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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

Offline Handwritten Mathematical Expression Recognition

R.Padmapriya[1], S. Karpagavalli[2]

Research Scholar, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India[1]

Associate Professor, Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India[2]

ABSTRACT: Recognition of handwritten mathematical expressions is helpful in writing technical documents as well as useful in converting handwritten documents with mathematical equations into electronic format. Symbol recognition in mathematical expressions is a complex task due to large character set and writer variability in size and style of symbols. In this work, mathematical expression recognition task carried out in different phases which include data collection, preprocessing, segmentation, feature extraction, symbol classification as well as mathematical expression. A set of 50 simple algebraic expressions written by 10 writers, each equation with 10 to 15 symbols converting 23 unique symbols are collected. The expressions are scanned and converted into image files. The images are preprocessed to remove noises, normalize the size and enhance. The symbols in each equation is segmented and features like, zonal, structural, skeleton based, directional are extracted. Multilayer Perceptron (MLP) and Support Vector Machine (SVM) classifiers are used to classify the symbols. The accuracy of symbol classification and whole algebraic expression recognition is analyzed. An interface to automatic mathematical expression recognition is developed with effective classifier.

KEYWORDS: Handwriting Mathematical expression; feature extraction; connected component analysis; classification.

I. Introduction

Most of the scientific and engineering technical documents consist of mathematical expressions, which demand the mechanism to convert handwritten mathematical expressions into electronic format, for speedy processing of articles. The recognition of handwritten mathematical expression can be on-line/off-line. In off-line recognition handwritten/printed expressions are given in the form of images i.e. static representation of the data. In on-line recognition from tablets/computers with pen devices recording and storing of data by digital ink i.e. a dynamic representation. In both cases preprocessing of data, segmentation of symbols/characters and symbol/character recognition tasks are involved using statistical, structural, skeleton based features [1].

Handwritten mathematical expression recognition has various applications like scientific documents digitalization information retrieval or accessibility for blind people. Automatic processing of bulk amount of handwritten data like bank cheques, application forms, credit card imprint and postal code reading are applications where efficient symbol classification is mostly useful.

II. Related work

Handwritten mathematical expression has been an active research field in the last years. Several approaches have been made by researchers to solve the problem of mathematical expression recognition. Some are

Nicolas D. Jimenez and Lan Nguyen have used effectively pyramids of oriented gradients (PHOG) features and one against one support vector machine to recognize handwritten mathematical symbols. They have used CHROME dataset, that contains 22000 character samples and the experiment is limited to 59 symbols. They demonstrated that SVM classifier gracefully generalize symbols with recognition rate 92% and new 75 handwritten symbols written by new writers also tested [2] .

DOI: 10.15680/IJIRCCE.2016. 0401011

B. Keshari, S. Watt has proposed hybrid mathematical symbol recognition system using both online and offline information. In offline they used intensity of the pixel at points and total number of pixels in the image as feature. A total of 137 unique symbols are classified using SVM write-up dependant and write-up independent tests were carried out and achieved high accuracy [3].

Fotini Simistira et. al carried out structural analysis of on-line handwritten mathematical symbols using SVM. They used CHROME 2012 dataset, collected 825 mathematical expressions. The occurrences of all candidate spatial relations are analyzed. The dataset consists of 1906 spatial relations between pair of symbols. The performance evaluation is made based on five aspects: stroke level classification rate, symbol segmentation rate, symbol recognition rate, MathML structure recognition rate, and expression level recognition rate. One-against-all, one-against-one SVM they used as classifier and overall mean error rate of 6.57% and 2.61% achieved [4].

Surendara P. Ramteke, Dhanashri V.Patil, Nilima P. Patil has proposed neural network approach to offline handwritten expression recognition. They used simple mathematical expression recognition task using MLP with back propagation to classify symbols and achieved 90% accuracy [5].

In the proposed work offline handwritten simple algebraic handwritten expression recognition task carried out using classifiers MLP and SVM.

III. HANDWRITTEN ARCHITECTURE

Recognition of mathematical symbols is an important pattern recognition problem. It typically consists of 4 major stages: preprocessing, segmentation feature extraction, symbol classification. The architecture of mathematical expression recognition is shown in figure 1.

