International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.3, 2015

RESEARCH ARTICLE

ISSN: 2321-7758

DESIGN AND PERFORMANCE EVALUATION OF CHARACTER AND NUMERAL RECOGNITION USING TEMPLATE MATCHING FOR MULTI FONTS / SIZE

SWATI MALIK[1] ,PREETI ANEJA[2]

[1]Student, Department of CSE, Hindu College of Engineering, India
Affiliated to DCRUST University, Murthal, Haryana, India.

[2]Assist Prof, Department of CSE, Hindu College of Engineering, India
Affiliated to DCRUST University, Murthal, Haryana, India.

Article Received:13/05/2015 | Article Revised on:21/05/2015 | Article Accepted on:31/05/2015

SWATI MALIK

ABSTRACT

The enormous amount of paper-based data in various other sectors challenges their ability to organize documents and records. Computers, working faster and more efficiently than human operators and performs large number of tasks required for efficient document and content management. Computers recognize characters for example ASCII code typed on a keyboard where every character or letter represents an identifiable code. However, these cannot distinguish characters and words from scanned images of paper documents. Therefore, information should be regained from scanned images like commercial or government documents, tax returns, passport applications and credit card applications, characters should first be converted to their ASCII equivalents and further they can be recognized as readable text. Optical character recognition system (OCR) or template matching allows converting a document into electronic text, which we can edited and searched etc. Character Recognition allows detection and recognition of characters from an input image and converts them into ASCII equivalent or machine editable form. The accuracy of character recognition depends upon the database used for the recognition process. It becomes very difficult to recognize printed characters due to poorly photocopied pages, broken and faded characters or due to ink and quality of papers used. So, it is not possible to achieve 100% accuracy under various conditions. The approaches that were previously used for recognizing English characters in printed text earlier cannot uniquely differentiate the similar looking characters and unable to give better recognition rate due to the complexity of the technique used. Hence, there is a need for a simpler approach for printed English character recognition to overcome from these problems. So, this paper proposes character recognition by using Template Matching. The templates formed, having variety of fonts and size .In this proposed system, Image pre-processing, Feature extraction and classification algorithms have been implemented so as to build an excellent character recognition technique for different scripts. Result of this approach is also shown in this paper. All the implementation and experimental work has been done in MATLAB R2008a using general MATLAB tool box and image processing tool box.

Keywords: Pattern recognition, Templates matching, Neural Networks, Handwriting recognition, Character Recognition.

I INTRODUCTION

Character Recognition is most powerful application in the area of pattern recognition and artificial intelligence. Optical Character Recognition deals with the process of recognizing the characters optically. It can be performed either off-line or on-line. Off-line Optical character recognition can be done after the written or printing work has been completed while in case of on-line Optical character recognition computer recognizes the characters as they are provided as input. In both the cases, characters can be recognized but the accuracy is directly dependent upon the type and quality of the input documents.

Character recognition is a system which loads a character (text) image, preprocesses the image, extracts proper image features, classify the characters based on the extracted image features (in the form of vector matrix) and the known features are stored in the image model library, and recognizes the image according to the degree of similarity between the loaded image and the image models. To recognize character firstly, the input images are acquired containing English text as an input image. Images are then stored in some picture file such as BMP, JPG etc. This image subsequently passes through preprocessing, segmentation, feature extraction and classification steps.

Preprocessing operations include image processing, binarization, noise reduction and skew detection & correction of a digital image so that subsequent algorithms along the road to final classification can be made simple and more accurate. Segmentation includes line segmentation-extract lines from a paragraph, and character segmentation-extract character from a line. After completing preprocessing and segmentation some features are extracted from the character image. Various feature extraction algorithms are there according to the behavior of the character.

Handwriting recognition can be defined as the task of transforming text represented in the spatial form of graphical marks into its symbolic representation. Since the early days of computers, automated recognition of text has been an active subject of research. This is useful for making digital copies of handwritten documents, and also in many automated processing tasks, such as automatic mail sorting or cheque processing. In automated mail sorting, letters are directed to the correct location by recognition of the handwritten address. Within the background of script recognition, it possibly will be connotation to study the distinctiveness of a variety of writing systems and the structural properties of the characters used in most significant scripts of the globe. There is only some degree of achievement in handwriting recognition, virtually for isolated and neatly hand-printed characters and word for limited vocabulary. The major difference between Online and Offline Character Recognition is that Online Character Recognition has real time contextual information but offline data does not. This difference generates a significant divergence in processing architectures and method. Special attention is given to the off-line handwriting recognition since this area requires more research to reach the ultimate goal of machine simulation of human reading.

The history of character recognition can be traced as early as 1900, when the Russian scientist Tyuring attempted to develop an assist for the visually handicapped. The first character recognizers developed in the middle of the 1940s with the growth of digital computers. The early work on the automatic recognition of characters has been concentrated either upon machine-printed text or upon a small set of handwritten text or symbols. Machine-printed Character recognition systems normally used template matching. For handwritten text, low level image processing techniques have been used on the binary image to extract feature vectors, which are then fed to statistical classifiers. Successful, but inhibited algorithms have been implemented mostly for Latin characters and numerals. However, some studies on Japanese, Chinese, Hebrew, Indian, Cyrillic, Greek, and Arabic characters and numerals in both machine printed and handwritten cases were also initiated. The commercial character recognizers were available in the 1950s. Historical review of character recognition research and growth during this period can be found in and for off-line and on-line cases, respectively. In the early 1990s, image processing and pattern recognition techniques were efficiently united with artificial intelligence (AI) methodologies. Nowadays, in addition to the more powerful computers and more accurate electronic equipments such as scanners, cameras, and electronic tablets, we have

efficient, modern use of methodologies such as neural networks (NNs), hidden Markov models (HMMs), fuzzy set reasoning, and natural language processing.

II CHARACTER RECOGNITION METHODS

Character recognition systems extensively use the methodologies of pattern recognition, which assigns an unknown sample to a predefined class. Many techniques for character recognition are investigated by the researchers and character recognition approaches can be classified as Template matching, Statistical techniques, Syntactic or structural, Neural network, Hybrid or Combination approaches.

Template matching approach

Template matching is the technique in which pixel definition of presorted patterns are sought in an image. This is the simplest way of character recognition, based on matching the stored data against the character to be recognized. The matching operation determines the degree of similarity between two vectors i.e. group of pixels, shapes curvature etc. a gray level or binary input character is compared to a standard set of stored data set. According to similarity measure (e.g. Euclidean, Yule similarity measures etc.), a template matcher can combine multiple information sources, including match strength and k-nearest neighbor measurements from different matrices. The recognition rate of this method is very sensitive to noise and image deformation. Template matching is one of the Optical Character Recognition techniques. Template matching is the process of finding the location of a sub image called a template inside an image. Once a number of corresponding templates is found their centers are used as corresponding points to determine the registration parameters. Template matching determines the similarities between a given template and windows of the same size in an image and identifying the window that produces the highest similarity measure. The recognition rate of Template Matching is mainly depending on noise and image deformation. For improved classification Deformable Templates and Elastic Matching are used.

This technique is different from the others in that no features are actually extracted. Instead the matrix containing the image of the input character is directly matched with a set of prototype characters representing each possible class. The distance between the pattern and each prototype is computed, and the class of the prototype giving the best match is assigned to the pattern. The technique is simple and easy to implement in hardware and has been used in many commercial OCR machines. Template Matching is as one of the solution to overcome the problem of character recognition.

