Commercial systems use more natural alphabets, but they tend to have limitations on writing characters when separating words from each other. Still, there are products that try to recognize writing without constraints, but these systems have little recognition [2].

In the context of recognizing Arabic and Persian peripheral handwriting, little research has been carried out.

The purpose of this study is the recognition of peripheral Persian handwritten characters. One hundred and eighty Handwriting characters have been used for recognition. It has been assumed that the words written on paper are almost horizontal. This system is independent of the author. Computing device is a personal computer. The processing speed is such that the recognition rate of the system can be computed according to the real time used. For this purpose, a simple algorithm is designed and implemented appropriately.

## 2- HANDWRITING RECOGNITION AND ONLINE AND OFF-LINE TEXT RECOGNITION AND COMPARISON

Text recognition is one of the main branches of pattern recognition which extensive research in this field has been and still is under investigation. Handwriting Recognition has two branches based on the method of obtaining information: offline and online .Offline recognition consists of 29and writing text recognition and typed text recognition .Online recognition is only used with handwriting.

In offline recognition, the input is a scanned document, but in online recognition the text input is the coordinates of the stylus path. In this mode means of communication is either a

pen or a computer which usually has a digital screen. Figure 1-1 shows how the two modes of data entry.

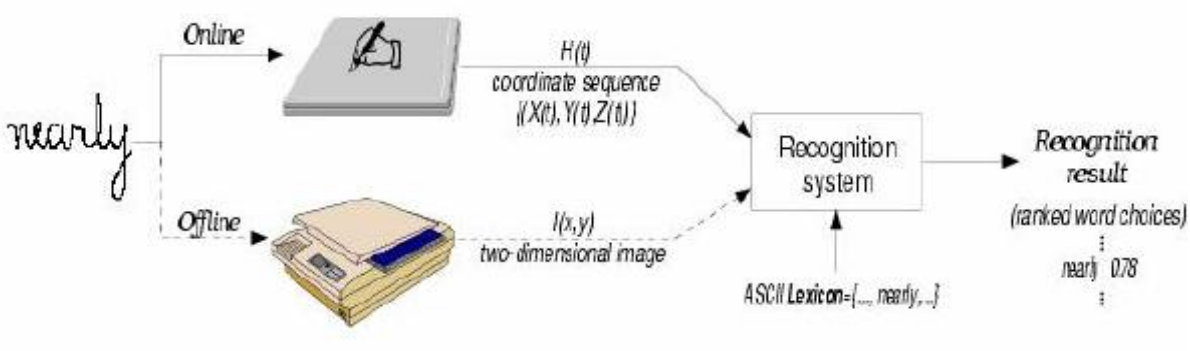



**Figure 1: Online and offline text recognition**

Extensive research in the field of online and offline text recognition has been done. At the highest level, handwriting recognition can be divided into two categories online and offline. Off-line recognition of handwritten documents are written on paper. The Information is given to the system by scanning the documents. In recognition of online handwriting the stylus path points are coordinated as a string and entered into the computer. This requires special equipments, such as a digital plate to see the movement of the pen when writing. The stylus Path is recorded in fixed time intervals, and the coordinates (x,y) is stored in the sample. In online recognition, information can be obtained at the time of writing, but in offline recognition, the writing is received.

However, the method often used in online articles can be recognized in real-time. However, there may be applications in which appropriate recognition is postponed to a later date. In online recognition dynamic information is available. This information includes the number of movements, the direction and the speed of pen strokes.

In the Online mode, the data is simply a coordination number of the points available. While in the offline mode, pre-processing is needed to extract text from the background. Recognition improves when information is online, but sometimes due to features that are not apparent in static images, the recognition becomes more complex.

In On-line recognition system it is possible for the author to adapt himself to the recognition system. Some of these systems are able to adapt themselves to the author as well.