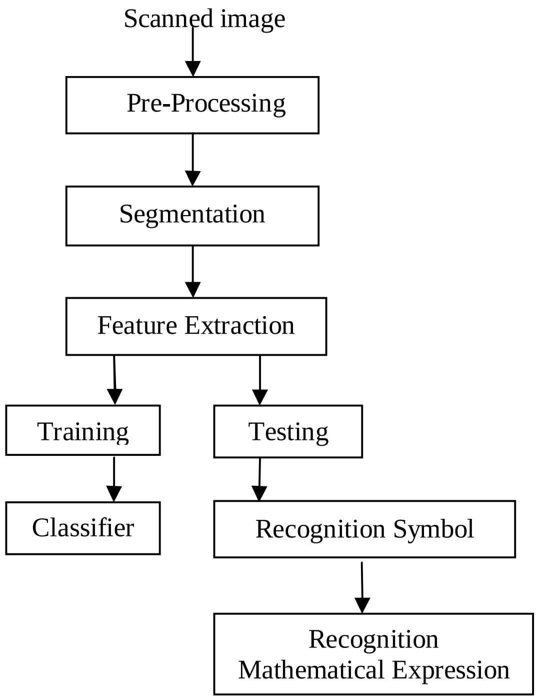

Fig 1: Architecture of Handwritten Mathematical Expression Recognition

DOI: 10.15680/IJIRCCE.2016. 0401011

IV. PREPROCESSING

Image pre-processing can significantly increase the reliability of an optical inspection. It commonly involves disposing of low-frequency , back-ground noises, normalizing the deepness of the individual images, repairing reflections and protecting images. Image pre-processing is the technique of enhancing data images prior to computational processing. In this work pre-processing of hand-written mathematical equation involves four operations that are:

A. Binarization

Image binarization is commonly done in the pre-processing part of the many different image processing related applications for example optical character recognition (OCR) and document image retrieval. It converts a gray level letter image into a binary letter image.

B. Noise reduction

The noise reduction smears an image, and the fine details of the image, if they were obscured by the noise, become even less visible after the noise reduction. The smearing is an unavoidable consequence of noise reduction, if no additional information about the noise or the image structure is available.

C. Size-Normalization

Normalization is a process that changes the range of pixel intensity values. Applications include images with poor contrast due to glare, for example. Normalization is sometimes called contrast stretching or histogram stretching. In more general fields of data processing, such as digital signal processing, it is referred to as dynamic range expansion.

D. Skew detection and correction

The skew of the scanned document image specifies the deviation of its text lines from the horizontal or vertical axis. The skew of the document image can be a global (all document's blocks have the same orientation), multiple (document's blocks have a different orientation) or non uniform (multiple orientation in a text line).

Correction in the skewed scanned document image is very important, because it has a direct effect on the reliability and efficiency of the segmentation and feature extraction stages.

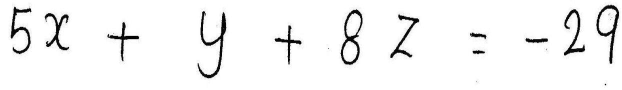

Fig. 2 (i). Scanned Handwritten Expression

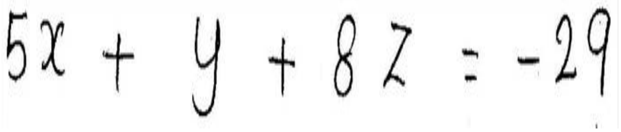

Fig. 2(ii) After Binarization

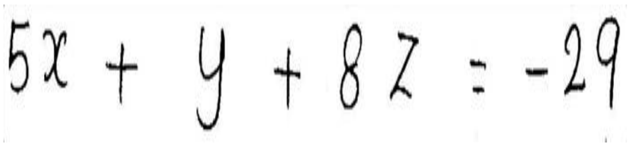

Fig. 3 Preprocessed Handwritten Expression

V. SEGMENTATION

Segmentation is the basic step in symbol recognition which is to segment the input image into individual symbols as there are many segmentation methods are available in this work connected components analysis technique is adapted [6].

Let I denote the input binary image. A connected component analysis algorithm is applied to the foreground region of 1 to produce the set of connected components. Then, for each connected components, its associated bounding box. The smallest rectangle box which circumscribes the component-is calculated, a bounding box can be represented by giving the coordinates of the upper left and lower right corners of the box.

