

# OPTICAL CHARACTER RECOGNITION TECHNIQUE ALGORITHMS

<sup>1</sup>N. VENKATA RAO, <sup>2</sup>DR. A.S.C.S.SASTRY, <sup>3</sup>A.S.N.CHAKRAVARTHY, <sup>4</sup>  
KALYANCHAKRAVARTHI P

<sup>1</sup>Research Scholar, Department of ECE , K.L.University, Vaddeswaram, Andhra Pradesh, India

<sup>2</sup>Professor, Department of ECE, K.L.University, Vaddeswaram, Andhra Pradesh, India

<sup>3</sup>Professor, Department of CSE , University college of Engineering, Vizianagaram, Andhra Pradesh, India

<sup>4</sup>Assistant Professor, Department of ECE, GMR Institute of Technology, Rajam, India Pradesh, India

E-mail: <sup>1</sup>venkatnekkanti@rediffmail.com, <sup>2</sup>ascssatry@kluniversity.in,  
<sup>3</sup>chakravarthy.cse@jntukucev.ac.in, <sup>4</sup>kalyanecebujji@gmail.com

## ABSTRACT

In this paper, we present a new neural network (NN) based method for optical character recognition (OCR) as well as handwritten character recognition (HCR). Experimental results show that our proposed method achieves increased accuracy in optical character recognition as well as handwritten character recognition. We present through an overview of existing handwritten character recognition techniques. All the algorithms describes more or less on their own. Handwritten character recognition is a very popular and computationally expensive task; we describe advanced approaches for handwritten character recognition. In the present work, we would like to compare the most important once out of the variety of advanced existing techniques, and we will systematize the techniques by their characteristic considerations. It leads to the behaviour of the algorithms reaches to the expected similarities.

**Keywords-** OCR, HCR, Neural Network. Recognition Technique

## 1. INTRODUCTION

Character recognition is an art of detecting segmenting and identifying characters from image. More precisely, character recognition is a process of detecting and recognizing characters from input image and converts it into American Standard Code for Information Interchange (ASCII) or other equivalent machine editable form [1], [2], [3]. It contributes immensely to the advancement of automation process and improving the interface between man and machine in many applications [4]. Character recognition is one of the most interesting and fascinating areas of pattern recognition and artificial intelligence [5], [6]. Character recognition is getting more and more attention since last decade due to its wide range of applications.

Conversion of handwritten characters is important for making several important documents related to our history, such as manuscripts into machine editable form, so that, it can be easily accessed and pres independent work is going on in Optical Character Recognition that is the processing of printed/computer generated document, handwritten

and manually created document processing i.e. handwritten character recognition. Figure 1 (a) and 1(b) represents the offline and online character recognitions.

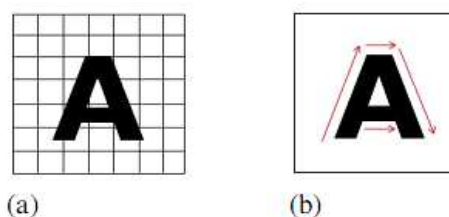


Figure 1. (a) Offline character recognition, (b) Online character recognition.

Offline character recognition system generates the document first, digitalizes, and stored in computer and then it is processed. Whereas, in case of online character recognition system, character is processed while it was under creation. External factors like pressure, speed of writing have any influence in case of offline system but they have great impact on online system. Again, offline or online system can

be applied on optical character (Fig 2. (a)) or handwritten characters (Fig 2.(b)). Based on that, systems can be classified as OCR or HCR respectively. Online methods are superior to their counterparts i.e. offline methods due to the temporal information present in the character generation [4].

Accuracy of HCR is still limited to 90 percent due to presence of large variation in shape, scale, style and orientation etc., [8]. Character processing systems are domain and application specific, like the other systems it is not possible to design generic systems which can process all kinds of scripts and languages. Lots of work has been carried out on European languages and Arabic (Urdu) language. Whereas domestic languages like Hindi, Punjabi, Bangla, Tamil and Gujarati etc., are very less explored due to limited usage. In this paper, our focus is to carry out in depth literature survey on handwritten character recognition methods.

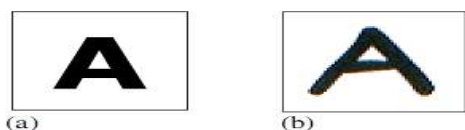


Figure 2. (a) Optical character (b) Handwritten character.