The Fig. 1 shows the workflow of the template matching algorithm.

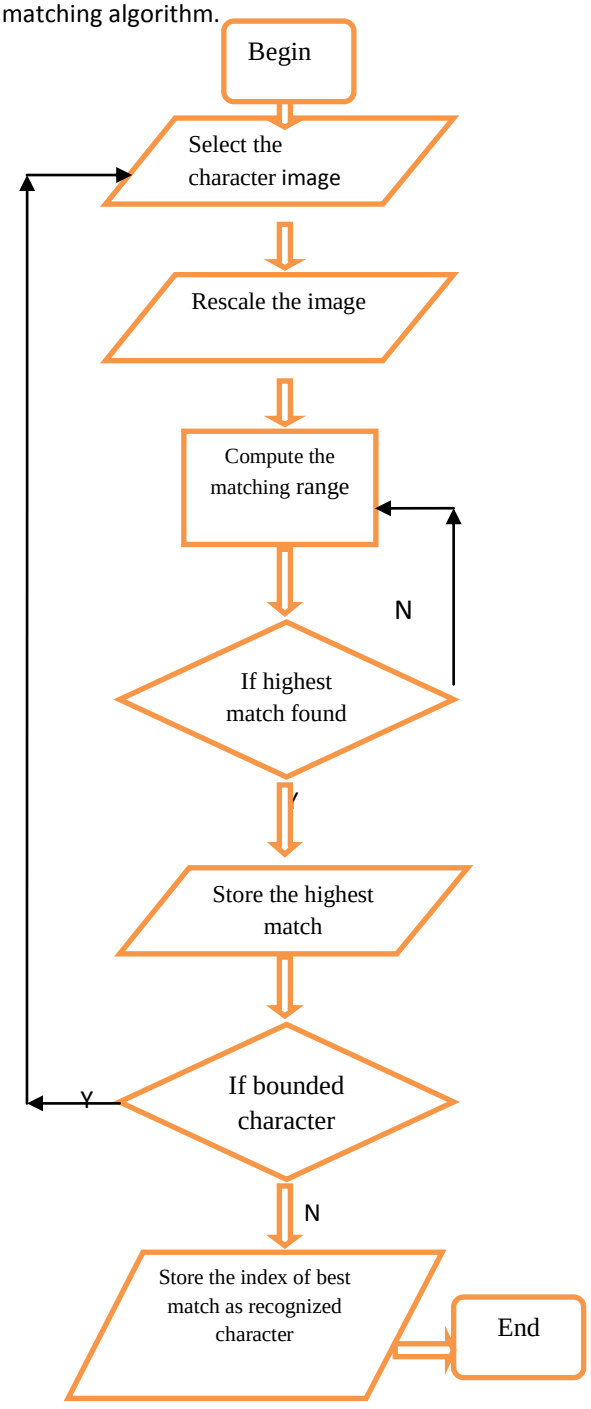


Fig. 1 Workflow of the Template Matching Algorithm

Statistical Techniques

Statistical decision theory is concerned with statistical decision functions and a set of optimality criteria, which maximizes the probability of the observed pattern given the model of a certain class. Statistical techniques are based on the assumptions such as Distribution of the feature set, statistics available for each class, collection of images to extract a set of features which represents each distinct class of patterns. The measurements taken from n-features of each word unit can be thought to represent an n-dimensional vector space. The major statistical methods applied in the character recognition field are Nearest Neighbor Likelihood or Baye's classifier, clustering Analysis, Hidden Markov Modeling, Fuzzy Set Reasoning, Quadratic classifier etc.

Syntactic or Structural Approach

In Syntactic Pattern recognition a formal analogy is drawn between the structure of pattern and syntax of a language. Structural pattern recognition is intuitively appealing because in addition to classification, this approach also provides a description of how the given path constructed from the primitives. Flexible structural matching is proposed for identification of alphanumeric characters.

Neural Networks

Various types of neural networks are used for character recognition classification. A neural network is a computing architecture that consists of massively parallel interconnection of adaptive neural processors. Because of its parallel nature, it can perform computations at a higher rate compared to classical techniques. Because of its adaptive nature, it can adapt to changes in the data and learn the characteristics of input signal. Output from one node is fed to another one in the network and final decision depends on the complex interaction of all nodes. Several approaches exist for training of neural networks like error correction, Boltzmann, Hebbian and competitive learning. Neural network architectures can be classified as, feed-forward, feed-back and recurrent networks. The most common neural networks used in the character recognition systems are the Multi Layer Perceptron (MLP) of the feed forward networks and the Kohonen's Self Organizing Map of the feedback networks.

Hybrid or Combination Classifier

We may have different feature sets, different training sets, different classification methods or different training sections, all resulting in set of classifiers, whose outputs may be combined, with the hope of improving overall classification accuracy. If this set of classifiers is fixed, the problem focuses on the combination function. It is also possible to use a fixed combiner and optimize the set of input classifiers. A typical combination scheme consists of a set of individual classifiers and combiner which combines the results of the individual classifiers to make the final decision. Various schemes for combining multiple classifiers can be grouped into three main categories according to their architecture parallel, cascading, and hierarchical.

III PROPOSED METHODOLOGY

1. Application of some pre-processing operation on input image i.e. conversion into grayscale format, so as to make it applicable for further process. Calculation of thresholding value so, as to apply binarization technique on gray scale image to convert it into Then take a complement of binary image matrix, which means 0 become 1 and 1 become 0 i.e. black and white are reversed.
2. To make the binary image matrix unwanted component free, application of the "BWAREAOPEN" function. This function will remove all connected components (objects) that have fewer than 30 pixels and will produce new matrix.
3. Creation and conversion of characters of data base into matrices. Also, loading of templates (character data base matrices) into main program. Resizing of templates in 42 x 42.
4. Extraction of the region of interest from the matrix and application of line segmentation using connected component analysis method. This operation will row wise identify the region of interest.
5. Application of character segmentation using connected component analysis method on each identified row and getting of each character.
6. Resizing of each identified character into a particular size i.e. 42 X 42, the size has also been chosen for templates.
7. Recognition of each character using template matching algorithm. It compares each character matrix with each template matrix one by one using 2 dimensional cross correlation operations.

8. If output of this function is greater than 0.9 than character is identified otherwise it goes for further comparison.

IV EXPERIMENTAL RESULTS

All the simulation work for the proposed method is implemented in MATLAB R2008a using MATLAB image processing tool box and generalized toolbox. Figure below shows the results for implementation of proposed work. Figure 2 is a snapshot of orignal printed character image with noise. After inputting of this image binarization is applied. Figure 3 is snapshot of binarized image. Figure 4 is snapshot of denoised binarized image through thresholding. After denoising of image line segmentation is applied results of which have been shown in figure 5, containing 4 snapshot of segmented lines of characters. After line segmentation, character segmentation is applied to each line. Figure 6 is the snapshot of the same, conataining 9 snapshot segmented characters. Figure 7 is the snapshot of the recognized characters with proper sequence.

