A Simplified Research for Mathematical Expression Recognition and Its Conversion to Speech

Punith Kumar, T Shreekanth, Shashank N S, Sneha S

Abstract: ***The number of visually impaired people appearing for various examination is increasing every year while on the other hand, there are several blind aspirants who are willing to enrich their knowledge through higher studies. Mathematics is one of the key language (subject) for those who are willing to pursue higher studies in science stream. There is a lot of advanced Braille techniques and OCR to speech conversion software's made available to help visual impaired community to pursue their education but still the number of visually impaired students getting admitted to higher education is less. This is not because most of the data is on paper in the form of books and documents. So, there is a great need to convert information from the physical domain into the digital domain which would help the visually impaired people to read the advanced mathematics text independently. Optical Character Recognition (OCR) systems for mathematics have received considerable attention in recent years due to the tremendous need for the digitization of printed documents. Existing literature reveals that, most of the works concentrated on recognizing handwritten mathematical symbols and some works revolve around complex algorithms. This paper proposes a simple, yet efficient approach to develop an OCR system for mathematics and its conversion to speech. For Mathematical symbol recognition, Skin and Bone algorithm is proposed, which proved its efficiency on a variety of data set. The proposed methodology has been tested on 50 equations comprising various symbols such as integral, differential, square, square root and currently achieving recognition rate of 92%.***

Keywords— Skin and Bone Algorithm, connected component labelling, Projection profile, Segmentation, 2D Correlation.

I. INTRODUCTION

Visual Impairment has become one of the serious issues in today's world. As per the statistics from World Health Organization (WHO), 253 million people are visually impairmed where 36 million of those are completely blind, out of which over 15 million blinds are present in India. Every year, several blind people appear for various examinations across the country, which means there are blind aspirants who are willing to enrich their knowledge through higher studies. Mathematics has been one of the key subjects included in most of the streams of higher studies and hence studying mathematics has become a necessity. In order for the blind to study advanced mathematics, they have to depend on the tutors most of the times. Even though advanced Braille techniques are available, still it is difficult for them to study mathematics independently. The proposed method provides an opportunity for the blind aspirants willing to study mathematics independently for their examinations.

II. REVIEW OF LITERATURE

Christopher Malon et al. evaluated the use of Support Vector Machines (SVM) on InftyReader, a free OCR system of scientific documents including math symbols, to enhance the classification and to reduce the misrecognition of mathematical symbols. As a result, the misrecognition was reduced by 41% [1].

Francisco Alvaro et al. developed a statistical framework through which symbols in printed mathematical equations can be recognized with improved efficiency. The results were tested on InftyCDB-1 [2] and it is evident that the structural analysis of mathematical equations resulted in improved recognition of mathematical symbols [3].

Pavan Kumar et al. concentrated on structural analysis of mathematical expressions and did a thorough analysis of structural errors in an isolated way. To analyze errors, a database of noise free expression images was created which can also be used for comparison of various approaches of structural analysis [4].

Richard Zanibbi, and Dorothea Blostein did a comprehensive survey of both retrieval and recognition methods of mathematical symbols [5].

Jianshu Zhang et al. presented a novel deep learning methodology for recognition of handwritten mathematical expressions. A new approach has been developed where two-dimensional mathematical expression is converted to a character sequence which can successfully avoid inherent problems of segmentation of symbols [6].

Fotini Simistiran et al. in 2015 developed an efficient system for online recognition of mathematical formulas using probabilistic approach. SVM classifier is used to distinguish between mathematical symbols/sub expressions. Output is obtained in MathML format by employing CYK based algorithm. Also, an assumption is made that symbols that comprise mathematical expression are correctly recognized [7].

Revised Version Manuscript Received on August 19, 2019.
Dr. Punith Kumar, M B Associate Professor, Dept. of ECE PESCE,Mandya, Karnataka, India
Dr.T Shreekanth, Mysore(email: speak2shree@gmail.com)
Shashank N S, Dept. of ECE, SJCE, Mysore, Karnataka, India
Sneha S, Dept. of ECE, SJCE, Mysore, Karnataka, India

Kenny Davila et al. in 2014 developed a unique and novel methodology for online recognition of mathematical symbols. In this methodology, offline features were converted to on-line counterpart by characterizing the shapes of the mathematical symbols. They have achieved a significant accuracy of 89.87% on MathBrush dataset [8].

Sachin Kulkarni et al. in 2014 proposed a novel approach for Text to Speech conversion of Mathematical equations and developed a tool called “MathsSays”. The module consists of three blocks namely Text Scanner, Text Extractor, and Speech Synthesizer. To handle the problem of recognizing challenging equations, a mapping table has been prepared which have textual representation of those particular equations which can be mapped accordingly [9].

