iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VI **Month of publication:** June 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Optical Character Recognition

Rubal Chugh[1], Ms Ashima Arya[2]
[1,2]Department of Computer Science B.M. Instituite of Engineering And Technology (BMIET), Guru Gobind Singh Indraprastha University (GGSIPU), Delhi

Abstract: ***Character Recognition (CR) is the process of recognizing handwritten characters. It is an active area of research and it is used in various applications such as reading license plate numbers, document processing, reading cheque numbers etc. The character recognition is classified into online and offline handwritten characters. In this paper, we mainly focus on the off-line character recognition of scanned handwritten English characters. First we receive the scan copy of handwritten characters, then we perform some pre processing to filter noise from the image. The features of handwritten characters are extracted and finally template matching concept is used to match the characters with the stored database of characters.***
Keywords: ***Character Recognition, Feature extraction, Template Matching***

I. INTRODUCTION

Character recognition [1] is one of the most successful applications of neural network technology. With character recognition technique we transform spatial form text representation into equivalent symbolic representation. Once characters of printed documents are transformed into ASCII computer format then we can use these texts for editing & compress the text for compact storage in computer hard disk. This tasks is not trivial as printed documents received for character recognition can be noisy & distorted. Also characters of printed documents are of different font style and sizes. There are different types of character recognition technique in different types of languages available [2]. Earlier attempts for handwritten character recognition involved the extraction of representative features from the training data. These features were designed manually and special templates were created to detect them. Later on, efforts were made to automatically generate these features. This automatic generation of features improved the recognition rate and made design of a neural network for character recognition easier. Today, many researchers have been done to recognize characters, but the problem of interchanging data between human beings and computing machines is a challenging task. Even today, many algorithms have been proposed by many researchers so-that these characters can be easily recognize [3]. But the efficiency of these algorithms is not satisfactory. The character recognition is classified into online and offline handwritten characters. With online character recognition we recognize characters which are written using PC, PDA or special digitizer. On the other hand off line character recognition recognizes the characters which are printed on paper documents using ones own handwriting. In this paper, we mainly focus on the off-line character recognition of scanned handwritten English characters.

II. APPLICATIONS

The most common application with example for the handwritten character recognition is as follows [5]:

1. Documents processing: People wish to scan in a document and have the text of that document available in a word processor. For example HCR can be used for form processing. Forms are normally used for collecting the public information. Replies of public information can be handwritten in the space provided e.g. Tax Form, automatic accounting procedures used in processing utility bills.
2. Reading license plate numbers: Character recognition system is used widely for recognizing license plate numbers by traffic police e.g. Chinese traffic police uses OCR for reading license plate numbers while vehicle is in moving state.
3. Reading zip-codes for Post Office: Handwritten recognition system can be used for reading the handwritten postal address on letters. Offline handwritten recognition system used for recognition handwritten digits of postcode. OCR can be used to read this code and can sort mail automatically e.g. United State Postal Services (USPS) uses Multi line Optical Character Reader (MLOCR) locates the address block on a mail piece, reads the whole address, identifies the ZIP+4 code generates 9-digit bar code. The character classifier recognizes up to 400 fonts and the system can process up to 45,000 mail pieces per hour.

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

4. Reading Cheques in Banks: The OCR system is used for cheque reading in banks. Cheque reading is the very important commercial application of offline handwritten recognition. Handwritten recognition system plays very important role in banks for signature verification and for recognition of amount filled by user.

5. Signature Verification: OCR can be also used for signature verification to identify the person. Signature identification is the specific field of handwritten identification in which the writer is verified by some specific handwritten text. Handwritten recognition system can be used for identify the person by handwriting, because handwriting may be vary from person to person.

