iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** IX **Month of publication:** September 2014

DOI:

www.ijraset.com

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

ISSN: 2321-9653

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AND ENGINEERING TECHNOLOGY (IJRASET)

Character recognition through image processing

Khushboo[#1], Smriti Srivastava[#2]

[1,2]*ECE, GGSIPU*

Northern India Engineering College, FC-26, Shashtri Park Delhi-110052

Abstract*—Recognition of alphabetic characters is a basic need in incorporating intelligence to computers. Machine intelligence involves several aspects among which optical recognition is a tool, which can be integrated to text recognition. To make these aspects effective character recognition with better accuracy is important. However, handwritten character recognition is still a difficult task because of the high variability in the character shapes written by individuals. Input is paragraphs of running text, which is pre-processed to segment it into normalized individual words. A neural network is trained onto the dataset containing 55 samples for each of the 26 alphabets for recognition. The paper entitled "Character Recognition" is comparable to optical character recognition software where input is fed through a camera, from internet or some scanned image as well and the processing is done at the instant of taking input. The goal is to recognize only the text component from the image stream which may have text-characters buried in the colour background, complex patterns and text like structures.

***Index Terms*—Character, OCR, Programming, Graphics, Normalization.**

I. INTRODUCTION

During the past half century, significant research efforts have been devoted to character recognition to translate human readable characters into machine-readable codes. It is one of the active research areas waiting for accurate recognition solutions and the accuracy of the recognition solutions is predominantly depends on proper features extraction methods. Firstly, an effective feature need to be invariant with respect to character shape variation caused by various writing styles of different individuals and maximize the separability of different character classes. It also needs to represents the raw image data of character through a reduced set of information which are most relevant for classification (i.e., used to distinguish the character classes) to increase the efficiency of classification process. The main goal of feature selection is to construct linear or non-linear decision boundaries in feature These features are invariant to character deformation and writing style to some extent. Some of the commonly used statistical features for character recognition are projection histograms, crossings, zoning and moments etc. The programming platform that we used is MATLAB with the support for image processing toolbox and supports following types of images:

.png: Portable Network Graphics, a bitmap image file format, is a raster graphics file format that supports lossless data compression. PNG supports palette-based images (with palettes of 24-bit RGB or 32-bit RGBA colours), grayscale images (with or without alpha channel), and full-colour non-palette-based RGB images (with or without alpha channel).

.jpeg: In computing, JPEG - named after its creator the Joint Photographic Expert Group (seen most often with the .jpg extension) is a commonly used method of loss compression for digital photography (i.e. images). The degree of compression can be adjusted, allowing a selectable trade-off between storage size and image quality.

.bmp file: This is a paintbrush extension file in which images of characters can be painted and can be used. The BMP file format, also known as bitmap image file or device independent bitmap (DIB) file format or simply a bitmap, is a raster graphics image file format used to store bitmap digital images, independently of the display device (such as a graphics adapter), especially on Microsoft Windows operating systems.

IJRASET: ISSN: 2321-9653 Volume II, Issue IX, September 2014

ISSN: 2321-9653

II. CHARACTER RECOGNITION & CHALLENGES

Character recognition is becoming more and more important now a days in the world. The ability to identify characters in an automated or a semi-automated manner has obvious applications in numerous fields. The field of character recognition is a multidisciplinary field which forms the foundation of other fields, as for instance, Image Processing, Machine Vision, and Artificial Intelligence.

There are different challenges faced while attempting to solve this problem. The handwritten digits are not always of the same size, thickness, or orientation and position relative to the margins. The characters may be written on a cluttered background, on crumpled paper or may even be partially occluded. "Character recognition is the study of how machines can observe the environment, learn to distinguish characters of interest from their background, and make sound and reasonable decisions about the categories of the character."

