

Optical Character Recognition using Grid Infrastructure

Santosh Kasaudhan¹ Subham Saxena² Varun Upadhayay³ Lucknesh Kumar⁴

^{1,2,3}PG Student ⁴Assistant Professor

^{1,2,3,4}Department of Computer Science & Engineering

^{1,2,3,4}Galgotias College of Engineering and Technology, Greater Noida, India

Abstract— As in the running world, there is a lot of demand for the software systems to recognize characters and words in computer system when information is scanned through paper documents as we know that we have number of newspapers and books which are in printed format related to different subjects. These days there is a huge demand in “storing the information available in these paper documents in to a computer storage disk and then later reusing this information by searching process”. Compared to various existing available character recognition systems it improves the accuracy of recognizing the characters during document processing. Since our character recognition is based on a grid infrastructure, heterogeneous characters of different universal languages with different font properties and alignments are recognized easily. We proposed a novel algorithm in this paper to extract text/characters from a scanned image using neural networks by using a method for combining independently trained networks to achieve higher performance. Proposed system consists of the following steps 1) Image Processing 2) Image Training 3) Image Recognition 4) Image Editing and 5) Image Searching. In our paper it is shown that, the proposed system is better than the existing systems and try to improve the efficiency and accuracy of recognizing the characters from a scanned image.

Key words: Optical Character Recognition, Neural Network, Grid Infrastructure, Kohonen Neural Network

I. INTRODUCTION

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. Digital Image Processing is a swiftly developing field with the growing applications in science and engineering. Image Processing has the capability of developing systems which can perform all the visual functions of living things. The first step in image processing is to get an image. Basically an image an arrangement of picture elements (pixels) thus a digital image is an array of real & complex numbers represented by finite number of bits. In the present world, there is demand growing for the software systems to recognize characters in computer system when information is scanned through paper documents as we know that we have number of newspapers and books which are in printed format related to different subjects.

II. RELATED WORK

Automatic character recognition improves the interaction between man and machine in many applications like mail sorting, cheques, verification, office automation and a large variety of banking. Character recognition methods include statistical, semantic, neural network, pattern recognition, etc.

A. Statistical Approach

A statistical approach looks for a typical spatial distribution of the pixel values that are represented by a spatial configuration of “0”s and “1”s which characterize each character. In this approach it searches for the statistical characteristics of various characters. Statistical methods ignore is that the pixels also form lines and contours. Statistical based character recognition consists of following steps. 1) Calculate the statistics of the character image and 2) Compare them with predefined statistics of the characters.

B. Semantic Approach

A semantic approach recognizes the way in which the contours of the characters are reflected in the pixels that represent them and try to find out typical characteristics for each character. Semantic data: open. The following steps are involved in Semantic approach for character recognition. 1. Identify contour's starting point, 2. Trace the contour, 3. Identify the characteristics of the contour, 4. Compare the obtained characteristics with the similar descriptions in the database.

C. Hybrid Approach

It is clear that statistical and semantic approaches to character recognition have specific advantages and disadvantages. Hybrid approach overcomes the problems associated with the statistical and semantic methods when utilized independently. The idea of Hybrid method in order to compensate their In Hybrid approach training time is very quick receiving the high accuracy rate.

III. PROPOSED SYSTEM

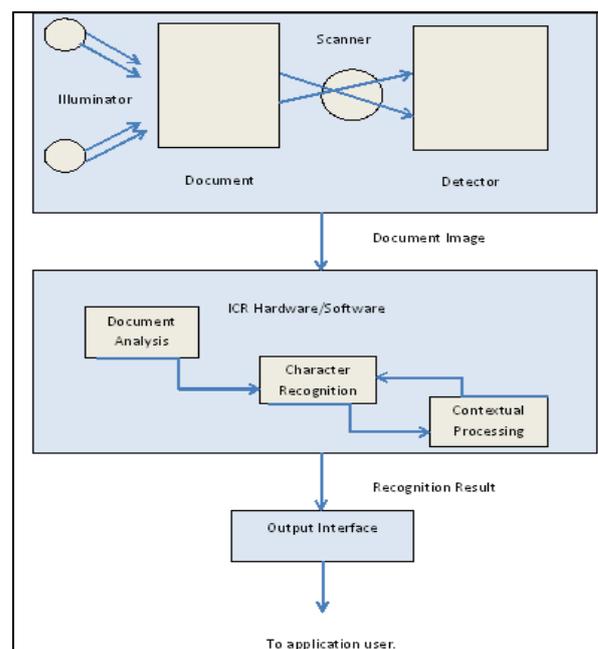


Fig. 1: Architecture of the proposed system

Our proposed system is grid infrastructure based Intelligent Character Recognition which is a system that supports recognition of the characters of multiple languages. The multiple functionalities include editing and searching too whereas the existing system supports only editing of the document. In this context, Grid infrastructure means the infrastructure that supports group of specific set of languages. Thus ICR on a grid infrastructure is multi-lingual. This feature is what we call grid infrastructure which eliminates the problem of heterogeneous character recognition and supports multiple functionalities to be performed on the document. The benefit of proposed system that overcomes the drawback of the existing system is that it provides heterogeneous characters recognition and supports multiple functionalities such as editing and searching.

IV. WORKING OF PROPOSED SYSTEM

Our proposed system is grid infrastructure based Intelligent Character Recognition which is a system that supports recognition of the characters of multiple languages.