III. TYPES OF CHARACTER RECOGNITION

Character recognition can be classified into following two categories (Figure 1 below) [6]:

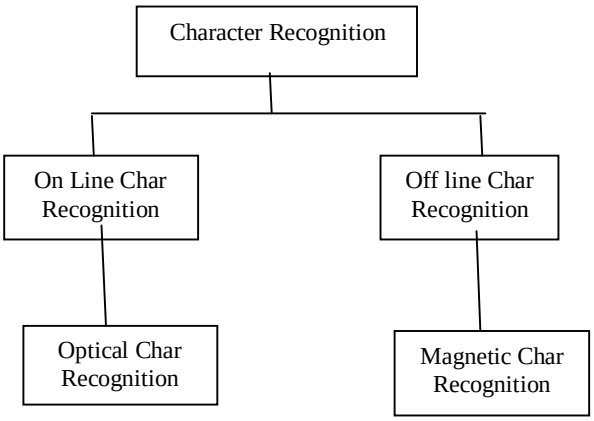


Figure 1: Character Recognition Classification

A. On-line Character Recognition System

On-line character recognition refers to the process of recognizing handwriting recorded with a digitizer as a time sequence of pen coordinates. In case of online handwritten character recognition, the handwriting is captured and stored in digital form via different means. Usually, a special pen is used in conjunction with an electronic surface. As the pen moves across the surface, the two-dimensional coordinates of successive points are represented as a function of time and are stored in order. It is generally accepted that the on-line method of recognizing handwritten text has achieved better results than its off-line counterpart. This may be attributed to the fact that more information may be captured in the on-line case such as the direction, speed and the order of strokes of the handwriting [7].

The on-line handwriting recognition problem has a number of distinguishing features which must be exploited to get more accurate results than the online recognition problem

1) It is adaptive: The immediate feedback is given by the writer whose corrections can be used to further train the recognizer.

2) It is a real time process: It captures the temporal or dynamic information of the writing. This information consists of the number of pen strokes, the order of pen-strokes. The direction of the writing for each pen stroke and the speed of the writing within each

pen stroke.

3) Very little preprocessing is required. The operations such as smoothing, and feature extraction operations such as the detection of line orientations corners loops are easier and faster with the pen trajectory data than on pixel images.

4) Segmentation is easy: Segmentation operations are facilitated by using the pen lift information particularly for hand printed characters.

5) Ambiguity is minimal: The discrimination between optically ambiguous characters may be facilitated with the pen trajectory informationOn the other hand, the disadvantages of the on-line character recognition are as follows:

6) The writer requires special equipment which is not as comfortable and natural to use as pen and paper.

7) It cannot be applied to documents printed or written on papers punching is much faster and easier than handwriting for small size alphabet such as English or Arabic.

B. Off-line Character Recognition System

Off-line handwriting recognition refers to the process of recognizing words that have been scanned from a surface (such as a sheet of paper) and are stored digitally in grey scale format. After being stored, it is conventional to perform further processing to allow superior recognition [8].

The offline character recognition can be further grouped into two types:

1) Magnetic Character Recognition (MCR)

2) Optical Character Recognition (OCR)

In MCR, the characters are printed with magnetic ink. The reading device can recognize the characters according to the unique magnetic field of each character. MCR is mostly used in banks for check authentication. OCR deals with the recognition of characters acquiring by optical means, typically a scanner or a camera. The characters are in the form of pixelized images, and can be either printed or handwritten, of any size, shape, or orientation.

The OCR can be subdivided into handwritten character recognition and printed character recognition. Handwritten Character Recognition is more difficult to implement than printed character recognition due to diverse human handwriting styles and customs. In printed character recognition, the images to be processed are in the forms of standard fonts like Times New Roman, Arial, Courier, etc [9].

3) The *drawbacks* of the off-line recognizers, compared to on-line recognizers are summarized as follows:

a) Off-line conversion usually requires costly and imperfect pre-processing techniques prior to feature extraction and recognition stages.

b) They do not carry temporal or dynamic information such as the number and order of pen-on and pen-off movements, the direction and speed of writing and in some cases, the pressure applied while writing a character.

c) They are not real-time recognizers.

Table 1 shows the comparison between online and offline handwritten character recognition.