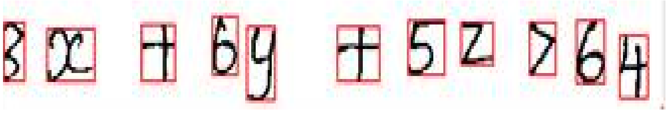

Fig. 4. Symbol Segmented Handwrittten Expression

VI. FEATURE EXTRACTION

The main aim of feature extraction is to extract appropriate characteristics which enable the classifiers in distinct classification of each symbol. The extraction of the features of the characters is done such a way that the complete portion of binary image covered and there is a distinct property associated with each position. It is one of the most important parts of any system using pattern recognition.

A. Zoning

The image is split into window frames of similar dimension and feature extraction is used to each individual part as compared to the whole image. In the task image was partitioned into nine same sized windows. Suggestions, Intersections and Key Beginners: To reduce various line segments in a specific region, the total skeleton in such a zone require traversed. Hence, chosen pixels in the image skeleton are considered for beginners.

B. Skeletonisation

The Skeletonisation method was connected with binary pixel image. The extra pixels which are unable to work for the strength of the symbol were cancelled and the wide-ranging strokes were reduced to one pixel fine lines. This translates into uniformity in most the testing and the training data.

C. Directional Features

The primary step needed to improve any character's boundary line from identity of separate stroke and also line segments in the image. Then , simply to present a normalized input vector to the neural network identification ways, the new innovative symbol representation was split into a number of windows of equivalent size (zoning) where in the number, size and different types of lines available in a single window considered

D. *Determining Directions*:

The boundary parts that could be decided in each individual symbol image were classified into 4 types:

- Vertical type,
- Horizontal type,
- Right diagonal and
- Left diagonal.

From these specific of these four area representations, besides it built intersection elements between any kinds of area. To help the extraction of direction features, the following methods needed to create the symbol model.

- Starting level with intersection point of view location
- Distinguish separate dotted line segments
- Labelling area segment information
- Line kind of normalization

Following the methods highlighted above, separate strokes in the symbol images are identified by one of these four numerical directions idea.

VII. CLASSIFICATION

Two aspects machine learning algorithms, Multilayer perceptron, and then Support Vector Machine classifier were useful for developing the classification model.

A. Multilayer perceptron

Multilayer Perceptron (MLP) network is one of widely used neural network classifier. MLP coverages are typically in practiced key purpose, multiple functional, nonlinear varieties which includes several units focused into several layers. The effort of the MLP signal can be re-structured by various the amounts of ranges and massive sets in each layer. Due to the fact adequate hidden units and adequate record, it happens to be frequently introduced that MLPs can certainly count on a overall performance to most exact accurateness. In the main, MLPs are large approximates. MLPs are completely reliable attributes of difficulties while some-one provides little if any understanding of the type of the relationship between input vectors and their related outputs.

B. Support Vector Machine

Support Vector Machines (SVM) is a pair of supervised learning methods which can be used for both classification and regression. Given a pair of training samples , each considerable as belonging to a couple of categories , an SVM classification training algorithm methods to build a decision model able of predicting whether a new sample falls into 1 category or the other . If the patterns are represented as points in space, a SVM model can be interpreted as a division of this space so that the examples belonging to separate categories are divided by a clear gap that is as wide as possible.

VIII. EXPERIMENT AND RESULTS

In this work a set of 50 simple algebraic expressions are prepared and given to 10 people to write the expressions in their own handwriting. Each equation approximately has 10 to 15 symbols covering 23 various symbols as shown in Table I. The expressions are scanned and converted into image files. The images are pre-processed to remove noise, to normalize the size and to enhance. The symbols in each mathematical equation is segmented and features like, zonal, structural, skeleton based, directional are extracted.

Table 1: Symbols in dataset

Digits	0,1,2,3,4,5,6,7,8,9
Alphabets	x,y,z,a,b,c
Operators	+,-,*,/,>,<,=

Totally 515 features for each symbol is extracted and Multilayer Perceptron (MLP) and Support Vector Machine (SVM) classifiers are trained for mathematical symbol classification. The models are tested with test data to measure their accuracy in symbol classification as well as equation recognition.