Image processing and pattern recognition plays significant role in handwritten character recognition. Rajbala et al [10], have discussed various types of classification of feature extraction methods like statistical feature based methods, structural feature based methods and global transformation techniques. Statistical methods are based on the planning of how the data should be selected. It uses the information of statistical distribution of pixels in image, it can be mainly classified in three categories: 1). Partitioning in regions, 2). Profile generation and projections 3) distances and crossing. Structural features are extracted from structure and geometry of character like number of horizontal and vertical lines, aspect ratio, number of cross points, number of loops, number of branch points, number of strokes and number of curves etc. Global transformation features are calculated by converting the image in frequency domains like Discrete Fourier Transformation (DFT), Discrete Cosine Transformation (DCT), Discrete Wavelet Transformation (DWT), Gabor filtering, and Walsh-Hadamard transformation etc.

Features extracted can be either low level or high level. Low level features include width, height, curliness and aspect ratio etc., of the character. These alone cannot be used to distinguish one character from another in the character set of the languages [11]. So, there are the number of other high level features which includes number and position of loops, straight lines, headlines and curves etc. Tirthraj Dash et al discussed HCR using associative memory net (AMN) in their paper [12], represents the direct work at pixel level. Dataset was designed in MS Paint 6.1 with normal Arial font size 28(twenty eight). Dimension of image was kept 31X39. Once characters are extracted, their binary pixel values are directly used to train AMN. I. K. Pathan et al have proposed offline approach for handwritten isolated Urdu characters in their work [13]. Urdu character may contain one, two, three or four segments, in which one component is known as primary (generally represents large continuous stroke) and rest of all are known as secondary components (generally represents small stroke or dots). Authors have used moment invariant (MI) feature to recognize the characters. MI features are well known to be invariant under rotation, translation, scaling and reflection. MI features are the measure of pixel distribution around the centre of gravity of character and it captures the global character shape information. If character image is single component than it is normalized in 60X 60 pixels and horizontally divided into three (3) equal parts. 7 MIs are extracted from each zone and 7 MIs are calculated from overall image, so, total of 28 features are used to train support vector machines (SVM), if image is having multi components then 28 MIs are extracted from primary component (60 X 60) and 21 MIs are extracted from secondary component (22 X 22). Separate SVMs are trained for both and decision is taken based on the rules satisfying some criteria. Proposed system claim to get highest accuracy of 93.59 %. In paper [4], Pradeep et al has proposed neural network based classification of handwritten character recognition system. Each individual character is resized to 30 X 20 pixels for processing; they have used the binary features to train the neural network. However, such features are not robust. In post processing stage, recognized characters are converted to ASCII format. Input layer has 600 neurons equal to number of pixels. Output layer has 26 neurons as English has 26 alphabets. The proposed ANN uses back propagation algorithm.

## 2. COMPARISION OF OCR TECHNIQUES

Various techniques used for the design of OCR by their characteristics.

**Matrix Matching**

Matrix Matching converts each character into a pattern within a matrix, and then compares the pattern with an index of known characters. Its recognition is strongest on monotype and uniform single column pages.

**Fuzzy Logic**

Fuzzy logic is a multi-valued logic that allows intermediate values to be defined between conventional evaluations like yes/no, true/false, black / white etc. An attempt is made to attribute a more human-like way of logical thinking in the programming of computers. Fuzzy logic is been used when answers do not have a distinct true or false values and there are uncertainly involved.

**Feature Extraction**

This method defines each character by the presence or absence of key features, including height, width, density, loops, lines, stems and other character traits. Feature extraction is a perfect approach for OCR of magazines, laser print and high quality images.

**Structural Analysis**

Structural Analysis identifies characters by examining their sub feature shapes of the image, sub-vertical and horizontal histograms. Its character repair capability is great for low quality text and newsprints.

**Neural Networks**

This strategy simulates the way the human neural system works; it samples the pixels in each image and matches them to a known index of character pixel patterns. The ability to recognize the characters through abstraction is great for fixed documents and damaged text. Neural networks are ideal for specific types of problems, such as processing stock market data or finding trends in graphical patterns. In all these approaches Neural Networks are efficient than others.

