ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

Comparative Study of Optical Character Recognition Techniques

Farhan Inamdar[1], Prof. S. B. Bagal[2]

M.E Student, KCT's Late G. N. Sapkal College of Engineering Nashik, India[1]

Principal, KCT's Late G. N. Sapkal College of Engineering Nashik, India[2]

ABSTRACT: The Optical Character Recognition is the electronic transformation of image of typewritten or printed text into machine-encoded text. It is common method of digitizing printed texts. It has become an important and widely used technology for pattern recognition. In this study, two different methods are implemented and compare for OCR . First is back propagation network algorithm is combined with genetic algorithm to achieve both accuracy and training swiftness for recognizing alphabets .In second method correlation method is combine with genetic algorithm to save time and increase accuracy. The objective is to develop, optimise and identify user friendly application which performs conversion of image into editable and searchable data. The OCR takes image as the input, gets text from that image and then converts it into editable document.

KEYWORDS: OCR, Backpropogation network, correlation, Genetic algorithm.

I. INTRODUCTION

Character recognition is the transformation process which can classify the input character according to the predefined character layout or class. With the ongoing digital revolution in the world, there has been increasing demand of application through use of which all documents of importance can be converted into editable document and can be stored online. This may be done with the purpose of storing valuable data online, so that it cannot be lost, or for saving time and labor demanded in converting handwritten or typed documents into editable document. Therefore modern society needs the input text into computer readable form. Character Recognition is a common method of digitizing printed texts so that it can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as machine translation, text-to-speech, key data and text mining.

The most popular method used in optical character recognizing is nevertheless backpropagation network. This method weakness is the required time to achieve the best result for recognizing alphabets tends to be long. Backpropagation itself could do the preprocessing phase for alphabet recognition less complex than genetic algorithm (Negnevitsky, 2005). Genetic algorithm would be used to optimize what a standard backpropagation network lacks, architecture and initial weights. This algorithm is often used to find an optimal solution in complex problems (Matic, 2010) by adapting the law of natural selection and natural genetics called survival of the fittest (Malhotra*et al*., 2011). To achieve both better accuracy and less training time needed, genetic algorithm is being used to optimize the backpropagation network. Genetic algorithm focused on exploring the best architecture possibilities and the best weight initial values to be then inputted for the backpropagation network's structure.

Quadri and Asif (2009)developed a system to identify the characters on the numberplate of the vehicle. The optical character system developed by them compares eachindividual character against the complete alpha numeric database. The system uses thecorrelation method to match the individual characters and finally the number isidentified. To increase the accuracy and save the time this correlation method is optimised with the help of genetic algorithm.

II. LITERATURE REVIEW

The first OCR was designed in 1965 based on technology proposed primarily by the Jacob Rainbow which was used by the United States Postal Services. Then in 1970s, Dr. Sinha of Indian Institute of Technology, Kanpur made efforts to propose pattern analysis system. In 1974, Ray Kurzwell developed Omni-font OCR which could recognize text printed in virtually any form. In 2000s, OCR was made available online as a service (WebOCR). OCR system has been

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

designed for most common writing system which includes Latin, Arabic, Indic, Bengali, Devanagri, Chinese, etc using most common programming languages MATLAB, ANN, LABVIEW, TESSERACT.

Claudiu et al. (2011) [1] has investigated using simple training data pre-processing gave us experts with errors less correlated than those of different nets trained on the same or bootstrapped data. Hence committees that simply average the expert outputs considerably improve recognition rates. Our committee-based classifiers of isolated handwritten characters are the first on par with human performance and can be used as basic building blocks of any OCR system (all our results were achieved by software running on powerful yet cheap gaming cards).

Badawy, W. et al. (2012) [6] has discussed the Automatic license plate recognition (ALPR) is theextraction ofvehicle license plate information from an image ora sequence of images. The extracted information can be used with or without a database in many applications, such as electronic payment systems (toll payment, parking fee payment), and freeway and arterial monitoring systems for traffic surveillance. The ALPR uses either a color, black and white, or infrared camera to take images.