## 3- BRIEF HISTORY

### 3-1- offline text recognition

First patent registered in this area was done in 1929 and 1933, which was recognized by matching typographic characters with stereotypes. Mechanical masks cross the characters image, and shine light on one side and, is received by a light detector. When a perfect match is made, light is not sent to the detector and input is detected. This patent has not been applicable since opto-mechanics technology was applied in it. The magnetic ink character recognition has also been used. This work was done by writing, with magnetic ink. And a special pen for this work was intended as well [3]. Figure 2 shows the word MICR which has been written in magnetic ink.

**Figure 2: the word MICR written in magnetic ink**

Other preliminary work has been done on typed texts which special pens are used for writing the text. OCR-A and OCR-B fonts are some examples which are shown in Figure 3. In the mid-1950s, OCR was considered to be a very active field for research and development. Today, typed text recognition software is available in the field of Latin for less than 100 dollars.

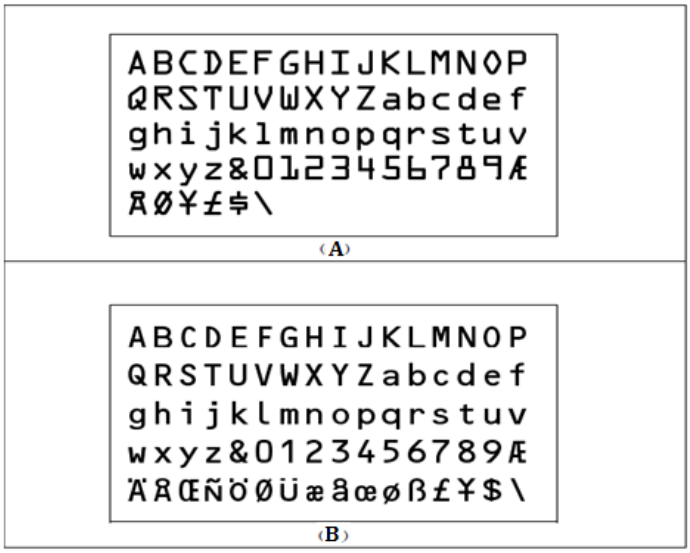



**Figure 3.1: defined fonts for the primary system**

**A: OXP-B and B:OXP-A**

In the context of recognizing offline handwritten characters and words, much research has been done. The preliminary restrictions have been investigated in this matter, to determine how they write. Nowadays, preliminary research on individual characters should be characters that are written into rectangles so they can be easily separated and recognized.

**3-2- Online Text Recognition**

Entering information into the computer by hand or speech is preferred in some situations for the reasons mentioned in section 1. Therefore, recognition of handwritten information when entered in the computer is necessary. The first research in this area is related to Japanese and Chinese languages [4]. In the context of the Latin online recognition, extensive research has been done and is still ongoing. For better recognition, a special set of characters are defined in some circumstances which have a fundamental difference with the main characters. A good example of this special set, the Graffiti collection of characters that can be named. In Figures 4 and 5 characters in this series can be observed. There are commercial systems that use this alphabet. Recognition of such alphabets tends to be over 98% which is a good rating [5].

However, these systems are not applicable when speed writing (for example taking notes) is under consideration. Other Commercial systems use more normal alphabets, but they can cause some restrictions when writing the individual characters and words separated from each other. Of course, there are also products that try to recognize writing without these reservations, but these systems have little recognition. Figure 4 shows some examples of sentences which are written with no restriction on Cross Pad. It also shows the recognition results that are writer-independent [6].

