Identification of Handwritten Simple Mathematical Equation Based on SVM and Projection Histogram

Sanjay S. Gharde, Pallavi V. Baviskar, K. P. Adhiya

Abstract—***Recognition of simple mathematical equation can applied on off-line handwritten samples. For smooth implementation database is prepaid with total 237 symbols which are collected from 28 different simple mathematical equations. The dataset 1and dataset 2 are training using most popular classifier named Support Vector Machine. In particular, this work tries to spotlight on evaluation of various methods used for feature extraction and recognition system. Moreover, some essential issues in simple mathematical handwritten equation recognition will be addressed in deepness.***

This paper discusses various steps of recognition process for simple mathematical handwritten equations. In that, pre-processing, segmentation, feature extraction, classification and recognition for handwritten mathematical symbol as well as for simple expression is described. Among the different phases applied in recognition system, features extraction and classification method may influence the overall accuracy and recognition rate of the system. Therefore, various techniques applied in this context are studied and comparative analysis is prepared. This evaluation study suggests projection histogram most suitable feature extraction technique and support vector machine is appropriate classification technique for implementation. Using projection profile and support vector machine two different dataset are recognized then 97.58% and 98.40% (as an average it resulted into 98.26%) recognition rate is achieved for simple handwritten mathematical equation.

Index Terms— Classification, mathematical expression, projection histogram, support vector machine.

I. INTRODUCTION

Character Recognition(CR) term is extended day by day, it's not only limited for recognition of English text, numbers and digits or any special language like Arabic, Tamil, Devanagari, Gurumukhi, Farsi, Chinese but also so much work is carried out for some special symbols belongs to mathematical branch. Mathematical equation recognition (ME) or mathematical symbol recognition (MS) is very challenging and interesting in optical character recognition. This work is in the direction of simple mathematical equation recognition which is simply combination of operators and operand. Symbols or expressions of a mathematical expression are arranged from left to right. In mathematical equation basic mathematical symbols are included special symbols, Latin/Arabic/Greek letters, operators and English letters, digits [1].

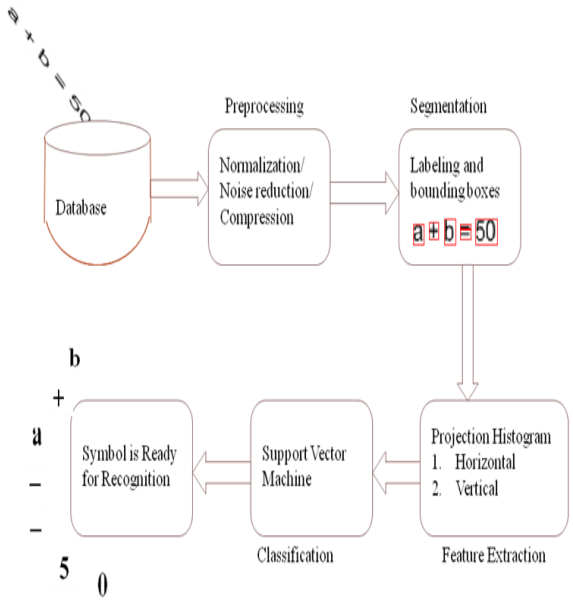

Fig. 1. Architecture of simple mathematical equation recognition.

"Fig. 1" indicates, simple mathematical equation recognition process in that, first handwritten equation is given to recognition system. Here the input samples are collected as database. Next step unwanted data is removed from image like dots, loops, curves, lines by using noise remove algorithm. Skew correction and Binarization are explained methods used for pre-processing which is apply on image to clean the image. Then apply segmentation for separating each symbol from simple equation using bounding box and labeling algorithm. It is useful to apply feature extraction method on individual symbols. Then the features are extracted from segmented symbols applying horizontal and vertical projection histogram. Ten features are extracted from each symbol. Then apply classification technique useful to make most accurate decision for obtained feature vector. In recognition step training and testing are apply on samples to separate it into error samples and accurate samples.

II. RELATED WORK

The related work found that a lot of work has been done in the recognition of isolated handwritten characters and numerals in different languages by different researchers but the work done in the field of handwritten Mathematical equation is increased day by day.

Manuscript received on May, 2013.