Ntirogiannis et al. (2013) [7] has studied that the document image binarization is of great importance in the document image analysis and recognition pipeline since it affects further stages of the recognition process. The evaluation of a binarization method aids in studying its algorithmic behaviour, as well as verifying its effectiveness, by providing qualitative and quantitative indication of its performance. This paper addresses a pixel-based binarization evaluation methodology for historical handwritten/machine-printed document images. In the proposed evaluation scheme, the recall and precision evaluation measures are properly modified using a weighting scheme that diminishes any potential evaluation bias.

Genetic algorithm is an algorithm for optimization and machine learning based loosely on several features of biological evolution. It can reduce badn components in programming and replace it with a better one. These components are called genes, as in biology, in genetic algorithm. Genetic algorithms are a class of parallel adaptive search algorithms based on the mechanics of natural selection and natural genetic system. It can find the near global optimal solution in a large solution space quickly. It has been used extensively in many application areas, such as image processing, pattern recognition, feature selection and machine learning (Majida*et al.*, 2010). On a population of chromosomes (individuals) in genetic algorithm, each chromosome has its own genes. From that population, a selection process will occur to find a chromosome with the best fitness/survival rate to be crossover-ed to produce a new chromosome with better fitness. The new chromosome will then replace the chromosome with the lowest survival rate. This process will be iterated until the desired error rate is achieved.

III. RELATED WORK

A) Backpropagation Network**:** Backpropagation network is a multi layer neural network. Backpropagationis more like the learning/training algorithm rather than the network itself. Backpropagation use supervised training algorithm for multi layer network, therefore the input and target output has been prepared for the training process. A data error in output layer is counted using network output and target output. The data errors then back propagated to thehidden layer, resulting in weight change for the synapses heading to the hidden layer (Pinjare and Kumar, 2012).Basically, backpropagation network consists of two phases, feedforward phase and backward phase.

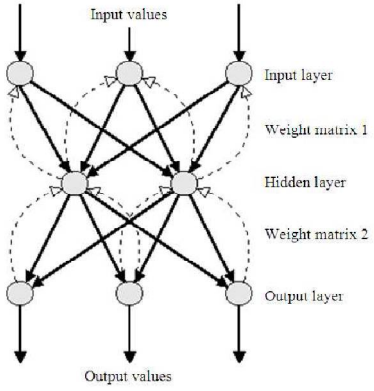


Figure 2: Backpropogation network

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

1. Feed forward Phase: In this phase, there are no cycles. Information only goes one way, from the input layer to the hidden layer (if there is any) and then to the output layer. Each node in output layer sum up the input nodes' weights Equation and use the activation function to measure the output(Sutojo*et al*. (2011))

2. Backward Phase :In the backward phase, weight readjusments are to be done. Backward phase algorithm (Rahajaan, 2011). Calculate factor δ output unit based on errors in every output unit. δ_k = Error unit that will be used for readjustment on the lower layer: Calculateweight adjustment with learning rate.Calculate factor δ output unit based on errors in every hidden unit then Calculate weight adjustment value.Calculate every weight readjustments.

3. Genetic Algorithm Optimized Neural Network: Genetic algorithm runs twice in this system, first is to determine the network architecture and sec, to determine the weight for the network synapses. Basic processes in genetic algorithm, such as population initialization, fitness calculation, selection, crossover and mutation, executed as much as the number of desired generations. In this optical character recognition system, the number of generations has been set to 10 before hand to find the optimal architecture and optimal weight. Each generation consist of five training data sets. Each set consists of 26 alphabetical characters, from 'A' to 'Z'. Population Initialization is the first step in geneticalgorithm process. This starts with initializing randomnumbers as much as the multiplication of chromosomenumbers and chromosome base for each chromosome.For the architecture, each chromosome consists of 9chromosome bases and there are 50 chromosomes asthe initial population. Base for the architecture is abinary number, 0 or 1.Each chromosome represents the number of nodes inthe neural network's hidden layer. With the length of 9bases per chromosome, the architecture of hidden layercould consist of 270 to 405 nodes. Outside thoseboundaries, there will be a chromosome re-initializationbased on Panchal*et al*. (2011) that optimum nodes inhidden layer range from 2/3 to 3/3 of the sum of inputand output nodes.Fitness calculation needs to be done before we enter theselection phase. Fitness is a comparison value to determinewhich chromosome should be eliminated and replaced.Selection concept in this system is a Tournamentselection. k random candidates will be selected (in thissystem, k = 8) and paired with each other, leaving 4pairs. Then a tournament selection will be executedand candidate with higher fitness value will survive.We do this until it is only a single candidate left withthe biggest fitness value (Sivaraj and Ravichandran,2011). This is held to find 10% or 5 pairs of parent tobe crossover-ed.Crossover will be used to create new chromosomesfrom the 5 pairs of parent. Every crossover will have arandom crossover point according to the chromosomelength. If the new chromosomes' fitness value is biggerthan the previous ones, the worse chromosomes will bereplaced by the new chromosomes.Mutation is executed on the new chromosomes(offspring). Mutation rate is decided by calculating theoffspring fitness value first. A random number of 0 and 1will be used to decide whether the offspring will bemutated or not.