A single hidden layer multilayer perceptron neural network was designed to classify the symbols. An input layer with 515 nodes, hidden layer with 650 nodes and output layer with 23 nodes is designed. The network trained with gradient-descent back-propagation algorithm with the learning rate 0 .1 with sigmoid activation function. To train the network 80% of the dataset utilized. To test the performance of the system 20% of dataset utilized.

An experiment with support vector machine is carried out to classify 23 mathematical symbols. Linear kernel support vector machine is implemented in Matlab and trained with 80% of dataset. The performance of the mathematical symbol classifier is tested with 20% test dataset.

Comparative results of two classifiers, Multilayer Perceptron and Support Vector Machine are summarized in Table II and pictorial representation shown in figure 5.

Table II: Performance of the MLP and SVM

Classifier	Learning Time (seconds)	Prediction Accuracy (%)
MLP	0.30	92
SVM	0.02	85

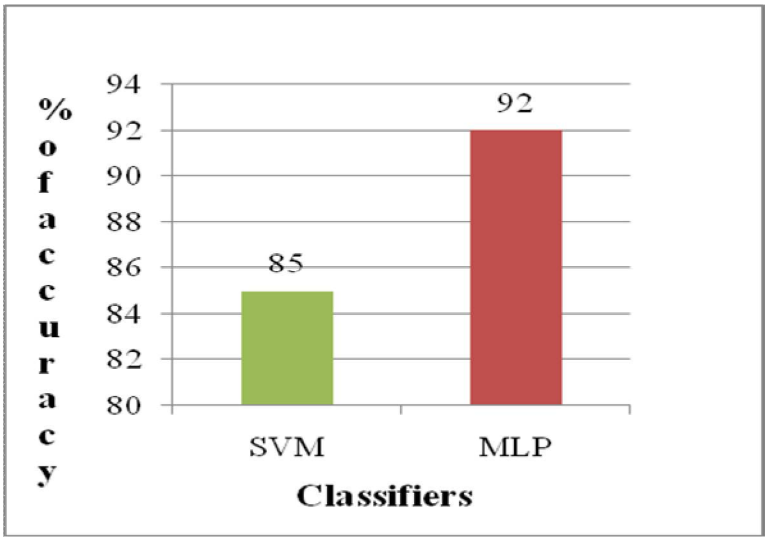

Fig.5 Classification accuracy of the models

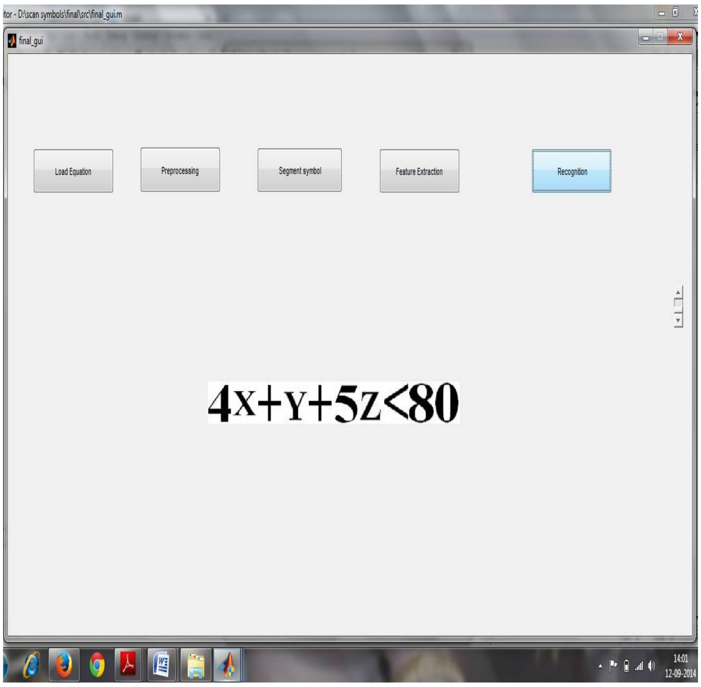

Fig. 6 Recognition of Hand-written equation

After assessing the performance of the classifier for symbol classification, it is observed that Multilayer Perceptron outperforms Support Vector Machine. An interface has been designed to recognize the given handwritten mathematical equation using MLP Classifier and shown in figure 6. Partial results of equation recognition are presented in table III.