Handwritten numeral recognition belongs to the field of pattern recognition, which is a hot field for a large number of researchers and also is a critical step in entry of information. It is widely used in public security, taxation, transportation, finance, education and other industries in the practical activities. For example, bank check systems have to take into account the great variability in the representation of a numerical amount, e.g., the number of components to be identified, which is not necessary for a zip code system since the number of digits is fixed and known a priori. Another important requirement from a bank check system is its reliability. It has been estimated that such a system becomes commercially efficient only when the error rate kept very low. Methods for courtesy amount recognition belong to the class of digit recognition techniques although in many cases these amounts include also some non- digit symbols such as commas, periods, strokes, currency names, etc.

CHARACTER RECONIZING IN IMAGES

Text can be detected by exploiting the discriminate properties of text characters such as the vertical edge density, the texture or the edge orientation variance. One early approach for localizing text in covers of Journals or CDs assumed that text characters were contained in regions of high horizontal variance satisfying certain spatial properties that could be exploited in a connected component analysis process Using a K-means algorithm, pixels are classified into three classes in the feature space. The class with highest energy in this space indicates text while the two others indicate non-text and uncertainty. Variation of edge orientation was computed in local area from image gradient and combined with edge features for locating text blocks.

III. OPTICAL CHARACTER RECOGNITION

Optical Character Recognition deals with the problem of recognizing optically processed Characters. Optical recognition is performed off-line after the writing or printing has been completed, as opposed to on-line recognition where the computer recognizes the characters as they are drawn. Both hand printed and printed characters may be recognized, but the performance is directly dependent upon the quality of the input documents.

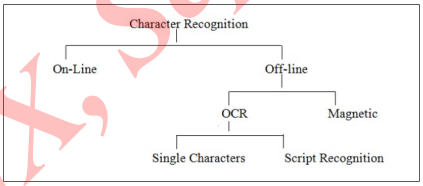


Fig 1. Different Areas of Character Recognition

In on-line character recognition, the computer recognizes the symbols as they are drawn. The most common writing surface is the digital tablet. Since characters are represented by line drawings, there is no need for skeletonization or contour extraction that is a relatively costly and imperfect process. Off-line recognition is performed after the writing or printing is completed. Handwriting Recognition enables a person to scribble something on a piece of paper and then convert it into text. OCR is one of the most fascinating and challenging areas within the broader area of pattern recognition.

OCR character recognition consists of the following procedures:

Learning - from an image file and corresponding text file or learning interactively. Extraction and isolation of individual characters from an image. Determination of the properties of the extracted characters. Comparison of the properties of the

IJRASET: ISSN: 2321-9653 Volume II, Issue IX, September 2014

ISSN: 2321-9653

learned and extracted characters. Additional operations on extracted characters if no good match is found.

BLOCK DIAGRAM

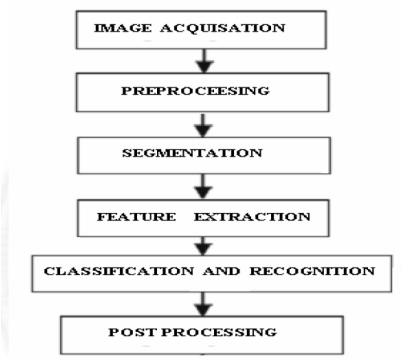


Fig 2. Block Diagram Showing Process Involved In OCR

The above block diagram shows the dataflow pattern according to which a character is recognized optically using the optical character recognition in MATLAB. The sensor used is basically a CMOS sensor which can be found in a VGA camera followed by an image acquisition into the MATLAB system where the image is acquired into the system by the method of using the various relay systems. The next step is feature extraction which is used to extract the features of the images using the characteristic loci method and then finally the image is recognized by the system via classification and comparing with the database and can be given as output to the user.

IV. IMPLEMENTATION PROCESS

The main principle in automatic recognition of patterns is first to teach the machine which classes of patterns that may occur and what they look like. In OCR the patterns are letters, numbers and some special symbols like commas, question marks etc., while the different classes correspond to the different characters. For handwriting recognition the input to our system is a scanned image containing handwritten text paragraph. The image should be in prescribed format such as jpeg, .png, .bmp etc.

