Characters Recognition based on Geometrical Features

Anwar Hassan Mahdy\ Computer Science Department\

College of Science\ University of Al -Mustansiryha

Abstract

Character recognition is one of the important subjects in the field of Document Analysis and Recognition (DAR). The general objective of DAR research is to fully automate the process of entering and understanding printed or handwritten data into the computer. The proposed pattern recognition system consists of two-stage process. The first stage is feature extraction and the second stage is classification. Feature extraction is the measurement on a population of entities that will be used in recognition process. This assists the recognition stage by looking for features that allows fairly easy to distinguish between the different classes. Several different features have been used for recognition process. The set of proposed features that are used makes up a feature vector. These set of features are: the first feature is represented the number of character pixels (the summation of pixels), the second features is represented the width of each character in pixels, and the third feature represented the height of each character in pixels. Finally, Pattern recognition system classifies each member of the population on the basis of information contained in the feature vector. The results show that the suggested features gives higher accuracy in text and character recognition.

المستخلص

التعرف على الحروف هو احد الموضوعات المهمة جدا في مجال تحليل وتمييز الوثائق. الهدف العام من تحليل الوثائق بعملية السيطرة الحاسوبية هو إدخال وفهم البيانات المطبوعة أو المكتوبة بخط اليد في الحاسوب.تمييز الحروف يمكن أن يتم على النص المطبوع أو النص المكتوب بخط اليد. القدرة على التعرف على الحروف المطبوعة أليا أو بطريقة شبه آلية هو تطبيق واضح في العديد من المجالات. وإن بناء خوارزمية التمييز بدقة 100 ٪ عادة ماتكون مستحيلة في عالمنا المليء بالضوضاء وأنماط الخطوط المختلفة، فمن المهم لتصميم خوارزميات التعرف على الحروف اخذ هذه الإخفاقات بنظر الاعتبار بحيث عندما تتم الأخطاء(لا محالة) ، فإنه على الأقل تكون مفهومة ومتوقعة للعاملين بهذا المجال. نظام تمييز الأنماط المقترح يتكون من مرحلتين . المرحلة الأولى هي استخلاص الخصائص والمرحلة الثانية هي عملية التصنيف والتمييز. استخلاص الخصائص أو السمات كقيم عددية تميز كل حرف عن غيره ولقد تم اعتماد ثلاث سمات أساسية هي : عدد نقاط الحرف, طول الحرف, وعرض الحرف بالبكسل. أما عملية التمييز فتمت بمقارنة سمات الحروف غير المعروفة مع السمات الرئيسية لكل حرف في قاعدة البيانات . ولقد تم الحصول على نتائج بدقة 100% في حالة الصور الخالية من الضوضاء والتشوهات.

Keywords:

Text recognition, Optical character recognition, Feature extraction, Pattern recognition, Classification.

Introduction

Character recognition, usually abbreviated to optical character recognition or shortened OCR, is the mechanical or electronic translation of handwritten, typewritten or printed text (usually captured by a scanner) into machine-editable text. It is a field of research in pattern recognition, artificial intelligence and machine vision [1]. This technology allows a machine to automatically recognize characters through an optical mechanism [2], where OCR is the process of converting scanned images of machine-printed or handwritten text (numerals, letters, and symbols), into a computer process able format [3]. Most traditional OCR applications have been designed to be highly automated and used on desktop machines. These recognition engines perform well but usually require high quality input text images that are reliably obtained [4].

OCR is one of the oldest ideas in the history of pattern recognition using computers. However, it remains a challenge till now to develop an OCR system which can achieve such a high recognition rate, regardless of the quality of the input document. A lot of research has been done on OCR in last 50 years. Some books and many surveys have been published on the character recognition. Useful reviews and surveys in the field of OCR include the historical review of OCR methods and commercial systems, the survey by Impedovo et al. focuses on commercial OCR systems [5], while the work by Tian et al. surveys the area of machine-printed OCR [6]. Jain et al. summarized and compared some of the well-known methods used in various stages of a pattern recognition system. They have tried to identify research topics and applications which are at the forefront in this field [7]. Pal and Chaudhuri in their study summarized different systems for Indian language scripts recognition. They have described some commercial systems like Bangle and Devanagari OCRs. They reported the scope of future work to be extended in several directions such as OCR for poor quality documents, for multi font size and language OCR and bi-script/multi-script OCR development etc. khaly and Ahmed, Amin and Lorigo & Govindraju have produced a comprehensive survey and bibliography of research on the Arabic optical text recognition [3].