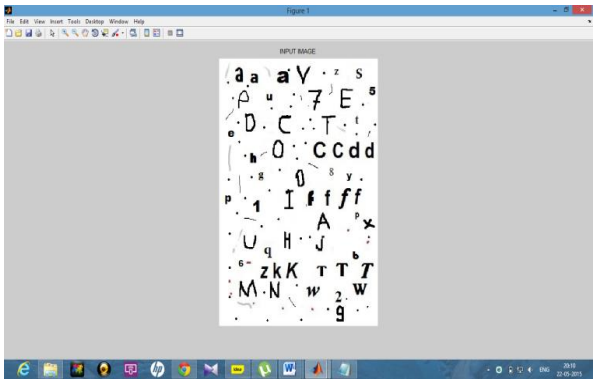


Fig.2. Snapshot of orignal printed character image with noise

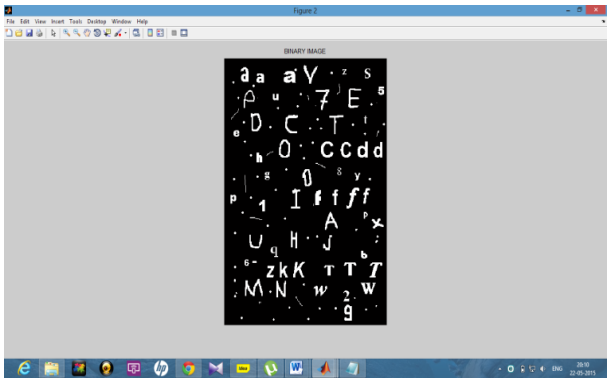


Fig.3 Snapshot of binarized image

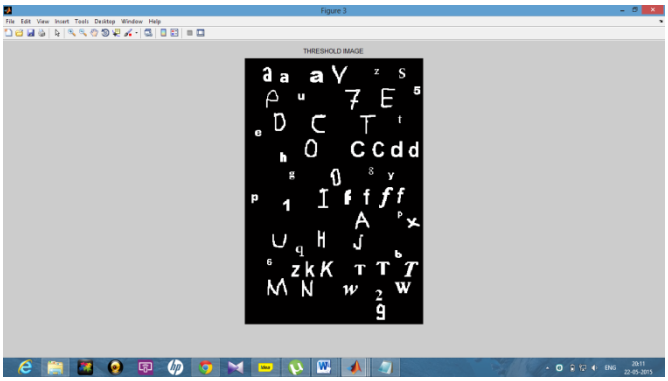
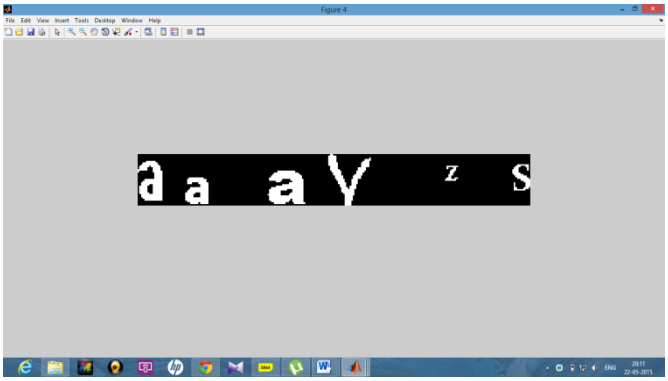
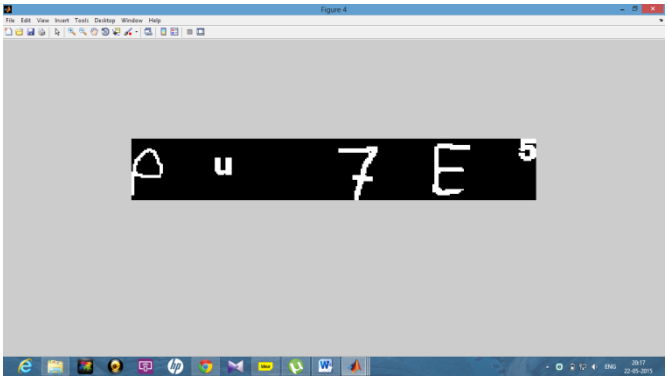
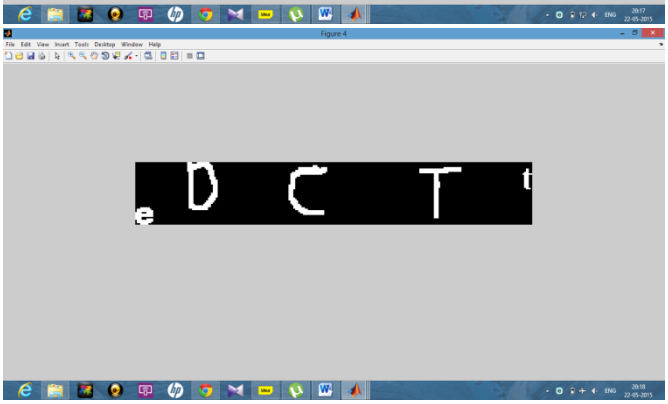