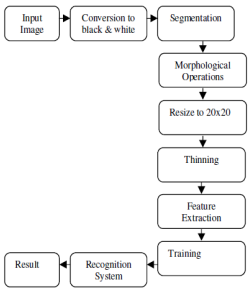


Fig 3. Implementation Process

A. SCANNING INPUT

Through the scanning process a digital image of the original document is captured. In OCR optical scanners are used, which generally consist of a transport mechanism plus a sensing device that converts light intensity into grey-levels. Printed documents usually consist of black print on a white background. The threshold process is important as the results of the following recognition are totally dependent of the quality of the bi-level image. Still, the threshold performed on the scanner is usually very simple. A fixed threshold is used, where grey-levels below this threshold is said to be black and levels above are said to be white. For a high-contrast document with uniform background, a pre-chosen fixed threshold can be sufficient.

B. PREPROCESSING DATA

Pre-processing is the major step in character recognition system. It employs several steps, that is, line segmentation, Image conversion from RGB to Grey, binarization of image, resizing the image, cropping, thinning etc. It takes input as a raw running text image and gives output as segmented words.



Fig 4. Com

IJRASET: ISSN: 2321-9653 Volume II, Issue IX, September 2014

ISSN: 2321-9653

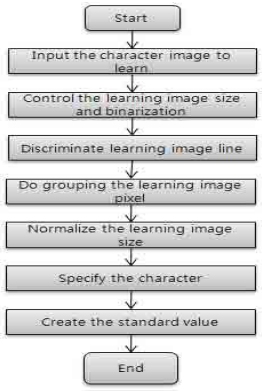


Fig 5. Pre-processing Steps

C. 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.

Applied to text, segmentation is the isolation of characters or words. The majority of optical character recognition algorithms segment the words into isolated characters which are recognized individually. Usually this segmentation is performed by isolating each connected component that is each connected black area. This technique is easy to implement, but problems occur if characters touch or if characters are fragmented and consist of several parts.

D. PREPROCESSING:

The image resulting from the scanning process may contain a certain amount of noise. Depending on the resolution on the scanner and the success of the applied technique for threshold, 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. In addition to smoothing, pre-processing usually includes normalization. The normalization is applied to obtain characters of uniform size, slant and rotation.

Filtering: After an RGB image is converted to gray, binarization is a technique by which any gray scale image is converted to binary image. Image threshold is advantageous as it is easier to manipulate images with only to levels of color, processing is faster, less computationally expensive and allows for more compact storage. Threshold value removes the noise from the image, if the intensity of that pixel is below threshold level.

Morphological Operation: The basic idea behind the morphological operations is to filter the document image replacing the convolution operation by the logical operations. Therefore, morphological operations can be successfully used to remove the noise on the document images due to low quality of paper and ink, as well as erratic hand movement.

Normalization & Resizing: Normalization is considered to be the most important preprocessing factor for character recognition. Normalization removes the un-necessary part from the character image and brings it into specific size, same as reference pattern. The size of the input image is fixed say 20X20 or 42X42 etc. pixels.

E. 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.

Feature extraction technique	Robustness 1 2 3 4 5	Practical use 1 2 3
Template matching	● ● ○ ○ ○	○ ● ○
Transformations	○ ● ● ● ●	○ ○ ●
Distribution of points: Zoning	○ ● ○ ○ ●	● ● ○
Moments	● ● ○ ● ●	○ ● ○
n-tuple	● ○ ● ○ ●	● ● ●
Characteristic loci	○ ● ● ● ●	● ● ○
Crossings	○ ● ● ● ●	● ● ○
Structural features	○ ● ● ● ●	● ○ ●

● High or easy ● Medium ○ Low or difficult

Fig 6. Comparison Of Various Feature Extraction Techniques

F. ZONING

The frame containing the character is divided into several overlapping or non-overlapping zones and the densities of object pixels in each zone are calculated. Zoning is partition of the control box of the pattern (i.e. the smallest rectangle

IJRASET: ISSN: 2321-9653 Volume II, Issue IX, September 2014

ISSN: 2321-9653

containing the pattern); the elements of such partition are used to identify the position in which features of the pattern are detected. Handwritten numerals are first normalized to some fixed pixel size and according to the zones of the control box, each feature is labelled with the name of the zone in which it has been detected.

