Complex Mathematical Expressions Recognition using Support Vector Machine as a Classifier

Sagar Shinde[1], Mukil Alagirisamy[2], Daulappa Bhalke[3], Lalitkumar Wadhwa[4]

[1] *Post Doctoral Fellow (PDF) in E & C Engineering, Lincoln University College, Malaysia & Assistant Professor, JSPM NTC, Pune, India*

[2] *Faculty of Engineering, Lincoln University College, Malaysia*

[3] *Professor, AISSMS COE, Pune, India*

[4] *Principal, NMIET, Pune, India*

Email: [1]sagar.shinde5736@gmail.com

Abstract: Mathematics is almost unavoidable in daily life as well as in mathematical model with analysis in recent technology. It is today's need to implement math recognition system to participate in digital world. Each and every element in the equations has been segmented using morphological operations and subsequently the features entropy, mean, variance, standard deviation, skewness, kurtosis, correlation, contrast are extracted and used appropriate classifier- support vector machine to improve the accuracy . The more complex handwritten equations have been considered for the recognition. The implemented algorithm can be used in offline recognition of equations, digits and symbols on postal and bank documents, university answer sheets, handwritten notes of mathematics as well as in blind math applications. The efficiency of the system is measured using confusion matrix and ROC (Receiver Operating Characteristics) under AUC (Area under Curve). Basically accuracy is depends on extracted features and classifier used.

***Keywords*: ROC, AUC, Support vector machine, complex equation recognition, features extraction**

1. Introduction

The mathematical equations, notations are well known and used throughout the world. The basic need to implement math recognition system for correct recognition of scan documents such as answer sheets of mathematics, banking and postal documents, handwritten notes of mathematics etc as well as for blind math applications. The automatic recognition of mathematical expressions aims to fill that gap between a person's knowledge and the input accepted by the organizers. The printed documents containing formulas could be digitized automatically, and the writing could be used to directly introduce mathematical notation into electronic devices. Generally, the foremost common manner for the elements identification within the math equation is administered in either on-line mode or offline mode. Just in case of on-line mode, the element is identified as presently because it is entered. Whereas with offline mode, the equations to be identify are written on paper and scanned with facilitate of scanner then the identification of has been carried out. Mathematical equation identification is one of the exploration subjects from a very long while however it is as yet a region of research theme in light of the fact that there are heaps of difficulties in this framework. It is vital with respect to scientific archive expression image investigation and these frameworks have applications like scientific report digitization, data recovery or openness for visually impaired individuals to understand mathematics and other numerical articulations. As far as there is an unavailability of standard database for offline handwritten mathematical equations recognition systems, it is an open door for the researchers to make such database for the reason of experimentation. The literature concluded that recognition of handwritten math equations are still area of research as existing systems has some drawbacks in terms of accuracy, efficiency, throughput, performance and lack in recognition of complex equations The distinctive individuals having diverse strokes, variety in composing style by various individuals, complex structure, contacting characters, overwriting, inherent ambiguities, complex semantics and 2D spatial course of action. The real time database is generated by writing the various mathematical equations and symbols from different people including college students, servicemen, ex-servicemen etc. The various kinds of mathematical equations are considered incorporating simple straight line equations, slope of line, law of indices, standard form of quadratic equations, quadratic formula which is the solution of quadratic equations, convolution sum, convolution integral, law of gravity. The captured input image has been segmented using morphological operations and features viz. entropy, mean, variance, standard deviation, skewness, kurtosis, correlation, contrast are extracted and support vector machine (SVM) has been utilized as a classifier to get maximum efficiency with good recognition rate.

2. Methodology

The proposed Support Vector Machine (SVM) algorithm uses statistical and complex features extracted from the handwritten mathematical equations and symbols along with the depth information. Recognition of equations has to be done with SVM classifier. The ten different classes of

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

equations have been considered with measurable database for each class of equation. The database has been created with help of surrounding people including students, ex-servicemen and many others which includes various writing style- size, shape , stroke while writing etc.

A. Segmentation

Basically, segmentation is the process of dividing a digital image into numerous components. The morphological segmentation with boundary box has been obtained. The segmentation of complex mathematical equations is not an easy task and complete equation is segmented or separated in individual elements and counting total number of elements in the equation. The depth data contains complementary information.

First of all, the preprocessing operation is carried out and then image is converted into binary using below equation.

$$B_i(x,y) = \begin{Bmatrix} 1\text{ , if } P_i(x,y) > 1 \\ 0, \text{if } P_i(x,y), 1 \end{Bmatrix} \quad (1)$$

Where P_i = pre-processed and B_i = Binary Image.

Sum of column and row gives the location of the element or component in the image as the other points are zero except element or component.

B. Feature Extraction

Actually, the recognition rate and efficiency is dependent on the features extracted and classifier used. In proposed system, statistical as well as complex features viz. entropy, mean, variance, standard deviation, skewness, kurtosis, correlation, contrast are extracted and support vector machine (SVM) used as a classifier to get maximum accuracy with higher efficiency.

C. Classifier

The varieties of classifiers are available viz. Support vector machine (SVM), k-nearest neighbor (K-NN), Bayes classifier, artificial neural network etc. The literature background has concluded that SVM classifier is good for the complex equations and it is a supervised machine learning algorithm and discriminative classifier which can be used for both classification and regression challenges. Classifiers compare inserted feature with saved sample or exemplar and perceive the best matching class for input. The best classifier will find the best hyper plane that has a maximum distance from all the classes. It is robust, accurate and effective one even in cases where numbers of training samples are small. It works really well with a clear margin of separation and effective in high dimensional spaces also effective in cases where the number of dimensions is greater than the number of samples. It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient. Due to high classification rate, it is used in pattern recognition applications as to achieve good performance with no prior knowledge of the data.

SVM Provides maximized margin in data. It always provides single solution. It finds a separating hyperplane which separates data with large margin for linearly separable data.

$$x \in R^1 \Phi(x) R^H \quad (2)$$

Where $\Phi(x)$ is kernel function, used to find hyper plane. SVM does not require training data again and again. Data is separated in two groups as +1 or -1 for two dimensions. Also it requires two hyper planes to separate data points in three dimensional data.

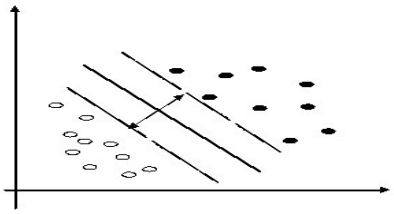

Fig. 1. Hyper plane with SVM

3. Experimental Results And Discussion

The equation for convolution sum and quadratic have gone through segmentation to separate each element or component and then numbering or labeling to each and every element or component has assigned to get the total number of elements or components in the equations to be recognized. The following figure 2 through figure 5 is shows the process of segmentation and labeling. The extracted features with their values from the each and every equation are shown in table I.

Fig.2. Segmentation of convolution summation

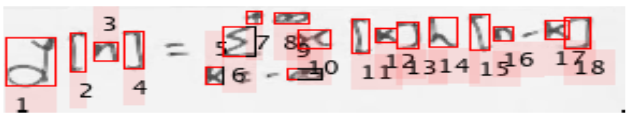

Fig.3. Labeling to convolution summation

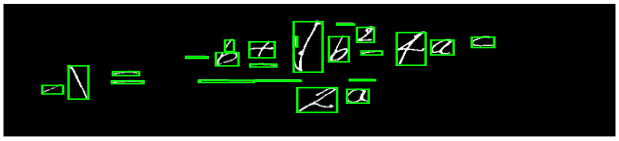

Fig.4. Segmentation of Quadratic Equation

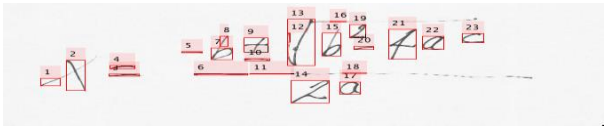

Fig.5. Labeling to Quadratic Equation

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

As shown in figure 2 and figure 4, after performing preprocessing operation the equation for convolution sum and quadratic equation have gone through morphological segmentation process in which each element of equations are separated or segmented with boundary box and finally total number of elements in convolution sum and quadratic equation have been calculated as shown in figure 3 and figure 5 respectively. The calculated values of extracted features viz. statistical as well as complex features from various mathematical equations are shown in below table 1.

Table I. Statistical and Complex features for Various Equations.

Sl. No	Type of equation	Entropy	Mean	Variance	Standard Deviation	Skewness	Kurtosis	Correlation	Contrast
1	Straight line	3.53	0.93	0.00010956	0.05832	-2.9559	14.72	0.2190	693.516
2	Law of Gravity	3.851	0.9320	0.00017898	0.0695	-3.1523	15.13	0.15633	693.2467
3	General Quadratic	12.28	0.928	0.00031137	0.079488	-2.9589	16.2576	0.17073	695.5428
4	Quadratic Equation for roots	9.5483	0.95399	3.9957e-05	0.046919	-5.0122	43.4196	0.038443	686.1132
5	Convolution Sum	5.3152	0.91823	0.00020847	0.071231	.3.0502	14.8096	0.1017	692.375
6	Convolution Integral	4.1991	0.92769	8.29297e-05	0.057685	-2.6293	12.0229	0.16559	688.102
7	Slope of line	3.53	0.93	0.00010956	0.05832	-2.9559	14.72	0.2190	693.516
8	Law of Indices	3.67	0.9362	0.0001831	0.069805	-2.0124	11.9319	0.095476	687.04
9	Area of Circle	3.6827	0.91829	0.00020781	0.072464	-2.4175	11.7392	0.15906	689.9321
10	Pythagorean Theorem	3.5094	0.92099	0.00024178	0.076506	-2.3895	10.0946	0.098802	690.1442

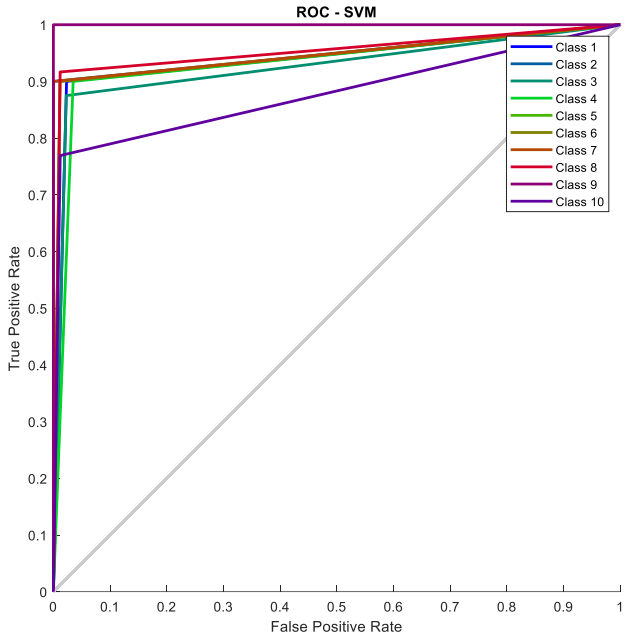

Fig. 5. Receiver Operating Characteristics

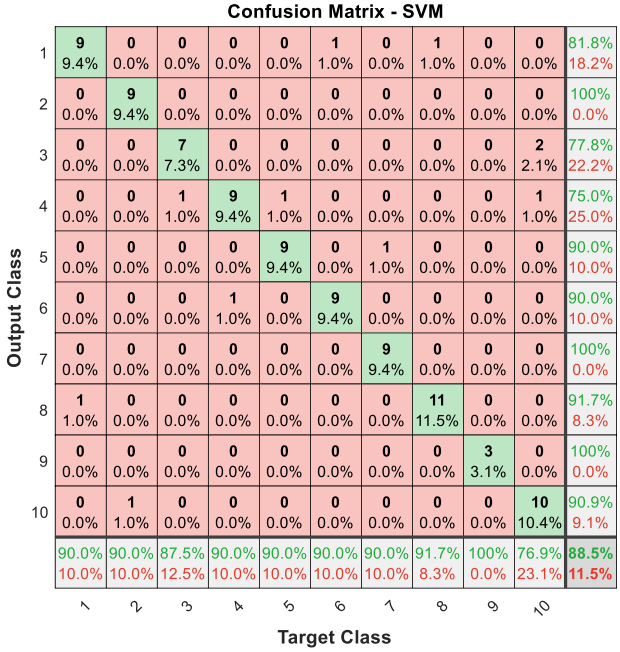

Fig. 6. Confusion Matrix

The figure 5, shows receiver operating characteristics for each class of equations with support vector machine has been explained with true positive rate and false positive rate. Each class of equations showing different true positive rate and false positive rate.

TP =True Positive - It is an outcome where the classifier correctly predicts the positive class.

FP =False Positive – It is an outcome where the classifier incorrectly predicts the positive class.

The recognition rate is basically depending on the classifier used and features has extracted. The figure 6, shows confusion matrix with target class and output class with ten classes of equations. As shown in above figure, it has been concluded that the accuracy of class nine is highest among the other classes of equations while class ten shows lowest accuracy among the other classes of equations. The overall accuracy obtained for proposed approach is 88.5 %. There are

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

10 set of classes and each class having set of equations through available database.

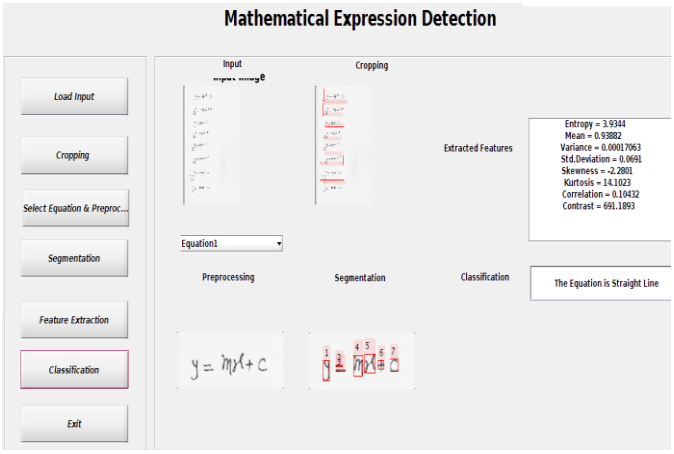

Fig. 7 . Graphical User Interface

The graphical user interface (GUI) is shown above in figure 7, which denotes the complete process from loading of the input image through classification to get the recognized equation. The bank of handwritten equation has acquired as a input and cropped it to identify particular equation as per choice and the selected equation has undergone through preprocessing to get noise free clean and clear image to get it passed through segmentation and finally statistical and complex features are extracted to get classification by using support vector machine. The above figure shows the recognition of straight line equation. Through this GUI, it have been noted that throughput and efficiency of the proposed system is at measurable level.

4. Conclusions

The simple through complex handwritten mathematical equations have been identified or recognized with accuracy through confusion matrix is about 89 percent and the efficiency and execution time is extents at a significant level by extracting statistical and complex features with utilization of support vector machine as a classifier..

References

[1] Ting Zhang, Harold Mouchere, Christian Viardgaudin, "Using BLSTM for Interpretation of 2D Languages–Case of Handwritten Mathematical Expressions"-DOI10.1109/TMM.2018.2844689, IEEE Transactions on Multimedia.

[2] Yassine Chajri and Belaid Bouikhalene, 2016, "Handwritten Mathematical Expressions Recognition"-International Journal of Signal Processing, Image Processing and Pattern Recognition Vol.9, No.5 (2016), pp.69-76

[3] Miroslav Hrdina, 2012, thesis on "Recognition of mathematical texts", Masarykova Univerzita Fakulta Informatiky.

[4] Scott MacLean, 2014, thesis on "Automated recognition of handwritten mathematics", University of Waterloo, Waterloo, Ontario, Canada.

[5] Ignasi Mas Méndez, 2016, thesis on "Mathematical expression recognition", Universitat Politècnica de Catalunya, Barcelona.

[6] Koschinski, M., Winkler, H.-J., and Lang, M. (n.d.), "Segmentation and recognition of symbols within handwritten mathematical expressions", 1995 International Conference on Acoustics, Speech, and Signal Processing, 0-7803-2431-5, doi:10.1109/icassp.1995.479986.