In this paper an approach on recognizing scanned text is proposed. The process of recognizing scanned text can be considered by classifies a given input as belonging to a certain class. In this phase the text is scanned first with a scanner and converted into an image format. Then applying several techniques the characters in the text are separated. These separated characters then applied a classifier to recognize the characters by using minimum distance technique.

Feature Extraction and Recognition Process

The steps were involved in character recognition after an image scanner optically captures text images to be recognized is given into four stages: Digitization, Segmentation, Feature extraction, and finally Classification.

Digitization refers to the process of converting a paper or text document into electronic form. The electronic conversion is accomplished through imaging a process whereby a document is scanned and an electronic representation of the original, in the form of a bitmap image, is produced. The imaging process involves recording changes in light intensity reflected from the document as a matrix of dots. The light/color value(s)of each dot is stored in binary digits. One bit would be required for each dot in a binary scan, whereas up to 24 bits could be required per dot for a color scan. Digitization produces the digital image, which is fed to the segmentation phase [3].

Segmentation is one of the most important phases in character recognition process. Segmentation errors are more serious and cannot usually be corrected by manual classification [4]. Segmentation is the process of segmenting the whole document image into recognizable units for feature extractor and classifier. Text area from the document is extracted and the segmentation step is followed by segmenting the text region into individual lines. Further each line is segmented into individual words, and finally, each word is segmented into individual characters. Character segmentation is fundamental to character recognition approaches which rely on isolated characters. It is a critical step because most recognition errors are due to the incorrect segmentation of the characters [3].

Feature extraction plays an important role in the successful recognition of machine-printed and handwritten characters [5]. Feature extraction can be defined as the process of extracting distinctive information from the matrices of digitized characters. In OCR applications, it is important to extract those features that will enable the system to differentiate between all the character classes that exist. Many different types of features have been identified in the literature that may be used for character and numeral recognition [3].

The final step of OCR engine is recognition stage. The objective of the classification is to give an accurate output from the input features, which extracted in the previous stage to identify the character segment [8]. Classification is concerned with making decisions concerning the class membership of a pattern in question. The task in any given situation is to design a decision rule that is easy to compute and will minimize the probability of misclassification relative to the power of feature extraction scheme employed [3]. List of various classification methods includes template matching (minimum distance), syntactic method, statistical methods, artificial neural networks, kernel methods and hybrid classifiers [7].

The Minimum Distance Classifier

The minimum distance classifier (MDC) is derived from Bayesian decision theory and shown to be the optimal classifier when assumed isotropic Gaussian distributions for the classes and with equal prior probabilities. The MDC is also shown to be equivalent to template matching when the templates are derived from the target reactivity data [9]. After feature vector have been computed for each character, a simple minimum distance classification has been implemented. The Euclidean distance of this feature vector to each of the features derived from the feature extraction stage is computed. The character is assigned the closest feature [10]. The advantage of this classifier is that it not only is a mathematically simple and computationally efficient technique, but also provides better accuracy than other classification techniques, in the case when the number of training samples is limited. Minimum distance classifier has no knowledge of the fact that some classes are naturally more variable than others, which consecutively can lead to misclassification [11]. The minimum distance algorithm is also more attractive since it is a faster technique than other classification [12].

The Suggested System Architecture

Following Figure 1 represent the System block diagram. It is shown that system architecture is consists of four blocks: Digitization, Segmentation, Feature extraction, and finally Classification.The text document is first scanned by a scanner than stored in digital image format. The histogram threshold technique is used for its better result, where the text regions are separated from non-text regions. Then the lines are extracted from the text image. Hence each line is segmented into words and finally the words are separated into constituent. These characters are then fed for Feature Extraction Operation. Feature extraction is the process of getting useful information from each character to be used for classification purposes.