IV. PROPOSED SYSTEM

A. *CORRELATION METHOD WITH GENETIC ALGORITHM:*

1. *Database generation (templates):* Database for the OCR system are collected from various sources and are saved for matching with the characters which are to be recognized. Templates are actually images of characters, which are recognized by their class and structure and are used for matching and comparison as shown in figure 3
2. *Converting colour image to gray scale image :*In today‟s world, almost all scanning and image capturing devices use color. Color images matrices are labeled as red(R), green (G) and blue (B). Techniques provided up there in this proposed system are based on grey scale images and therefore, there is need of initially converting scanned or captured color images to grey scale.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

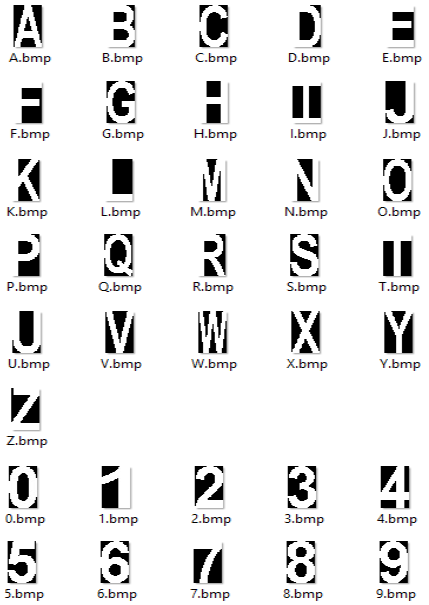


Figure 3: Templates of Characters

3. *Binarisation:* The process of converting a gray scale image into binary image (0 and 1pixel values) as shown in figure 4.
4. *Converting image to array :*Image is converted into array with the help of mat 2 cell command, which form the array of 0 and 1"s as shown in figure 4.

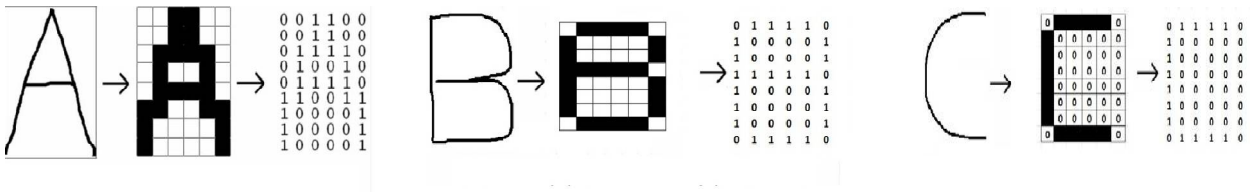


Figure 4: Binarisation of image

5. *Application of GA to System:* Furthermore Roulette Wheel Selection method used as parents to crossover that we have applied to a set of binary numbers (0,1 encoding) representing capital English alphabets. The alphabets ranges from A to Z represent in a matrix of 8 * 6 array dimension (see figure 4) which behaves as initiated value for comparing. Each sample (character) considered a chromosome, which has 48 genes presented as Table 1as below.