Table 1: Comparison between online and offline handwritten characters

Sr. No.	Comparisons	Online Char	Offline Char
1	Availability of Pen strokes	Yes	No
2	Raw data requirements	# samples/ seconds	# dots/ inch
3	Ways of writing	Digital Pen	Paper Document
4	Recognition Rate	Higher	Lower
5	Accuracy	Higher	Lower

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

IV. PHASES OF HANDWRITTEN CHARACTER RECOGNITION

The process of handwritten character recognition can be divided into phases as shown in the figure 2 below

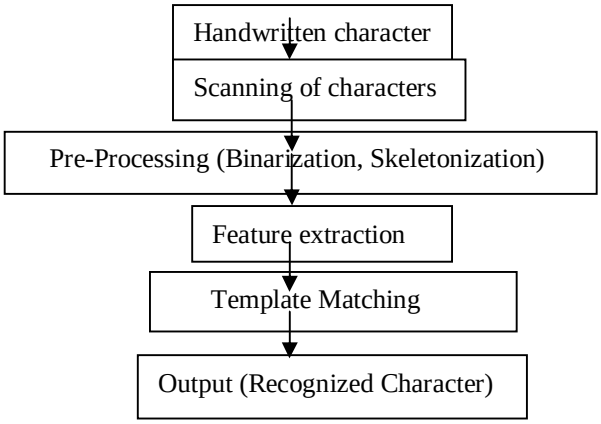


Figure 2: Block diagram for handwritten character recognition

A. Pre-Processing

Pre-processing is the name given to a family of procedures for smoothing, enhancing, Filtering, cleaning-up and otherwise massaging a digital image so that subsequent algorithm along the road to final classification can be made simple and more accurate. Various Pre-processing Methods are explained below:

1) *Binarization:* Document image Binarization (thresholding) refers to the conversion of a gray-scale image into a binary image.

a) *Two categories of thresholding:*

i) Global, picks one threshold value for the entire document image which is often based on an estimation of the background level from the intensity histogram of the image.

ii) Adaptive (local), uses different values for each pixel according to the local area information

2) *Noise Removal:* The major objective of noise removal is to remove any unwanted bit-patterns, which do not have any significance in the output. Various filtering operations can be applied to remove noise e.g. Median Filter, Weiner filter etc.

3) Skeletonization : Skeletonization is also called thinning. Skeletonization refers to the process of reducing the width of a line like object from many pixels wide to just single pixel. This process can remove irregularities in letters and in turn, makes the recognition algorithm simpler because they only have to operate on a character stroke, which is only one pixel wide. It also reduces the memory space required for storing the information about the input characters and no doubt, this process reduces the processing time too**.**

4) *Smoothing:* The objective of smoothing is to smooth shape of broken and/or noisy input characters.

5) *Thresholding:* In order to reduce storage requirements and to increase processing speed, it is often desirable to represent grey scale or color images as binary images by picking some threshold value for everything above that value is set to 1 and everything below is set to 0. Two categories of thresholding exist: Global and Adaptive. Global thresholding picks one threshold value for the entire document image. Adaptive thresholding is a method used for images in which different regions of the image may require different threshold values.

6) *Normalization:* Normalization is a linear process. If the intensity range of the image is 50 to 180 and the desired range is 0 to 255 the process entails subtracting 50 from each of pixel intensity, making the range 0 to 130. Then each pixel intensity is multiplied by 255/130, making the range 0 to 255. Auto-normalization in image processing software typically normalizes to the full dynamic range of the number system specified in the image file format. The normalization process will produce iris regions, which have the same constant dimensions, so that two photographs of the same iris under different conditions will have characteristic features at the same spatial location. Normalization methods aim to remove all types of variations during the writing and obtain standardized data. For example Size normalization is used to adjust the character size to a certain standard.

Methods of character recognition may apply both horizontal and vertical size normalizations [10][11].

B. Feature Extraction

Each character has some features, which play an important role in pattern recognition. Feature extraction describes the relevant shape information contained in a pattern so that the task of classifying the pattern is made easy by a formal procedure. In pattern recognition and in image processing, feature extraction is a special form of dimensionality reduction. The main goal of feature extraction is to obtain the most relevant information from the original data and represent that information in a lower dimensionality space. When the input data to an algorithm is too large to be processed and it is suspected to be redundant (much data, but not much information) then the input data will be transformed into a reduced representation set of features (also named features vector). Transforming the input data into the set of features is called feature extraction. If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input [12][13].

1) *Feature extraction techniques: The* techniques that are used for feature extraction are:
2) *Principal component analysis (PCA)*: PCA is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has as high a variance as possible (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (uncorrelated with) the preceding components. Principal components are guaranteed to be independent only if the data set is jointly normally distributed. PCA is sensitive to the relative scaling of the original variables. Depending on the field of application, it is also named the discrete Karhunen–Loève transform (KLT), the Hotelling transform or proper orthogonal decomposition (POD).
3) *Independent Component Analysis (ICA:* ICA is a statistical technique that represents a multidimensional random vector as a linear combination of non-gaussian random variables ('independent components') that are as independent as possible. ICA has many applications in data analysis, source separation, and feature extraction.