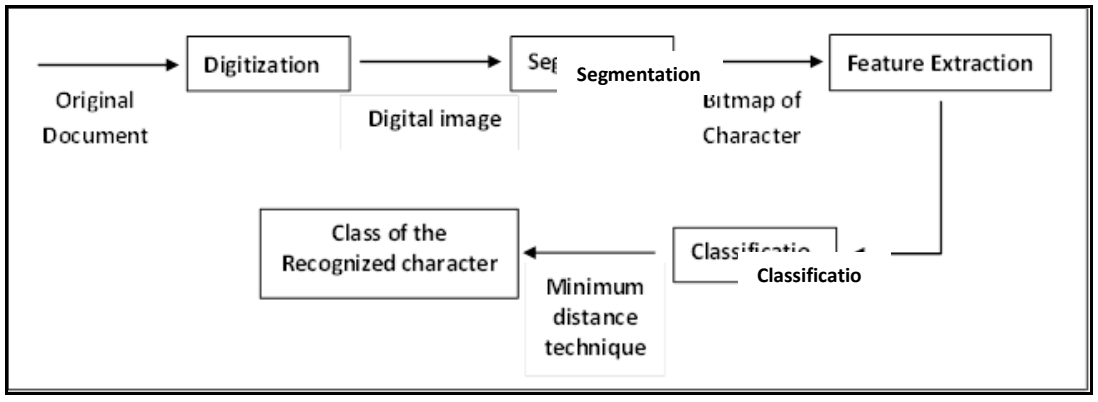

Figure 1: The adopted system block diagram

Digitization Process

Digitization process is a technique by which any text image is converted to digital image. The most common method is to select a proper threshold for the image and then convert all the intensity values above the threshold intensity to one intensity value representing either "black" or "white" value. Image thresholding is advantageous as it is easier to manipulate images with only two intensity levels, processing is faster less computationally expensive and allows for more compact storage. Threshold value removes the noise from the image, if the intensity of that pixel is below threshold level.

Segmentation Process

After converting the text document into electronic form, it comes to line, word and character separation before extracting features from each character.

1. Line Segmentation

The OCR system should first identify and separate different lines of from the image file. For this purpose, the digitized image file is scanned horizontally to detect black pixels in order to find out the starting (left top corner) point and the ending (bottom right corner) point of each line. Then lines will obviously be separated with vertical span of spaces. As shown in figure 2.

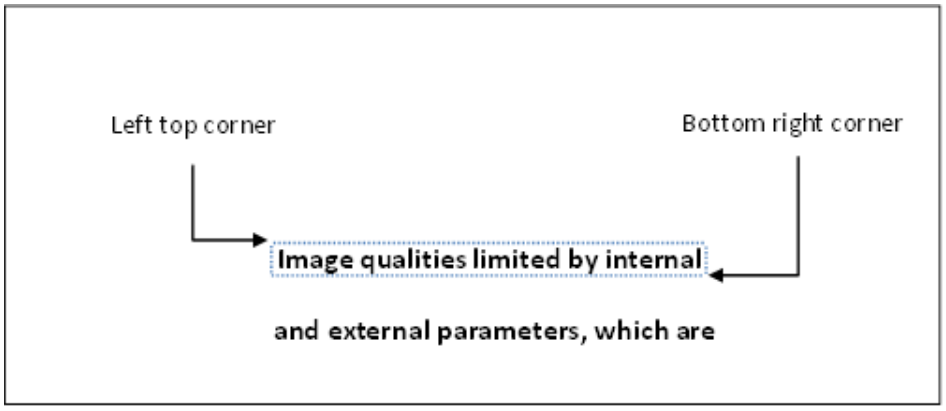

Figure 2: Line Segmentation

2. Word Segmentation

Finding the span of each word is the next essential step. For this purpose, each line is scanned vertically to detect black pixels. Wherever there are continuous black pixels, that portion of the line is considered to be a word in that line. Otherwise, if no black pixel is found in some vertical scan that is considered as the spacing between words. Thus different words in different lines are separated. So the image file can now be considered as a collection of words. As shown in figure 3.