Table III: Partial Results of Equation Recognition

Expression	Total No. of symbols	Correctly recognized symbols	Recognition rate (%)
x+y+z=10	*8*	*8*	*100*
2x-y+z=60	*10*	*9*	*88*
x-2y+2z=7	*9*	*8*	*88*
4x+3y+5z=80	*11*	*10*	*90*
x+8y-2z=18	*10*	*9*	*90*
6x+4y+7z=14	*11*	*11*	*100*
7x+2y+9z=20	*11*	*11*	*100*
8x+6y+5z=64	*11*	*11*	*100*
x+3y+2z=4	*9*	*9*	*100*

From the whole equation recognition results, it is observed that the MLP classifier recognizes the given algebraic equation approximately 92% accurately.

IX. CONCLUSION

Handwritten mathematical expression recognition of simple algebraic expression process performed successfully. The work included several phases which include data collection, preprocessing, segmentation, feature extraction, classification and symbol recognition. The accuracy of symbol classification and whole algebraic expression recognition is tested. An user interface to automatic mathematical expression recognition is developed with effective classifier. Hence, it is observed that MLP classifier produces higher accuracy. In future, the work can be extended to recognize on-line mathematical expression recognition. The work can be carried out with local, global features, ink related features, geometrical features. Contextual information can be used in recognition to overcome the problem of similarities and variability in handwritten characters.

REFERENCES

1. Qi Xiangwei, Xinjiang, “The study of mathematical expression recognition and the embedded system design”, Journal of Software, Vol. 5, No.1, January 2009.
2. Nicolas D. Jimenez and Lan Nguyen “Recognition of Handwritten Mathematical Symbols with PHOG Features”, http://cs229.stanford.edu .
3. B. Keshari, S. Watt, “Hybrid Mathematical Symbol Recognition using Support Vector Machine”, Analysis and Recognition – volume 02 Pages859-863, 2007.
4. Fotini Simistira, Vassilis Papavassiliou, Vassilis Katsouros and George Carayannis, “Structural Analysis of On-line Handwritten Mathematical Symbols based on Support Vector Machines”, DPR, 2013.
5. Surendra P. Ramteke, Dhanushri V. Patil, Nilima P. Patil, “Neural Network Approach to Mathematical Expression Recognition system”, International journal of engineering research and technical ISSN-2278-0181 Vol. Issue 10 December 2012.
6. Firoj Parwej, Ph.D., “A Perceptive method for Arabic Word Segmentation using Bounding Boxes by dilation”, International Journal of Computer Applications(0975-8887) Volume 71, No 1, June 2013.
7. Francisco Alvaro, Joan-Andreu Sanchez, Jose-Miguel Benedi, “Unbiased Evaluation of Handwritten Mathematical Expression Recognition”, International Conference on Frontiers in Handwriting Recognition, 2012.

8. Sanjay S. Gharde, Vidya A Nemade, K. P. Adhiya, " Design and Implementation of Special Symbol Recognition System using Support Vector Machine", 2013.

9. Shailedra Kumar Shrivastava, Sanjay S. Gharde, "Support Vector Machine for Handwritten Devanagari Numeral Recognition", 2010 .

10. Saula, J. and Pietikainen, M., " Adaptive document image binarization", Pattern Recognition, 2000.

11. Bozinovic, R.M., and Srihari, S.N., "Off-Line Cursive Script Word Recognition", 1989.

12. Hsin-Chia Fu, Member and Yeong Yuh Xu Multilinguistic "Handwritten Character Recognition by Bayesian Decision-Based Neural Networks", 1998

13. Ian H. Witten, Eibe "Data Mining-Practical Machine Learning Tools and Techniques", 2nd edition, Elsevier, 2005.

14. Sanjay S. Gharde, Pallavi V. Baviskar, K. P. Adhiya, "Identification of Handwritten Simple Mathematical Equation Based on SVM and Projection Histogram", IJSCE, Volume -3, Issue-2, 2013.

15. Dipak D. Bage, K.p. Adhiya, Sanjay S. Gharde, " A New approach offline handwritten mathematical symbols using character geometry", IJIRSET, vol. 2, Issue 7,2013.