A neural network is a powerful data modelling tool that is able to capture and represent complex input/output relationships. Motivation for the development of neural network technology stemmed from the desire to develop an artificial system that could perform "intelligent" tasks similar to those performed by the human brain. Neural networks resemble the human brain in the following two ways: (1) A neural network acquires knowledge through learning; (2) A neural network's knowledge is stored within inter-neuron connection strengths known as synaptic weights.

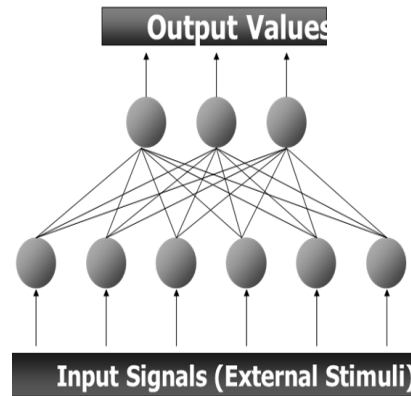


Figure 3: MLP Network

The most common neural network model is the multilayer perception (MLP), this type of neural network is known as a supervised network because it requires a desired output in order to learn. The goal of this type of network is to create a model that correctly maps the input to the output using historical data so that the model can be used to produce the output when the desired output is unknown. A graphical representation of an MLP is shown below.

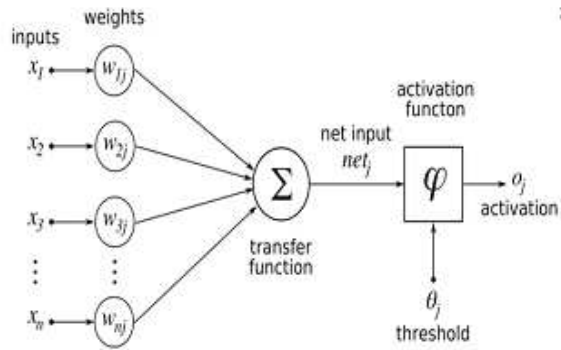


Figure 4: Structure of ANN

**3. ARTIFICIAL NEURAL NETWORK**

**4. COMPONENTS OF OCR SYSTEM**

**Optical scanning**

Through the scanning process is the digital image of the original document is captured. Whereas OCR optical scanners generally consist of a transport mechanism plus a sensing device that converts light intensity into gray-levels. Printed documents usually consist of black print on a white background; hence, when performing OCR, it is common practice to convert the multilevel image into a bi-level image of black and white. Often, this process is known as thresholding, is performed on the scanner to save memory space and computational effort.

### Location and segmentation

Segmentation is a process that determines the constituents of an image, it is necessary to locate the regions of the document where data have been printed and distinguish them from figures and graphics. For instance, when performing automatic mail-sorting, the address must be located and separated from other print on the envelope like stamps and company logos, prior to recognition.

### Pre-processing

The image resulting from the scanning process may contain a certain amount of noise depending on the resolution of the scanner and the success of the applied technique for thresholding, the characters may be smeared or broken. Some of these defects, which may later cause poor recognition rates, can be eliminated by using a pre-processor to smooth the digitized characters.

### Feature extraction

The objective of feature extraction is to capture the essential characteristics of the symbols, and it is generally accepted that this is one of the most difficult problems of pattern recognition. The most straight forward way of describing a character is by the actual raster image. Another approach is to extract certain features that still characterize the symbols, but leaves out the unimportant attributes.