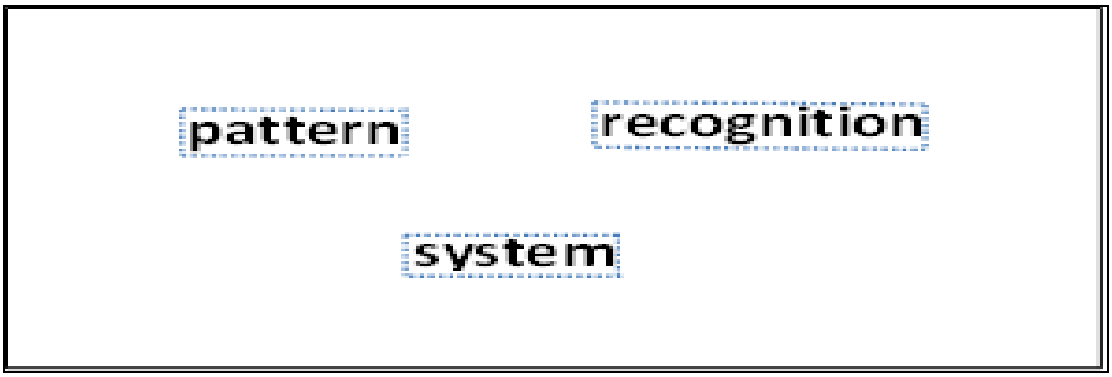

Figure 3: Word Segmentation

3. Character Segmentation

Once the line and word segmentation has been achieved, we have to segment the characters for the purpose of recognizing them. Character segmentation is an extremely important step in a text recognition system and the accuracy of the text recognition system heavily depends on this step. For this purpose, each word is scanned vertically to detect black pixels. Wherever there are continuous black pixels, that portion of the word is considered to be a character in this word. Otherwise, if no black pixel is found in some vertical scan that is considered as the spacing between adjacent characters. Thus different characters in different words are separated. So the image file can now be considered as a collection of characters.

Algorithm (1): Line, Word, and Character Segmentation Algorithm:

Input: the input of the algorithm is a binary text image img(i,j) of sized (r×c) with pixel values either 0 or 1 (background black (0), and foreground white(1)).

Step1: separate text image into lines by scan the image for totally white rows.

Step2: for each separated line, scan for totally white column with first threshold (th1) for separated space between the words.

Step 3: for each separated word scan for totally white column with second threshold (th2) for separated space between the characters in the word.

Note: th1 is higher than th2.

Output: array of separated characters of the text image.

Feature Extraction

The sub image of each character is analyzed to estimate character features. Because of the varying nature of character writing and image quality, it is best to search for invariant features

that uniquely define the character. The feature extraction stage analyses a character segment and selects a set of features that can be used to uniquely identify the character segment. The selection of a stable and representative set of features is the heart of pattern recognition system design. Among the different design issues involved in building an OCR system, perhaps the most significant one is the selection of the type and set of features.

In this paper there are three features extracted:-

1. Feature1: represented the number of character pixels.
2. Feature2: represented the width of each character in pixels.
3. Feature3: represented the height of each character in pixels.

Classification

Classification is performed by comparing an input character features with a set of templates (or prototype) from database of features of all characters (from a-z in low case) features class, each comparison results in a similarity measure between the input characters with a set of templates. One measure increases the amount of similarity when a feature in the observed character is identical to the same features in the template image. If the features differ the measure of similarity may be decreased. After all templates have been compared with the observed character image, the character's identity is assigned the identity of the most similar template. Template matching is a trainable process as template characters can be changed

RESULTS AND DISCUSSION

The system performs character recognition by quantification of the character into a mathematical vector entity using the geometrical properties of the character image. The scope of the proposed system is limited to the recognition of a single character. As such, for a suggested OCR system, segmentation process involves the following steps:

1. Segmentation of a text region into individual lines.
2. Segmentation of a text line into individual words.
3. Segmentation of a word into individual characters.

These steps have been shown in the fallowing figures below: Figure 4 contains an input text document.