The basic steps involved in Grid Infrastructure based ICR are:

- 1) Image Processing
- 2) Image Training
- 3) Image Recognition
- 4) Image Editing
- 5) Image Searching.

A. Image Processing

The activities involved in this step are:

- 1) Scanning an Image
- 2) Saving the image
- 3) Pre-processing the image
- 4) Recognizing the characters in the image
- 5) Generating grid infrastructure data structure.

In the recognition process, it uses the ICR method in support of grid infrastructure data structure.

B. Image Training

The next step after Image processing is Image Training. Before converting the printed images in to editable and searchable documents, the mandatory step is providing training to the system. Training is to make system understand the font followed in the scanned document. All the characters that are required for recognition from the scanned document should be typed and saved as an image file.

1) Kohonen Neural Network:

The Kohonen neural network contains only an input and output layer of neurons, but there is no hidden layer in a Kohonen neural network. The input patterns presented to the Kohonen neural network are the dot image of the character that was hand written. The Kohonen neural network should classify the input pattern into one of the 26 input patterns. The 26 output neurons correspond to the 26 English alphabets. As the Kohonen neural network is trained the weights should be adjusted so that the input patterns are classified into the 26 output neurons.

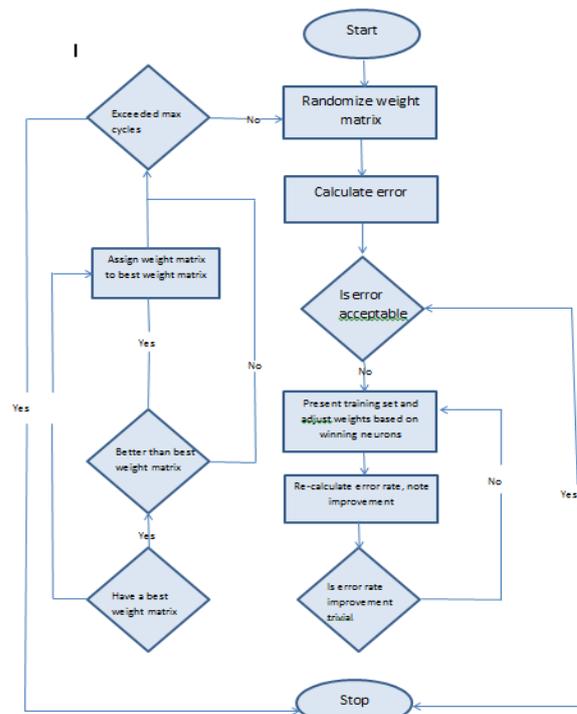


Fig. 2: Working of Kohonen neural network

C. Image Recognition

In this step two types of recognitions are possible. They are Hand written and Scanned document. In Hand written recognition the characters of any language can be recognised due to the adaptation of grid infrastructure. In handwritten recognition, the handwriting of the user in any language trained to the system only for the first time.

V. CONCLUSION

To conclude, a robust algorithm has been described for extraction of text from scanned images, measuring and correcting the skew and shift values that are present in a scanned image. Subsequently, three techniques, viz. (i) pre-processing techniques, (ii) Recognition using Kohonen Neural Networks and (iii) Recognition using Feed Forward Neural Networks, that together comprise the handwritten character and scanned image extraction process have been presented. The impact and necessity of these methods on the overall performance of the ICR system has been illustrated by examples. It is observed from all the above explanations, that the character recognition is a very small part of a very vast field of Digital Signal Processing, it is considered to be a boon for many institutions for different purposes. The other applications like automated entry of data by OCR is one of the most attractive, labour reducing technology, pattern recognition of the scanned digital images from the satellite and comparing them with the previous images is also one of the application. The prediction about the climate can be made effectively by doing this. It can be implemented in many branches like banking systems, satellite communications where there is a necessity for extraction of handwritten and scanned images.

REFERENCES

- [1] Shamik Sural,P.K.Das,Recognition of an Indian Script using Multilayer Perceptrons and Fuzzy Features Sixth

- International Conference on Document Analysis and Recognition (ICDAR2001), Seattle, 2001, pp. 1120-1124.
- [2] Mamta Maloo, Dr. K.V. Kale, Gujarati Script Recognition: A Review, IJCSI International Journal of Computer Science Issues, Vol . 8, Issue 4, No 1, July 2011 ISSN (Online): 1694-0814
- [3] Sujata S. Magare and Ratnadeep R. Deshmukh, Offline Handwritten Sanskrit Character Recognition Using Hough Transform and Euclidean Distance, International Journal of Innovation and Scientific Research ISSN 2351-8014 Vol. 10 No. 2 Oct. 2014, pp. 2- 302
- [4] Rajiv Kapoor, AmitDhamija, A New Method for Identification of Partially Similar Indian Scripts, International Journal of Image Processing (IJIP), Volume (6) : Issue (2) : 2012
- [5] Swapnil A. Vaidya, Balaji R. Bombade, A Novel Approach of Handwritten Character Recognition using Positional Feature Extraction, IJCSMC, Vol. 2, Issue. 6, June 2013, pg.179 – 186
- [6] Iman Yousif, Adnan Shaout, Off-Line Handwriting Arabic Text Recognition: A Survey, Volume 4, Issue 9, September 2014 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering
- [7] Aradhana A Malanker, Prof. Mitul M Patel, Handwritten Devanagari Script Recognition: A Survey, IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE) e-ISSN: 2278-1676,p-ISSN: 2320-3331, Volume 9, Issue 2 Ver. II (Mar – Apr. 2014), PP 80-87