Methodology for zoning:

The rectangle circumscribing the character is divided into several overlapping, or non- overlapping, regions and the densities of black points within these regions are computed and used as features.

Break the box into equal parts using for loop

Compute the Density of Pixels in each zone.

Store these features in a matrix for further reference.

CROSSINGS

In the crossing technique features are found from the number of times the character shape is crossed by vectors along certain directions. This technique is often used by commercial systems because it can be performed at high speed and requires low complexity. When using the distance technique certain lengths along the vectors crossing the character shape are measured.

MOMENTS

The rectangle circumscribing the character is divided into several overlapping, or no overlapping, regions and the densities of black points within these regions are computed and used as features.

DENSITY

Density is calculated by finding the number of object pixels in each zone and dividing it by total number of pixels. The number of ON pixels (1 =ON; 0=OFF) divided by the total number of points in the grid evaluates density.

$$D(m, n) = \frac{\text{Number of ON pixels (White) in the image denoted by } (m, n)}{\text{Total points in the grid}}$$

While computing the density of each zone, the ON pixels are needs to be calculated and divided by the total pixels (ON as well as OFF). It is possible that some zones may have a value zero because of no ON pixels in that zones and which evaluates to zero density. This process is repeated sequentially for each zone and features are noted. The number of zones is kept on increasing and so as the features for the character image.

G. CLASSIFICATION

The classification is the process of identifying each character and assigning to it the correct character class. In the following sections two different approaches for classification in character recognition are discussed. First decision-theoretic recognition is treated. These methods are used when the description of the character can be numerically represented in a feature vector. For instance, if we know that a character consists of one vertical and one horizontal stroke, it may be either an “L” or a “T”, and the relationship between the two strokes is needed to distinguish the characters. A structural approach is then needed.

H. DECISION MAKING MATCHING

Matching covers the groups of techniques based on similarity measures where the distance between the feature vectors, describing the extracted character and the description of each class is calculated. Different measures may be used, but the common is the Euclidean distance. This minimum distance classifier works well when the classes are well separated, that is when the distance between the means large is compared to the spread of each class.

When the entire character is used as input to the classification, and no features are extracted (template-matching), a correlation approach is used. Here the distance between the character image and prototype images representing each character class is computed.

I. RECOGNITION PROCESS

For recognizing the text and the corresponding characters all the features are computed and input to the neural network is done. Output of which is classified to be among the domain of characters (here Alpha-Numeric). Neural network approach is tried to implement and deriving the appropriate results for the same.

NEURAL NETWORK

Recently, the use of neural networks to recognize characters (and other types of patterns) has resurfaced. Considering a back-propagation network, this network is composed of several layers of interconnected elements. A feature vector enters the network at the input layer. Each element of the layer computes a weighted sum of its input and

IJRASET: ISSN: 2321-9653 Volume II, Issue IX, September 2014

ISSN: 2321-9653

transforms it into an output by a nonlinear function. During training the weights at each connection are adjusted until a desired output is obtained. A problem of neural networks in OCR may be their limited predictability and generality, while an advantage is their adaptive nature.

Creating the First Neural Network

We use a feed forward neural network set up for pattern recognition with 25 hidden neurons. Since the neural network is initialized with random initial weights, the results after training vary slightly every time the example is run.

