IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Segmentation Of Symbols Using Image Processing

[1] Shivabasayya Kulkarni, [2] Doddamani Basavaraj

[1,2] Lecturer in Department of Computer Science and Engineering

[1,2] Government Polytechnic for Women, Hubli, Dharwad, Karnataka, India

Abstract: The recognition of handwritten characters, numbers, special characters and mathematical expressions is difficult and challenging task and encourages a several applications, for example, text reader, bank check preparing system and many more. The framework capacities by performing four operations: preprocessing, feature extraction, segmentation and classification. Converts an input image into gray scale image, and then resize the image to standard size, and finally getting binary image of resized image. Segmentation of the preprocessed image by morphological operations like thinning should be done. In this work, for recognition of handwritten characters, number and mathematical symbols using zone based statistical feature extraction technique is proposed. The zone based statistical technique works in two stages. During training stage, the zone based statistical features are extracted from training tests and knowledge base is constructed. During testing stage, the image is prepared to acquire zone based statistical features and recognition of mathematical symbols is done using nearest-neighbour (K-NN) classifier. The strategy is robust and insensitive to variety in style, skew and different corruptions. The recognition exactness of 90% is accomplished.

Index Terms – Segmentation, Image Processing, Normalization, symbol extraction.

I. INTRODUCTION

Detecting and segmenting of characters, digits, symbols and equations from handwritten is the main intention of our project. We have a wide range of handwritten symbols such as algebraic equations; differential equations etc., and also have different types of symbols like theta, summation, infinity and so on. The extraction of symbols from images involves pre-processing, segmentation, feature extraction, classification and recognition. In this project i am concentrating mainly on segmentation and recognition. Pre-processing incorporates the process that are required to shape the input image into a form that is appropriate for segmentation.

Recognition of symbols from the image is another sort of issue in the field of image processing. Data which appear as equation in images may differ from each other in terms of size, style, font, orientation, contrast, which makes it an extremely makes task to extract the symbol contained in an image with higher accuracy. While universal study of related problems such as online handwritten symbols recognition, offline character recognition etc., the problem of symbol information extraction from an image is not well surveyed. Many strategies have been followed in the past but the problem is still the challenging one. In consequence, the extraction of equation in mathematical symbols is another crucial engineering science in computer vision. The information available in images which contained mathematical symbols is considered to be an important aspect of overall image understanding.

Segmentation is most influential part in image processing. In the segmentation stage, the input image is divided into separate characters and after that, every character is again resized into m x n pixels. Segmentation of an entire image into different parts which are more significant and easier for additional steps. These fragmented parts can be rejoined will become the entire image. The main intention of the segmentation process is to discover more knowledge in the influenced region in an image which motivates the annotation of the object. The main target of segmentation is to mark difference between the object of image and the background of the image. Image that contains different features that are extracted from segmentation. Segmentation depends on number of features in the image. This features may be either color,

texture or special symbols etc. The choice of a segmentation technique over another and the level of segmentation are picked by the particular kind of picture and characteristics of the issue being considered. Basically image segmentation techniques are divided into five types based on the following two properties and they are:

Detecting Discontinuity: Detecting discontinuity meaning is that, the image can be partitioned based on random changes in the intensity. Image segmentation algorithm such as edge detection segmentation is a good example of detecting discontinuity.

Detecting Similarity: Detecting similarity meaning is that, the image can be portioned into regions which are similar kind of predefined paradigm. Image segmentation algorithm such as thresholding, region growing, region splitting and merging are good examples of detecting similarity.

II. PROBLEM STATEMENT

The goal of work is to plan and add to another system which takes care of the issue of symbol extraction from images which contains handwritten symbols.

Scope and difficulties of work:

We can usually write mathematical symbols such as such as integration, exponents or arithmetic operator by hand. But there is no human suitable route to enter these symbols into computer. Mathematical formula contain two dimensional information so there is two problems need to be solved and they are symbol segmentation and symbol recognition.

In the field of computer vision, images are one of the most important channel of transferring information. The information can be extracted from images by understanding the image. Now we need a technique to understand the image and extract the information or objects. Image segmentation fulfills the above requirements. That is why, the segmentation is the important step in recognition of handwritten symbols.