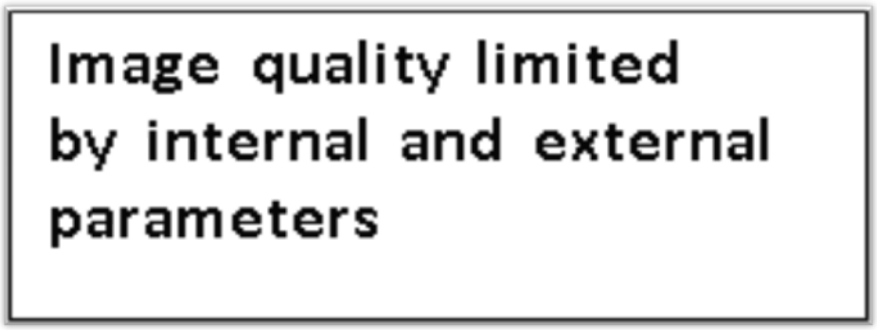

Figure 4: The input text document

A text image region has been segmented into lines based on using the horizontal text dots histogram h(y) see Figure 5. As shown in this figure the first graph was represented the histogram of the first line, the second graph was for the second line, and the third graph for the last line.

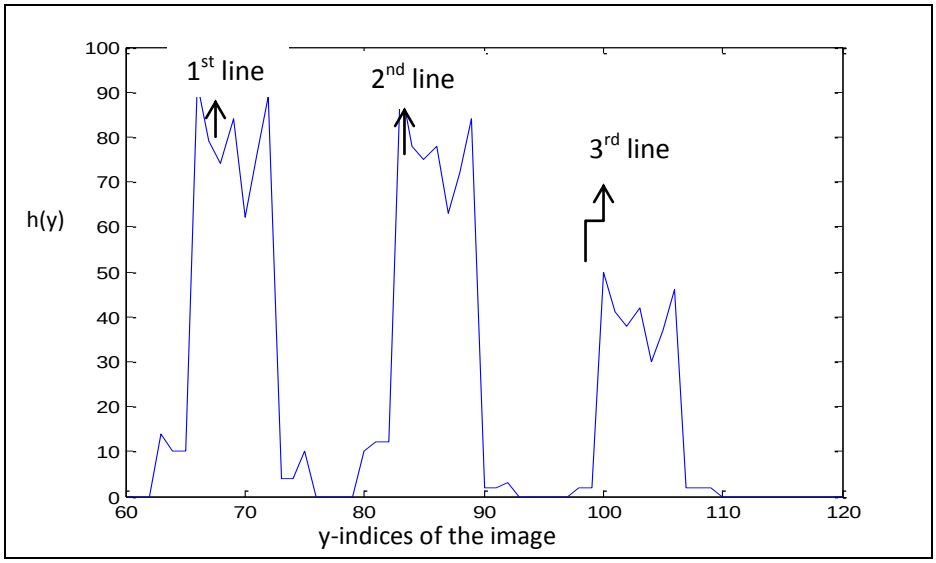

Figure5: The horizontal text dot histogram

Figure 6 contains the graph of histogram of words, point in the image. As shown in this figure, the first line contains three words; the second line contains four words, while the third line contains only one word.