Fig.4. snapshot of denoised binarized image through thresholding







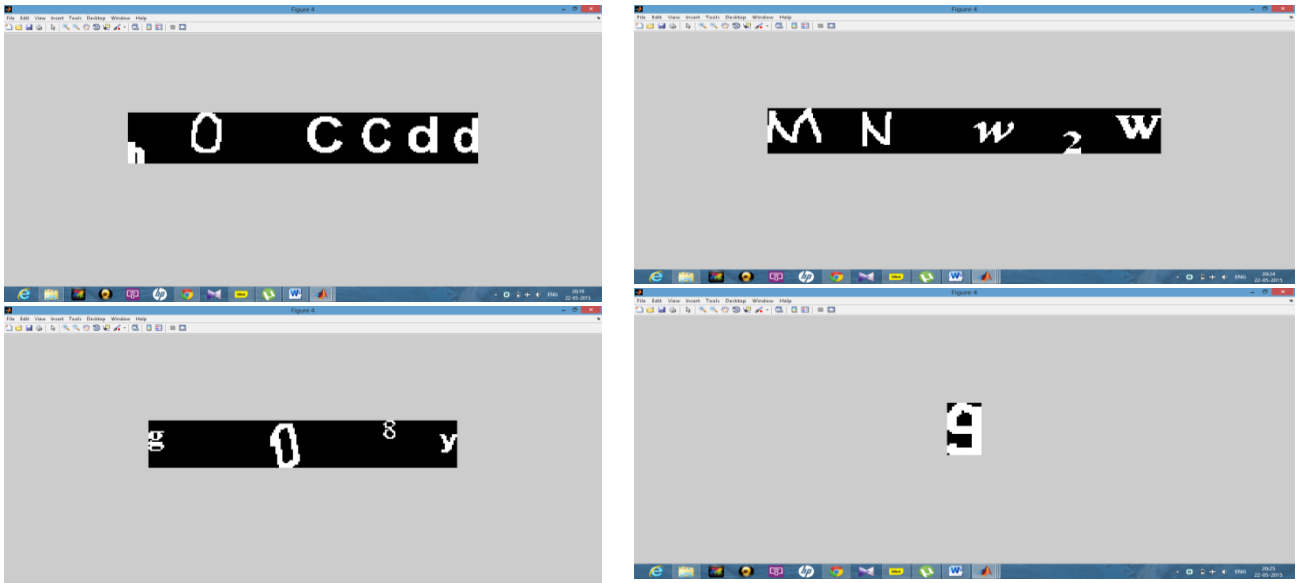
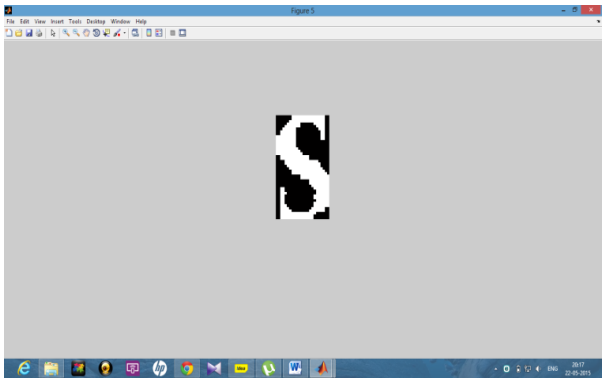
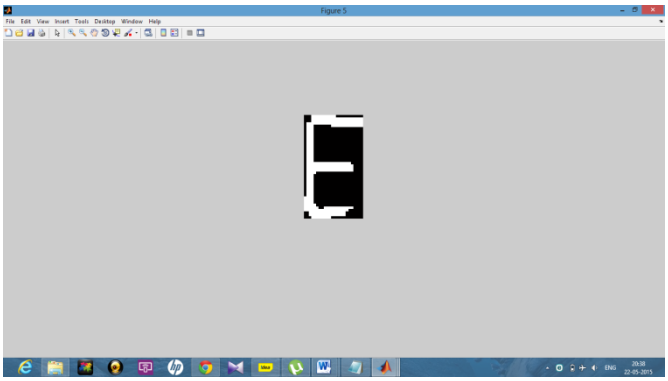
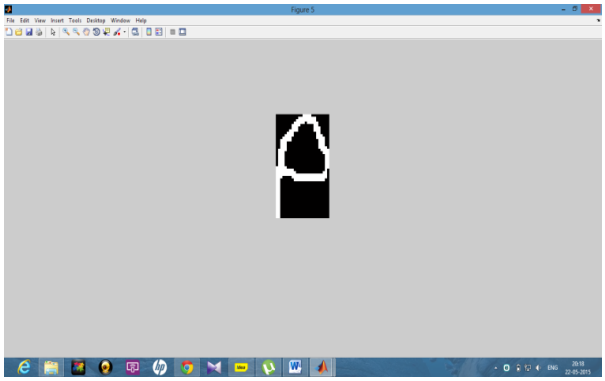
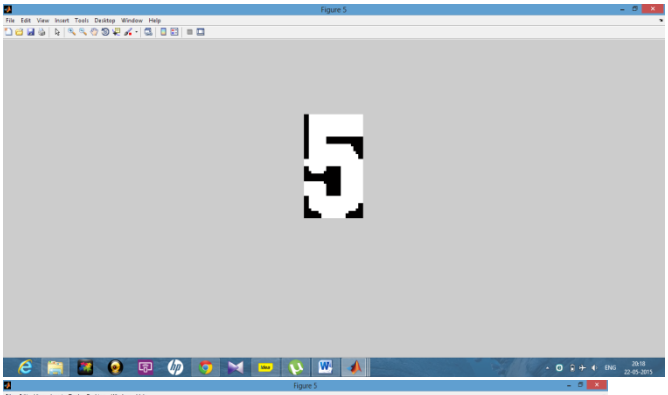