ISSN: 2278-6252

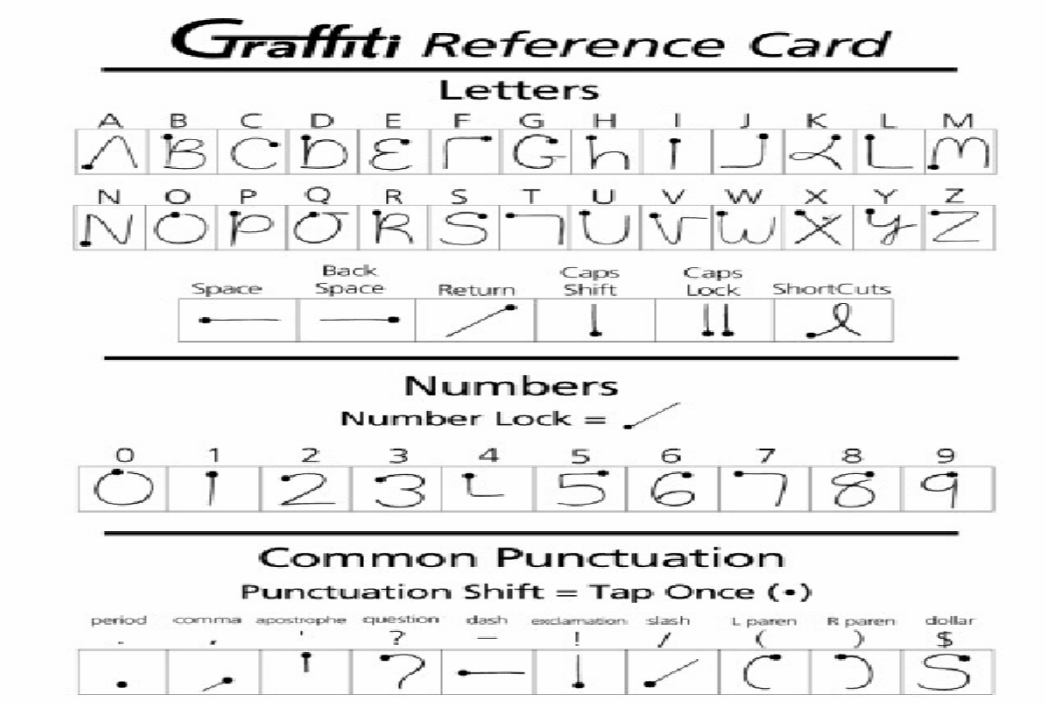



**Figure 4: Character Set**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **a** | | **l** | | **w** | | double quote | |
| **b** | | **m** | | **x** | | tab | |
| **c** | | **n** | | **y** | | space | |
| **d** | | **o** | | **z** | | backspace | |
| **e** | | **p** | | period | • or \ * | new line | |
| **f** | | **q** | | comma | | cut | |
| **g** | | **r** | | aspostrophe | | copy | |
| **h** | | **s** | | question | | paste | |
| **i** | | **t** | | exclamation | | undo | |
| **j** | | **u** | | ampersand | | command | |
| **k** | | **v** | | at | | | |

* • or •• is written in the writing area.
\ or \\ is used when writing on the display.

Write numbers and the following symbols to the right of the division marks.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **0** | | **6** | | dash | | ( | |
| **1** | | 7 | | tilde | | ) | |
| **2** | | **8** | | + | | = | |
| **3** | | **9** | | * | | backspace | |
| **4** | | period | • or \ * | / | | | |
| **5** | | comma | | \ | | | |

Write accent marks to the right of the division marks after writing an upper or lower case letter.

| | | | | | |
|---|---|---|---|---|---|
| **à** | \ | **ã** | | **ä** | •• or \\ * |
| **á** | / | **ã** | | **å** | |

**Figure 5: Character Set**

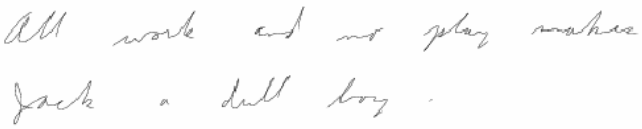
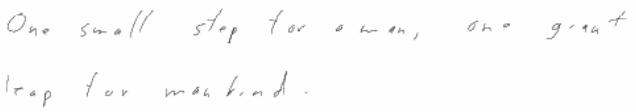
| Written Sentence | Recognition Results |
|---|---|
|  | All work and no play play makes Jack a dull **try**. |
|  | One **Sin all** step for a **in an**, one **9-9 it** leap for man kind. |

**Figure 6: Examples for the character sets without reservation**

**4- Author-independent recognition**

**1-4- Different sections of a text recognition system**

Different parts of a document recognition system are shown in Figure7. Word recognition process in this system consists of the following steps:

- ✓ Receive input
- ✓ primary processing
- ✓ extracting features
- ✓ Recognition with the use of one or more classifiers

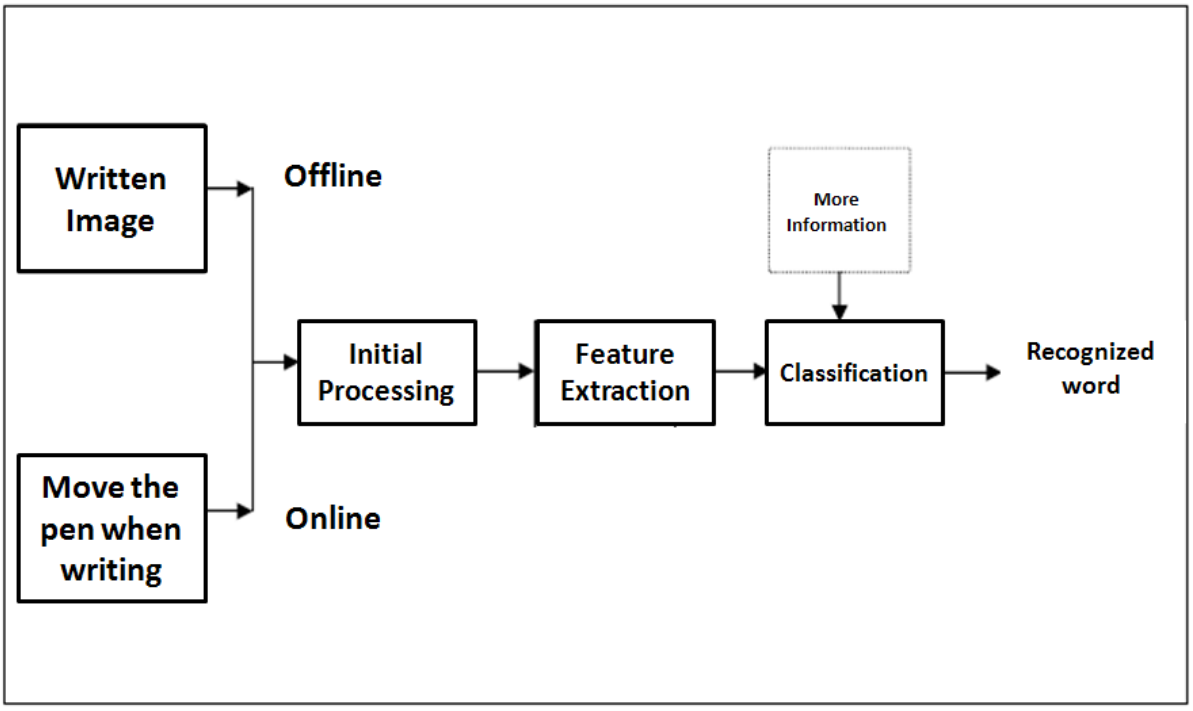



**Figure 7: Recognition system sectors**

**4-1- Receiving Input**

In the Offline mode, text input is done by imaging in gray or black and white with an appropriate degree of separation. The Grade of separation distance is usually between 200 to 300 dpi.

**4-2- Primary Processing**

Bedeviling Grayscale images is done by using a global threshold and a local threshold in the offline mode. Curvature correction, slope correction, (based on methods of segregation), and the size normalization, are preprocesses which are done in both online and offline cases. Bedeviling Grayscale images using a global threshold and local threshold is used in the offline mode.

**4-3- Feature Extraction**

In any pattern recognition system, the feature extraction is one of the main parts. Features in use can be divided into two categories as the following:

A) Static properties: These properties are related to the shape and structure of the paper and doesn't have any information related to time. These features can be used in online and offline recognition.

B) Dynamic features: These features include information about time. These data are available in online texts only. Although some research has been done to provide dynamic information to extract information from offline data as well.

**4-4-Methods of Classification**

**4-4-1 statistical methods**

In these methods, the distribution model and the density probability functions for each class are estimated. Unknown samples with minimum standards or minimum risk of error are classified. Bayes classifiers and classifications using Hidden Markov Models are placed in this category. Hidden Markov Model has been used in many studies to document recognition.