	Chromosome consist 48 genes
A	0 0 1 1 0 0 0 0 1 1 0 0 0 1 1 1 1 0 0 1 0 0 1 0 0 1 1 1 1 0 1 1 0 0 1 1 1 0 0 0 0 1 1 0 0 0 0 1
C	0 1 1 1 1 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 1 1 1 1 0
C	0 1 1 1 1 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 1 1 1 1 0

Table 1: Initial chromosome of alphabet

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

In GA, For each population two or more chromosomes are selected to be parents to crossover. The chromosomes with higher fitness have a higher possibility to be selected to produce offspring for the next generation. After many generations of evolution, the optimal solution of the problem is hopefully to be found in the population. The selected fitness function that we have utilized in our research is as follows:

$$F.F = \frac{1}{\sum_{i=1}^{n}(\alpha_i - \beta_i)^2 + 1}$$

Where:

α : Desired gene from the Actual data;

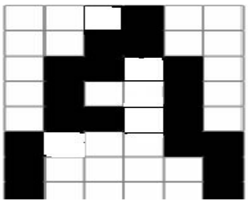
β : Fault level of a chromosome computed by genetic algorithm and

n : the number of genes in chromosome.

Using the above fitness function, we have trained our system like if the fault level of a chromosome is equal to the desired the fitness value of that chromosome will be equal to 1 or if generation number is more than thirty (30) the process of GAs will be terminated.

The system structure can be described as below, Process the character before recognition since procedures takes the unknown character and then return the unknown character without empty or points in to array. We have used 30 generations, for each generation 10 chromosomes and for each chromosome has 48 genes. The genes of all individuals consist of either 0 or 1(assuming a binary encoding for simplicity).

- **The First Generation:**Firstly, we generate chromosomes by combining all features extracted from unknown character in array.
- **Test for unclear A:**For example if we have the following image Character of unclear A



Chromosome consist 48 genes																																															
0	0	0	1	0	0	0	0	1	1	0	0	0	1	1	0	1	0	0	1	0	0	1	0	0	1	1	0	1	0	1	0	0	0	1	1	1	0	0	0	0	1	1	0	0	0	0	1

Figure 5: Test for unclear A

Choosing two chromosomes with a higher fitness and applying crossover and mutation operation to get a new chromosome, which is considered an initial for a second offspring and so on (see bold rows below) and because of the great fitness function values of these selected chromosomes which mean that character have enough recognition from the first steps and hence the process stop.

Chromosomes consist 48 genes																																																Fitness F.
0	0	0	1	0	0	0	0	1	1	0	0	0	1	1	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0	1	0.143
0	0	1	0	0	0	0	0	1	0	0	0	1	1	1	1	0	0	0	1	0	0	1	0	1	0	1	0	0	0	1	0	0	0	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0.1
0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	0	1	0	1	0	1	0	0	0	1	0	0	0	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0.125
0	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**1**	**1**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**1**	**0**	**1**	**0**	**1**	**0**	**0**	**0**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**0**	**1**	**0.25**
0	0	1	0	0	0	0	0	1	0	0	1	0	1	1	1	0	0	0	1	0	0	1	0	1	0	1	0	0	0	0	1	0	0	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0.11
0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	1	0	0	0	1	0	0	0	1	1	1	0	0	0	0	1	1	0	0	0	0	1	0.11
0	**0**	**1**	**1**	**0**	**0**	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**1**	**1**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**1**	**1**	**0**	**1**	**0**	**1**	**0**	**0**	**0**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**0**	**1**	**0.33**
0	0	1	0	0	0	0	0	1	0	1	0	0	1	1	1	0	0	0	1	0	1	1	1	1	0	1	0	0	0	1	0	0	0	1	1	1	0	1	0	0	1	1	1	0	0	0	1	0.08
0	1	1	0	0	1	1	0	1	0	0	0	0	1	1	1	0	0	0	1	0	0	1	0	1	0	1	1	0	0	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	0	0	1	0.09
0	0	1	1	0	0	0	0	1	0	0	0	0	1	1	1	0	0	0	1	0	0	1	1	1	0	1	0	0	0	1	0	0	0	1	1	1	0	0	1	0	1	1	0	0	0	0	1	0.11