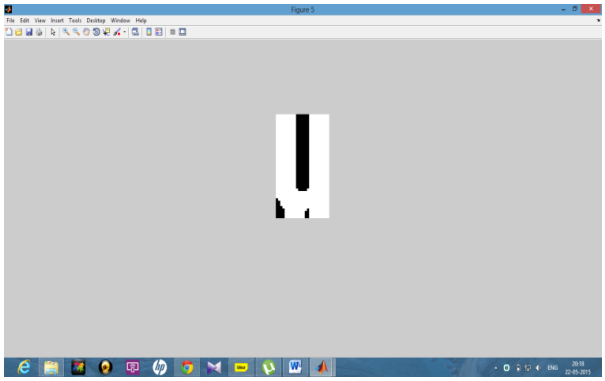
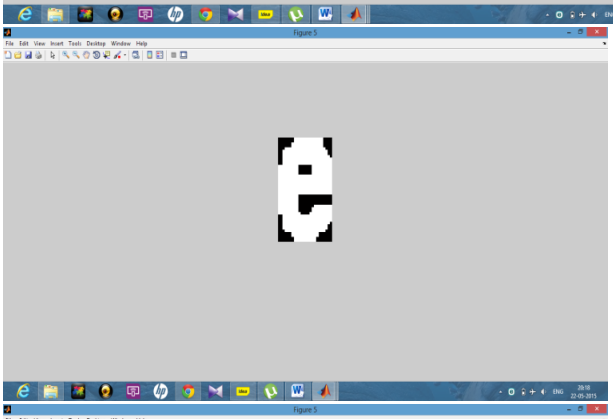
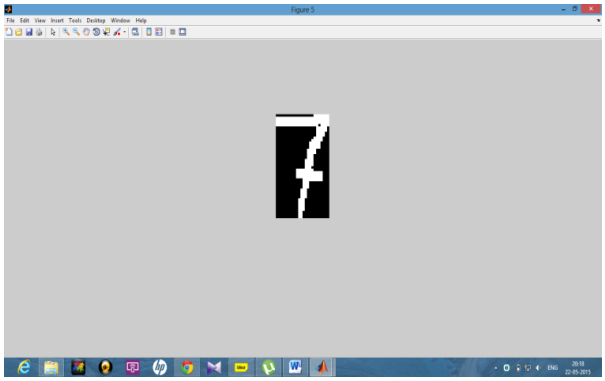
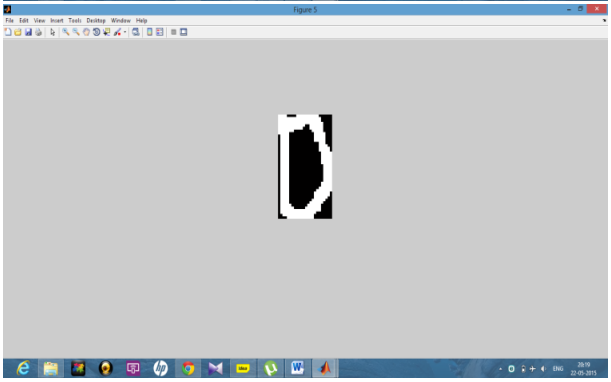
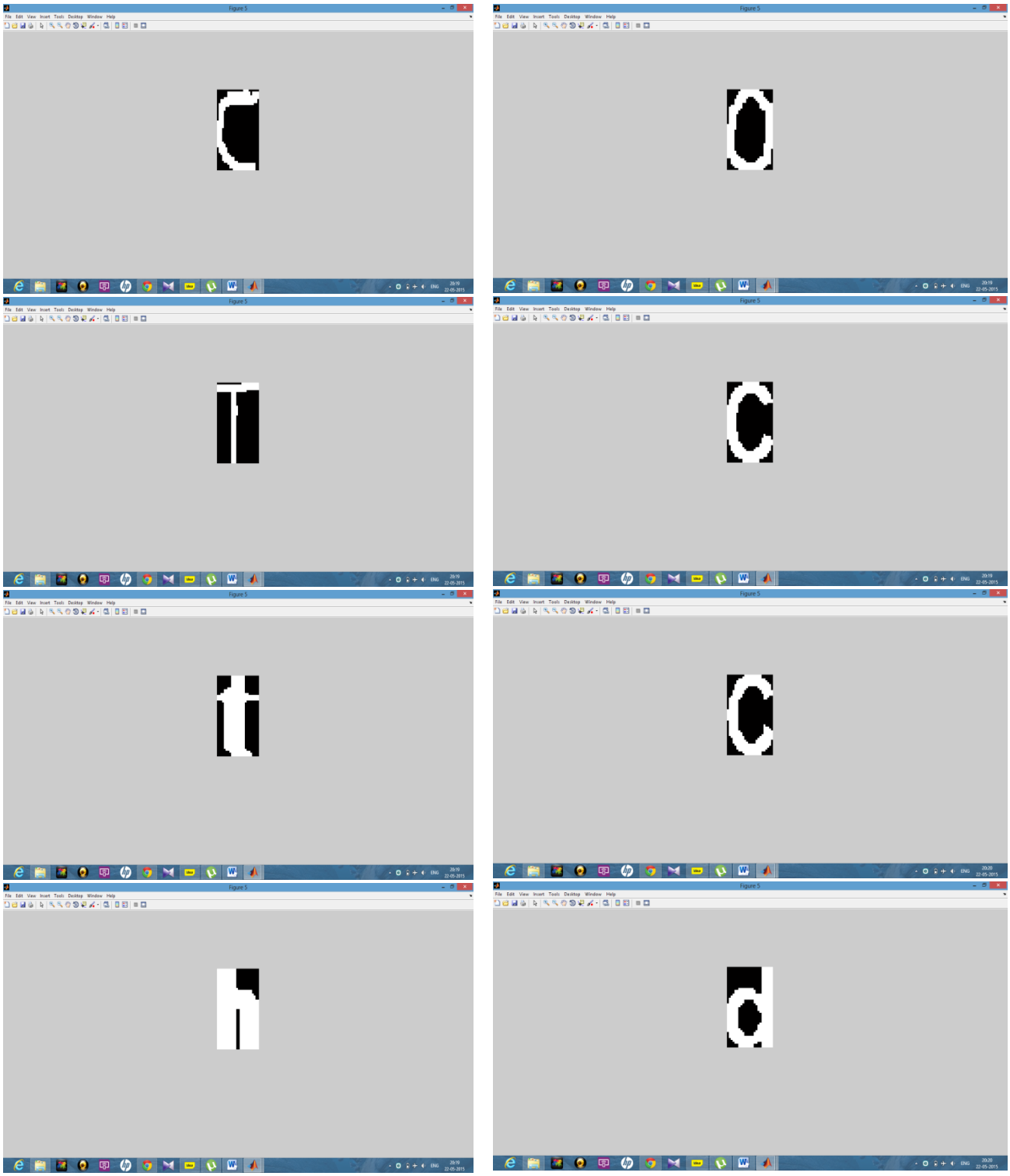
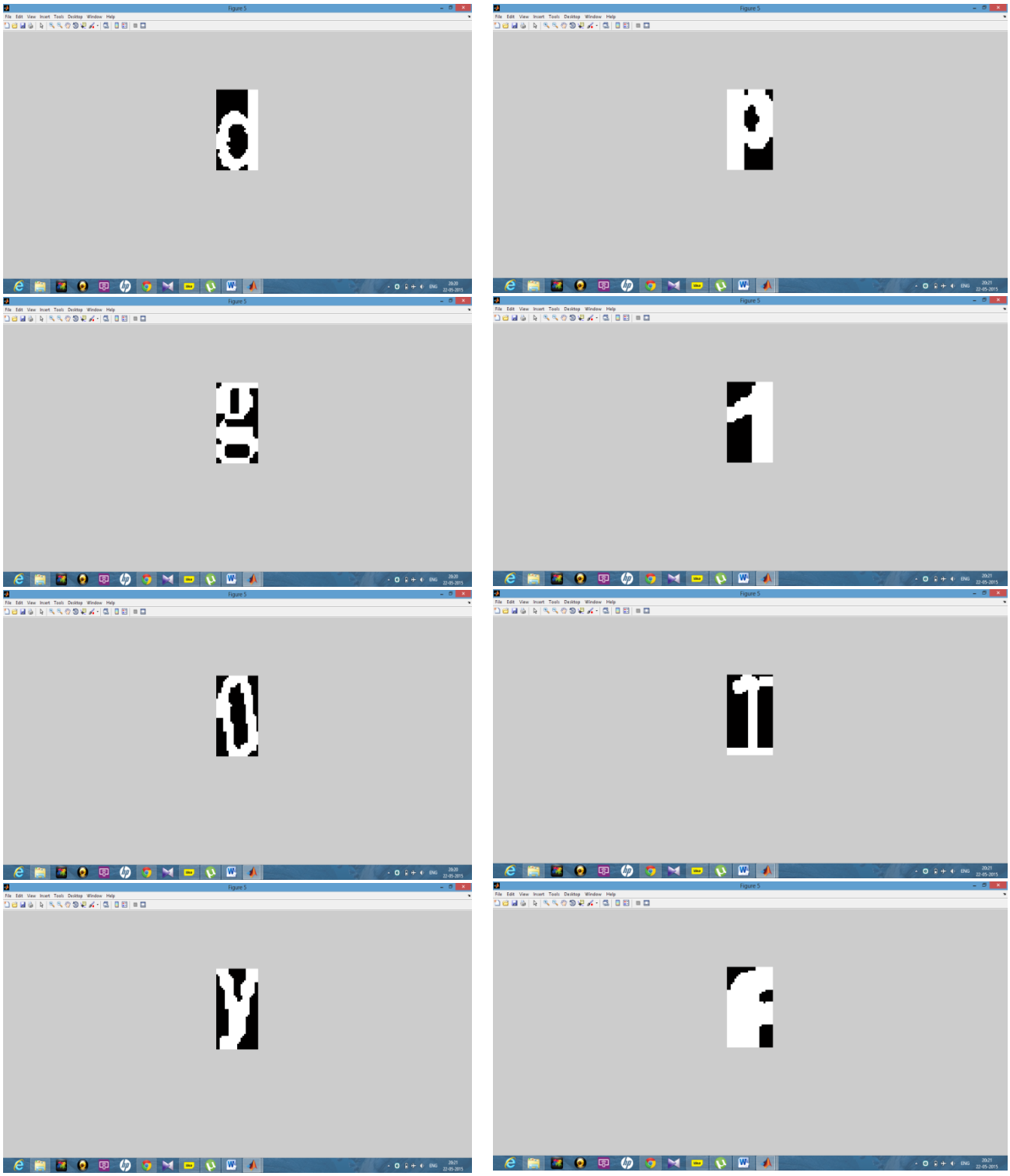
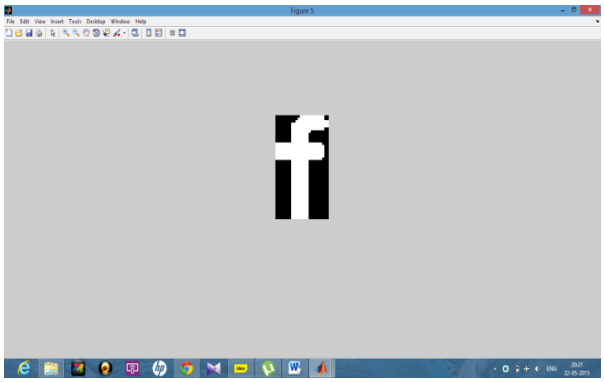