Shinpei Yamazaki et al.in 2011 developed an OCR system for Mathematical formulas and embedded the module into OCRopus, which is an open source, stable OCR system. This work enabled the digitization of scientific document which contains myriad number of mathematical equation in it [10].

Peter Caky et al. in 2009 discussed about the normalization steps in Text to Speech conversion of Mathematical formulas. The paper delineates the application of Speech Synthesis Markup language (SSML) in processing the text [11].

Wei Wu et al. developed a system for recognition of Mathematical formulas by imparting novel technique for component labelling and image extraction of each symbols. A neural network based character recognition algorithm has been adopted and the recognition outcomes are converted to LaTeX file [12].

This extensive survey indicates that there are many matured methods for hand written mathematical character recognition that are computationally complex for recognition of simple mathematical equations. This work proposes a simplified approach based on skin and bone concept for recognition of selected group of simple mathematical expressions and its conversion to speech. The reminder of the paper is segmented as follows, Section II presents the proposed method, Section III provides the results and discussion and Section IV provides the conclusion and scope for future enhancements.

III. METHODOLOGY

Several steps used in the implementation of proposed method are listed as follows.

A. Segmentation

Segmentation of Mathematical expressions into sub images of characters is a crucial stage in Mathematical expression recognition. Segmentation can be either achieved by connected component analysis [13] or by projection profile cutting method [14].

Connected component analysis method labels each component by assigning a number to all the individual connected component, thereby we can extract each symbol through their respective bounding boxes, on the other hand, projection profile method gives the histogram of number of black pixel values piled up along parallel lines of the image. Using this we can plot the horizontal and vertical projection profile of the image and extract the symbols. The mathematical expression for vertical and horizontal projection profiles are given by (1) and (2).

$$VPP(y) = \sum_{1\le x\le m} f(x,y) \quad ......(1)$$

$$HPP(x) = \sum_{1\le y\le n} f(x,y) \quad ......(2)$$

Where, m and n are the rows and columns of the image respectively.

The horizontal and vertical projection profile for the image in Figure .1 is shown in Figure. 2.

Figure 1. Sample Equation and its binarized representation

Each symbol of Figure.1 is stored as an individual image and is later used for symbol recognition. Both projection profile and connected component analysis methods yields the same result. Flowchart of segmentation process is shown in the Figure.3.

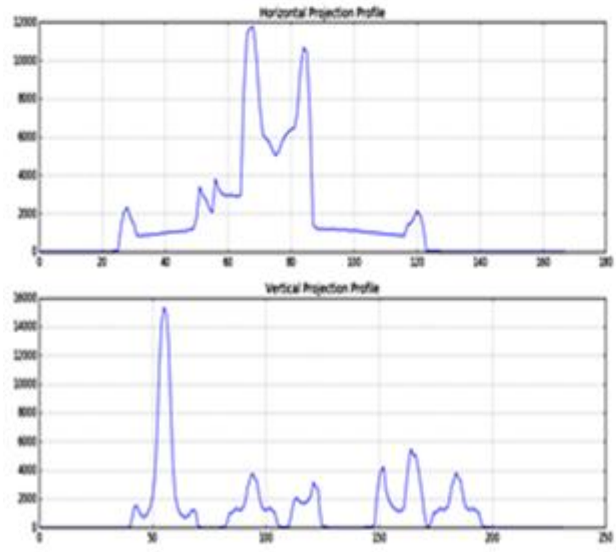

Figure 2. Horizontal and Vertical projection profile

B. Symbol Recognition

Mathematical symbol recognition is one the most challenging areas in pattern recognition because of the eclectic number of symbols involved in different areas of mathematics. Also, the convention in which each symbol is used in different contexts makes the recognition even more complex. In this work, we started with the rudimentary phase of the mathematical symbol recognition by using relatively simpler methods to get the desired output. The approach followed for symbol recognition involves, creation of an organized database, application of skin and bone algorithm on this database and finally application of 2D correlation for

International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-2S8, August 2019

identification of the corresponding symbol from the database.