Table 2 : fitness function value for generation 1

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

Two chromosomes with the high fitness function value are as follows

Chromosomes consist 48 genes	Fitness
0 0 1 1 0 0 0 0 1 1 0 0 0 1 1 0 1 0 0 1 0 0 1 0 0 1 1 0 1 0 1 0 0 0 1 1 1 0 0 0 0 1 1 0 0 0 0 1	0.33
0 0 0 1 0 0 0 0 1 1 0 0 0 1 1 0 1 0 0 1 0 0 1 0 0 1 1 0 1 0 1 0 0 0 1 1 1 0 0 0 0 1 1 0 0 0 0 1	0.25

Table 3 : Chromosome for with best fitness value

Now we will choose two chromosomes with a higher fitness and applying crossover and mutation operation to get new chromosomes, the two selected chromosomes are as bellow.Using fitness calculation function we have found that a function values are greater than the previous two (02) bold chromosomes hence a crossover for these chromosomes are given as follows.

Table 4 : Crossover

Moreover the all configured chromosomes and then crossover process or mutation as follows:

Table 5: Mutation

After account of fitness value that genetic algorithm is able to identify the character in the 30th generation.

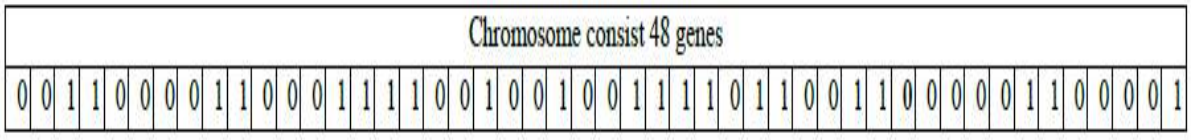


Table 6: Final chromosome after 30th generation

Above Table consists final chromosome, which specify character A after testing with different only in one gene (in the 37th position) comparing with Table 6 which specify normal (clear) character A as target.In the similar way, we can recognise all alphabet and numbers as well.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

V. RESULT

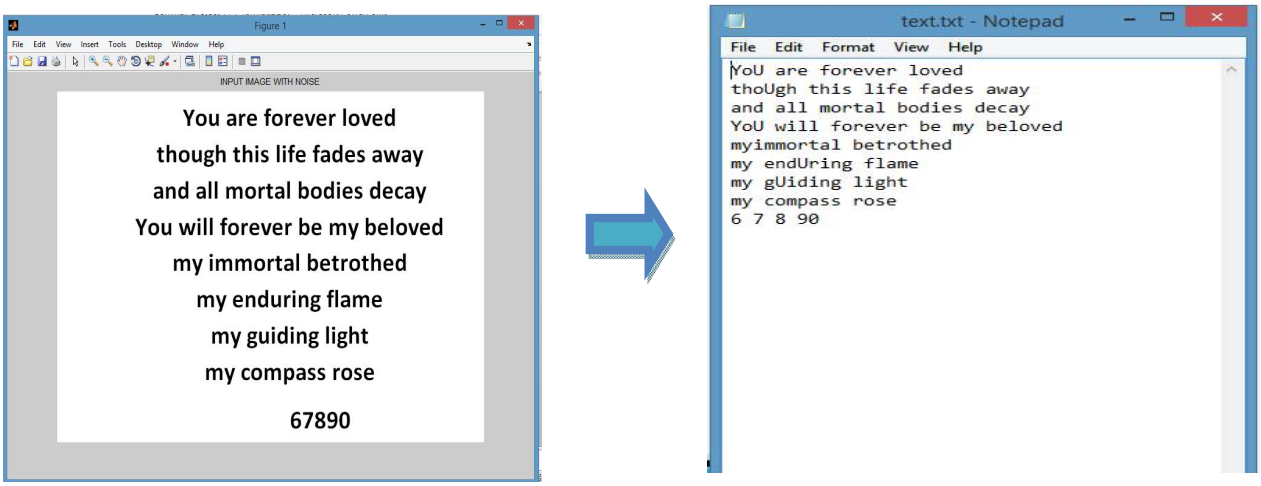


Figure 6 :Image sample and its output

Experiments have been performed to test the proposed method.The experiments were performed on alphabetical characters. Table 7Gives the average results of recognition accuracy by standard Backpropagation Network (BPN) , by optimized Genetic Algorithm Neural Network (GABPN) also by Correlation method using GA alphabetical characters. The backpropagation optimized with genetic algorithm showed significant improvement than BPN, moreover correlation with GA gives you the very high accuracy. By using genetic algorithm, we get high accuracy within less time.