C. Representation of Character Features

After extracting the features, the data should be represented in one of two ways, either as a boundary or as a complete region. When the focus is on external shape characteristics such as corners and variation then boundary representation is appropriate. While regional representation is appropriate when the focus is on internal properties such as textures or skeletal shape. In some applications like character recognition these representations coexist, which often require algorithm based on boundary shape as well as skeletons and other internal properties. In terms of character recognition descriptors such as holes and bays are powerful features that help differentiate one part of the character from another. This description also called feature selection, deals with extracting features which results in some quantitative information of interest or features that are basic for differentiating one class of objects from another.

In this work we are using improved template matching technique for feature extraction [14]. It is used to find the pixel level and matches the character boundaries in the segmented image. Many template matching techniques are used in image processing, which are used to find the exact character. One such technique is normalized cross correlation.

Normalized Cross Correlation computes the correlation coefficient between A and B, where A and B are matrices or vectors of the same size. It computes the correlation coefficient using following equation:

$$r = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \bar{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \bar{B})^2\right)}}$$

where A'= mean2(A), and B' = mean2(B)

D. Recognition

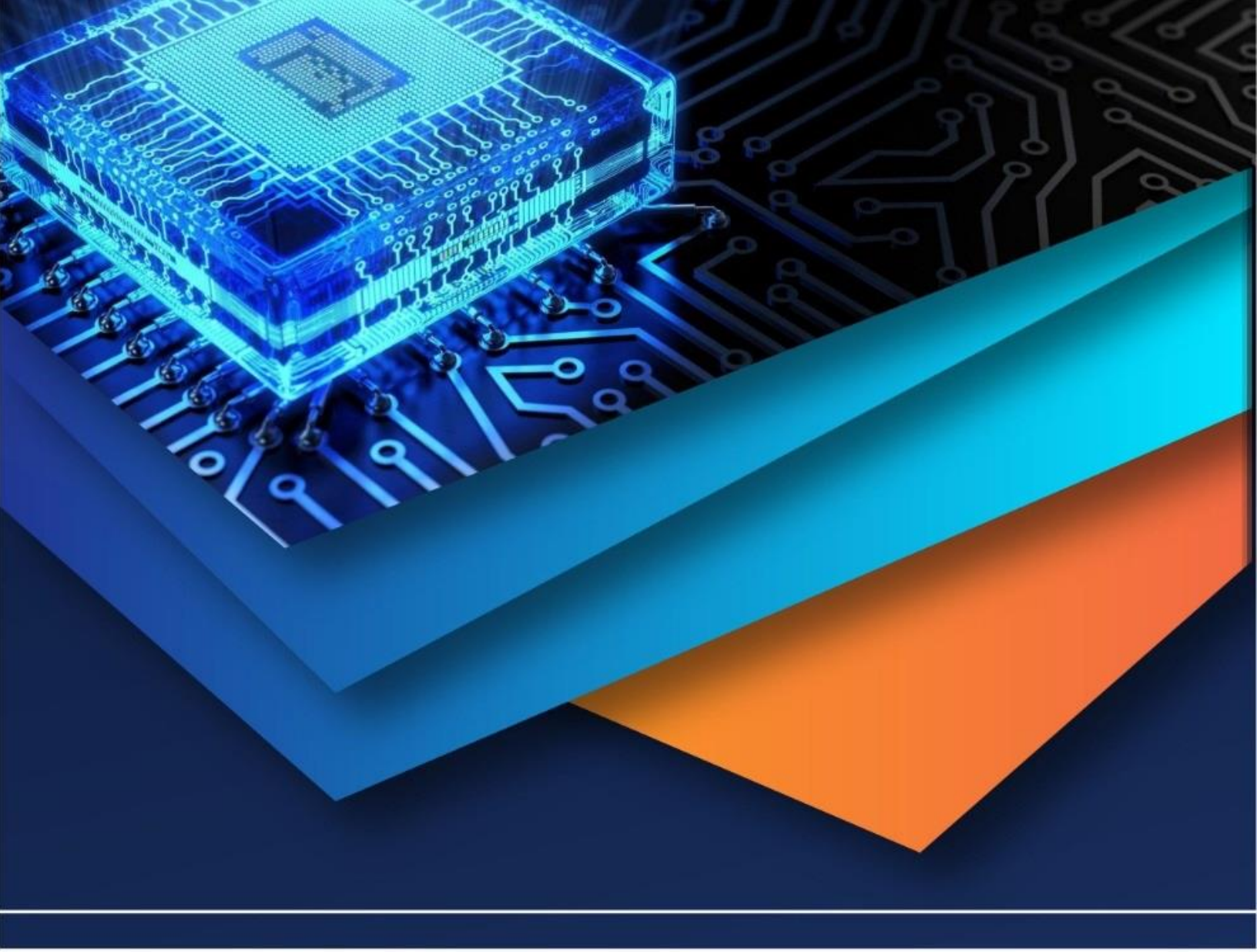
The image from the extraction stage is correlated with all the templates which are preloaded into the system. Once the correlation is completed, the template with the maximum correlated value is declared as the character present in the image.