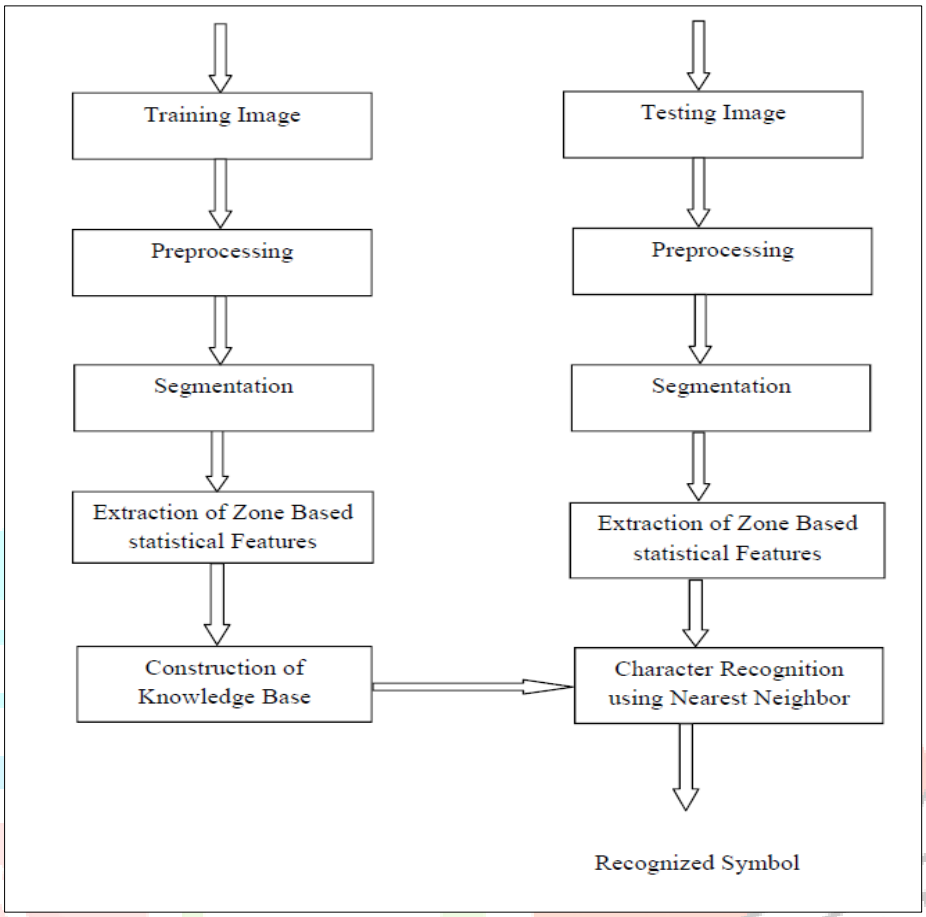
Handwritten mathematical symbols is frequently required to be naturally detected and handled. The investigation field of written by hand character recognition gets a growing consideration because of the variety of potential applications. Such applications incorporate bank check perusing, postal location perusing, image record perusing, insurance, signature perusing, and so on.

Objectives:

1. The proposed work concentrates on the recognition of disconnected hand written characters with no modifiers.
2. A zone based statistical feature extraction is utilized for recognition alongside other sorts of recognition strategies.
3. Two assessment parameters are to be utilized as a part of the relative investigation specifically recognition rate and handling
   time.

PROPOSED SYSTEM:

The proposed model uses zone based statistical feature extraction is used for recognition of handwritten symbols. This strategy includes distinctive stages, for instance, preprocessing, feature extraction, knowledge base construction and nearest neighbor classifier for mathematical symbol recognition. The block figure of the proposed model is presented in below Figure. The explanation of each and every step is displayed in going with subsections below.



PREPROCESSING:

Pre-processing can be characterized as cleaning the input picture and improving it suitable for feature extraction. The mathematical symbol image is preprocessed for binarization, removal of noise, and resized to a consistent determination of size 30 x 30. Further, the image is thinned and bounding box is produced. Real strategies, for example, binarization, crop to edge, distortion evacuation, normalization, and thinning are considered under pre-processing.

Binarization: The work of binarization is to discrete the foreground from the background. The threshold value is given somewhere around 0 and 255. Supplant each one of the pixels with gray level lower than or identical to T with black (0) and the rest with white (1). Decision of the best possible edge quality is picked all globally or locally. This approach is called thresholding, in which gray scale picture is changed over into binary picture.

Normalization: The process of translating the unequal sized equation or picture into standard sized image is known as normalization. The inter class variation among characters are eliminated by this normalization. Extra white spaces (background space) present in the image are removed, before applying normalization. Finally, the given input image is normalized to a standard resolution.

Thinning: Thinning is a image preprocessing operation. This operation performed to make the image crisper by decreasing the twofold esteemed image region to lines that estimated the basic structure of the region.

SEGMENTATION:

In the segmentation phase, a sequence of an image of characters are divided into sub images of individual characters. Segmentation is the process in which we can give pre-processed input image and get isolated characters. Using a labeling process, we can assign a label to each isolated characters. We can get the information about the characters in an image by labeling process. In this stage input picture is segmented and trained using K-NN technique as a part of request to perceive the condition contained in an input image. The obtained picture will be preprocessed with a specific end goal to change over that picture to grayscale,

grayscale picture will be resized to a standard size lastly the resized picture is then changed over to binary. Each individual character is uniformly resized. At that point the parallel picture will be supplemented to subject morphological operations, draw a bounding box for all region of interest.