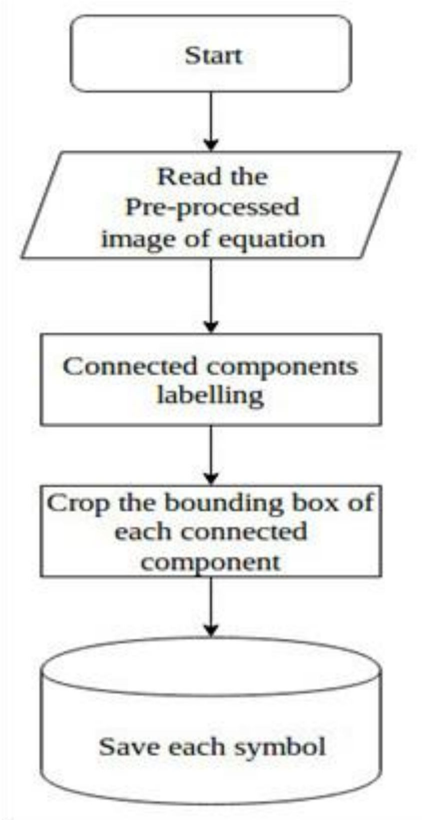

Figure 3. Flowchart of Segmentation

1. Creation of database:

Using the segmented symbols, an organized database has been created with 5 images of each symbol. The cluster of 5 images has to be created in such a way that they have a different orientation to make it heterogeneous. Creating the database is an arduous process which requires patience and should be done with utmost care.

2. Skin and Bone Algorithm:

Skin and bone algorithm is a simple yet efficient algorithm for symbol recognition. In this algorithm, the images of each symbol in the created database are used to generate skin image and bone image. Skin is obtained by performing the binary-AND operation on the images of symbols in the created database and bone is obtained by performing the binary-OR operation on the images of symbols. So, there will be two images (Skin image and Bone image) for each symbol and it will be used in subsequent comparisons. Figure 4 shows the skin and bone image.

Figure 4. Skin and Bone images (Skin to the left and Bone to the right)

$$\text{Skin/Bone Probability} = \frac{\text{Number of pixels lying outside skin or bone}}{\text{Total number of pixels}} \quad \ldots(3)$$

3) 2D correlation:

2D correlation is a technique for finding areas of an image that match (are similar) to the image which is to be compared (patch). Here, the Source image (I) is the character segmented image which must be recognized and the Template image (T) is the Skin and Bone image which is to be compared against the input image. At each comparison of segmented character with skin and bone of each glyph, a correlation coefficient is calculated so that it represents how good the match is with respect to skin and bone. For each comparison of T over I, the correlation coefficient has to be stored in a data structure and obtain the best match for the segmented character. The correlation coefficient is obtained by (4),

$$R(x,y) = \sum_{(x\prime,y\prime)} (T'(x',y'), I(x+x', y+y')). \quad \ldots.(4)$$

Where,

$$T'(x',y') = T(x',y') - \frac{1}{(w.h)} \cdot \sum_{(x\prime\prime,y\prime\prime)} T(x',y')$$

And

$$I(x+x',y+y') = I'(x+x',y+y') - \frac{1}{(w.h)} \cdot \sum_{(x\prime\prime,y\prime\prime)} T(x',y')$$

After matching each symbol to its respective Unicode, a textual output of that symbol has to be mapped to construct the meaningful pronunciation of the equation. The flowchart of symbol recognition is shown in Figure. 5.