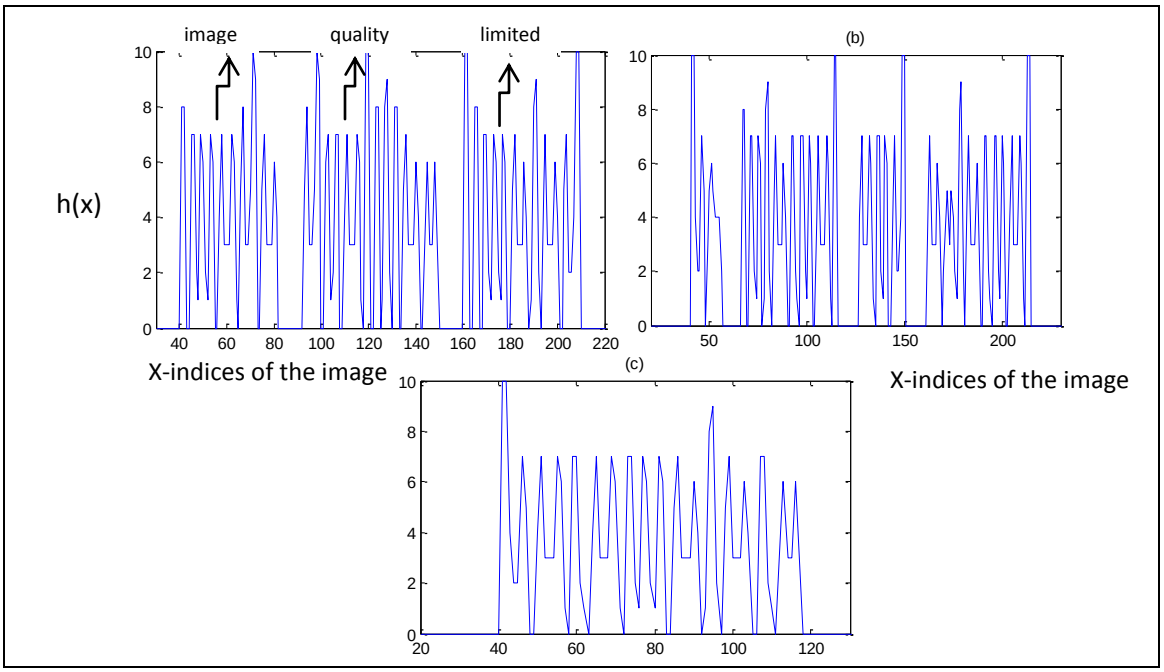

Figure 6: The vertical dots histogram of the text word,(a), (b), and (c) were represented first, second, and third lines dots histogram respectively.

Also, in Figure 7, words have been segmented into characters. For example the word "parameters" is separated into ten characters: parameters.

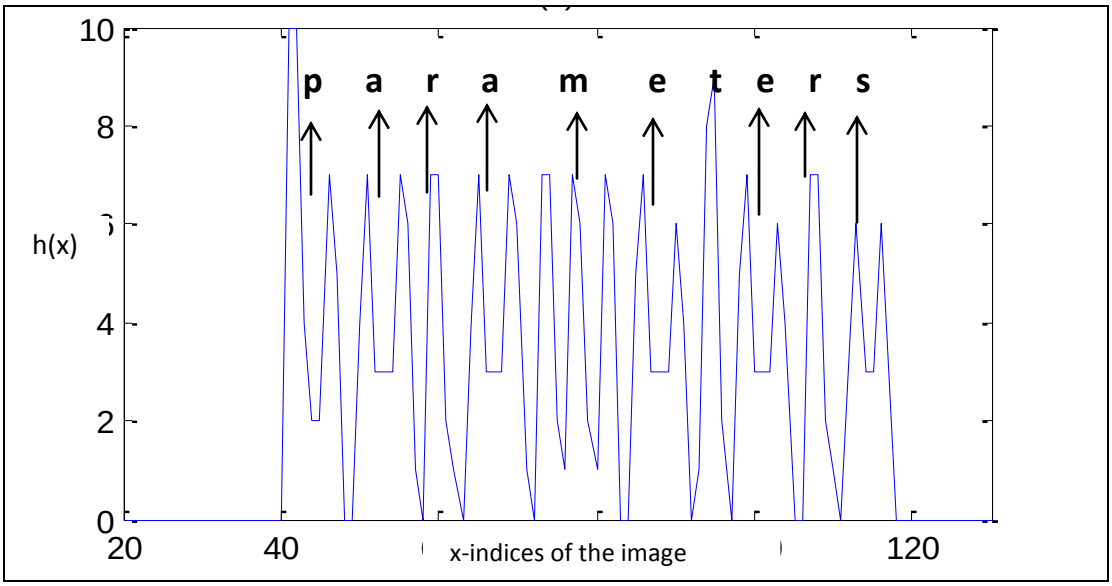

Figure 7: The vertical dots histogram of the " **parameters** " word.

Table 1 shows the results of extraction features stage for each segmented character, where feature 1, feature 2, and feature 3 were represented the sun of pixel, the length, and the high of each character respectively.