### Post processing

They are two types of post processing,

1. Grouping
2. Error-detection and correction.

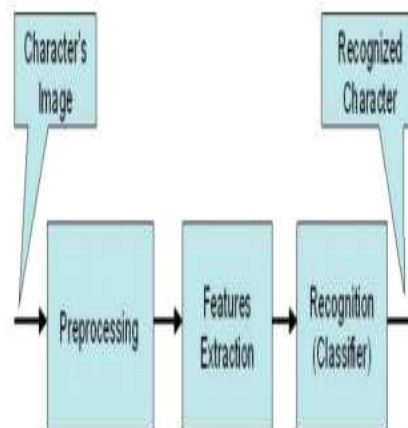


Figure 5: Component of OCR

## 5. PROPOSED SYSTEM FUTURE SCOPE OF WORK

Aim of the proposed system is to develop a neural network based method for accurate optical character recognition. The algorithm will learn from the training data set and will provide 100 percent accurate optical character recognition result. Here we develop a neural network based method for accurate hand written character recognition. OCR is the acronym for Optical Character Recognition; this technology allows a machine to automatically recognize characters through an optical mechanism. Human beings recognize many objects in this manner our eyes are the "optical mechanism." But while the brain "sees" the input, the ability to comprehend these signals varies in each person according to many factors. By reviewing these variables, we can understand the challenges faced by the technologist developing an OCR system. The ultimate objective of an OCR system is to simulate the human reading capabilities so the computer can read, understand, edit and do similar activities it does with the text. Block diagram of the typical OCR system shown in fig 4. Each stage has its own problems and effects on the overall system's efficiency. Thus, to tackle all the problems, either by solving each particular problem. OCR system by integrating all stages to one main stage, and this is what our research proposes. This paper presents new structure of OCR system which relies on the powerful proprieties. The algorithm is designed and tested in the related sections.

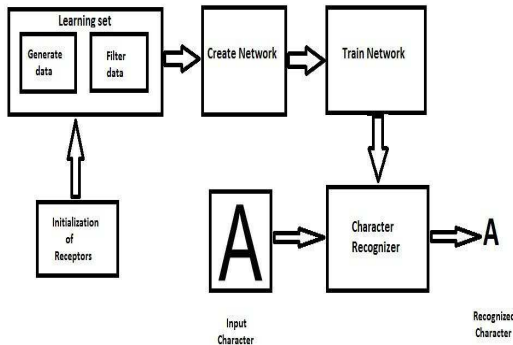


Figure 6: Structure of the proposed system.

Creation network consists of the following structure for our proposed algorithm

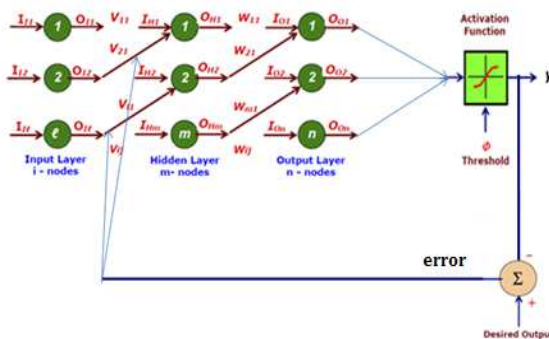


Fig 7: Network used for proposed algorithm.

6. PROPOSED ALGORITHM

- The input pattern is presented to the input layer of the network.
- These inputs are propagated through the network until they reach the output units.
- This forward pass produces the actual or predicted output pattern, because, back propagation is a supervised learning algorithm, the desired outputs are given as part of the training vector.
- The actual network outputs are subtracted from the desired outputs and an error signal is produced.
- This error signal is the basis for the back propagation step, whereby the errors are passed back through the neural network by computing the contribution of each hidden

processing unit and deriving the corresponding adjustments needed to produce the correct output.

- The connection weights are then adjusted and the neural network has just “learned” from an experience. Once the network is trained, it will provide the desired output for any of the input patterns.
- The network undergoes supervised training, with a finite number of pattern pairs consisting of an input pattern and a desired or target output pattern.
- An input pattern is presented at the input layer. The neurons here pass the pattern activations to the next layer neurons, which are in a hidden layer.
- The outputs of the hidden layer neurons are obtained by the weights and the inputs, these hidden layer outputs become inputs to the output neurons, which process the inputs using an optional bias and a threshold function.
- The final output of the network is determined by the activations from the output layer.
- A similar computation, still based on the error in the output, is made for the connection weights between the input and hidden layers.
- The procedure is repeated with each pattern pair assigned for training the network. Each pass through all the training patterns is called a cycle or an epoch. The process is then repeated as many cycles as needed until the error is within a prescribed tolerance.
- The adjustment for the threshold value of a neuron in the output layer is obtained by multiplying the calculated error in the output at the output neuron and the learning rate and the momentum parameter used in the adjustment calculation for weights at this layer.

$$\Theta_k = \alpha e$$

$$Y = f(I) = f \left\{ \sum_{i=1}^n X_i W_i - \Theta_k \right\}$$

- After the network has learned the correct classification for a set of inputs from a training set, it can be tested on a second set of inputs to see how well it classifies untrained patterns.

## 7. APPLICATIONS

Optical character recognition has been applied to number of applications. Some of them have been explained below.