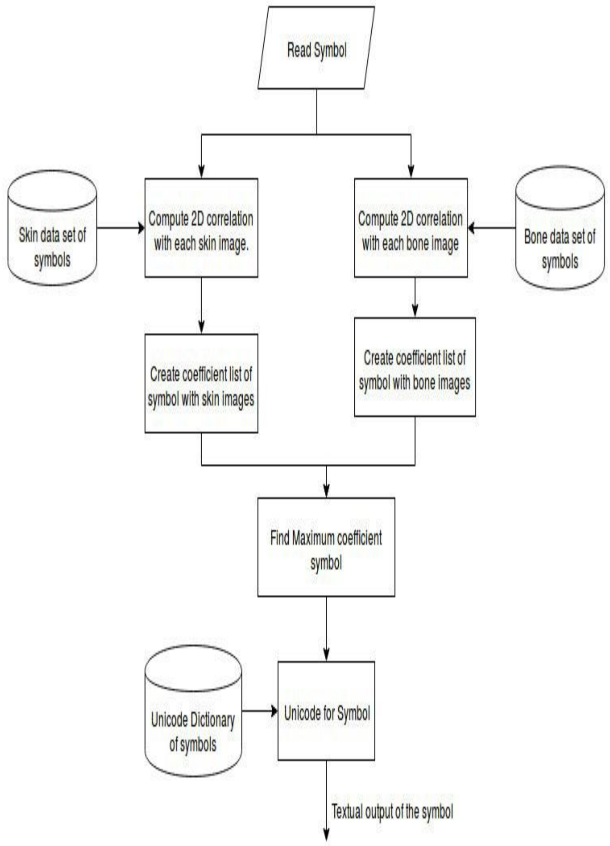

Figure 5. Flowchart of Symbol Recognition

C. Post Processing

The complete implementation has been carried out using OpenCV library and Python. The textual output is constructed by mapping the recognized symbol to its corresponding text and final output is obtained by joining the text of each symbol. Equations with powers and square roots are segmented using connected component analysis by labeling the connected components and rewritten on a separate mask by storing the x-y coordinate value and it is used to construct textual output of the equation. Differential and Integral calculus equations consist of trigonometric functions (sin, cos, tan). So, in order to accommodate English alphabets which, occur in equations, Unicode of all English alphabets are included in the database. Whenever trigonometric functions like sin, cos, tan occurs in the equation, they are mapped as sine, cosine, and tangent in textual output. Figure. 6 shows output of some sample equations.

Figure 6. Textual output of sample equations

To convert these textual outputs to speech, a Python package called pyttsx which supports common text-to-speech engines on Windows, Linus and Mac OSX is used.

However we may not get accurate textual output for some equations just by joining the output of each recognized symbol but as a future work this study can be extended to improve the existing system to get most accurate textual output of mathematical expressions. Flowchart of the entire methodology is shown in Figure. 7.

IV. RESULTS AND DISCUSSION

The proposed methodology is implemented in Python using OpenCV library and tested on 50 different equations. The results are analyzed statistically and efficiency of segmentation and symbol recognition is calculated. Symbol recognition efficiency is calculated using (5).

$$Efficiency = \frac{\text{Total number of symbols in all equations} - \text{Incorrectly recognized symbols}}{\text{Total number of symbols in all equations}} \times 100 \quad ......(5)$$

Average recognition efficiency of 92% has been achieved and most of the misrecognition is due to the occurrence of

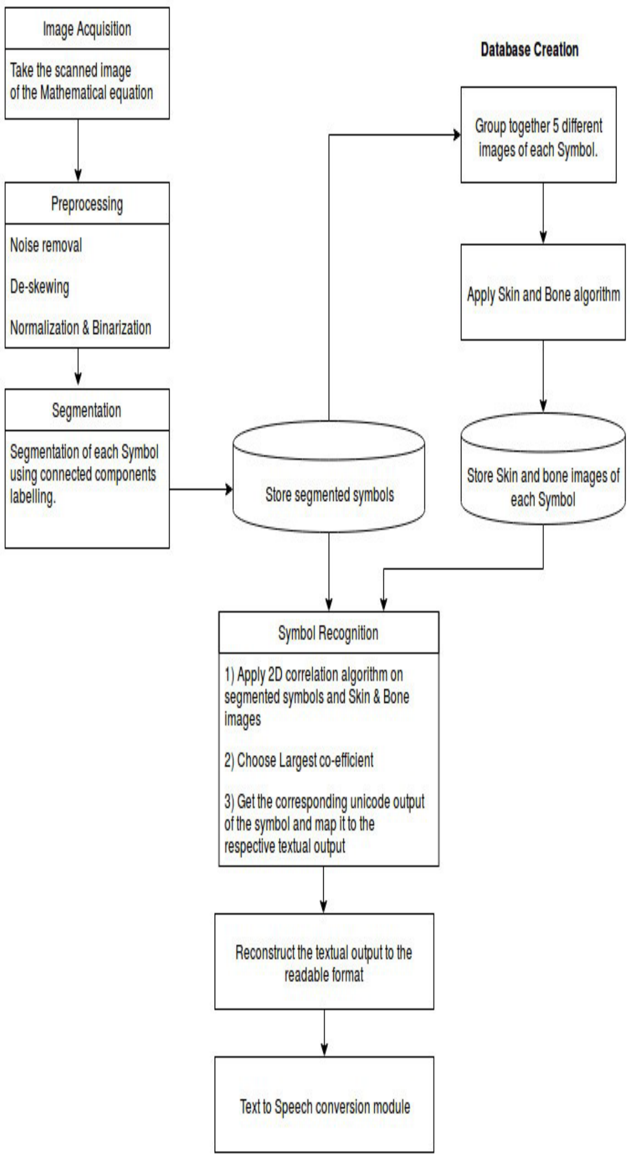

Figure 7. Flowchart of entire methodology.

closely matching symbols. Besides this some symbols were not segmented accurately because both projection profile method and connected component analysis method failed to segment some integral and summation equations with lower and upper limits. Table.1 depicts the evaluation of the proposed method on few dataset, where A is the total number of symbols in the equation, B is the number of symbols correctly recognized and B is the number of symbols miss classified.