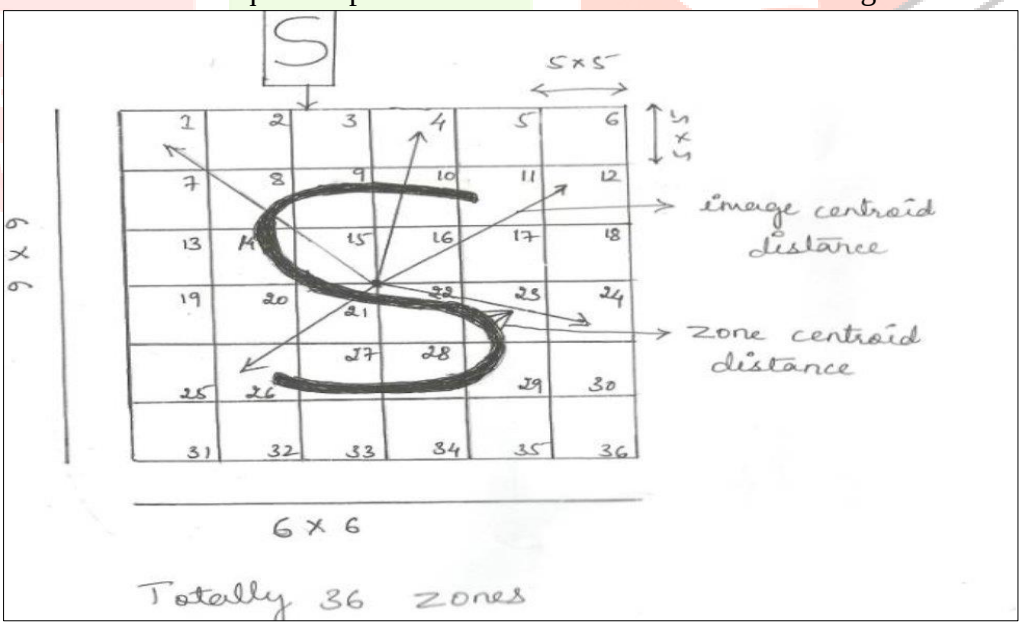
FEATURE EXTRACTION

Statistical distribution of points are called statistical features. This method provide high speed and low complexity. In this method image centroid and zone based distance metric feature extraction and zone centroid and zone based distance metric feature extraction are used. Here statistical features are used in terms of number of zones being used, distances from both image centroid and zone centroid measures and the area of applications.

The strategy zoning is used to calculate the number of black pixel present in every zone. The rectangle box surrounding the character is divided into many regions. The densities of black pixels within particular region are calculated and used as a feature set. This can be shown in figure below.

Feature extraction is special type of phase in image processing is used to reduce the dimension. A set of features is to be extracted, which maximizes the recognition rate with the least amount of elements is the major goal of feature extraction. When the input data is too large and it is doubt to be redundant that it is converted into a reduced representation of feature sector feature vector. Converting input image into a set of feature vector is called feature extraction. The important phase in recognition of mathematical symbol is feature extraction in which feature is used to indicate a small amount information that is pertinent for finding the solution of computational task. For the ease of implementation and good quality recognition, feature extraction is vital strong in recognition of handwritten mathematical symbols. Recognition of handwritten mathematical symbols contains different types of information such as characters, numbers and variety of symbols. We can build feature vector during or after segmentation. In a recognition system feature vector play an important role. A feature vector contains characteristics of class that is used to distinguish from other classes.

In this stage, the preprocessed image of size 30x30 is separated into 36 zones each of size 5x5. Here the entire image is divided into 36 zones. The centroid of image or character is processed. At that point the average distance from the image centroid to each pixel present in the zone/grid/box is computed. Totally 36 features of every image is subsequently acquired and stored in feature set X. Zones which are empty are considered to be zero. This technique is repeated for all zones in the character image.



The individual zone is again divided into 5 x 5 equal zone. The centroid of every zone is calculated. Next we can compute average distance from the zone centroid to each pixel present in the zone. The technique is continuously repeated until for all the pixels present in the zones/grids/boxes. There are some zones that empty. And that specific zone image value must store as zero in feature vector. At that point the statistical features are figured from all zones and put away into a feature vector X. The sum of all pixels, zone centroid and image centroid in every zone is used as a feature value. Totally 36 features are stored in feature vector X.

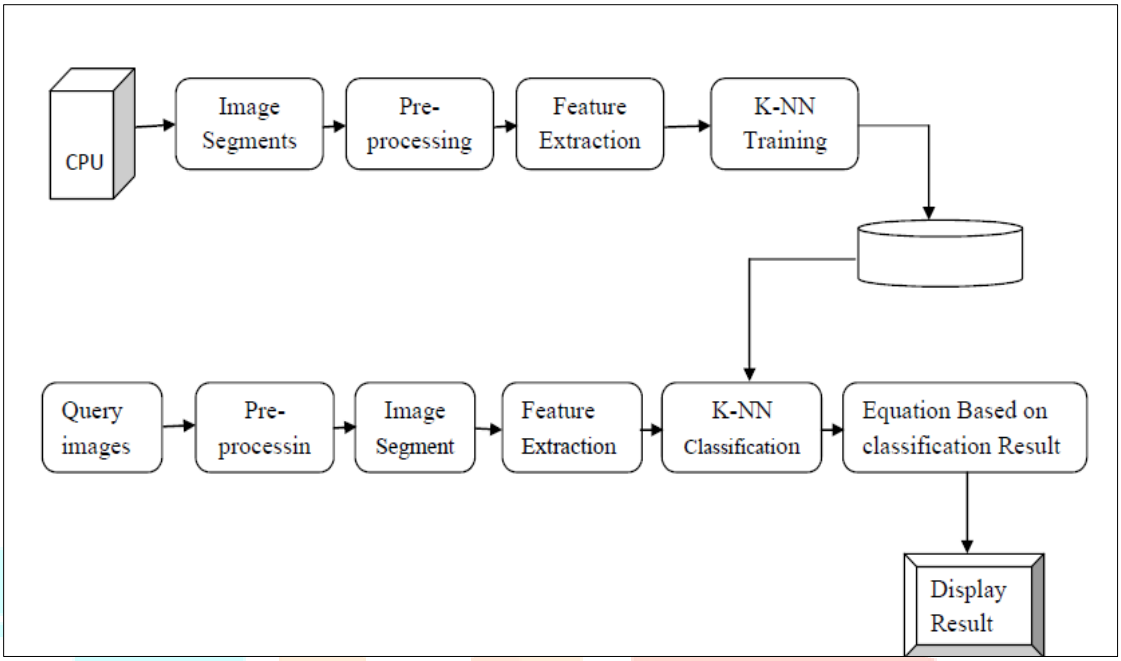
KNOWLEDGE BASE FOR CHARACTER RECOGNITION