[7] Lapointe, A., and Blostein, D. (2009)' "Issues in Performance Evaluation: A Case Study of Math Recognition", 2009 10th International Conference on Document Analysis and Recognition, 978-1-4244-4500-4 , doi:10.1109/icdar.2009.247

[8] Toumit, J.-Y., Garcia-Salicetti, S., and Emptoz, H. (1999)," A hierarchical and recursive model of mathematical expressions for automatic reading of mathematical documents", Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR '99 (Cat. No.PR00318), 0-7695-0318-7 , doi:10.1109/icdar.1999.791739.

[9] Noboru Takagi, 2010, "A Pattern Classification Method Of Mathematical Figures With Broken Lines, 2010 World Automation Congress, Kobe, Japan, 978-1-889335-42-1, **INSPEC Accession Number:** 11699875.

[10]. Genoe, R., and Kechadi, T. (2010), "A Real-Time Recognition System for Handwritten Mathematics: Structural Development", IEEE 12th International Conference on Frontiers in Handwriting Recognition. 978-1-4244-8353-2, , doi:10.1109/icfhr.2010.97

[11]. Toyozumi, K., Yamada, N., Kitasaka, T., Mori, K., Suenaga, Y., Mase, K., and Takahashi, T, (2004), " A study of symbol segmentation method for handwritten mathematical formula recognition using mathematical structure information", Proceedings of the IEEE 17th International Conference on Pattern Recognition, ICPR 2004, 0-7695-2128-2, doi:10.1109/icpr.2004.1334327

[12]. Toyozumi, K., Suzuki, T., Mori, K., and Suenaga, Y. (n.d.), 2001, " A system for real-time recognition of handwritten mathematical formulas", Proceedings of IEEE Sixth International Conference on Document Analysis and Recognition, 0-7695-1263-1 , doi:10.1109/icdar.2001.953948 .

[13]. Shi Y., LI H. and Soong F. K. , (2007), " A Unified Framework for Symbol Segmentation and Recognition of Handwritten Mathematical Expressions", IEEE Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 0-7695-2822-8, doi:10.1109/icdar.2007.4377036

[14]. Guo Y., Huang L., Liu C.and Jiang X. (2007), "An Automatic Mathematical Expression Understanding System", IEEE Ninth International Conference on Document Analysis and Recognition, (ICDAR 2007), 0-7695-2822-8, doi:10.1109/icdar.2007.4377009

[15]. Li Z. and Tian X. , (2010), " An Improved Analysis Approach of Overbrace/Underbrace Structure in Printed Mathematical Expressions", 2010 IEEE International

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Conference on Innovative Computing and Communication and 2010 Asia-Pacific Conference on Information Technology and Ocean Engineering, 978-1-4244-5635-2, doi:10.1109/cicc-itoe.2010.22

[16]. Alvaro F., Sanchez J.-A., and Benedi J.-M. , (2013), "Classification of On-Line Mathematical Symbols with Hybrid Features and Recurrent Neural Networks", IEEE 12th International Conference on Document Analysis and Recognition , 978-0-7695-4999-6 , doi:10.1109/icdar.2013.203

[17] Smirnova E. and Watt S. M., (2008), "Communicating Mathematics via Pen-Based Interfaces", 2008 , IEEE 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, 978-0-7695-3523-4 , doi:10.1109/synasc.2008.94

[18] Nomura A., Michishita K., Uchida S., and Suzuki M. (n.d.), 2003, "Detection and segmentation of touching characters in mathematical expressions", IEEE Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings 0-7695-1960-1 , doi:10.1109/icdar.2003.1227645 .

[19] Celik M. and Yanikoglu B, (2011), " Handwritten mathematical formula recognition using a statistical approach", 2011, IEEE 19th Signal Processing and Communications Applications Conference (SIU), 978-1-4577-0463-5, doi:10.1109/siu.2011.5929696.

[20] Chen Y., Shimizu T., and Okada M. (n.d.), 1999, " Fundamental study on structural understanding of mathematical expressions", IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.99CH37028), 0-7803-5731-0 , doi:10.1109/icsmc.1999.825383

[21] Xinyan C., Hongli Y. and Xin, W. (2013), " Handwritten Mathematical Symbol Recognition Based on Niche Genetic Algorithm", 2013, Third International Conference on Intelligent System Design and Engineering Applications, 978-1-4673-4893-5, doi:10.1109/isdea.2012.191

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731