ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Comprehending Handwritten Mathematical Formulation Using SVM Classifier

Abirami M, Rajashri S , Ramya R

Assistant Professor, Dept. of CSE., St.Joseph's Institute of Technology, OMR Chennai, India
UG Student, Dept. of CSE., St.Joseph's Institute of Technology, OMR Chennai, India
UG Student, Dept. of CSE., St.Joseph's Institute of Technology, OMR Chennai, India

ABSTRACT: Handwritten Character Recognition remains one of the most fascinating and exciting areas in the field of Image Processing and Pattern Recognition. Recognition of handwritten characters is a demand for many fields. It is the process of conversion of handwritten text into machine readable form. Handwritten mathematical expressions recognition is yet a challenging task due to its intricate spatial structure, tangled semantics and 2-dimensional layout of the characters. The challenge in comprehending handwritten expressions is the variability of symbols from one writer to another. Differences may occur in shape, size and position of characters even when the same person writes them. In this project we will explore one of the efficient methods to interpret handwritten characters using concepts of Segmentation, Classification and Latex Conversion.

KEYWORDS: Handwritten, Recognition, Segmentation, SVM, Latex

I. INTRODUCTION

For centuries, Handwriting is one of the most basic tools for communication as it forms as an integral part of the learning process. Today, computers and the internet are the crucial way of modern communication, which is turning the world into a small town.

The input of mathematical symbol recognition constitutes an essential part in most scientific and engineering disciplines. The input of mathematical expressions into computers is generally difficult compared to that of plain text, because math expressions will contain numbers, special symbols and Greek letters. With these large number of characters and symbols, the frequently used keyboards has to be modified in order to accommodate all the keys needed and also to make use of math symbols a unique keyboard can be designed along with the normally used keys. But making use of this specially designed keyboard is a challenging task because it requires meticulous training and practice.

Recent advances of the usage of digital pens and touch screen allows us to use handwritten input tools which is an interesting task to input math expressions in digital document. Hence, it is important to develop a system in order to convert natural handwritten data to digital format. The recognition systems are widely classified into two types namely, Online systems in which the user writes the math expression using Tablet PC's or some pen based technologies which thereby generates a sequence of points. While, Offline systems, user writes the expression in a piece of paper and scan the document which generates two dimensional arrays of pixels. The mathematical expressions can be of any type namely Greek letters, algebraic expressions, functions, relations, matrices. Recognizing these types of expressions is a resilient task which should be taken immense care while recognizing it.

Many challenges arise in recognizing math symbols. Symbols such as dot and commas, are frequently used and are critical to meaning of notation, which is very difficult to distinguish from one another. These symbols play various roles according to situations like a dot operator can represent a decimal point, a multiplication operator or a symbol annotation which leads to ambiguity problems. The ambiguity of spatial relationships is extensively increased, due to free placement and alignment of symbols that includes superscripts, subscripts which consists of symbols with different fonts and typefaces and it has to be dealt carefully in order to avoid problems.

Our research focuses on recognizing handwritten math expressions using a desktop or laptop system. The advantage of our proposed model is to develop a math system which recognizes the input expressions thereby perform segmentation and simultaneous classification of symbols among various classes namely Latin variables, Greek letters and Special Symbols. Specifically, the classifier is used to recognize symbols based on SVM algorithm and thus the classified results are converted to Latex format which is extensively used for documentation purposes.