The extracted features are utilized to manufacture the knowledge base. In the training stage, all the features are extracted from all the training images and knowledge base in constructed as a dataset of feature vector X.

Dataset consists of 47 different folders containing a collection of characters, numbers and mathematical samples.

For the knowledge base construction purpose, the characters, numbers, special symbols and mathematical equations were collected from different writers with variation in style, and size. The image database consists of more than 800 handwritten characters, numbers, special characters and mathematical symbols. Then from the database, samples are used for training.

ARCHITECTURE DIAGRAM OF PROPOSED MODEL



As we can see from the architecture diagram, the images which contains mathematical symbols will be given as input to proposed system, they will be segmented by applying morphological operations such as thinning and then pre-processed to convert to gray scale and resized to standard size finally we get binary image.

Features from those segmented images will be collected and then we will train using KNN training to make a knowledge base, we will also call this as database of training.

In the next step we will take user input which contains mathematical symbol, we follow the sequence of steps like pre-processing, image segmentation, feature extraction and then K-NN classifier training and finally equations will be displayed as output.

III. LITERATURE SURVEY

This section introduces the related work of the image segmentation by various authors with different perceptions regarding to the segmentation and recognition of symbols.

[1] Probabilistic Mathematical Formula Recognition Using a 2D Context-Free Graph Grammar.

Mehmet Celik and Berrin Yanikoglu have proposed a paper Probabilistic Mathematical Formula Recognition Using a 2D Context-Free Graph Grammar. This research work presented a probabilistic framework problem for the recognition of mathematical expression. In this paper its grammer can be extended easily because the developed system is flexible. So we are expressing thanks to its graph grammar which removes the necessity for specifying rule precedence. In this paper author tested this manually tokenizing the expressions and entering the exact OCR results. Author obtained result up to 85% for correct structure recognition. Here typically long expressions failed only by time out.

[2] A Structural Analysis Approach for Online Handwritten Mathematical Expressions.

B.Q. Huang and M-T. Kechadi have presented this paper. A structural analysis approach for mathematical expressions is introduced in this paper. This approach is based on the Attribute String Grammar and the Baseline Tree Transformation approaches. The proposed approach in this paper is consists of geometrical feature extraction, parsing structure and expression analysis steps. Structure parsing algorithm uses baselines. Baseline represented by geometrical features to continuously decompose the hierarchical
levels. It predicts the relationships along the symbols, during a hierarchy decomposition of an expression. Next groups the symbols into a number of sub-expressions and a set of basic units. In the experiments author used two datasets, the first dataset is used to train the spatial functions (the SVM and MLP) and the second is used to verify the whole structural analysis approach.

[3] Recognizing Offline Handwritten Mathematical Expressions based on a Predictive approach of Segmentation using K-NN Classification.

Sachin Naik and Pravin Metkewar have proposed this paper. Recognition of Hand written mathematical expressions has been a critical theme for some analysts for quite a long time. It stays a standout amongst the most difficult and energizing areas in pattern recognition. In the recognition procedure of offline hand written mathematical expressions, segmentation is the most vital procedure. Issues in ambiguities of distinguishing superscript and subscript in complex offline mathematical expressions stay a standout amongst the most critical issue. To the best of our insight little work has been done in the segmentation of offline hand written mathematical expressions as for superscript and subscript. In this paper a proficient segmentation procedure for superscript, subscript and fundamental characters inside offline hand written mathematical expressions has been proposed. This method depends on the generation of forecasts for superscript, subscript and fundamental characters inside hand written mathematical expressions, which helps for the recreation of mathematical expressions amid the recognition procedure with their spatial interrelationship. The proposed framework was led as an examination with a database of 300 examples of checked mathematical expressions that contained 2,000 symbols out of which there were 31 diverse sorts of Mathematical Symbols. The grouping of the components was done by the K-NN-classifier taking into account thickness highlights. This examination demonstrates wonderful results.

[4] Segmentation Method for Handwritten Character Recognition.