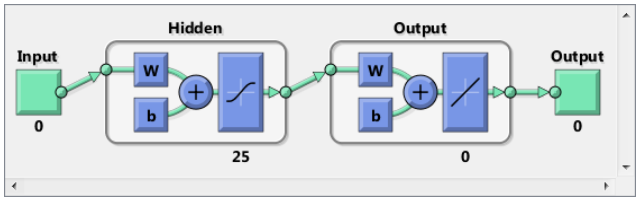


Fig 7. Neural Network

Training the first Neural Network

The function train divides up the data into training, validation and test sets. The training set is used to update the network, the validation set is used to stop the network before it over-fits the training data, thus preserving good generalization. The test set acts as a completely independent measure of how well the network can be expected to do on new samples.

Training stops when the network is no longer likely to improve on the training or validation sets.



Fig 8. Neural Network Steps

J. DISPLAYING THE OUTPUT:

The output of the recognized character is displayed in a text window (Notepad window). The detected lines, Alphabets and Numerals are displayed in the window. Subsequently, the accuracy of the system is checked and can be seen that for some input, how the outputs are generated and improvements to the same can be brought, if necessary. This becomes user friendly as an input corresponds to some output and further inspection becomes easier. The respective MATLAB code calls for the text window in which output needs to be displayed.



Fig 9. Output Image

of a table footnote. *(table footnote)*

Fig. 1. Example of a figure caption. *(figure caption)*

V. CONCLUSIONS

The coding used for noise reduction was also successful and capable of removing stray marks on the sheet of paper being used as well as the other noises that came while taking the picture was also removed automatically. Today's uses of OCR are still somewhat limited to the scanning of the written word into useable computer text. The uses include word processing, mail delivery system scanning, ticket reading, and other such tasks. OCR is already used to detect viruses, or unfortunately create them, and to stop spammers. Anti-spamming applications continue to be improved, as the need for increased technology is a constant. OCR is used to prevent viruses by detecting codes hidden in images. While the difficulty of transferring information from legacy systems to modern operating systems can be cumbersome, OCR is able to read screenshots. This can facilitate the transferring of information between incompatible technologies. One of the current hopes for OCR is the chance of developing OCR software that can read compressed files. Text that is

IJRASET: ISSN: 2321-9653 Volume II, Issue IX, September 2014

compressed into an image and saved as ASCII or Hexadecimal data could be read by OCR and transferred back into readable text.

Finally, it is anticipated that OCR will be used in the development of more advanced robotics. The eyes of a robot are essentially a camera meant to input information. If OCR is used to help the robot comprehend text, the uses could be almost endless. All of these advancements and more are expected to keep the technology of OCR in use and continue to expand its current capabilities.