### A. Invoice Imaging

Invoice imaging is widely used in many business applications to keep track of financial records and prevent a backlog of payments from piling up. In government agencies and independent organizations, OCR simplifies data collection and analysis, among other processes. As the technology continues to develop, more and more applications are found for OCR technology, including increased use of hand writing recognition. Furthermore, other technologies related to OCR, such as barcode recognition used daily in retail and other industries.

### B. Legal Industry

Legal industry is also one of the beneficiaries of the OCR technology. OCR is used to digitize documents, and directly entered to computer database. Legal professional scan further search documents required from huge data bases by simply typing a few keywords.

### C. Banking

Another important application of OCR is in banking, where it is used to process cheques without human involvement. A cheque can be inserted into a machine where the system scans the amount to be issued and the correct amount of money is transferred. This technology has nearly been perfected for printed checks, and is fairly accurate for hand written checks as well reducing the waiting time in banks.

### D. Healthcare

Healthcare has also seen an increase in the use of OCR technology to process paperwork. Healthcare professionals always have to deal with large volumes of forms for patients, including insurance forms as well as general health forms. To keep up with all of this information, it is useful to input relevant data into an electronic database that can be accessed as necessary. Form processing tools, powered by OCR, are able to extract information from forms and put it into databases, so that every patient's data is promptly recorded.

### E. Captcha

A CAPTCHA is a program that can generate and grade tests that human can pass but current computer programmers' cannot. Hacking is a serious threat to internet usage, now a day's, most of the human activities like economic transactions, admission for education, registrations, travel bookings etc., are carried out through internet and all this requires a password which is misused by hackers. They create programs to like dictionary attacks and automatic false enrolments which lead to waste of memory and resources of website. Dictionary attack is attack against password authenticated systems where a hacker writes a program to repeatedly try different passwords like from dictionary of most common passwords. In CAPTCHA, an image consisting of series of letters of number is generated which is obscured by image distortion techniques, size and font variation, distracting backgrounds, random segments, highlights and noise in the image. This system can be used to remove this noise and segment the image to make the image tractable for the OCR (Optical Character Recognition) systems.

### F. Institutional Repositories and Digital Libraries

Institutional repositories are digital collections of the outputs created within a university or research institution. It is an online locale of intellectual data of an institution, especially a research institution where it is collected, preserved and aired. It helps to open up the outputs of an institution and give it visibility and more impact on worldwide level. Enables and encourages inter disciplinary approaches to research and facilitates the development and sharing of digital teaching materials and aids. It is basically a collection of peer reviewed journal articles, conference proceedings, research data, monographs, books, theses and dissertations and presentations. Their first role is to provide the Open Access literature. Practical implementation of this includes setting up a system which consists of scanner which scans the documents. This scanned document is then fed as an input to an Optical Character Recognition system where information is acquired and retained in digitized form.

### G. Optical Music Recognition

Automated learning system extract information from images and is part of major researches. Optical music recognition (OMR) born in 1950's is a developed field and initially was aimed towards recognizing printed sheets which can be edited into playable form with the help of electronic and electrochemical methods. An OMR system has many applications like processing of different classes of music, large scale digitization of musical data and also it can be used for diversity in musical notation. Image enhancement and segmentation is the basic step and hence the paper focuses on it.

## H. Automatic Number Recognition

Automatic number plate recognition is used as a mass surveillance technique making use of optical character recognition on images to identify vehicle registration plates. ANPR has also been made to store the images captured by the cameras including the numbers captured from license plate. ANPR technology own to plate variation from place to place as it is a region specific technology. They are used by various police forces and as a method of electronic toll collection on pay-per-use roads and cataloguing the movements of traffic or individuals.

## I. Handwriting Recognition

Handwriting recognition is the ability of a computer to receive and interpret intelligible handwritten input from sources such as paper documents, photographs, touch-screens and other devices. The image of the written text may be sensed "off line" from a piece of paper by optical scanning (optical character recognition) or intelligent word recognition. Alternatively, the movements of the pen tip maybe sensed "on line", for example by a pen-based computer screen surface.

## 8. RESULT

We take 3 characters with its 5 pattern and examine that character set.