Namrata Dave proposes Segmentation Method for Handwritten Character Recognition. Hand written Character Recognition is range of examination since numerous years. Automation of existing manual framework is need of most commercial enterprises and government zones. Recognition of hand written characters is an interest for some fields. In this paper we have examined our methodology for written by hand character segmentation. This paper talks about different procedures to segment a text based image at different levels of segmentation. This paper serves as a helper for people managing the content based picture segmentation scope of Computer Vision. To begin with, the prerequisite for segmentation is shielded as to content based information recognition. By then, the distinctive segments impacting the segmentation methodology are discussed. Taken after by the levels of content segmentation are researched. Similarly, the available frameworks with their central focuses and inadequacies are evaluated, close by headings for fast referral are suggested. At last, we have given our approach to manage content division in a word.

[5] Image Feature Extraction Techniques and Their Applications for CBIR and Biometrics Systems.

Ryszard S. Chora's has proposed this paper. Visual components such as shape, shading and synthesis are isolated in CBIR (Content-Based Image Retrieval), to portray picture. Using one or more component descriptors are addressed by each of the elements. Components and descriptors of the query are stand out during the retrieval from those of the pictures in the database. This can be done to deciding objective to rank each listed picture as showed by its partition to the request. In the systems of biometric pictures used as illustrations (for e.g. one of a kind imprint, iris, hand et cetera.) are in like manner addressed by highlight vectors. The candidate's cases are then recuperated from database by taking a gander at the division of their component vectors. The element extraction systems for this applications are talked about. The major duties of CBIR are recognizing verification of the issues existing in CBIR and Biometrics structures. This depicting picture substance and picture feature extraction. This paper analyzed the use of different unmistakable shading, surface and shape highlights for picture recuperation in CBIR and Biometrics structures.

IV. EXPERIMENTS AND RESULTS

In this part the outcomes of the proposed model have been plot by considering the sample tests of the images which contains handwritten symbols like math symbols, notations and graphical notations. Here we will see test examination overseeing diverse issues in seeing numerical expressions in an image. With the end goal of testing, 30 of tests are prepared. The proposed framework has been surveyed for 30 examples. An exactness of 90% is accomplished.

AN EXPERIMENTAL ANALYSIS FOR HANDWRITTEN SYMBOLS

The equation of size with uneven thickness, style, and different corruptions in the equation image is first pre-processed for binarization, removal of noise and resized to a consistent determination. Further, the picture is partitioned into 36 zones, each of size 5x5. We have utilized zone based statistical feature extraction procedures on this project and repeated on various size of picture. We also used nearest neighbour classification to recognize mathematical symbols. Extracting features from test input image, next we have to

compute the distance between feature vector T of test sample and each stored record in knowledge base KB is determined using Euclidean distance measure. The zone based statistical features are used to extract features and stored in feature vector T. The test values are clearly mentioned the distribution of pixels in different segments or primitives of the character image. And, these dispersions are not quite the same in character to character because of varying in positions and shapes of segments of fundamental mathematical symbols and characters.

AN EXPERIMENTAL ANALYSIS DEALING VARIOUS SNAP SHOTS

For the recognition of handwritten symbols, the database is collected from different writers of up to 50 members with variation in style and size. The database contains characters a-z, numbers 0- 9 and special characters. Totally 800 samples contained in our database, and stored in jpeg format. The entire set is separated into two data sets namely training set and testing set, in that 50% images are used for training purpose. Another 50% images are used for testing purpose. Amid tests it is seen that, the zone based features made test images divisible in the feature space.

Thus, the proposed work is healthy and accomplishes a normal recognition accuracy of 90%. The general execution of the framework after conducting the experimentation on the dataset is recorded in below table. Previews of realizing mathematical symbols applying K-NN are indicated beneath.