II. Related work

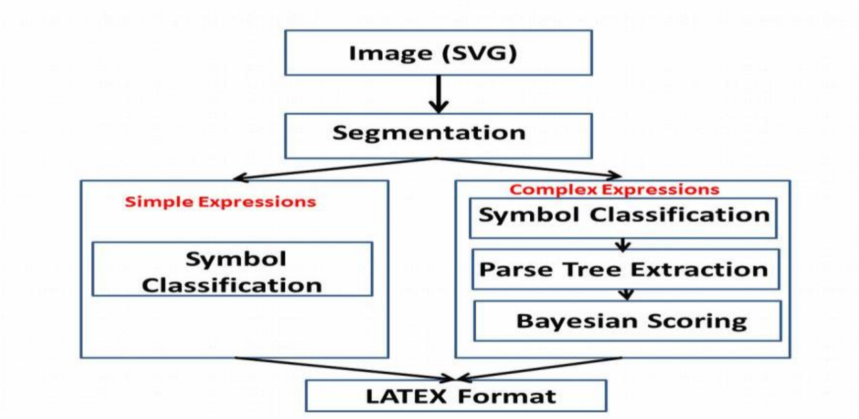
Namarta Dave [1], “Segmentation Method for Hand Written Character Recognition”, (2015).The paper is mainly focused on two stages, pre-processing of document image followed by segmentation phase.The segmentation methodologies used in character recognition are pixcel counting approach and histogram approach. In pixel counting approach the line separation procedure consists of scanning the image row by row. Histogram approach is a method to automatically identify and segment the text line regions of a handwritten document. The pixel counting algorithm is simple to implement and we can conclude that it excels only for the printed text document. Also, additional overhead like skew correction module is required. Monica Patel et al [2] , “Handwritten Character Recognition in English”, (2015). The structure of handwritten character recognition consists of three steps preprocessing, segmentation, feature extraction and classification. In preprocessing Noise removal, Binarization, Skew correction are performed. In segmentation there are three parts line segmentation character segmentation, word segmentation. Next in feature extraction statistical feature and structural features are extracted. Finally the classification is based on two types of learning supervised and unsupervised learning.Separate characters givegood accuracy but word recognition is affected bydifferent writing style. Holistic method eliminates the complicate segmentation but they use limited vocabulary.
Sachin Naik et all [3], ”Recognizing Offline Handwritten Mathematical Expressions (ME) Based On A Predictive Approach Of Segmentation Using K-NN Classification” (2016). The main challenge is to generalize segmentation techniques to accommodate a large set of mathematical expressions for recognition. The input consists of superscript and subscript components. The output consists of 3×n matrix with appropriate location for superscript, subscript and main characters within HME. After the extraction process, the features extracted from each of the components are sent to a classifier. The objective here is to interpret the sequence of components of ME taken from the test set.This is K-NN classification method.After this symbol recognition and reconstruction of mathematical expression takes place.
Sagar Shinde et al [4] ,“Handwritten Mathematical Expressions Recognition using Back Propagation Artificial Neural Network” (2016). In this paper, feed-forward back propagation neural network is implemented to achieve both character recognition and mathematical structure recognition with upgrade in effective performance in addition to accuracy of the experimental results including lessen efforts. System proves its potency by recognizing expressions in analysis of math documents. Centroid and bounding box are the key features that are extracted from each character and uprightness of this system is achieve using back propagation neural network for the recognition of mathematical equations. Chuanjun Li et al [5] , “Online Recognition of Handwritten Mathematical Expressions with Support for Matrices” (2016). This paper presents an online system for recognizing handwritten mathematical matrices in the context of an interactive computational tool called MathPaper. Automatic segmentation and recognition of multiple expressions are supported based on a spacing algorithm that leverages recognized symbol identities, sizes, and relative locations of individual symbols. Both well and non-well-formed matrices can also be recognized. Matrix elements can be any general mathematical expressions including imbedded matrices. Our recognizer also addresses the poor column alignment problem of handwritten matrices, and allows for slight horizontal overlaps between elements in neighboring columns and different rows.

III. Proposed System

The conventional systems used algorithms like pixel counting algorithm, KNN classification, back propagation artificial neural network, matrix recognition, Multilayer Perceptron Neural network(MLP), Structural analysis using Context Free Grammar and many such methods. However they lacked efficiency and ease of operation. In this project we propose a unique method for the computer to comprehend offline handwritten mathematical and logical expressions using the SVM Multiclass algorithm. The important modules are Segmentation, Classification and

Latex conversion. The tools used in this project are, Inkscape, which is used in mathematical expression recognition to edit vector graphics such as illustration, diagrams, complex paintings etc. Classification is executed using the Rstudio, a parse tree is generated, which represents the syntactic structure of a string according to some context free grammar. Finally, the result of the comprehended mathematical expression is viewed in LATEX.

