Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Conceptual Studies of Intelligent Character Recognition Identified by ICR Algorithms

[1]Bhupeshwar Kumar Sahu, [2]Dr. Vishnu Mishra, [3]Nagendra Kumar Sahu
[1]Research Scholar, Dr. C. V. Raman University, Bilaspur (C.G.) India
[2]Associate Prof. BCET, Durg (C.G.) India
[3]Head, Dept. of Computer Sc. Netaji Subhas College, Abhanpur, Raipur (C.G.) India

Abstract— The Image Processing is nowadays considered to be a favourite topic in the IT industry. It is a field under Digital Signal Processing. One of its major applications is Intelligent Character Recognition (ICR). Intelligent character recognition, usually abbreviated to ICR, is the Mechanical or electronic conversion of scanned images of handwritten, typewritten or printed text into mach encoded text. ICR enables the computer or machine to visualize an image and extract text from it such that it can edit the text, store it, display or print without any scanning and apply techniques like text to speech and text mining to it. In this paper we propose a novel algorithm to extract text/characters from a scanned form image. Our system consists of various stages like 1) uploading a scanned image from machine/computer 2) Extraction of text zone from the image 3) recognition of the text and 4) applying post processing techniques(error correction and detection methods). In addition, discussed the form image registration technique image masking and image improvement techniques are implemented in our system as part of the character image extraction process. In our experiment we show that, the proposed system will get good results then the existing systems and trying to improve the efficiency and accuracy of recognizing the characters from a scanned form image.

Keywords— Form based ICR, Character Image Extraction, and algorithm

I. INTRODUCTION

The Digital Image Processing is a rapidly evolving field with the growing applications in science & engineering. Image Processing holds the possibility of developing an ultimate machine that could perform visual functions of all living beings. The term Digital Image Processing generally refers to the processing of a two dimensional picture by a digital computer i.e. altering an existing image in the desired manner. Since the image processing is a visual task, the foremost step is to obtain an image. An image is basically a pattern of pixels (picture elements) thus a digital image is an array of real & complex numbers represented by finite number of bits. Manual data entry from hand printed forms is very time consuming more so in offices that have to deal with very high volumes of application forms (running into several thousands). A form based Intelligent Character Recognition (ICR) System has the potential of improving efficiency in these offices using state Of The art technology. An ICR system typically consists of several sequential tasks or functional components viz. form designing, form distribution, form registration, field image extraction, feature extraction from the field image, field recognition (here by field we mean the handwritten entries in the form). At the Centre for Artificial Intelligence and Robotics (CAIR), systematic design and development of methods for the various subtasks has culminated into complete software for ICR. The CAIR ICR system uses the neural networks for recognition. For all the other tasks such as form designing, form registration, field image extraction etc. algorithms have been specially designed and implemented. The NIST neural networks have been trained on NIST's Special Database. The classification performance is good provided the field, i.e. the handwritten character entry in the form, is accurately extracted and appropriately presented to the neural network classifiers. Good pre-processing techniques preceding the classification process can greatly enhance recognition accuracy. It is always fascinating to be able to find ways of enabling a computer to mimic human functions, like the ability to to read, to write, to see things, and so on. ICR enables the computer or machine to visualize a image and extract text from it such that it can edit the text, store it, display or print without any scanning & apply techniques like text to speech and text mining to it.

In this paper, we propose a novel algorithm to extract text from a scanned form based image, which is used to extract text information from the scanned copy of form and also discussed about Image registration, Image masking and Improvement techniques. We concentrated on Skewing methodologies also. In next coming sections are articulated as related work, proposed system, conclusion and reference

II. RELATED WORK

A digitized image is, after all, just a collection of numbers. For binary images, every point or pixel is assigned a value of either 0 or 1, for gray level images pixel values range from 0 to 255, and for color images pixel values usually consist of three numbers, each in the range of 0 to 255. While the ensuing discussion is valid for any type of image, for the sake of simplicity, only binary images will be addressed. Recognition technologies may be classified as

1. Statistical
2. Semantic and
3. Hybrid.

Statistical Approach

Since every electronic image of a digit consists of pixel values that are represented by a spatial configuration of "0"s and "1"s, a statistical approach to image character recognition would suggest that one look for a typical spatial distribution of the pixel values that characterize each digit. In general, one is searching for the statistical characteristics of various digits. These characteristics could be very simple, like the ratio of black pixels to white pixels, or more complex, like higher order statistical parameters such as the third moments of the image. The general flow of statistic based character recognition algorithms is as follows:

1. Compute the relevant statistics for a digitized image
2. Compare the statistics to those from a predefined database.

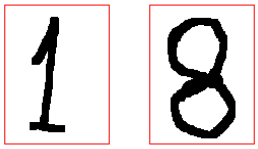

Fig.-1

Continuing the same approach, cursory analysis shows that the ratio of height to width for the digit "0" is less than the same ratio for the digit "6

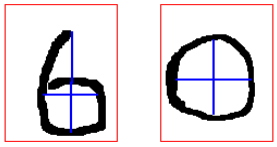

Fig.-2

In the following, these methodologies are discussed. Only handwritten numeric characters or digits will be considered, such that the character recognition algorithms return only the values 0, 1, 2...9 or "reject".

The Semantic Approach

Digitized images of handwritten characters indeed consist of pixels. However, a fact that most statistical methods ignore is that the pixels also form lines and contours. This is the essential point of the semantic approaches to character recognition: first recognize the way in which the contours of the digits are reflected in the pixels that represent them and then try to find typical characteristics or relationships for each digit. As is seen in the following examples, this is also the main advantage of semantic methods versus statistical ones.

Consider the following case:

Fig.-3

The steps of a semantic based classifier for character recognition are as follows:

- Find the starting point of a contour.
- Start tracing the contour.
- Identify the characteristics of the contour while tracing it: "up", "down", "diagonal up", "arc", "loop", etc.
- Search the database for a description similar to the one obtained. Technically, this would be executed by representing the descriptions as a logic tree (graph) and then by matching the graph against the graphs contained in the database.

Hybrid Methods and Voting Algorithms

It is clear that statistical and semantic approaches to character recognition have specific advantages and disadvantages. The obvious question: Is it possible to combine the best of both methods? The answer: To a certain extent, yes, it is possible to develop algorithms that are part statistical and part semantic in an effort to leverage the advantages of both. Top Image System's proprietary T.i.S. ICR/OCR reflects such a hybrid approach and, in many cases, overcomes the problems associated with the statistical and semantic methods when utilized independently. There is another step which can be taken, given extant technologies and methodologies, to obtain the best of all available recognition algorithms. Today, there is substantial number of good ICR recognition engines available. Each of these engines has its own specific

strengths and weakness. Each engine, on a particular type of image or document, performs better than its peers, on another, worse. Recognizing this flaw common to all ICR engines, T.i.S. analyses the relative strengths and weaknesses of the different recognition engines. This knowledge allows for the creation of unique "voting" algorithms which draw on the strengths of various engines optimizing recognition results. Character extraction problems have been solved by some researchers using variant methods. Neves [9] proposed the cell extraction method for Table Form Segmentation which consists of steps such as initially locating and extracting the intersections of table lines. The weakness of this method is that the process involved complicated table extractions. Chen and Lee [10] presented a novel approach using a gravitation based algorithm. However, in their work, some field data could not be extracted correctly, which led to mis extraction. Tseng and Change proposed the stroke extraction method [11] and then used it for the Chinese characters. However, the method did not solve the overlapping problems. Lilies et al [12] described a system for form identification based on power spectral density of the horizontal projection of the blank form to obtain the feature vectors. Here too, the overlapping problem has not been addressed.

III. PRAPOSED SYSTEM

The general scheme of the proposed system to extract text/characters from a scanned form image. The following sections are discussed about various steps involve in the proposed system.

Binarization

It converts an image from color or grayscale to black And white (called a "binary image" because there are two colours). It converts the acquired form images to binary format, in which the foreground contains logo, the form frame lines, the pre-printed entities, and the filled in data.

a) Global Gray Thresholding

Global Thresholding algorithms use a single threshold for the entire image. A pixel having a grey level lower than the threshold value is labeled as print (black), otherwise background (white)

b) Local Gray Thresholding

Local Thresholding algorithms compute a designated threshold for each pixel based on a neighbourhood of the pixel. A pixel having a grey level lower than 14 the threshold value is labelled as print (black), otherwise background (white). It is used for slowly varying background. Some of the methods applied for this are Bersen method, Niblack method, etc.

(c) Feature Thresholding

Feature Thresholding is just like local grey Thresholding but is applied for abruptly changing background.