Fig: GUI based main window

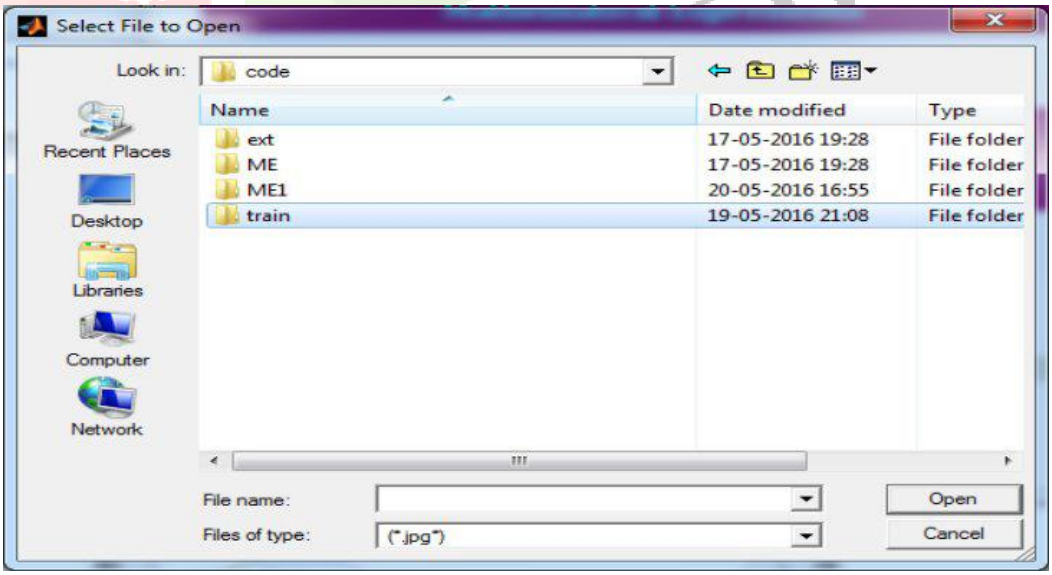


Fig: K-NN Trained expressions stored in a Train Folder

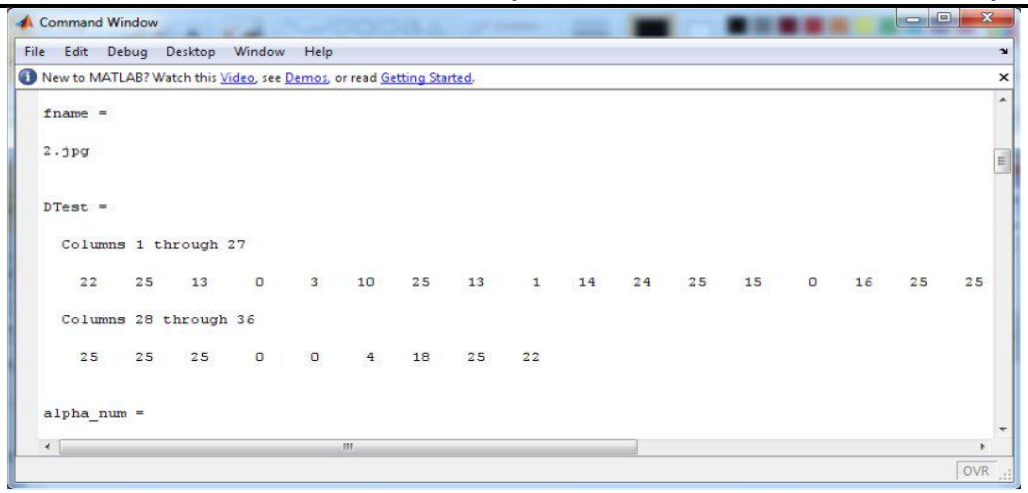


Fig: MATLAB Command Window classification

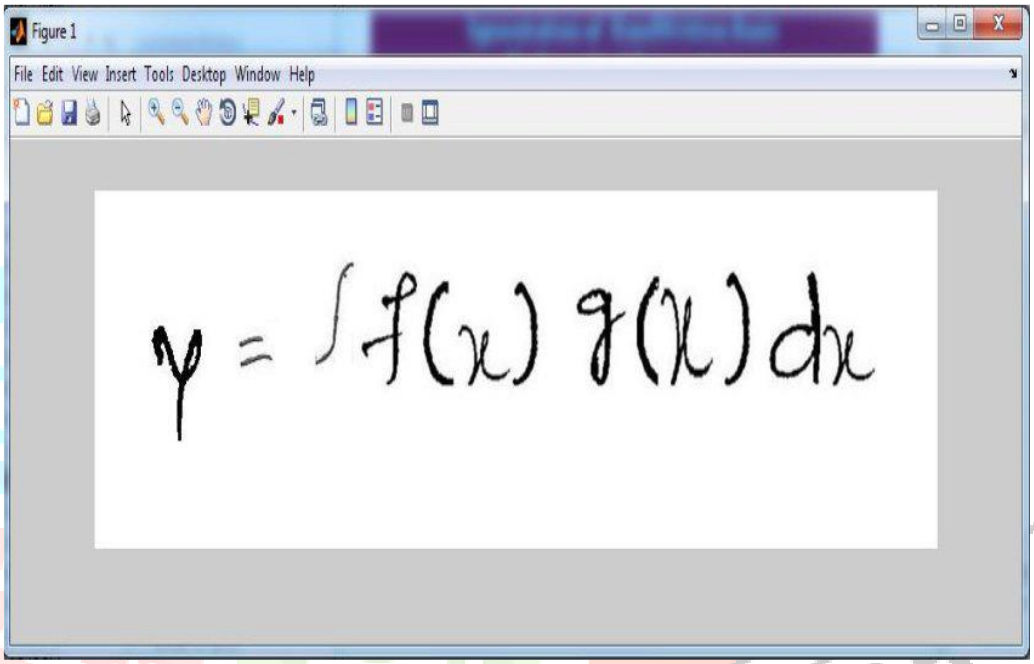


Fig: System shows the original image

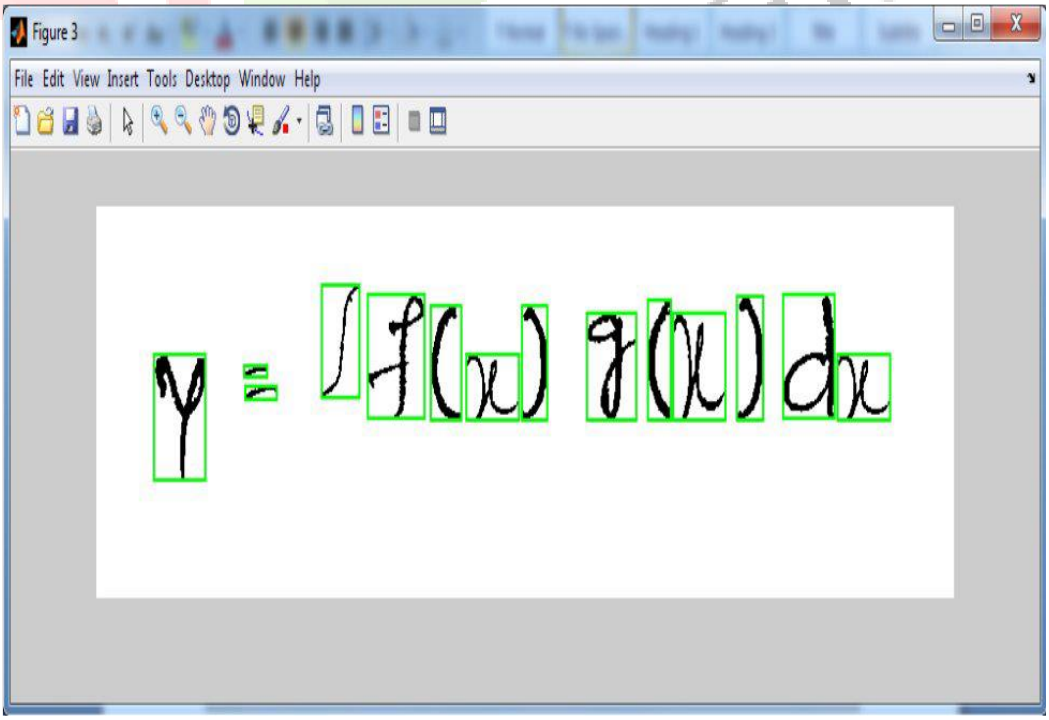


Fig: System shows segmented image

Fig: Segmented symbols t ,a and n

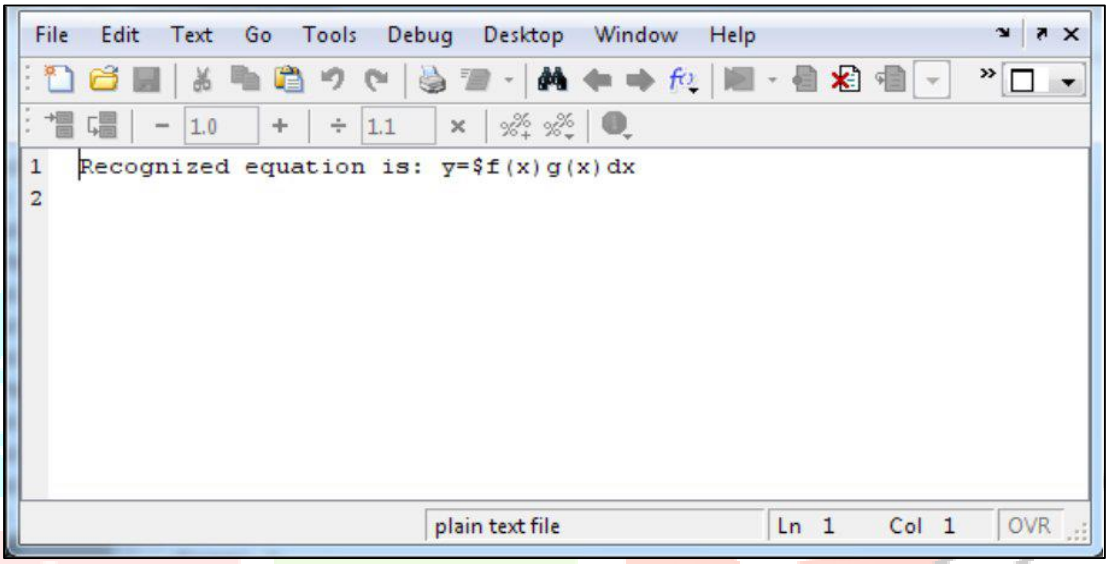


Fig: Recognized Expressions Stored in Text File

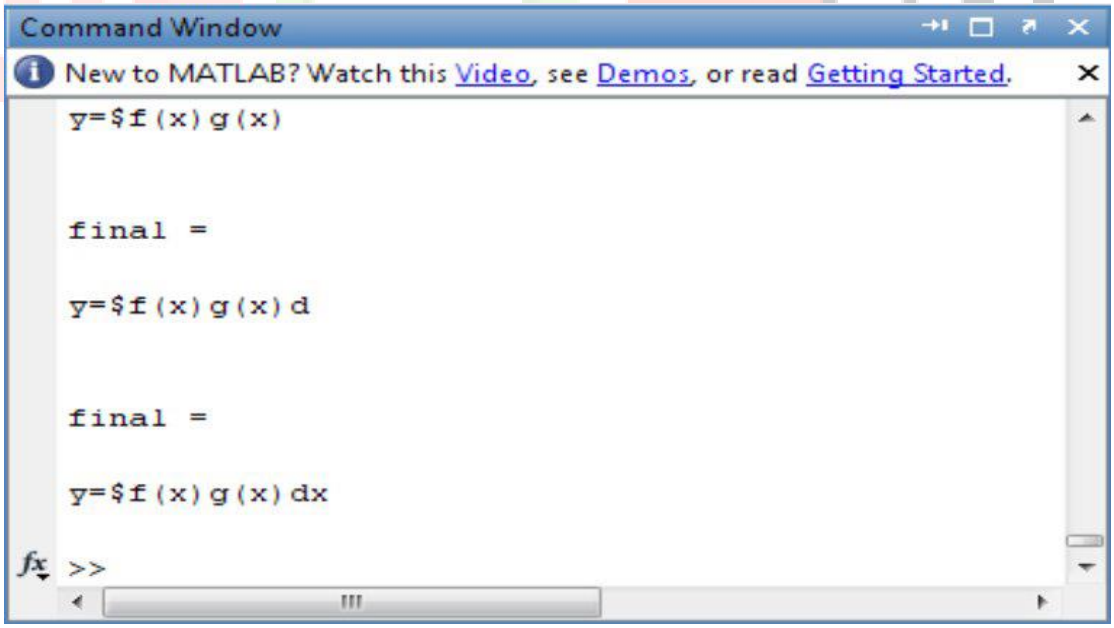


Fig: Command Window Output (result)

V. CONCLUSION

This project includes strategy for recognizing handwritten mathematical symbols, characters, numbers, special characters utilizing zone based statistical features. This zone based statistical feature combine image centroid zone and zone centroid zone. Thorough experimentations were to dissect the viability of zone based statistical feature. This zone based statistical feature utilizing k nearest neighbor classifier. K-NN classifier is used for classification and recognition. The detailed and complete experimentation showed

that the framework is strong and insensitive to size, style, skew, and so forth. This significant result, which makes this work appropriate for symbol understanding and interpretation system.

The trial results have shown that the proposed model is computationally powerful and it gives awesome and exact mathematical symbol extraction rate. The extraction of symbols exactness of 90% is accomplished.

VI. FUTURE ENHANCEMENT