SYSTEM ARCHITECTURE:

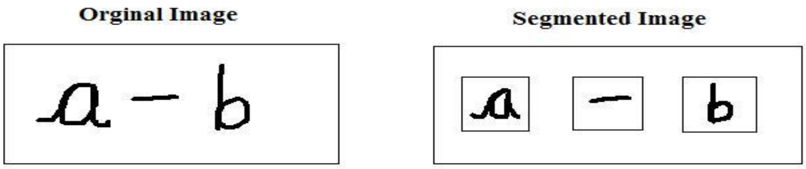


A. *Segmentation*

Symbol Segmentation is the first step in recognizing handwritten mathematical expressions. The segmentation process takes place in such a way that the SVG file is taken as an input and for each pen down and pen up, the symbols are segmented and stored as a SVG file along with the pen coordinates. The SVG (Scalable Vector Graphics) is a language for describing two-dimensional graphics in XML. SVG allows for three types of graphic objects: vector graphic shapes (e.g., paths consisting of straight lines and curves), images and text.

PSEUDO CODE

```
Step 1: The handwritten expression is written by the user using Inkscape, a vector Graphics editor.
Step 2: The math expression is stored in *.svg format.
Step 3: The expression is segmented based on various pen ups and pen downs and a separate SVG file is created for each pen movement.
Step 4: Once the SVG files are extracted for a single expression, the SVG files are converted to PNG format to perform recognition of symbols.
Step 5: The PNG images are used for performing classification.
```



B. *Symbol Classification*

Classification is the most important step in recognizing the characters in the handwritten mathematical expressions. The various class of characters and symbols that are to be recognized are, Latin Variables, Greek Letters, Operators, Functions, Matrices. These characters and symbols are segmented as the individual symbols based on the input. This recognizes the entire handwritten expressions along with the spatial layout of symbols. (eg. Adjacent, subscript). In this project we use the SVM classification to classify these characters. In the Classification phase, initially, the test and train data set is converted into matrix values. These converted values are stored as CSV files. CSV is a simple file format used to store tabular data, such as a spreadsheet or database. Files in the CSV format can be imported to and exported from programs that store data in tables, such as Microsoft Excel or OpenOffice Calc. CSV stands for "comma- separated values". This matrix conversion is done by executing a R code named Matrix.R. Finally we classify the test data by comparing their matrix values with that of the train data set. If a good match is found, the class of the specific character is returned by the code. From this we can obtain the actual mathematical expression written by the user. This works perfectly for simple mathematical expressions. Inorder to recogonise complex expressions we need to proceed with developing a parse tree.

C. *Parse tree generation*

Parse Tree is an ordered, rooted tree that represents the syntactic structure of a string according to some context-free grammar. The parse tree is constructed starting from the root node of the parse forest. Everynode in the parse forest is equipped with a priority queue ordered by parse tree score. Every valid parse tree represents a string generated by the grammar. We use parse trees to provide scoring to the expressions. Based on this score, the corresponding Latex snippet is executed. In order to group the symbols, we generate a parse tree and assign scores to each character. Based on the total score, we group the characters that fall within a particular range. Scoring functions are used to differentiate between the network structures. In this project we use Bayesian Probability to perform the scoring. This is because Bayesian Classification is very efficient in machine learning.

D. *Latex generation*

LaTeX is a document preparation system for high-quality typesetting. It is most often used for medium-to-large technical or scientific documents but it can be used for almost any form of publishing. LaTeX is not a word processor. Instead, LaTeX encourages authors not to worry too much about the appearance of their documents but to concentrate on getting the right content. LaTeX is based on the idea that it is better to leave document design to document designers, and to let authors get on with writing documents. Based on the value generated by the parse tree, the corresponding format of LaTex is executed. Thus the final result is viewed in digital format with the help of LaTex.