IV. DIFFERENT ICR ALGORITHM

1. Template Matching Algorithm

The template matching algorithm has been fully implemented and tested. An input character is Size-normalized to a l6x16 grid and compared by a Hamming distance to a set of size-normalized prototypes. The N classes or the N' prototypes that most closely match an input character are then Determined. Up to 18,000 prototypes have been used at one time with this technique. Experiments have shown that perfonnance steadily improves as more prototype data is added. This is because the large number of variations in hand printed text are better represented as some of the more obscure prototypes are added to the training data. The prevenance of this algorithm has been detennined with a training set of 18,000 characters And a test set of 2,000 characters. The results of this analysis are shown in Table 1. The percentage Correct is shown for a given number of classes. The error rate is 100 minus this value. It is seen that the technique is 95.8 percent correct at guessing the input is among four of the 40 classes.

2. Structural Analysis Algorithm

The structural analysis algorithm has been fully implemented and tested. It partitions a character with a 5x5 grid and detentions the presence or absence of a horizontal or vertical stroke, a hole, a Cross point, an endpoint, or a small or large concavity in each grid cell. These 175 features are supplemented with five more that describes global features of the character. This feature vector of 180 components is then input to a Bayesian classifier that detentions the top N classes that most closely match the input character

3. Contour Analysis Algorithm

A contour analysis method has been fully developed for digit recognition and is being modified for Character recognition. This method calculates the curvature at every point along the inner and outer Contours of a binary image. Eigbt types of features are defined based on the amount of present at any point the features are similar to those described in [I]. Three features are used for concave Curvature, and five for convex curvature. Each feature is also associated with its direction and Location the feature string extracted from an unknown character is matched against a rule base to achieve recognition.

V. CONCLUSION

Top Image Systems is a leading innovator of enterprise solutions for managing and validating the flow of information between an enterprise and its customers and employees. Whether originating from mobile, electronic, paper or other sources, TiS solutions deliver the content to applications that drive the organization. TiS' eFLOW Unified Content Platform is a common platform for the Company's products - Integra, Freedom and Mobile Our ICR system has been

successfully deployed for recruitment in an Indian government office .Approximately 700 forms were processed. The form designed had three pages. All examples in this paper have been taken from filled application forms received in the above mentioned recruitment exercise. Our ICR system proved to be efficient and reduced the time required for processing. Our mission is to enable enterprises to integrate external information, improving business processes, in order to profitably deliver products and services to customers and employees.

REFERENCES

[1] P. J. Grother. Karhunen feature extraction for neural handwritten character recognition. *NIST Internal Report 4824*, April 1992.

[2] Patrick J. Grother. Nist special database 19 hand printed forms and characters database. Technical report, NIST, March 1995.

[3] M. D. Garris, J. L. Blue, G. T. Candela, D. L. Dimmick, J. Geist, P. J. Grother, S. A. Janet, and C. L. Wilson. Nist form-based handprint recognition system. *NIST Internal Report 5469 and CD-ROM*, July 1994.

[4] P. J. Grother. Hand printed forms and characters database, nist special database 19. *NIST Technical Report and CDROM*, March 1995

[5] Tseng, L.Y. and C.T. Chuang, 1992. An Efficient Knowledge-based stroke extraction method for multipoint Chinese character. Pattern Recognition, 25: 1455-1458. DOI: 10.1016/0031-3203(92)90119-4

[6] Liolios, N., N. Fakotakis and G. Kokkinakis, 2002. On the generalization of the form identification and skew detection problem. Pattern Recognit. 35: 253-264. DOI: 10.1016/S0031-3203(01)00030-9

[7] Pizano, A., 1992. Extracting line features from images of business forms and tables. Proceedings of the 11th IPAR International Conference on Pattern Recognition, Aug. 30- Sep. 3, IEEE Xplore Press, The Hague , Netherlands, pp: 399-403. DOI:10.1109/ICPR.1992.202008

[8] Boatto, L., V. Consorti, M. De Buono, S. Di Zenzo and V. Eramo *et al*., 1992a. An interpretation system for land register maps. Computer, 25: 25-33. DOI:10.1109/2.144437

[9] Wang, D. and S.N. Srihari, 1994. Analysis of form images. Int. J. Pattern Recognit. 8: 1031-1031.

[10] Casey, R.G. and D.R. Ferguson, 1990. Intelligent forms processing. IBM Syst. J., 29: 435-450. DOI: 10.1147/sj.293.0435