1. Extraordinarily composed textual styles (characters) are not considered.
2. Running Handwriting is not considered.
3. Attached characters are not recognized.
4. All the special characters are not considered such as square root, sigma, etc.
5. Preformatted paper is utilized to gather the information tests.
6. In handwritten the meaning of some characters little bit confusion because they look like same such as C and (, O and 0.

REFERENCES

[1]. Wichian Premchaiswadi, 2009, "A mobile Image search for Tourist Information System", Proceedings of 9th international conference on SIGNAL PROCESSING, COMPUTATIONAL GEOMETRY and ARTIFICIAL VISION, pp.62-67.

[2]. Te´ofilo E. de Campos and Bodla Rakesh Bab, 2009, "Character Recognition in Natural Images", Computer Vision Theory and Applications, Proc. International Conf. volume, pp. 273- 280 (2009)

[3]. Onur Tekdas and Nikhil Karnad, 2009, "Recognizing Characters in Natural Scenes: A Feature Study", CSCI 5521 Pattern Recognition, pp. 1-14 (2009)

[4]. Sangame S.K., Ramteke R.J., and Rajkumar Benne, 2009, "Recognition of isolated handwritten Kannada vowels", Advances in Computational Research, ISSN: 0975–3273, Volume 1, Issue 2, pp 52-55 (2009)

[5] A. Belaid and J. P. Haton. A syntactic approach for handwritten mathematical formula recognition. *IEEE Trans. PAMI, Vol 6.*, pages 105–111, 1984.

[6] S. K. Chang. A method for the structural analysis of 2-d mathematical expression. *Information Sciences 2(3)*, pages 253–272, 1970.