**4-4-2 Non-statistical methods**

In these methods, certain statistical properties for the model will not be considered beforehand. Usually distance functions or discriminate functions are used. In classification methods, Minimum distance, conformity stereotypes and k nearest neighbor methods is used. Conformity stereotypes in many studies, especially in the online recognition are used as the compatibility chart. Superposition of the linear and the non-linear curve is also used.

**4-4-3- structural methods**

In these methods, a model based on a set of components and the relationships between them are considered.

Each class is defined by a set of syntactic rules. These rules help define each pattern class as a combination of basic elements.

**4-4-4- neural network techniques**

Neural networks have found many applications in pattern recognition. Usually, multilayer neural networks recognize the error back propagation when learning algorithm is used.

## 5. WRITER-DEPENDENT AND WRITER-INDEPENDENT RECOGNITION

Handwriting recognition can be divided into two categories, depending on the author and writer freelance. A system for writer independent handwriting recognition is taught with large variations in writing styles. While in a writer-dependent system recognition of a particular person's handwriting is taught. Data in writer dependent systems work with less variation and therefore renders a higher recognition rate. A writer independent system, meticulous recognition which can be achieved by the writer is sacrificed in order to figure out further changes in the writing style.

On the other hand, since many people can be used to collect data, lots of data for training can be gathered easily in the freelance system which is totally different than the author dependent system. In order to have a writer-dependent system, the training phase can be done in two ways:

One method is to collect a lot of data from only one author and to train the system based on this data. Another way is to give training on a system dependent from the author.

Quantitative data from the author can raise recognition and the accuracy of the system for a certain author. Lots of research is ongoing to solve this problem.

## 6- COMPARING PERSIAN AND ENGLISH

- ✓ Handwritten letters of Latin are written in "succession", meaning that each letter comes after the previous, and are very similar to Persian handwritten letters. The following are some of their similarities:
- ✓ Letters are connected to each other.
- ✓ Based on the location of the letter (beginning, middle or end of the word) the form of the letter changes.

- ✓ A change in the form of letters based on the author or the author's orthography.

But there are differences that can be noted as follows:

- ✓ Unlike the Latin alphabet, many Persian words are characterized with different punctuation, some characters have the same structure and the only difference is the number of dots above or below the letter.
- ✓ Unlike Persian, Latin letters are written from right to left.
- ✓ Height of the Latin alphabet is constant, but the height of the Persian alphabet letters can differ.

## 7- CONCLUSION

The handwritten character recognition is a research topic in the fields of pattern recognition, machine learning, and image processing and computer vision. There are Many applications for character recognition such as car plate detection, keyword extraction, annotating images, Postcode and bank checks recognition and finally university test grading. Hence the need for a system which is able to solve these problems is necessary.

## 8- REFERENCES

[1] Sani R., "Recognition of Persian handwritten letters by using artificial neural network", Master of the letter, Faculty of Computer Engineering, Sharif University, 2008.

[2] Abulhaiba S. H., Mahmood S. A., Green R. J., "Recognition of handwritten cursive Arabic characters", IEEE Trans. On PAMI, Vol. 16, No. 6, pp. 664-671, 1994.

[3] Arica N. and Yarman-Vural F.T., "One dimensional representation of two dimensional information for HMM based handwritten recognition", Pattern Recognit. Lett., vol. 21, No. 6–7, pp. 583–592, 2009.

[4] H. Freeman, "Computer processing of line drawing images," Computer Survey, Vol.6, pp.57-97, 1999

[5] Arıca Nafiz, Vural Fatoş Yarman, "A new scheme for off-line handwritten connected digit recognition", 14th Int. Conf. Pattern Recognit., Brisbane, Australia, pp. 1127–1131, 2005.

[6] C. T. Zahn and R. Z. Roskies, "Fourier descriptors for plane closed curves," IEEE Transaction on Computer, C-21 (1), pp.269-281, 2003.