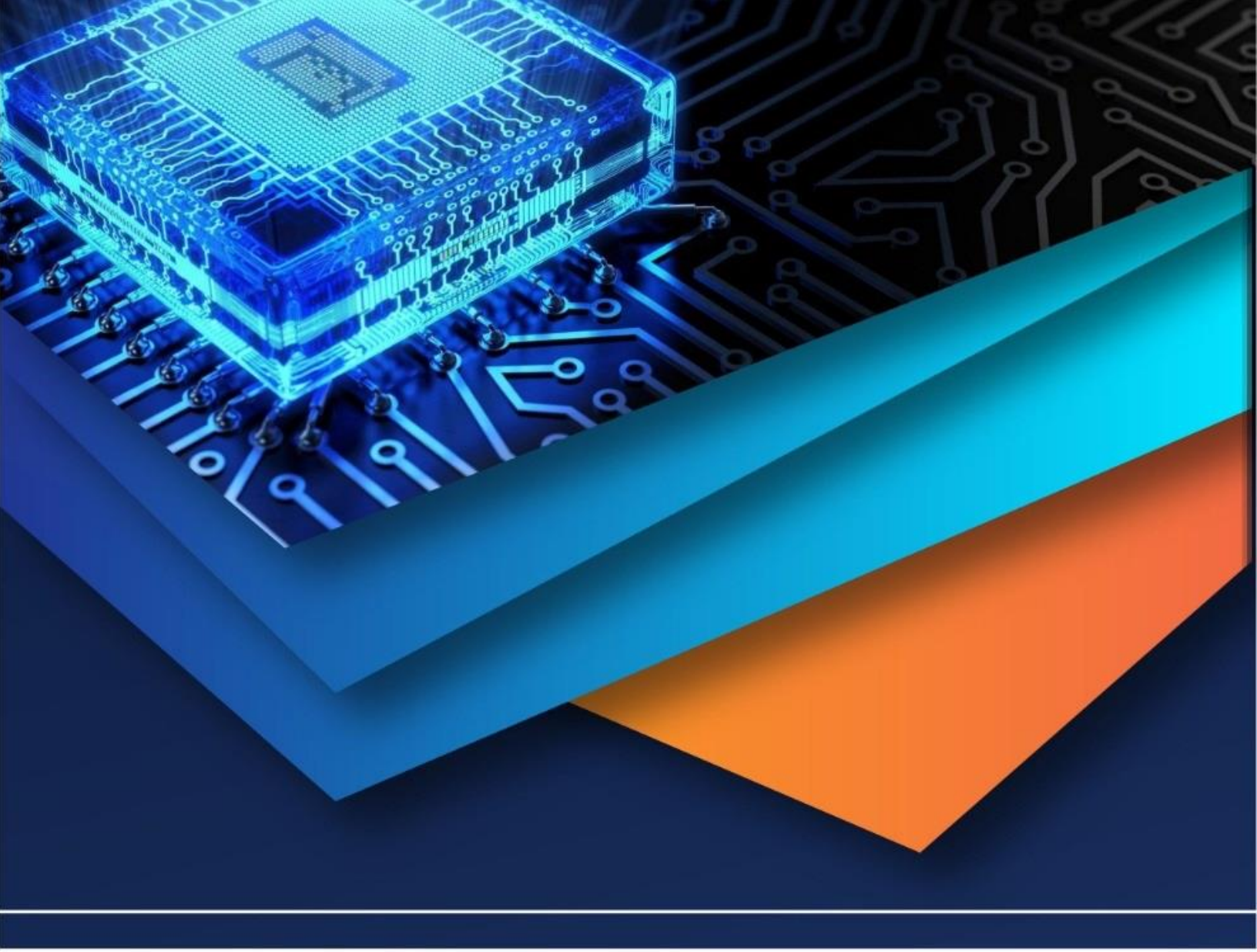
VI. ACKNOWLEDGMENT

I would like to thanks all people for providing us with additional information and for sharing personal experience to aid our project. Finally I would like to acknowledge for technical assistance and for sharing expertise with the advanced radio equipment.

REFERENCES

[1] Bahl, I. J and Bhartia, P; "Microstrip Antennas", Artech House, 1980.
[2] Garg, R and Ittipiboon, A; "Microstrip Antenna Design Handbook", Artech House, 2001.
[3] Zurcher, J-Francois and Gardiol, F; "Broadband Patch Antenna" Artech House, 1995.
[4] Kumar, G and Ray, K.P; "Broadband Microstrip Antenna", Artech House, 2003.
[5] Pozar and Schaubert; "Microstrip Antennas", Proceedings of the IEEE, vol. 80, 1992.
[6] Brown, S; "Microstrip Patch Antennas for PCS Applications", Department Electrical and Electronics Engineering; The University of Auckland, 1997.
[7] AscomSystec AG; Miniature Broadband Antennas, Datasheet: Model MBA-5, http://www.art-solutions.ch, [online], 10/04/2003.
[8] Khatri, N; "Directional Antennas for indoor Wireless Communications", Department of Electrical and Electronic Engineering; The University of Auckland, 2002.
[9] Khatri, N; "Directional Antennas for indoor Wireless Communications", Department of Electrical and Electronic Engineering; The University of Auckland, 2002.
[10] Rogers Corporation: Techtip #3. [Online]. Available: http://www.rogerscorp.com/mwu/techtip4.htm [2003 September 5].

IJRASET: ISSN: 2321-9653 Volume II, Issue IX, September 2014



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)