Table 1: Experimental Result of OCR System

Character	No. of patterns given	Recognition	Not Recognized	Rate (%) of Recognition
L	5	5	0	100%
M	5	5	0	100%
O	5	5	0	100%

Table 2: Printed And Handwritten Document Results

TWDB		
Test Set	Train Set	Recognition Rate
2793 chars	11173 chars	95.44%
HWDB		
Test Set	Train Set	Recognition Rate
1351 chars	5407 chars	94.62 %

## 9. CONCLUSION

In this paper we surveyed a large number of methods of optical character recognition. We analyzed the advantages and drawbacks of various OCR methods. We also proposed a modified back propagation method. It is widely used in neural network. The proposed method computes error rate efficiently, it results in increasing the accuracy of neural network. Our proposed neural network based method is providing 100 % accuracy in OCR.

## REFERENCES

- [1] Kai Ding, Zhibin Liu, LianwenJin, Xinghua Zhu, A Comparative study of GABOR feature and gradient feature for handwritten 17hinese character recognition, International Conference on Wavelet Analysis and Pattern Recognition, pp. 1182-1186, Beijing, China, 2-4 Nov. 2007
- [2] Pranob K Charles, V.Harish, M.Swathi, CH. Deepthi, "A Review on the Various Techniques used for Optical Character Recognition", International Journal of Engineering Research and Applications, Vol. 2, Issue 1, pp. 659-662, Jan-Feb 2012
- [3] Om Prakash Sharma, M. K. Ghose, Krishna Bikram Shah, "An Improved Zone Based Hybrid Feature Extraction Model for Handwritten Alphabets Recognition Using Euler Number", International Journal of Soft Computing and Engineering, Vol.2, Issue 2,



- pp. 504-58, May 2012
- [4] J. Pradeepa, E. Srinivasana, S. Himavathib, "Neural Network Based Recognition System Integrating Feature Extraction and Classification for English Handwritten", International journal of Engineering, Vol.25, No. 2, pp. 99-106, May 2012
- [5] Liu Cheng-Lin, Nakashima, Kazuki, H, Sako, H.Fujisawa, Handwritten digit recognition: investigation of normalization and feature extraction techniques, Pattern Recognition, Vol. 37, No. 2, pp. 265-279, 2004.
- [6] SupriyaDeshmukh, Leena Ragha, "Analysis of Directional Features - Stroke and Contour for Handwritten Character Recognition", IEEE International Advance Computing Conference, pp.1114-1118, 6-7 March, 2009, India
- [7] AmrithaSampath, Tripti C, Govindaru V, Freeman code based online handwritten character recognition for Malayalam using Back propagation neural networks, Advance computing: An international journal, Vol. 3, No. 4, pp. 51-58, July 2012
- [8] RajibLochan Das, Binod Kumar Prasad, GoutamSanyal, "HMM based Offline Handwritten Writer Independent English Character Recognition using Global and Local Feature Extraction", International Journal of Computer Applications (0975 8887), Volume 46 No.10, pp. 45-50, May 2012
- [9] Bhatia, N. and Vandana, Survey of Nearest Neighbor Techniques, International Journal of Computer Science and Information Security, Vol. 8, No. 2, (2001),302-305.
- [10] RajbalaTokas, ArunaBhadu, "A comparative analysis of feature extraction techniques for handwritten character recognition", International Journal of Advanced Technology & Engineering Research, Volume 2, Issue 4, pp. 215-219, July 2012
- [11] AmrithaSampath, Tripti C, Govindaru V, "Freeman code based online handwritten character recognition for Malayalam using backpropagation neural networks", International journal on Advanced computing, Vol. 3, No. 4, pp. 51 - 58, July 2012
- [12] Tirtharaj Dash, Time efficient approach to offline hand written character recognition using associative memory net., International Journal of Computing and Business Research, Volume 3 Issue 3 September 2012
- [13] Imaran Khan Pathan, Abdulbari Ahmed Bari Ahmed Ali, Ramteke R.J., "Recognition of offline handwritten isolated Urdu character ", International Journal on Advances in Computational Research, Vol. 4, Issue 1, pp. 117-121, 2012
- [14] Ashutosh Aggarwal, Rajneesh Rani, RenuDhir, "Handwritten Devanagari Character Recognition Using Gradient Features", International Journal of Advanced Research in Computer Science and Software Engineering (ISSN: 2277-128X), Vol. 2, Issue 5, pp. 8590, May 2012
- [15] Dan ClaudiuCires,an and Ueli Meier and LucaMaria Gambardella and Jurgenschmidhuber, "Convolutional Neural Network Committees for Handwritten Character Classification", 2011 International Conference on Document Analysis and Recognition, IEEE, 2011.