Table 1: The extraction for each segmented character

Class	character	Feature1	Feature2	Feature3
1	A	8	9	53
2	B	8	12	60
3	C	6	9	33
4	D	8	12	62
5	E	6	9	37
6	F	6	12	44
7	G	8	10	68
8	H	9	12	68
9	I	4	11	36
10	J	5	15	51
11	K	10	12	65
12	L	4	12	39
13	M	14	9	88
14	n	9	9	59
15	o	7	9	46
16	p	8	13	67
17	q	8	13	65
18	r	7	9	38
19	s	5	6	36
20	t	5	12	35
21	u	9	9	58
22	v	8	8	35
23	w	13	8	61
24	x	8	7	41
25	Y	8	9	44
26	z	7	7	41

Table 2 shows the recognition accuracy, it can be observed that we get maximum accuracy (100%) where all the characters have been well recognized. For example let's consider the input text document "ghrtuvopqa", and then we can see the character "g" is recognized as class "7" and the character "h" as class "8" and so on.

Table 2: The recognition result for the text

Class	character	Feature1	Feature2	Feature3
7	g	8	10	68
8	h	9	12	68
18	r	7	9	38
20	t	5	12	35
21	u	9	9	58
22	v	8	8	35
15	o	7	9	46
16	p	8	13	67
17	q	8	13	65
1	a	8	9	53

Conclusions

Form the introduced results can be concluded the following points:-

1. The best feature for good recognition result is no. of the character pixels.
2. While the other features character width and height in pixels that complement each other to give best recognition results.
3. By using these 3-featyres can be recognize all English characters in 100% for the perfect image (with no noise and distortion).

REFRENCES

[1] Žiga Zadnik, "***Character Recognition Handwritten character Recognition: Training a Simple NN for classification using MATLAB***", pp.1-11 see in 2011.

[2] Report AIM, Inc., "***Optical Character Recognition (OCR)***", Published by: AIM, Inc. 634 Alpha Drive Pittsburgh, PA 15238-2802, USA, pp.1-10, 2000.

[3] Manish Kumar," ***Degraded Text Recognition of Gurumukhi Script***", PhD. Thesis, 2008.

[4] Michael Hsueh, " ***Interactive Text Recognition and Translation on a Mobile Device***", Electrical Engineering and Computer Sciences, University of California at Berkeley Technical Report, p.1-13, 2011.

[5] S. Impedovo, L. Ottaviano and S. Occhinegro, "***Optical character recognition- a survey***", International Journal Pattern Recognition and Artificial Intelligence, Vol. 5(1-2), pp. 1-24, 1991.

[6] Q. Tian. P. Zhang. T. Alexander and Y. Kim, "***Survey: omnifont-printed character recognition***", in the proceedings of Visual Communications and Image Processing SPIE, Vol. 1606, pp. 260-268, 1991.

[7] A. K. Jain, R. P. W. Duin and J. Mao, "***Statistical pattern recognition: a review***", IEEE Transactions on PAMI, Vol. 22(1), pp. 4-37, 2000.

[8] Ramzi Haraty and Catherine Ghaddar, "***Arabic Text Recognition***", The International Arab Journal of Information Technology, Vol. 1, No. 2, Lebanese American University, Lebanon ,pp.1-8, 2004.

[9] M.Robert, Jr.Taylor, Peter Smith, Denis Donohue, Raid Awadallah, "***Minimum Distance Classification of Airborne Targets using High Resolution Radar***", Johns Hopkins University, Applied Physics Laboratory, Work performed under contract at JHU/APL, 1999.

[10] Jan B¨ohm and Claus Brenner, "***Curvature based range image classification for object recognition***", Institute for Photogrammetry (ifp), University of Stuttgart, Germany , pp1-10, 1999.

[11] P. Mishra and D. Singh, "***Land Cover Classification of Pulsar Images By Knowledge Based Decision Tree Classifier and Supervised Classifiers Based on Sar Observables***", Progress In Electromagnetics Research B, Vol. 30, 2011.

[12] Aykut AKGÜNa A.Hüsnü ERONATb and Necdet TÜRKa, "***Comparing Different Satellite Image Classification Methods: An Application In Ayvalik District, Western Turkey***", 2004.