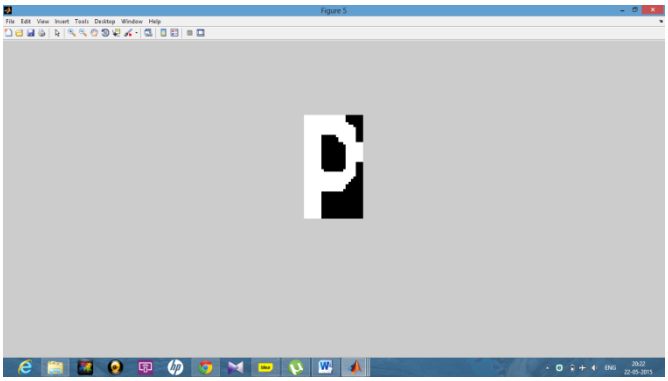
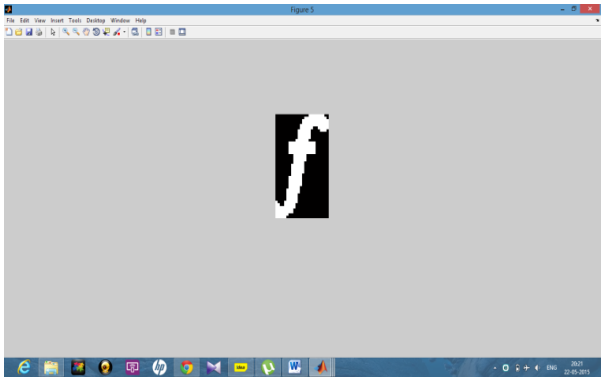
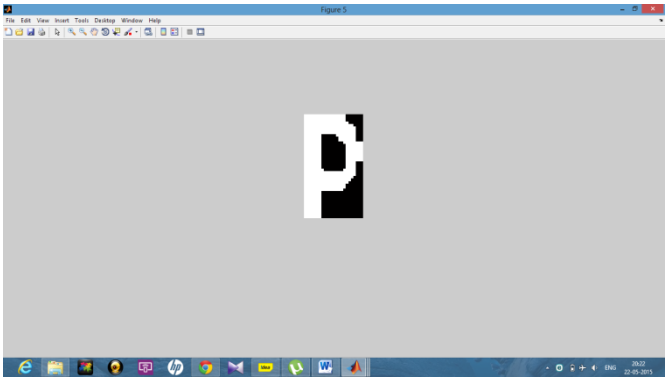
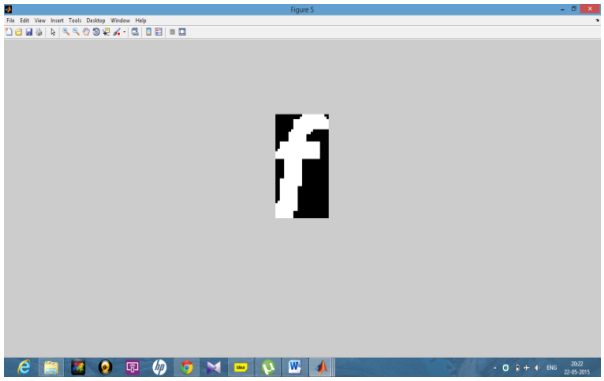
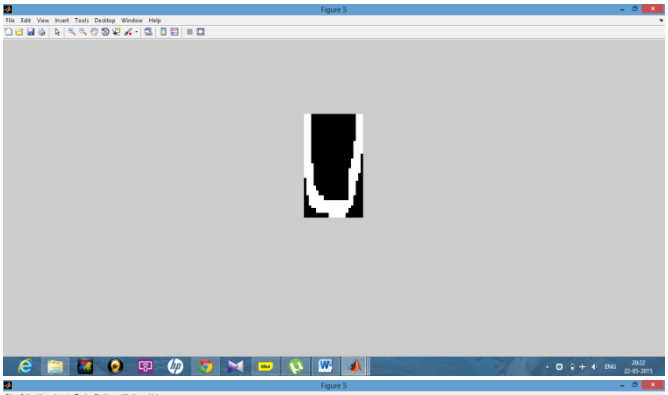
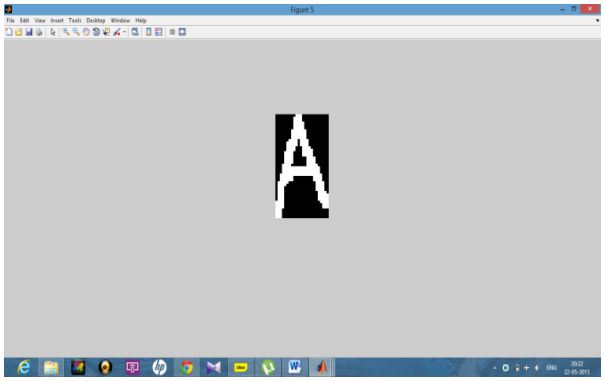
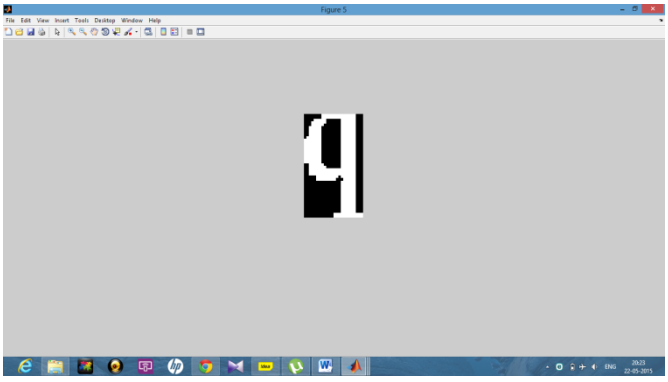