Alphabetical Character	BPN	GABPN	Correlation using GA
A – Z	80%	90.77%	95-97%

Overall, the system was then tested for most of 26 characters and it shows 95-97% of characters recognition accuracy which is more than the system result, GABPN has an accuracy rate of 90.77% while BP's 80%.. The genetic algorithms being trained using standard templates of the capital alphabets and then calculate a fitness function value.Training accomplish that genetic algorithm was able to identify the character in the 30th generation.

VI. CONCLUSION AND FUTURE WORK

The backpopagation is studied and successfully optimized with genetic algorithm. Optical character recognition application can be built with combination of genetic algorithm and neural network to achieve more accurate application and shorter learning time. This can help any similar recognizing application to accurately recognize with less error.
Also proposed method signifies important achievement of selecting and using Genetic Algorithms methods being an interesting field of AI. These algorithms have better capability not only in classification or recognition but also in responding of learning faster as compared with others algorithms which produces powerful function in optimization. Overall our study shows that Gas generate better results through training of system and testing in record time as compared with other algorithms.
It could be interesting to test the neural networks on the images we rejected to see how different the recognition rate would be. In this proposed method the chromosome encodings were only tested against optical character recognition problem with limited character set (English alphabets and numbers). There might be an interest in testing it against other character sets of different sizes.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

REFERENCES

[1] Dan ClaudiuCires¸an and Ueli Meier and Luca Maria Gambardella and JurgenSchmidhuber, "Convolutional Neural Network Committees for Handwritten Character Classification", *2011 International Conference on Document Analysis and Recognition, IEEE*, 2011
[2] Majida A., A.N. Ismail and Z.M. Hazi, 2010. Pattern recognition using genetic algorithm. Int. J. Comput. Electr. Eng., 2: 1793-8163
[3] 1Ankit Kumar Singh, 2Aman Gupta, 3Aman Saxena, 'Optical Character Recognition: A Review', *JETIR (ISSN-2349-5162)*, April 2016.
[4] Hendy Yeremia, NikoAdrianusYuwono, Pius Raymond and WidodoBudiharto 'Genetic Algorithm And For Optical Character Recognition', *Journal of Computer Science 9 (11): 1435-1442, 2013, ISSN: 1549-3636* , 2013
[5] Khaled M.G Noaman, JamilAbdulhameed M. Saif, Ibrahim A.A. Alqubati, 'Optical Character Recognition Based on Genetic Algorithms', *Journal of Emerging Trends in Computing and Information Sciences, Vol. 6, No. 4* April 2015
[6] Badawy, W. "Automatic License Plate Recognition (ALPR): A State of the Art Review." (2012): 1-1.
[7] Ntirogiannis, Konstantinos, Basilis Gatos, and IoannisPratikakis. "A Performance Evaluation Methodology for Historical Document Image Binarization." (2013): 1-1.
[8] RavinaMithe, SupriyaIndalkar, NilamDivekar, "Optical character recognition" published under International Journal of Recent Technology and Engineering (IJRTE), Vol. 2, Issue 1, March 2013, ISSN: 2277-3878.
[9] Neetu Bhatia, "Optical Character Recognition Techniques: A Review" published under International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol. 4, Issue 5, May 2014, ISSN: 2277 128X.
[10] Patra, S.R., R. Jehadeesan, S. Rajeswari, S.A.V.S. Murty and M.S. Baba, 'Development of geneticalgorithm based neural network model for parameter estimation of fast breeder reactor subsystem', *Int. J.'SoftComput. Eng., 2: 87-90*.
[11] Panchal, G., A. Ganatra, Y.P. Kosta and D. Panchal, "Behaviour analysis of multilayer perceptrons with multiple hidden" *Int. J. Comput. Theory Eng.,3: 332-337*. 2011
[12] Sivaraj, R. and T. Ravichandran,."A review of selection methods in genetic algorithm". Int. J. Eng. Sci. Tech., 3: 3792-3797. 2011
[13] Sutojo, T., E. Mulyanto, and V. Suhartono, 2011.Kecerdasan buatan. ANDI, Yogyakarta.