Sanjay S. Gharde, is currently working as assistant professor in Computer Engineering Department, North Maharashtra University/ SSBT College of Engineering and Technology, Jalgaon India.

Pallavi V. Baviskar is currently pursuing master's degree in computer science engieeringin from North Maharashtra University, Jalgaon, India.

K. P Adhiya is currently Associate Professor in Computer Engineering Department of SSBT College of Engineering and Technology,Jalgaon, India.

In 2004 U. Pal, B.B. Chaudhari [7] have provided a survey on all feature extraction techniques as well as training, classification and matching techniques used for recognition of machine printed and handwritten **Devanagari** characters and numerals state of the art from 1970s.

In 2007 M. Hanmandlu et.al [8] using Input Fuzzy Modeling for the Recognition of Handwritten **Hindi Numerals**. This paper presents the recognition of Handwritten Hindi Numerals based on the modified exponential membership function fixed to thefuzzy sets derived from normalized distance features obtain using the Box approach method . They have achieved 95% recognition rate.

In 2008 S.V. Rajashekararadhya and Dr P. Vanaja Ranjan [9] have used efficient zone based feature extraction algorithm for handwritten numeral recognition of four **South Indian** scripts. They used Nearest Neighbors, Neural Network as a classifier. They have obtained 98.5% of accuracy.

In 2010 Shailendra Kumar and Sanjay Gharde [6] have used Support Vector Machine for Handwritten **Devanagari** Numeral Recognition. Moment Invariant and Affine moment Invariant techniques are used as feature extraction. This linear SVM produces 99.48% overall recognition rate which is the highest among all techniques applied on handwritten Devanagari numeral recognition system.

In 2010, Puneet Jhajj and dharamveer Sharma [10] have work on **Gurmukhi** script and used a feature extraction technique of zoning using K-NN and SVM for character recognition. They have obtained maximum accuracy of 73.02% using SVM classifier with polynomial kernel.

In 2011 Mahesh Jangid have proposed these method for feature extraction: Zonal density, Projection histogram, Distance Profiles, Background Directional Distribution (BDD) available [11], and SVM for classification and they have got 98%, 99.1% and 99.2% of accuracy. They used **Persian** handwritten digits for experiment.

In 2011 R.Jayadevan, Satish R. Kolhe, Pradeep M. Patil and Umapada Pal have presented the literature survey [12] on different Indian language scripts is presented. It is found that a lot of work has been done in recognition of **Devanagari** and **Bangla** script characters, the two most popular languages in India.

In 2012, Anoop Rekha [13] has presented a complete survey on different feature sets and classifiers used in offline handwritten **Gurmukhi character** and numeral recognition. In a box approach is proposed for extracting the features of handwritten **Persian** digits to achieve higher recognition accuracy and decreasing the recognition time of Persian numerals. In classification phase, support vector machine with linear kernel has been employed as the classifier. By reviewing all researchers work, it is summarized that good results are collectively depend upon database, preprocessing, segmentation, feature extraction and classification.

III. PROPOSED SYSTEM

"Fig. 2" depicts overview of recognition process for simple mathematical equation. Original equation is slanted means at the time of writing image is tilt. After read the equation Binarization, noise reduction and skew correction is performed in preprocessing phase. Now system gets neat and clean equation for segmentation. Each character is separated and labeled by using labeling method. After-wards features are extracted from current image and train the samples. In recognition phase test the train samples with database and match it.

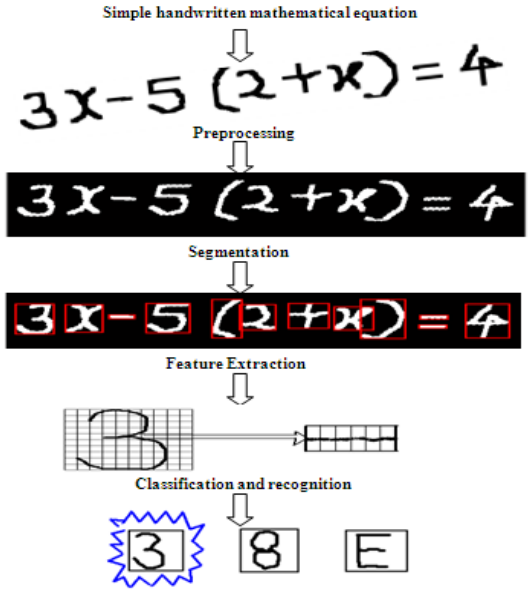

Fig. 2. Overview of equation recognition process.

A. *Preprocessing*

The raw data that is off line simple mathematical handwritten equation is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. Preprocessing is perform basic task like smoothing, enhancing, Filtering, cleaning-up of image then a digital image is provide to next steps to made classification simple and more accurate.

B. *Segmentation*

Segmentation is a process of separating a document image into homogenous units of images that contain pixel groups. The distribution of bounding boxes tells a great transaction about the correct segmentation of an image consisting of non-cursive characters. Labeling of the connected component is one of the fundamental operations in many intelligent vision systems. This operation works on binary images by allocating an individual values to pixels that belong to the same connected area. The original image is necessary to convert into the gray-scale image in case of color image. Then the segmented image can be achieved by means of automatic global thresholding and binarizing [13].

C. *Feature extraction*

The feature extraction stage is used to extract the most appropriate information from the text image which helps us to recognize the characters in the text. The selection of a constant and delegate set of features is the heart of pattern recognition system [14].

1) *Projection histogram*

Projection histograms calculate the number of pixels having value "1" in different direction. Basically there are three types of projection histograms.

a) Horizontal

b) Vertical

c) Left diagonal and right diagonal.

D. *Classification*

Classification is the main judgment making stage of OCR system. It uses the features extracted in the prior stage to recognize the text segment according to predetermined rules.

*1) **Support vector machine***

Support Vector Machines (SVM) is a set of supervised learning methods which can be used for both classification and regression. Given a set of training samples, each noticeable as belonging to one of two categories, an SVM classification training algorithm tries to build a decision model able of predicting whether a new sample falls into one category or the other. If the examples 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. Next samples are then predicted to belong to a grouping based on which side of the gap they fall on.

A. The optimal separating hyperplane

Consider the problem of separating the set of training vectors belonging to two separate classes with hyperplane,

$$D= \{(x^{1,} y^{1}),.., (x^{1,} y^{1})\}, x \varepsilon R^{n}, y \varepsilon \{ -1, 1\} \quad (1)$$

There are two approaches to generalizing the problem, which are dependent upon prior knowledge of the problem and an estimate of the noise on the data. In the case where it is expected (or possibly even known) that a hyperplane can correctly separate the data, a method of introducing an additional cost function associated with misclassification is appropriate.

IV. COMPARATIVE STUDY

A. Review Study on Recognition and Classification

Table 1. Depicts various approaches and methods used for mathematical symbol and expression recognition. These researchers mentioned segmentation rate, recognition rate, error rate of training samples and testing samples.

TABLE 1. COMPARATIVE STUDY FOR VARIOUS CLASSIFICATION METHODS [1]

Methods	No of Symbols	Recog. Rate	Others
Multi-layer perceptions neural network (MLP)	839 symbols: including digits, Roman letters, Greek letters, binary operators.	87.5%	Segmentation Rate 94.8% Exp.rate 29.2%
KNN, Euclidean dist. SVM- BEST WNN, HMM Gaussian distributions	2233 symbols: InftyCDB-1 database 25% for test and 75%for training	98.5% avg.	39% of the symbols, obtaining an accuracy
Without using any feature	227 symbols: including digits, Latin letters, Greek letters and mathematical operators	94.8%	-
Convolution neural Network (LeNet-5 CNN topology)	IFH-CDB test database (19840samples, test set with 9840 samples)	90.1% avg.	5.35% error rate
Gabor feature	100 Chinese mathematical literature	97.1%	_
CNN (with one hidden layer Perceptions)	two-digit strings, numbers from 00 to 99	94.6%	Error rate 1.34%
GREC symbol recognition	Collection of 1800 graphic symbols.	92.3% avg.	Query symbol =1800

MLP- Multi Layer Perception, CNN- Convolution Neural Network, HMM- Hidden Markov Model, WNN-Weighted Nearest Neighbor, KNN- K nearest Neighbor.

B. Review Study on Feature Extraction

TABLE 2. COMPARATIVE STUDY FOR VARIOUS FEATURE

Author	Feature Extraction	Classifier	Recog. rate	Details
Kartar, Siddharth Rajneesh Rani	Distance profile	SVM	98%	C= soft margin γ= RBF single kernal parameter 128 samples
Kartar, Siddharth Rajneesh Rani [14]	Projection histogram	SVM	99.2%	C=16, γ = 0.05-0.15 190 samples
Anoop Rekha [13]	Zoning	KNN	72.54%	For single character
	Zoning	SVM	73.02%	Polynomial Kernel
	Structural Features	NN	83.32%	For single character
	Transition	KNN	86.57%	
Anuj Sharma et al.	Strokes recognition &matching	EM	90.08%	
Munish Kumar	Diagonal & transition feature	KNN	94.12%	
Kartar Singh Siddharth et al	ZD,distance profiles, projection histograms ,BDD	KNN, SVM PNN	95.01%	190 samples
Gita Sinha, Anita Rani, Renu Dhir, R. Rani	Zonal based feature	SVM	95.11%	γ = 0.001 C = 4, 8,32, 64 and 500 7000 samples

SVM- Support Vector Machine, KNN- K nearest Neighbor, PNN- Probabilistic Neural Network, BDD- Background Directional Distribution, ZD- Zonal Density, EM- Elastic Matching.

The feature extraction stage is used to extract the most appropriate information from the text image which helps us to recognize the characters in the text. The selection of a constant and delegate set of features is the heart of pattern recognition system [14].

Table 2. Indicates comparative study for various feature extraction methods collected from different research papers published on feature extraction methods in the area of character recognition.

V. DATA COLLECTION

In mathematical equation basic mathematical symbols are included like basic Latin letters, mathematical operators, 0-9 digits, A-Z alphabets, Greek letters, parenthesis, general punctuation, arrows and special symbols. While create simple mathematical equation various fonts, size and styles can be apply on an equation.

For recognition of simple handwritten mathematical equation data is inputted in the form of simple equation which is combination of various operators and operand. Each equation is may be collection of digits, Latin letters, Latin symbols, basic mathematical symbols, brackets. Total 91 symbols are selected from different kind of mathematical symbols to write simple mathematical expression.

VI. RESULT

Table 3 and Table 4 indicates recognition rate for each mathematical symbol in dateset1and dataset2 respectively, in this work dataset1 and dataset2 is written by 3 different people from engineering background. Dataset1 having 12 simple handwritten mathematical equation, total 124 number of mathematical symbols, 121 recognized symbols produced 97.58 % recognition rate.

TABLE 3. RECOGNITION RATE OF DATASET 1

s. no	Expression Name	Total no. of Symbols	Recognized Symbols	Recognition Rate
1	Equ15	11	11	100
2	Equ16	11	11	100
3	Equ17	16	16	100
4	Equ18	16	16	100
5	Equ19	5	5	100
6	Equ20	5	5	100
7	Equ21	9	9	100
8	Equ22	9	9	100
9	Equ23	11	11	100
10	Equ24	11	10	90.90
11	Equ25	11	11	100
12	Equ26	11	11	100
13	Equ27	13	13	100
14	Equ28	13	12	92
15	Equ29	18	18	100
16	Equ30	18	17	94
		188	**185**	**98.40**

Dataset2 having 16 simple handwritten mathematical equation, total 187 number of mathematical symbols, 185 recognized symbols produced 98.40 % recognition rate. By combining the result of both dataset, system provides 98.26% Recognition rate obtained from 237 symbols, 28 different handwritten simple mathematical equations.

TABLE 4. RECOGNITION RATE OF DATASET 2

s. no	Expression Name	Total no. of Symbols	Recognized Symbols	Recognition Rate
1	Equ1	11	11	100
2	Equ2	11	10	90.90
3	Equ3	9	9	100
4	Equ4	9	9	100
5	Equ5	13	13	100
6	Equ6	13	12	92.30
7	Equ7	7	7	100
8	Equ8	7	7	100
9	Equ9	6	6	100
10	Equ10	6	6	100
11	Equ11	16	16	100
12	Equ12	16	15	93.75
		124	**121**	**97.58**

VII. CONCLUSION

Recognition of simple mathematical equation incorporates common steps in image processing. But due to complexity in symbols of handwritten equations, improvisation in recognition rate becomes more challenging. For that purpose, this work evaluates the various techniques which may improve the result of recognition. Using various research papers the comparisons of classification and feature extraction methods are studied. Hence it is observed that support vector machine is used as classifier and Projection Histogram method for extracting features from handwritten mathematical equation. The classification and recognition is performing on total 237 symbols which are extracted from 28 different equations. Dataset1 represents 121 symbols are recognized correctly from 124 with 97.58% recognition rate and dataset2 represent 184 re recognized from 188 symbols with recognition rate 98.40%This work achieved average 98.26% recognition rate for simple handwritten mathematical equations.

This work is nicely performed on simple handwritten mathematical equation, it can be proposed for complex equations like equation having capacity to create small letters. This proposed work can be useful to implement memory games for kids to develop an interest in mathematics. Because proposed work is recognized all included mathematical symbols and inputted equations correctly. So in future this implementation is useful to made computation on mathematical equation to find out answer of that equation.

REFERENCES

[1] Sanjay S.Gharde, Baviskar Pallavi V., K.P.Adhiya, "Evaluation of Classification and Feature Extraction Techniques for Simple Mathematical Equation", International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868.

[2] Anoop Rekha," *Offline Handwritten Gurmukhi Character and Numeral Recognition using Different Feature Sets and Classifiers - A Survey*" International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622,www.ijera.com Vol. 2, Issue 3, May-Jun 2012, pp. 187-191.

[3] U. Pal, B.B. Chaudhuri," *Indian Script Character Recognition: A Survey"* Pattern Recognition, Elsevier, pp. 1887-1899, 2004
[4] R.Jayadevan, Satish R. Kolhe, Pradeep M. Patil and Umapada Pal," *Offline Recognition of Devnagri Script: A Survey,* IEEE Transactions On Systems, Man, And Cybernetics-Part C:Applications And Reviews,Vol.41,No. 6, November 2011,
[5] M. Hanmandlu, J. Grover, V. K.. Madasu, S. Vasikarla *" Input Fuzzy Modeling for the Recognition of Handwritten Hindi Numerals "* International Conference on Information Technology (ITNG'07) 0-7695-2776-0/07 ,2007 IEEE.
[6] S.V. Rajashekararadhya, Dr P. Vanaja Ranjan, . 2008 "*efficient zone based feature extraction algorithm for handwritten numeral recognition of four popular south indian"* journal of theoretical and applied information technology
[7] Shailedra Kumar Shrivastava, Sanjay S. Gharde " *Support Vector Machine for Handwritten Devanagari Numeral Recognion "* International Journel of Computer Application (0975-8887) Volume 7-No. 11,October 2010
[8] Dharamveer Sharma, Puneet Jhajj," *Recognition of Isolated Handwritten Characters in Gurmukhi Script"* International Journal of Computer Applications (0975 – 8887) Volume 4– No.8, August 2010.
[9] Omid Rashnodi, Hedieh Sajedi, Mohammad Saniee Abadeh," *Using Box Approach in Persian Handwritten Digits Recognition"* International Journal of Computer Applications (0975 – 8887) Volume 32– No.3, October 2011.
[10] Seyed Mojtaba Mousavi, Seyed Omid Shahdi, and S.A.R. Abu-Bakar, "*Car Plate Segmentation Based on Morphological and Labeling Approach",* Advances in Computing, Control, and Telecommunication Technologies 2011
[11] Widad Jakjoud, Azzeddine Lazrek," *Segmentation method of offline Mathematical symbol",* 978-1-61284-732-0/11 2010 IEEE.
[12] Nafiz Arica, "*An Off-line character recognition system for free style Handwriting*" thesis submitted on Sep 1998
[13] Seyed Mojtaba Mousavi, Seyed Omid Shahdi, and S.A.R. Abu-Bakar, "*Car Plate Segmentation Based on Morphological and Labeling Approach",* Advances in Computing, Control, and Telecommunication Technologies 2011
[14] Kartar Singh Siddharth Renu Dhir Rajneesh Rani, "*Handwritten Gurmukhi Numeral Recognitionusing Different Feature Sets",* International Journal of Computer Applications (0975 – 8887) Volume 28– No.2, August 2011
[15] Ahmad-Montaser Awal, Harold Mouchère, Christian Viard-Gaudin," *Towards Handwritten Mathematical Expression Recognition"* 2009 10th International Conference on Document Analysis and Recognition.