IV. Results

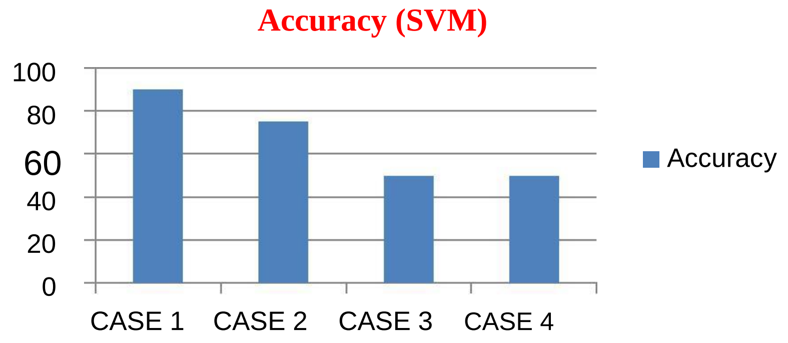
Classification using SVM

Support Vector Machines is a set of supervised learning methods which can be used for both classification and regression. Given a set of training samples, SVM classification training algorithm tries to build a decision model and able to predict which sample belongs to which category.

Case	Train Dataset	Test Dataset	No of classes in Test	Accuracy
i)	32	10	5	90%
ii)	32	10	7	75%
iii)	32	2	2	50%
iv)	32	2	1	50%

The SVM algorithm results are understood by considering 9 classes with train dataset consists of 32 rows and that of test dataset consists of 10 rows with varying number of classes which is depicted in the table above. It is evident from the obtained results that, for the same train and test data, accuracy increases significantly with decrease in the number of classes.

The recognition system has been implemented using R programming language using “e1071” package for SVM classification. These handwritten images are used as a train dataset. As the test dataset consists of the segmented images that are obtained from the math expression. The accuracy obtained for four different cases are illustrated in the below figure.



It is observed that maximum accuracy is obtained when the test data set is increased and the no.of classes for the same data set is maintained at a smaller value. Also note that when the test data set is reduced drastically, classification becomes difficult and the rate of accuracy decreases steadily.

V. CONCLUSION AND FUTURE WORK

In this paper, we have handled a new methodology to convert handwritten mathematical expressions into digital format so that they can be used for various purposes. We mainly focus on the application of this technique in document preparations and paper creations. SVM multiclass is efficient in machine learning and it also possesses ease of operation. Though immense work has been done in the field of pattern recognition, 100% accuracy has not been obtained yet. Thus giving way for future works and developments. This mechanism can be extended to recognise handwritten texts from pictures, textbooks etc. Further it can be used to create various customised font faces pertaining to an individual’s handwriting.

REFERENCES

1. Namarta Dave, Image Processing and Pattern Recognition, International Journal of Signal Processing, Vol. 8, No. 4, pp.155-164, 2015.
2. Monica Patel, Shital P. Thakkar, Handwritten Character Recognition in English: A Survey, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 2, February 2015.
3. Sachin Naik, Pravin Metkewar, Recognizing offline handwritten mathematical expressions based on a predictive approach of segmentation using K-NN classification, International Journal of Technology, 3: 345-354, 2015.
4. Sagar Shinde, Rajendra Waghulade.Handwritten, Mathematical Expessions Recognition using Back Propagation Artifiical Neural Network, Communications on Applied Electronics (CAE) – ISSN : 2394-4714 Foundation of Computer Science FCS, New York, USA Volume 4– No.7, March 2016.
5. R. Zanibbi, D. Blostein, and J. R. Cordy, Recognizing mathematical ex-pressions using tree transformation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(11):1455–1467, Nov. 2016.

BIOGRAPHY

Abirami M is an Assistant Professor in Computer Science And Engineering Department, St Joseph’s Institute of Technology.

Rajashri S is an Under Graduate Student, pursuing her B.E. Degree in Computer Science and Engineering at St Joseph’s Institute of Technology. Her Research interests are Big Data, Machine Learning etc.

Ramya R is an Under Graduate Student, pursuing her B.E. Degree in Computer Science and Engineering at St Joseph’s Institute of Technology. Her Research interests are Data Mining, Image Processing Algorithms etc.