V. CONCLUSION

Recognition approaches heavily depend on the nature of the data to be recognized. The recognition process needs to be much efficient and accurate to recognize the characters written by different users. As template matching technique is used here for recognition of offline English character images and it has been seen that recognition increases. For some characters like I & J are similar, so the recognition system gives sometimes bad results for similar character. Also it is based on the handwriting style e.g. G may be written as  or  . This may also create problem sometimes. For overcoming above problems we use large number of data set so that our work successfully recognizes large numbers of characters accurately.

REFERENCES

[1] Handwritten Optical character recognition http: // en. Wikipedia. Org / wiki / Optical_character_ recognition.
[2] \Anita Pal & Dayashankar Singh, “Handwritten English Character Recognition Using Neural Network” International Journal of Computer Science & Communication Vol. 1, No. 2, July-December 2010, pp. 141-144.
[3] Janusz k Starzyk and Nasser Ansari – “Feed forward Neural Network for Handwritten Character Recognition”, in IEEE symposium on circuit and systems, 2011.
[4] S. Nath, Afseena, “Handwritten Character Recognition – A Review”, International Journal of Scientific and Research Publications, Volume 5, Issue 3, March 2015.
[5] Manoj Sonkusare and Narendra Sahu, “A Survey on Handwritten Character Recognition (HCR) Techniques for English Alphabets”, Advances in Vision Computing: An International Journal (AVC) Vol.3, No.1, March 2016.
[6] Shalin A. Chopra, Amit A. Ghadge, Onkar A. Padwal, Karan S. Punjabi, Gandhali S. Gurjar, “Optical Character Recognition”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014
[7] Neetu Bhatia, “Optical Character Recognition Techniques: A Review”, India International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 5, May 2014.
[8] Surya Nath R S, Afseena S, “Handwritten Character Recognition – A Review”, International Journal of Scientific and Research Publications, Volume 5, Issue 3, March 2015.\
[9] N. Venkata Rao, A. Sastry, A. Chakravarthy, K. Chakravarthi, “Optical Character Recognition Technique Algorithms”, Journal of Theoretical and Applied Information Technology, Vol.83. No.2, January 2016.
[10] Kauleshwar Prasad, Devvrat C. Nigam, Ashmika Lakhotiya and Dheeren Umre, “Character Recognition Using Matlab’s Neural Network Toolbox”, International Journal of Science and Technology Vol. 6, No. 1, February, 2013.
[11] Ankush Acharyya, Sandip Rakshit, Ram Sarkar, Subhadip Basu, Mita Nasipuri, “Handwritten Word Recognition Using MLP based Classifier: A Holistic Approach”, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 2, March 2013.
[12] Gauri Katiyar and Shabana Mehfuz, “A hybrid recognition system for off line handwritten characters”, Katiyar and Mehfuz SpringerPlus (2016)
[13] Arindam Saha1, Nitupon Talukdar, “Typed Character Recognition using ANN”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 4, April 2016.
[14] S. Vijayarani, A. Sakila, “Template Matching Technique For Searching Words In Document Images”, International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 6, December 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY