SVM Classifier for Offline Handwritten and Printed Mathematical Expression Recognition

Manisha Bharambe[1(✉)], Kavita Kobragade[2], and Poonam Ponde[3]

[1] MES Abasaheb Garware College, Karve Road, Pune, India
mgb.agc@mespune.in
[2] Fergusson College, F.C. Road, Pune, India
[3] N. Wadia College, Pune, India

Abstract. The recognition of Mathematical Expressions (ME) constitutes a challenging problem in character recognition research. A very few studies of offline Mathematical expressions have been so far reported in the literature. This paper focuses on offline handwritten and printed mathematical logical expressions recognition using Support Vector Machine classifier (SVM). In the work of expression recognition, the expressions were segmented into individual characters. The feature extraction method with combination of Normalized chain code and zone based density was used to get the features of a character. The present work considers logical expressions with subscripts for recognition. The experimental results for recognition rates of handwritten and printed expressions are reported. The result shows that the recognition rate of handwritten expression is 84.1% and that for printed expression is 90.3%.

Keywords: Offline handwritten and printed · Mathematical expression recognition · logical expressions · SVM · recognition rate

1 Introduction

Research into recognizing handwritten Mathematical Expressions (ME) began in the late 1960's. The rate of progress was limited in this problem, with absence of benchmark datasets and standard evaluation tools. The input of ME into computers is more difficult than plain text. MEs are more complex in structures. The complexity occurs in the expression because math symbols can have different fonts such as regular roman, italic, bold and calligraphic, and can have different sizes. Hence, recognition problem becomes more challenging. In the sudden outbreak of COVID-19 pandemic, the application of handwritten interaction in education, remote work, and distance learning is the requirement [19].

S. Tamane et al. (Eds.): ICAMIDA 2022, ACSR 105, pp. 953–965, 2023.
https://doi.org/10.2991/978-94-6463-136-4_84

In the offline printed mathematical expression recognition, some mathematical symbols might be incorrectly recognized or missed, which leads to the insufficient context information. So it is necessary to make a deeper study of the handwritten and printed mathematical expression recognition. Surendra Ramteke et al. [7] had proposed offline handwritten ME recognition system using ANN classifier. The feature set was obtained by using area and centroid feature of bounding box. The recognition result was 90% of the simple mathematical equation. Sanjay S. Gharde et al. [10] have discussed the various steps of recognition process for simple off-line mathematical equations. The feature extraction methods such as zoning, skeleton based direction, projection histogram, boundary values from four directions (top, bottom, left and right) of symbols, and structural features like crossing points, end points and loops were used. ANN and SVM classifiers were used for recognition. Kazuki Ashida et al. [16] developed the performance evaluation of the mathematical formulae recognition system applied to printed formula images. Dong-Yu Zhang et al. [11] had proposed the method for segmentation of touching symbols in printed Mathematical Expressions. Francisco Álvaro and Richard Zanibbi [14] have proposed problem of classifying special relationship between symbols and expressions in online handwritten ME'S. Hans-Jurgen Winkler et al. [15] have proposed research on online segmentation and recognition of mathematical expressions. Dipak Bage et al. [8] had proposed the offline handwritten mathematical symbol recognition system, and adopted the number of feature extraction techniques to recognize the mathematical symbols.

From the literature it was observed that only few studies of ME recognition were reported. Due to large database of math symbols, the researchers used the subset of math symbols in their works. It was observed that some researchers worked on expression recognition only on simple expressions [7, 10] while others have not specified the class of expression under consideration. The segmentation methods used for MEs were not clearly described in the literatures. The expression recognition rate computed by Sanjay S. Gharde et al., [10] and Surendra Ramteke et al [7] was percentage of the number of symbols recognized from the expression.

2 Stages in the Recognition of ME

The Fig. 1 shows the phases in recognition of ME. The scanned expression was preprocessed by using binarization, filtering, and morphological operations. The preprocessed expression was segmented into isolated characters. The feature set was obtained for each segmented character using statistical, topological and moment features. There are various classification methods such as Artificial Neural Network, SVM, Naive Bayes, decision tree. The SVM classifier is commonly used for the classification problems such as handwriting recognition, face detection, email classification, gene classification, and in web pages. The paper uses SVM classifier for ME recognition.

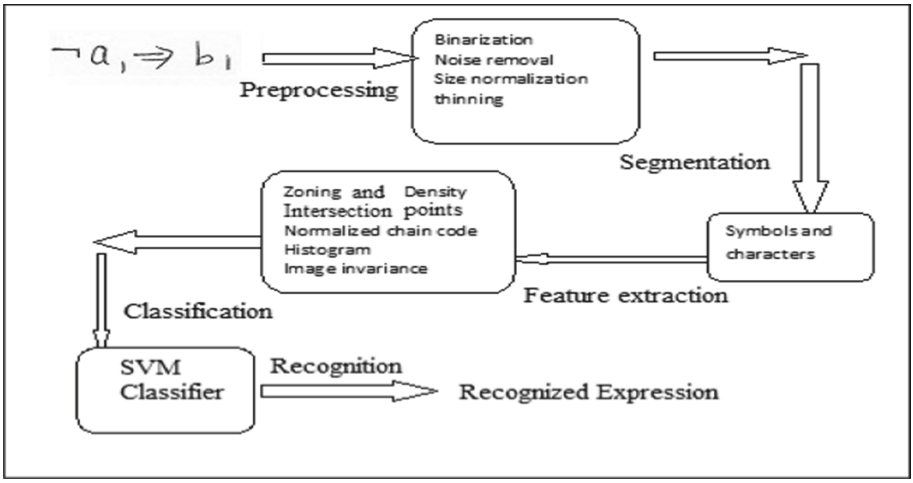


Fig. 1. Phases of Mathenmatical Expression (ME) recognition system

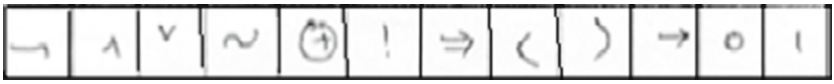


Fig. 2. Dataset of 12 characters (10 logical symbols and 2 digits)

The recognition was performed on 48 handwritten expressions. Each handwritten expression was written by 20 different writers, resulting in 960 expressions. The total number of characters of handwritten expressions after segmentation was 8640 (80% used for training and 20% used for testing). An experiment was carried out on 35 printed expressions. The number of occurrences of each expression varies from 5 to 15 times with different fonts, results in total 288 printed expressions. A number of 288 printed expressions were segmented into 3559 images (80% used for training and 20% used for testing). The dataset of 26 alphabets (a-z), 10 logical symbols (Fig. 2) and numerals 0 and 1 were used in the expressions. Hence there were 38 class labels.

The handwritten or printed training expressions were preprocessed by using filtering, binarization morphological operations and cropping. The preprocessed expression was segmented which results in the isolated characters. Segmentation is a difficult task in the handwritten ME due to the following reasons:

- ME structure is a 2-Dimentional structure, containing overlapping characters, i.e. the neighboring characters are written such that they share the area of others in a symbol block.
- Touching characters reduces the rate of segmentation.
- The characters in the ME have different spatial relationship like subscripts.

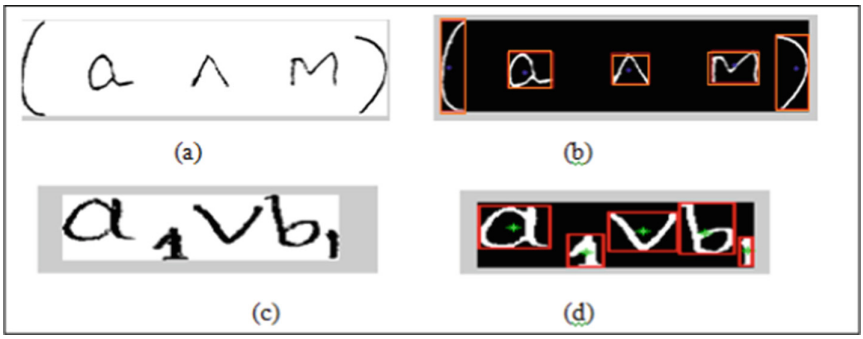


Fig. 3. (a) and (c) Original handwritten images, (b) and (d) Segmented image with BB and Centroid plot.

Two methods of segmentation are used to segment the characters from the image, namely, projection profile cutting and connected component labeling. Projection profile cutting works quite well for one dimensional structures. However, ME is a two dimensional structure that contains symbols having special relationship or one inside other symbol. Hence, Projection profile cutting is not well suited for the ME. The connected component labeling method was used for segmentation of ME [3]. Connected components of the image, Labeled L, were obtained for all the characters in the expression. The values of Bounding Box (BB) were obtained by using properties of image region. Centroid of each BB is found to identify the subscripts which is less than the half of the height of the expression image. Each subscript denoted with 'sb' and other characters with 'up'. Each segmented character had label structure contained three fields: class label, sb (subscript)/up (no subscript), and position. The proposed method is used to classify the characters into 38 different classes of characters (26 alphabets, 10 math symbols, 2 numerals (subscripts). A class label from the set C(yj) $\in${1, 2, 3,…, 38} was assigned to each character to recognize the characters in the expression (Fig. 3).

Once the character is segmented, the feature extraction methods: Normalized chain code (16 features), zone based density (26 features), diagonal and intersection points (128 features), Moment invariant features (7 features), Projection Histogram (127 features) were used to get the features of the characters of training expressions [4, 5]. The proposed system was carried out the classification with combinations of these features to get better recognition rate. It was observed that the performance of classifier was high using combination of zone based density features and normalized chain code features with feature vector of size 42. The SVM classifier was used to classify a character of the expression. The expression is reconstructed by using label structure. All the correctly recognized characters in the expression were used for training. The recognition process of test expressions using SVM classifier is shown in the Fig. 4.

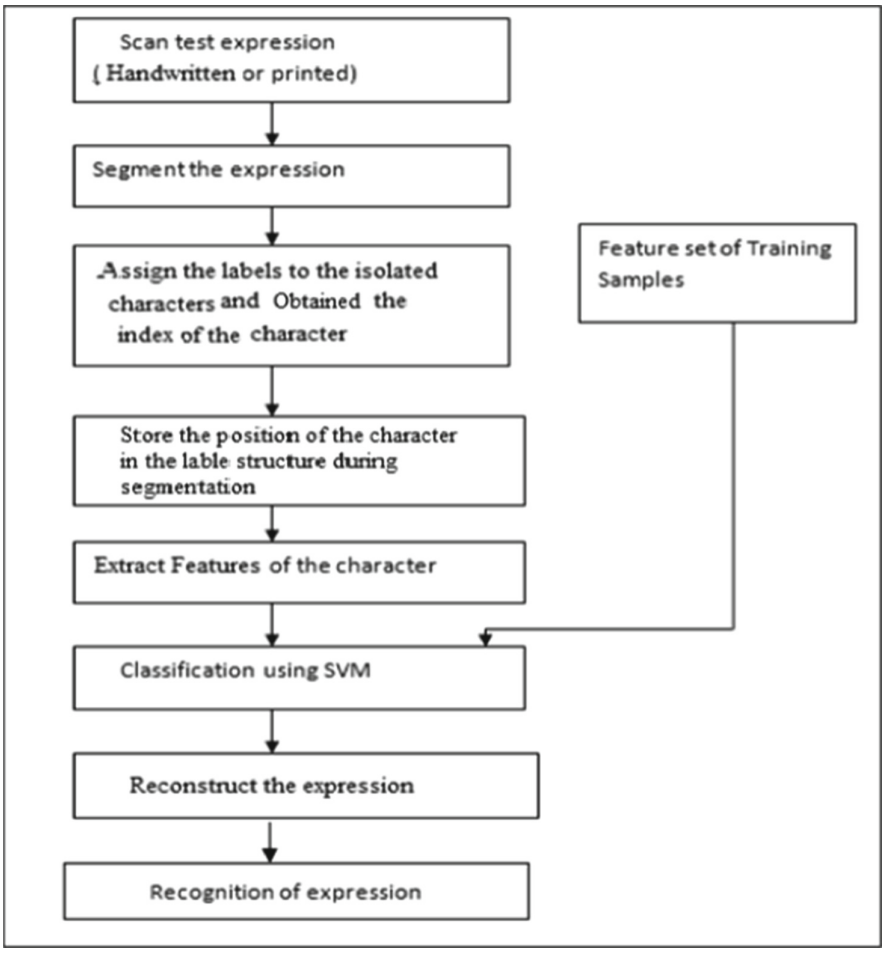


Fig. 4. Expression recognition system using SVM

3 Testing Process

The result of classification of each character was a class label from the set of class C(yj) ∈{1, 2, 3, …, 38}. The true label (label of the character with respect to its class) implies the correct recognition of the character. However, just the features and class labels do not give sufficient information to recognize the expressions. The information about the location of the character and grouping of the characters are also needed in expression recognition. The expression was reconstructed using label and relative position of subscripts and main character. An algorithm for expression recognition is discussed below.

3.1 Algorithm for Expression Recognition

Input: Set of Handwritten or Printed Expressions.
Output: Class labels of the characters in the expressions.
Method:
Step 1. Read scanned handwritten or printed expression.
Step 2. Preprocess the expression image.
Step 3. Segment the expression to separate characters using the segmentation algorithm and obtain the isolated characters and its position information.
Step 4. Apply the feature extraction algorithm to each image to get feature set of each character, and obtain the feature set of an expression.
Step 5: Input the feature set of an expression to the SVM classifier to obtain the recognized characters of an expression.
Step 6: Reconstruct the expression with the help of position numbers of the characters.
Step 7: Display the recognized expression.

4 Experimental Results

To predict the class of the test characters in the expressions, the unknown pattern vector of the test expressions and the feature set representing training images were fed to classifiers SVM.

Normalized chain code and zone based density method was used to extract the features of segmented characters from the test expression. A recognition accuracy of the SVM classifier was calculated. The expression recognition rate (ERR) was computed as percentage of correctly recognized expressions from the set of same expressions.

Define:

CR = Number of correctly recognized expressions, out of correctly segmented

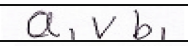
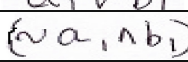
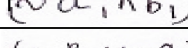
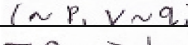
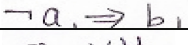
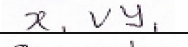
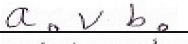
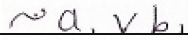
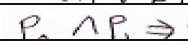
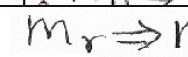
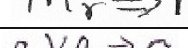
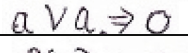
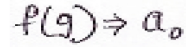
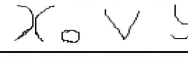
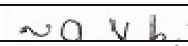
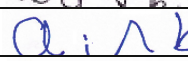
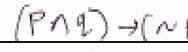
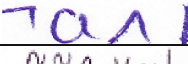
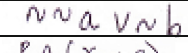
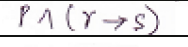
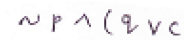
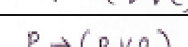
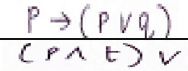
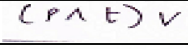
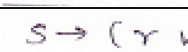
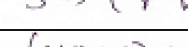
TE = Total number of occurrences of a single expression

Then, for each expression the recognition rate was computed as,

$$\text{ERR} = \frac{CR}{TE} * 100$$

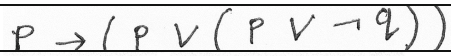
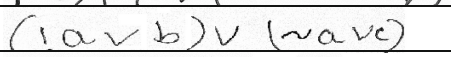
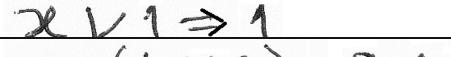
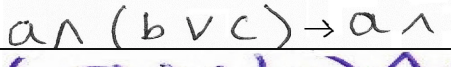
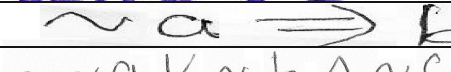
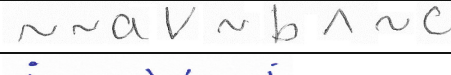
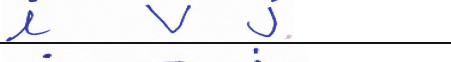
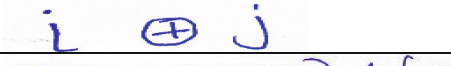
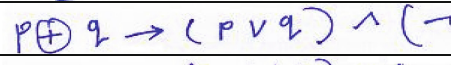
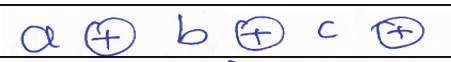
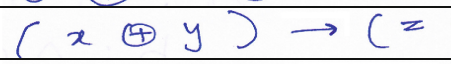
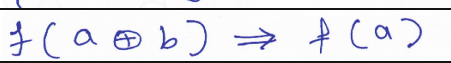
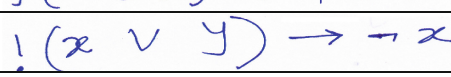
The average recognition rate of handwritten expressions and printed expressions was 84.1% and 90.3% respectively. It was observed that the recognition rate of printed expressions was higher than that of handwritten expressions. The experiment was performed using Matlab 8.1. The recognition of some expressions using SVM is shown in the Fig. 5 and Fig. 6. The numeral values in the output shows the class labels of the characters. The class labels for letters ‘a’ to ‘z’, 10 math symbols, numerals ‘0’ and ‘1’ were assigned as 1 to 26, 27 to 36, 37, and 38 respectively. The error rate in output defines the percentage of misclassified characters. The resultant recognized expression is shown as the output in the command window after reconstructing the expression using position of the characters (Tables 1, 2, 3).

Table 1. Handwritten Expression Recognition Rate (HERR) of handwritten test expressions

#Exp	Handwritten Expressions	HERR (%)
1		90
2		85
3		90
4		75
5		90
6		85
7		80
8		75
9		80
10		70
11		75
12		85
13		70
14		75
15		80
16		90
17		100
18		95
19		90
20		100
21		90
22		85
23		85
24		90
25		85
26		85
27		70
28		95
29		90
30		80
31		75

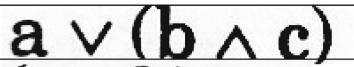
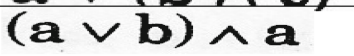
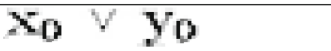
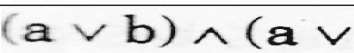
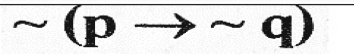
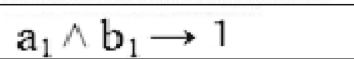
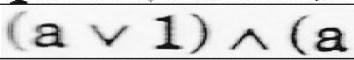
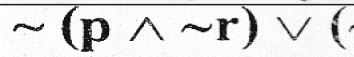
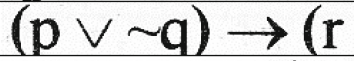
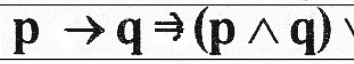
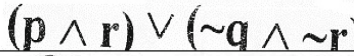
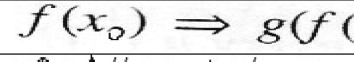
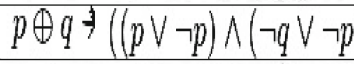
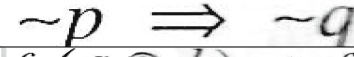
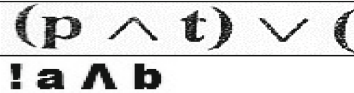
(*continued*)

Table 1. (*continued*)

32		80
33		75
34		85
35		85
36		85
37		90
38		95
39		85
40		95
41		85
42		85
43		85
44		75
45		80
46		85
47		75
48		80

From the experiments, it was observed that the recognition errors were due to misclassification characters of similar shape in handwritten expression recognition. The confusing pairs of letter ‘v’ and symbol ‘V’; ‘c’ and ‘(’ gives the incorrect recognition. The correct classification of handwritten expressions using SVM classifier was shown in the Fig. 5. The recognition rate of the expressions was poor due to the misclassified character ‘v’ in the Fig. 6. The resultant output of classification is consists of class label 22 represent character ‘v’ instead of symbol ‘or’.

Table 2. Printed expression Recognition rate (PERR) of printed test expressions.

#Exp	Printed Expressions	PERR (%)
1		100
2		100
3		90
4		90.9
5		83.33
6		80
7		90
8		100
9		87.5
10		90
11		100
12		86.6
13		90
14		70
15		80
16		75
17		80
18		70
19		100
20		100
21		90
22		100
23		85.7
24		100
25		100

(*continued*)

Table 2. (*continued*)

26	$(p \rightarrow r) \vee (\neg s \rightarrow \neg t) \vee (\neg u \rightarrow v)$	80
27	$((p \rightarrow (q \rightarrow r)) \rightarrow ((p \rightarrow q) \rightarrow (p \rightarrow r))$	80
28	$(p \wedge q) \rightarrow \sim (p \rightarrow \sim q)$	85.7
29	$(p \vee \sim q \vee \sim r) \wedge p \vee (q \wedge r)$	87.5
30	$\sim \sim a \vee \sim b$	100
31	$p \oplus q \rightarrow \neg((p \wedge q) \vee (\neg p \wedge \neg q))$	100
32	$((p \rightarrow q) \wedge (q \rightarrow r)) \rightarrow (p \rightarrow r)$	100
33	$(\neg q \wedge (p \rightarrow q)) \rightarrow \neg p$	100
34	$x \oplus (y \oplus z) \Rightarrow (x \oplus y) \oplus z$	100
35	$(p \oplus q) \rightarrow (p \oplus \neg q)$	100

Table 3. Average Expression Recognition Rate

Expression Type	#exp	Wrongly Segmented	Wrongly Recognized	Correctly recognized	ERR
Handwritten	960	105	48	807	84.1
Printed	288	20	8	260	90.3

In case of printed expression recognition, normally, the expressions were not correctly segmented due to touching characters. The correctly segmented expressions results in the high accuracy in recognition step. Most of the errors in the recognition occurs due to misclassified characters ‘a’ with ‘)’. The correct classifications of printed expression is shown in the Fig. 7.

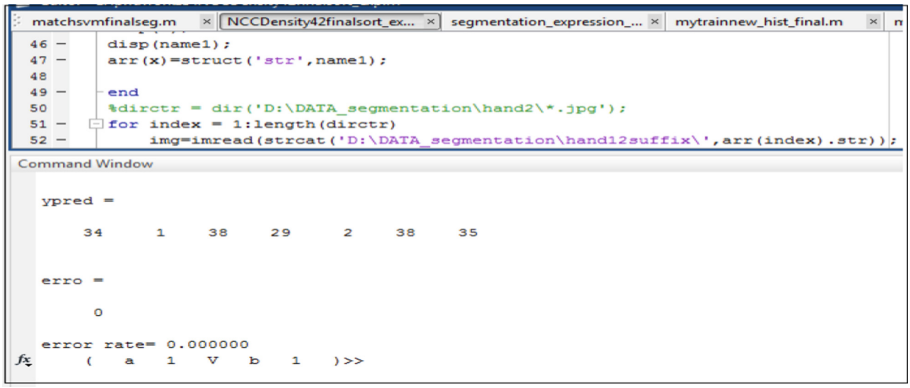


Fig. 5. Recognition of handwritten expression $(a_1 \vee b_1)$

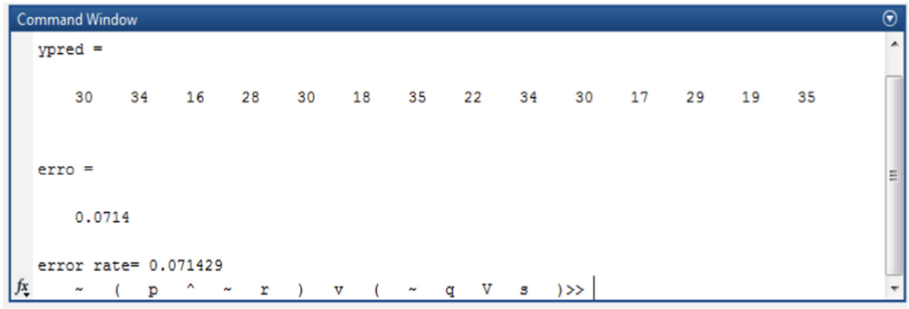


Fig. 6. Recognition of handwritten expression $\sim(p \wedge \sim r) \vee (\sim q \vee s)$

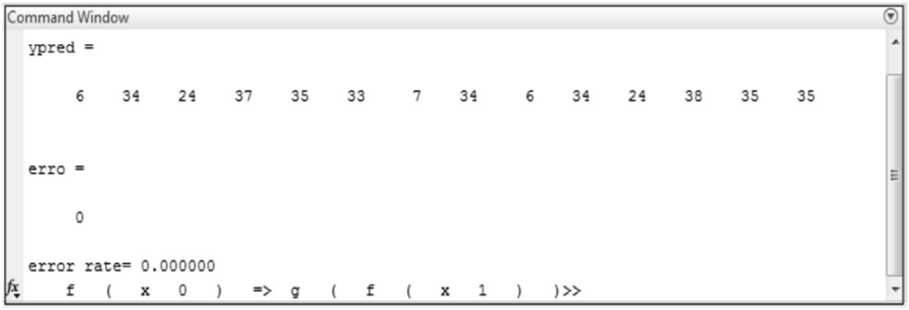


Fig. 7. Recognition of printed expression $f(x_0) \Rightarrow g\,(f(x_1))$

5 Conclusion

SVM classifier was used to classify the characters in the expressions. The accuracy of the expression was computed. The average recognition rate of handwritten was 84.1%

and printed expression was 90.3%. The recognition rate of printed expression was higher than that of handwritten expression.