[7] B. B. Chaudhuri and U. Garain. An approach for recognition and interpretation of mathematical expressions in printed document. *Pattern Analysis and Applications*, 3(2):120–131, 2000.

[8] A. K. Das. *Document Image Segmentation: A morphological approach*. PhD thesis, Bengal Engineering College (Deemed University), Sibpur, India, 1998.

[9] Qi Xiangweri Pan Yusup Wang Yang," The study of structure analysis strategy in handwritten recognition of general mathematical expression", 978-0-7695-3600-2/09 2009 IEEE.

[10] Nafiz Arica, "An Off-line character recognition system for free style Handwriting" thesis submitted on Sep 1998.

[11] J.Pradeep, E. Srinivasan, S.Himavathi, "Neural Network based Handwritten Character Recognition system without feature extraction" International Conference on Computer, Communication and Electrical Technology 2011 IEEE.

[12] http://www.scribd.com/doc/60245721/English-Character-Recognition-System-Usingmatlab.

[13] Xue-dong Tian, Li-na Zuo, Fang Yang, Ming-hu Ha," An Improved Method Based On Gabor Feature for Mathematical Symbol Recognition ", -4244-0973-X/07 2007 IEEE.

[14] Qi Xiangwei, Xinjiang, "The study of mathematical expression recognition and the embedded system design", Journal of Software, Vol. 5, No.1, January 2009.

[15] Nicolas D. Jimenez and Lan Nguyen "Recognition of Handwritten Mathematical Symbols with PHOG Features", http://cs229.stanford.edu .

[16] B. Keshari, S. Watt, "Hybrid Mathematical Symbol Recognition using Support Vector Machine", Analysis and Recognition – volume 02 Pages859-863, 2007.