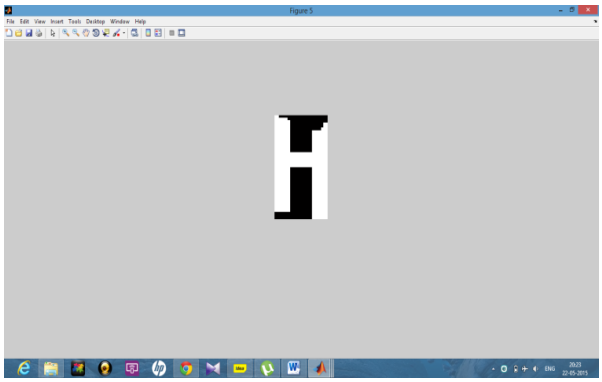
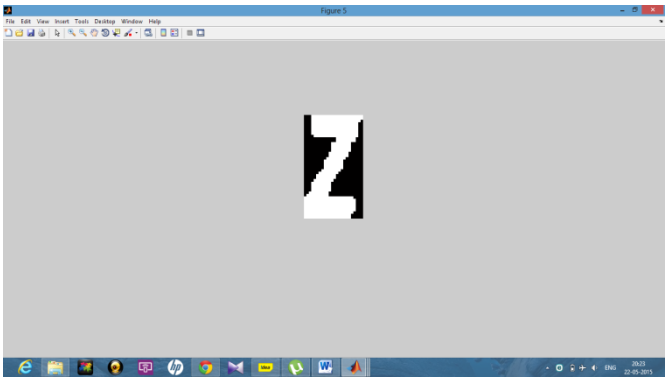
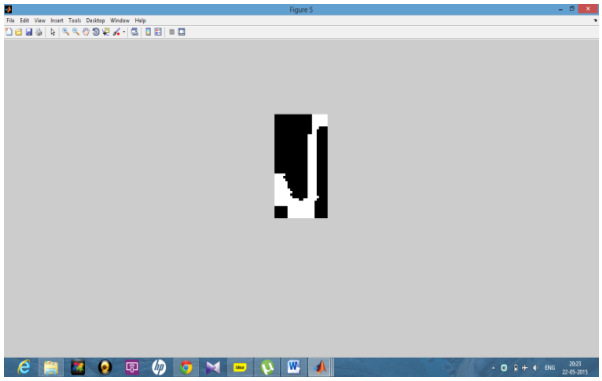
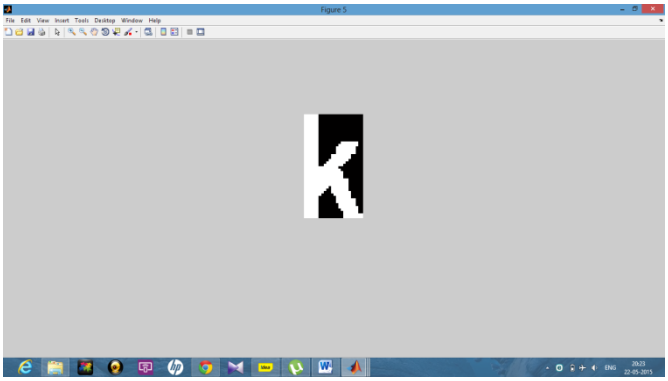
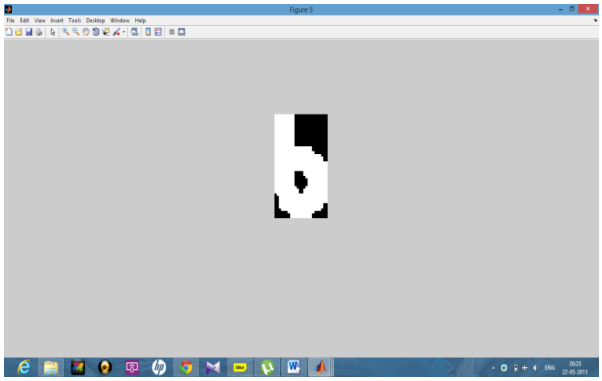
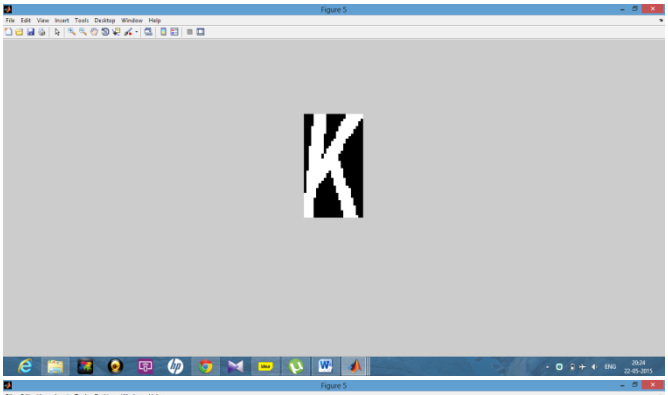
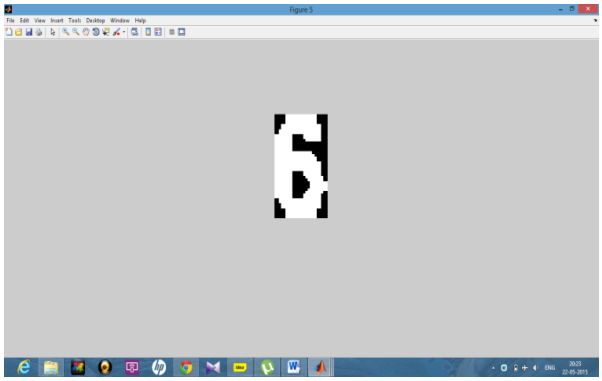
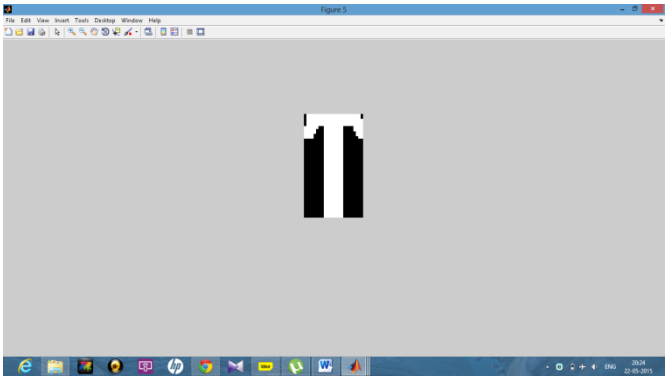
Fig. 5. Snapshot of segmented lines of characters