References

1. Ahmad M., Harold Mouchere, Christian Viard Gaudin, (2009), "Towards Handwritten Math ematicalExpression recognition", International Conference on Document Analysis and Recognition (ICDAR), IEEE 978-0-7095, pp:1046–50.
2. Ahmad Montaser Awal, Harold Mouchere, Christian Viard Gaudin, The problem of handwritten mathematical expression recognition, International Conference on Frontiers in Handwriting Recognition (ICOFHR), ISBN, 978-0-7695-4221-8.
3. Manisha Bharambe**,** "Segmentation of Offline Printed and Handwritten Mathematical Expressions" International Journal of Computer Applications (0975–8887) National Conference on Digital Image and Signal Processing 2016
4. Manisha Bharambe, "Recognition of Offline Handwritten Mathematical Expressions"International Journal of Computer Applications (0975–8887) National conference on Digital Image and Signal Processing, DISP 2016
5. Manisha Bharambe *"*Logical Symbol Recognition using Normalized Chain code and Density Features*",* International Journal of Engineering Research & Technology (IJERT), ISSN: 2278–0181, Vol. 3, Issue-12, December 2014, pp: 619–623
6. Kang kim, Taik Rhee, Jae LEE, (2009), "Utilizing consistency context for handwritten mathematical expression recognition", ICDAR, 978-0-7695-3725-2, IEEE, pp: 1051–1055.
7. Surendra Ramteke, Dhanashree Patil, Nilima Patil, (2012), "Neural network Approach to Mathematical Expression Recognition System", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 10, December-2012, ISSN: 2278–0181.
8. Dipak D.Bage, K.P. Aditya, Sanjay Gharde, (2013), "A new approach for recognizing offline handwritten mathematical symbols using character geometry", International Journal of Innovative research in science, engineering and Technology, vol. 2, Issue 7, ISSN: 2319–8753, pp: 2823–2830.
9. H. Mouchere, C. Viard-Gaudin, D. H. Kim, J. H. Kim, U Garain, (2012), "ICFHR 2012-Competition on Recognition of online mathematical Expressions", (CROHME 2012), 078-0-7695-4774-9/12, IEEE.
10. Sanjay S. Gharde, Baviskar Pallavi V, K. P. Adhiya, (2013), "Identification of Handwritten Simple Mathematical Equation Based on SVM and Projection Histogram", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231–2307, Volume-3, Issue-2, pp: 425–430.
11. Dong-Yu Zhang, Xue-Dong Tian, and Xin-Fu Li, (2010), "An Improved Method for Segmentation of Touching Symbols in Printed Mathematical Expressions", International Conference on advance Computer Control, IEEE, ISBN: 978-1-4244-5848-6, pp: 251–253.
12. Ernesto Tapia and Raul Rojas, (2004), "Recognition of On-line Handwritten Mathematical Expressions Using a Minimum Spanning Tree Construction and Symbol Dominance", Llados and Y.B.Kwon(Eds.): GREC 2003, LNCS 3088, pp: 329–340, 2004, Springer.
13. Ernesto Tapia and Raul Rojas, (2005), **"**Recognition of On-Line Handwritten Mathematical Expressions in the E-Chalk System - An Extension", Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05), 1520–5263/05, IEEE
14. Francisco Álvaro, Richard Zanibbi, (2013), "A Shape-Based Layout Descriptor for Classifying Spatial Relationships in Handwritten Math", DocEng'13, ACM 978-1-4503-1789-4/13/09

15. Hans Jurgen Winkler and Manfred Lang, (1997), "On-Line symbol segmentation and recognition in handwritten mathematical expressions", International conference on Acoustics, Speech and Signal Processing, ICASSP, 0-8186-7919-0/97, IEEE.
16. Kazuki Ashida, Masayuki Okamoto, Hiroki Imai, (2006), **"**Performance Evaluation of a Mathematical Formula Recognition System with a large scale of printed formula images", Proceedings of the Second International Conference on Document Image Analysis for Libraries (DIAL'06), 0-7695-2531-8/06, IEEE
17. Lei Gao, Shulin Pan, Shen Jiao, (2013), "An Analytic Hierarchy Process Based Method to Process Mathematical Expressions", Journal Of Theoretical And Applied Information Technology, Vol. 47 No.3, ISSN: 1992–8645.
18. Fanglin Wang , Fukeng He , Ning Bia, Ching Y Suen , Jun T., Offline Handwritten Mathematical Expression Recognition Based on Square Loss Sequence-to-Sequence Neural Network, https://doi.org/10.2139/ssrn.4212220
19. Dmytro Z., Viktor Z., And Olga R, Online Handwritten Mathematical Expression Recognition and Applications: A Survey, IEEE, Volume 9, 2021, https://doi.org/10.1109/ACCESS2021.3063413.
20. Riddhi A, Shila P, Anilkumar T, Gaurav H, Survey of Mathematical Expression Recognition for Printed and Handwritten Documents, https://doi.org/10.1080/02564602.2021.2008277
21. Yuging W, Zhengu W, Zhaokun Z, Shuajian Z., Yuesheng Z., Dual Branch Network Toswards Accurate Printed Mathematical Expression Recognition, ICANN (2022), pp 594–606
22. Lyzandra D'souza, Maruska ascarenhas, Offline Handwritten Mathematical Expression Recognition using Convolutional Neural Network, IEEE *Xplore***:** 15 November 2018, https://doi.org/10.1109/ICICET.2018.8533789