The average segmentation efficiency is 91% and among 50 sample equations 45 equations are segmented accurately. Among properly segmented 45 equations 41 equations are recognized precisely and reconstructed into textual output. Few properly segmented symbols from inaccurately segmented equations (Only few symbols are segmented improperly) are also getting recognized precisely but it is quite difficult to construct textual output for such equations

Table.1 Performance Evaluation on selected equations

Equation	A	C	B	Accuracy= (A-B)/A *100
$a^2 + b^2 = c^2$	08	08	00	100
$y=e^x$	04	04	00	100
$\int x^2 dx$	05	05	00	100
$\int (x^2 + 9x + 6)\, dx$	12	10	02	83.3
$\int \cot^{-1} x\, dx$	09	08	01	88.8
$\int tan^2\, x\, dx$	08	07	01	87.5
$A = \pi\, r^2$	05	05	00	100
$\int (\, sinx + \, cosx\,)\, dx$	14	13	01	92.8
$\sin^2\theta + \cos^2\theta = 1$	13	13	00	100
$\sin 30^0 = 1/2$	10	09	01	90

V. CONCLUSION

The main objective of this work is to develop a program by which information held in hard copies of journals, papers, and books which have been published in the Mathematics language to be converted to digital form. This would mean a flood of new data is available to knowledge seekers everywhere without having to track down the material physically. This was partially achieved by this work, which have given the means to directly convert pages together of information into digital format

In a matter of minutes. Through Skin and Bone algorithm, an efficiency of 92% has been achieved on 50 printed Mathematical expressions. This work can be expanded to integrate English and Mathematics OCR for providing a common platform for information conversion for a majority of the visually impaired population.

REFERENCES

1. C. Malon, S. Uchida, and S. Masakazu. Mathematical symbol recognition with support vector machines. Pattern Recognition Letters, 29(9):1326–1332, 2008.
2. The Infty Project. http://www.inftyproject.org/.
3. F. Álvaro, J. Sánchez, and J. Benedí. Recognition of printed mathematical expressions using two-dimensional stochastic context-free grammars. In Proceedings of ICDAR, 2011.
4. Kumar, P.P, Agarwal, A. & Bhagvati, C. Isolated structural error analysis of printed mathematical expressions. Pattern Analysis Application Springer London: 1433-755X ,2017.
5. Zanibbi R, Blostein D (2012) Recognition and retrieval of mathematical expressions. IJDAR 15(4):331–357.
6. Jianshu Zhang, Jun Du, Shiliang Zhang, Dan Liu, Yulong Hu, Jinshui Hu, Si Wei, Lirong Dai (2017). Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition. Pattern Recognition, Volume 71, 0031-3203, Elsevier 2017, pp. 196-206.

7. Fotini Simistira, Vassilis Katsouros, George Carayannis (2014). Recognition of online handwritten mathematical formulas using probabilistic SVMs and stochastic context free.
8. Kenny Davila, Stephanie Ludi, Richard Zanibbi (2014). Using Off-line Features and Synthetic Data for On-line Handwritten Math Symbol Recognition. 14th International Conference on Frontiers in Handwriting Recognition, 2167-6445/14, IEEE 2014, pp. 323-328.
9. Kulkarni, Sachin & Mukhopadhyay, Debajyoti. (2014). TEXT-TO-SPEECH CONVERSION FOR MATHEMATICAL DATA. 10.13140/2.1.4962.1122.
10. Yamazaki, Shinpei & Furukori, Fumihiro & Zhao, Qinzheng & Shirai, Keiichiro & Okamoto, Masayuki. (2011). Embedding a mathematical OCR module into OCRopus. 880 - 884. 10.1109/ICDAR.2011.180.
11. Caky, P & Boron, J & Klimo, Martin & Bachratá, Katarína. (2009). Mathematical formulas in text to speech system. 11. 49-53.
12. Wu W., Li F., Kong J., Hou L., Zhu B. (2006) A Bottom-Up OCR System for Mathematical Formulas Recognition. In: Huang DS., Li K., Irwin G.W. (eds) Intelligent Computing. ICIC 2006. Lecture Notes in Computer Science, vol 4113. Springer, Berlin, Heidelberg.
13. D. Průša and V. Hlaváč, "Mathematical formulae recognition using 2d grammars," International Conference on Document Analysis and Recognition, vol. 2, pp. 849–853, 2007.
14. X.-D. Tian, H.-Y. Li, X.-F. Li, and L.-P. Zhang, "Research on symbol recognition for mathematical expressions," First International Conference on Innovative Computing, Information and Control - Volume I (ICICIC'06).