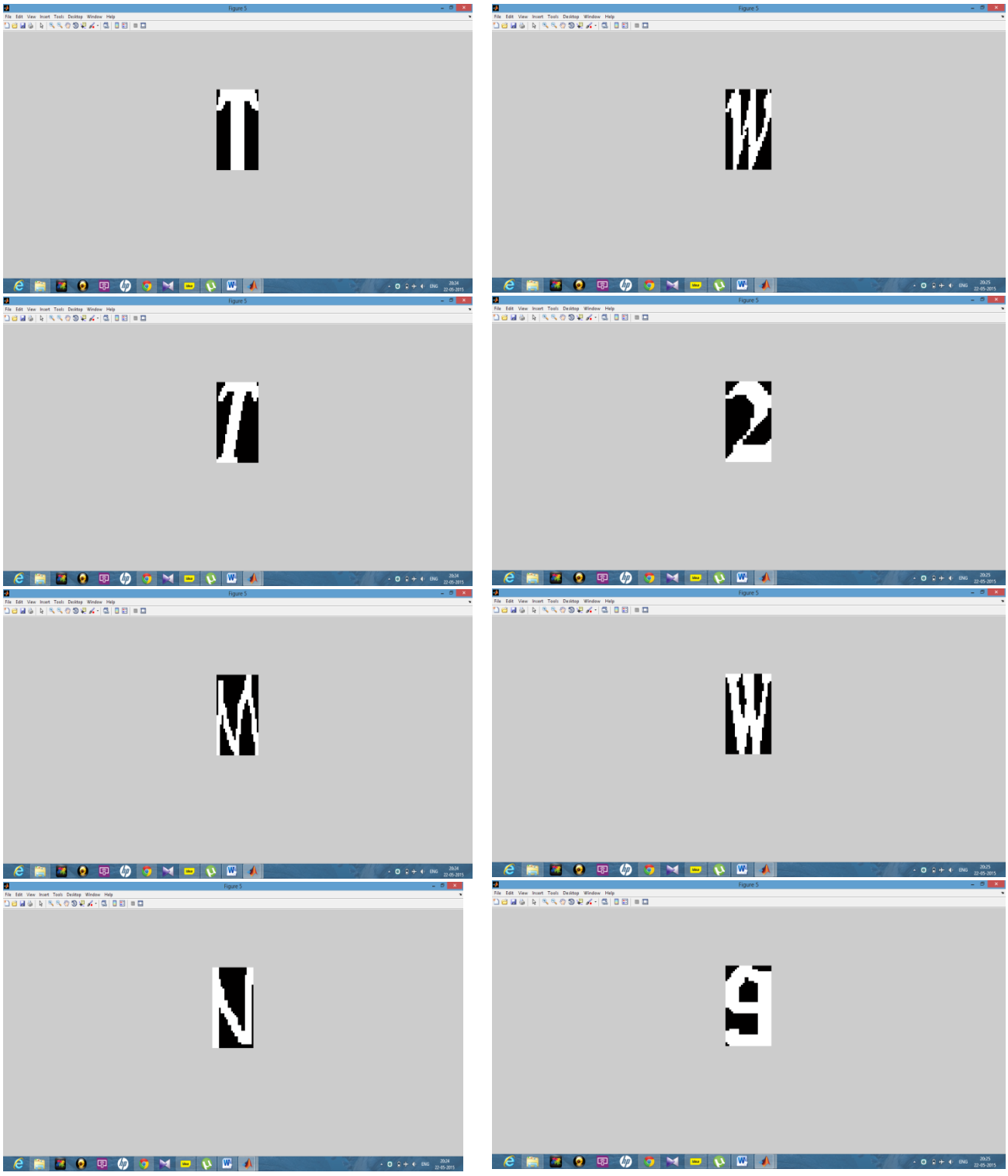


Fig.6. Snapshot of segmented characters

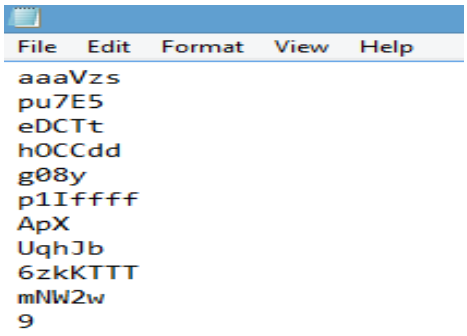


Fig.7 Snapshot of the recognized characters with proper sequence

CONCLUSION & FUTURE SCOPE

Recognition is one of the important tasks in Pattern Recognition. In the present work we have used template matching technique for character recognition. In this proposed work firstly we have done line segmentation & character recognition is done using the proposed technique. In our work we have got 100% accuracy in removal of noise & to resolve the problems related to various font sizes. The popularity and use of the Character Recognition is increasing day by day with the invention of new, fast and efficient hardware and software. But automatic character recognition for the Indian languages is still in preliminary stage and hence there is a need of lot of research to resolve the various issues and their complexities. There are many factors such as noise, various font sizes, broken lines or characters, quality of image, problems in segmentation that influence recognition process. As an overall view of the system prototype, it could be conclude that this proposed system prototype has been developed by using the technique that has mentioned and elaborated which is the Template Matching approach to recognize the character image. Besides, the interface of the system prototype looks user-friendly and makes the user of this system prototype easier to use it. As a result, the recognition process of this system become smoothly because of the steps that used in this proposed system while recognizing the character. A good text recognizer has many commercial and practical applications. In future we can also achieve 100% accuracy by improving more parameters like broken lines quality of image etc.

REFRENCES

[1] Nikhil Rajiv Pai, Vijaykumar S. Kolkure "Design and implementation of optical character recognition using template matching for multi fonts / size" International Journal of Research in Engineering and Technology Volume: 04, Issue: 02, | Feb-2015.

[2] Baheti M. J. Kale K. V. "Gujarati Numeral Recognition: Affine Invariant Moments Approach" 1st International Conference on Recent Trends in Engineering & Technology, Mar-2012 Special Issue of International Journal of electronics, Communication & Soft Computing Science & Engineering. PP. 140-146.

[3] Sanjay S. Gharde, Vidya A. Nemade, K. P. Adhiya "Evaluation of Feature Extraction and Classification Techniques on Special Symbols" International Journal of Scientific & Engineering Research Volume 3, Issue 4, April-2012. PP.1-5.

[4] Ms.M.Shalini1, Dr.B.Indira" Automatic Character Recognition of Indian Languages – A brief Survey" IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 2, April 2014. PP, 131-138.

[5] Amitava Choudhury, Joydeep Mukherjee "An Approach towards Recognition of Size and Shape Independent Bangla Handwritten Numerals" International Journal of Science and Research (IJSR), India.PP. 223-226.

[6] Nafiz Arica and Fatos T. Yarman-Vural "An Overview of Character Recognition Focused on Off-Line Handwriting" IEEE Transactions on Systems, Man, and Ybernetics Part C: Applications and Reviews, Vol. 31, No. 2, May 2001. Pp. 216-233.

[7] Chien-Hsing Chou , Wen-HsiungLin, FuChang "A binarization method with learning-built rules for document images produced by cameras" Pattern Recognition 43 (2010) 1518–1530.

[8] Rachit Virendra Adhvaryu "Optical Character Recognition Using Template Matching" International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) Vol. 3, Issue 4, Oct 2013, 227-232.

[9] Nadira Muda, Nik Kamariah Nik Ismail, Siti Azami Abu Bakar, Jasni Mohamad Zain

"Optical Character Recognition By Using Template Matching (Alphabet)" pp. 1-6.

[10] Jagruti Chandarana and Mayank Kapadia " Optical Character Recognition" International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 5, May 2014 pp. 219-223.

[11] Sunny Kumar and Pratibha Sharma "Offline Handwritten & Typewritten Character Recognition using Template Matching" International Journal of Computer Science & Engineering Technology (IJCSET). Vol. 4 No. 06 Jun 2013pp. 818-825.

[12] Jagruti Chandarana and Mayank Kapadia " Optical Character Recognition" International Journal of Emerging Technology and Advanced Engineering, Volume 4, Issue 5, May 2014 pp. 219-223.

[13] Jatin M Patil, Ashok P. Mane "Multi Font And Size Optical Character Recognition Using Template Matching" International Journal of Emerging Technology and Advanced Engineering Volume 3, Issue 1, January 2013 pp.504-506.

[14] Nadira Muda, Nik Kamariah Nik Ismail, Siti Azami Abu Bakar, Jasni Mohamad Zain "Optical Character Recognition By Using Template Matching (Alphabet)" pp. 1-6.

[15] Ning LI "An Implementation of OCR System Based on Skeleton Matching" Computing Laboratory University of Kent at Canterbury United Kingdom August 1991.

[16] Mohammed Ali Qatran "Template Matching Method for Recognition Musnad Characters Based on Correlation Analysis" Department of Computer Science, Amran University, Yemen.pp.1-4.