IDENTIFICATION OF SPATIAL RELATIONS IN MATHEMATICAL EXPRESSIONS

Sridevi Ravada
Dept.of Information Technology
Gayatri College of Engineering for Women
Visakhapatnam, India
srideviravada@gvpcew.ac.in

Sudheer Gopinathan
Dept.of Information Technology & Mathematics
Gayatri College of Engineering for Women
Visakhapatnam, India
g.sudheer@gvpcew.ac.in

D.Lalitha Bhaskari
Dept. of Computer Science & System Engg ,
Andhra University
Visakhapatnam, India
lalithabhaskari@yahoo.co.in

Abstract

The automatic recognition of mathematical expressions in digital content is a challenging task due to the complex spatial relationships between the symbols involved in the expression. The accuracy of the recognition is dependent on a variety of factors that includes nature of the input medium. The reliability of the performance of the system is dependent on the identification of the spatial relationships in an expression. Symbol recognition and structural analysis are the two important stages in the recognition process. In the present work these two stages are considered using the concepts of connected components and minimum spanning tree. For our analysis, we have created a database of 500 expression images drawn from standard databases and the experimental results are reported on them.

Keywords: minimum spanning tree, connected, spatial, segmentation.

I. Introduction

Continuous research and development over the last few decades have enabled optical character recognition (OCR) systems to achieve a sufficient level of maturity in the recognition and retrieval of information from digital content. However, recognition and retrieval of images, tables, diagrams and mathematical expressions (ME) are yet to reach a comfortable level of applicability [1]. The amount of work in the literature that is directed towards the extraction and recognition of ME is a mature field of study and the research efforts in this area have been surveyed in the works [2,3].The recognition of ME is a challenging pattern recognition problem that includes segmentation ambiguities, symbol recognition challenges and ambiguity of meaning. The problem finds application in several areas of science and engineering that include document searching / editing , computer algebra systems, tutoring systems, and mathematical information retrieval to name a few[1]. The existing OCR systems have difficulty in converting the ME present in scientific/technical documents into a corresponding digital/electronic form for recognition. Instead of developing OCR systems specifically for scientific documents, the emphasis has been on including mathematical /math OCR module into the existing OCR systems. The digital document analysis in OCR systems typically consists of the pre-processing stage followed by the determination of the physical layout and logical structure of the document. The pre-processing is basically concerned with the removal/correction of noise, artifacts, unwanted variations introduced during the document generation stage and is an essential part of the OCR system. The physical layout of a document is basically its geometric structure and its analysis aims to decompose the document into a hierarchy of homogenous/similar regions. The logical structure is the actual content of the document and its analysis is aimed at understating the logical/ functional entities in a document along with their interrelationships [4]. The document structure analysis that includes these tasks get complicated due to the variations in the source of the input documents that consists of vector graphics, historical documents, printed documents etc., each having distinct physical and layout structures.

In the development of math- OCR module, the math zones are to be segmented from the input documents either manually or through semi or fully automated segmentation logic. Though research into the segmentation of math zones from documents have been carried out in the past few decades, challenges still remain as the developed methods have not been at a sufficient level to be adopted in realistic application scenarios [5]. It is to be noted that, no one technique for layout analysis completely dominates another and improving these methods in an active area of research[1]. The present work is concerned with the main module of a math OCR which is the math expression recognition module that consists of the two stages: Component character recognition and structural analysis [6]. The recognition of mathematical symbols is a difficult problem due to the presence of a large character set with a variety of font styles and a range of font sizes coupled with a set of symbols having an enormous range of possible scales [7,2]. The symbols occurring in the ME are arranged around different operators that form different layouts, some of which are one dimensional, while others are two dimensional. The 2D structure induced by the operators appear in normal or nested modes increasing the complexity of the expression which is further increased by the number of horizontal lines on which the constituent symbols are arranged[8]. In the structural analysis part, the spatial relationships between symbols are analysed to capture the structure of the expression, together with the logical meaning to aid the recognition process. The complex structure of the ME makes structural analysis a challenging task even when all the symbols have been properly recognized [9].

Document mathematical expression recognition is generally considered to be printed mathematical expression recognition[10].Over the past few decades, many excellent methods have been proposed in the field of ME recognition, however the search for an optimum technique/method is far from being over. In recent years, the focus has shifted to handwritten mathematical expression recognition. Though the techniques for both printed and handwritten ME recognitions are almost similar, handwritten ME recognition is difficult and is more or less dependent on the strokes used by the individual person. In addition to segmentation and character classification, the spatial relation classification is one of the dominant problem associated with ME recognition[11].

The ME can be represented in many formats such as symbol layout tree, operated tree, label graphs, Tex and so on and the output of the ME recognition systems are to be put into these formats for reconstruction. A network representation of the ME has helped researchers to utilize graph theory concepts in tasking symbol segmentation and recognition. The minimum spanning tree (MST) problem is one of the famous mathematical problems in computer science that has been adopted suitably by many workers for the development of mathematical OCR[12]. In unconstrained handwritten documents, separation of text lines is a challenge because of the skewed, curved and non-uniform structure of the text .[13] developed an approach for text line segmentation in unconstrained handwritten Chinese document using an approach based on the minimum spanning tree .A variation of the MST problem was applied to mathematical OCR by [15] utilizing the notion of candidate selection and link-label selection. A structural analysis method for the recognition of online handwritten ME based on a MST construction and symbol dominance was presented in [14]

The present work is concerned with the recognition of ME. Two approaches are utilized in the process. The first approach considers bounding box, labeling matrix and minimum spanning tree to group similar component and study the spatial relationship between them. The second approach considers the separation of horizontal lines into five regions and allot the components of ME into these regions for the identification of base symbol subscript, superscript and sub expressions. The relationship between these components is then obtained utilizing a minimum spanning tree.

II. Characteristics of Mathematical expressions

Symbol recognition and structural analysis are the two involved activities that are crucial in understanding a mathematical expression. Pre-processing, segmentation, recognition, identification/recognition of spatial and logical relationships and construction of meaning of the symbol involved in the expressions are the processes involved in these two activities [16]. A mathematical expression that is printed/handwritten can be viewed as a collection of symbols with spatial relationships among them. The problem of searching for the most likely Interpretation for a given set of mathematical inputs reduces to the problem of searching for the most likely symbol identities with the likely spatial relationships [17].

Understanding the essential features of the ME aids the processes involved in recognition. The mathematical expression / formulae are represented with various kinds of entities such as

i) Arabic digits, Greek, Latin, Roman, calligraphic letters etc., in addition to the English characters.

ii) Mathematical symbols that include bracket symbols, arrow symbols, miscellaneous symbols etc.,

iii) Mathematical operators – logical, set and relational etc., function names

In a ME, the alphabetic characters occur with a variety of typefaces such as normal, bold, italic etc., and can be touching, broken, overlaid etc.,. In addition, the alphabetic letter can be of type ascender (b,d,h,k…..) or descender (g,j,p,y…) or normal (a,m,r,n....). In the case of operators, factors such as operator range, operator precedence, symbol identity, relative symbol size and case interact in a complex way and understanding these provide a clue for the structural analysis of ME[13]. Some of the features can be further extended as :

1. Relative symbol placement is important in the identification of operators. For e.g.; the commonly used basic spatial relationships: left (L), Right (R), Below(B), Above (A), above right(AR) and above left(AL), below right (BR), below left (BL) and can be visualized as shown in Figure 1. Identification of these relationships is an important aspect of a recognition system.

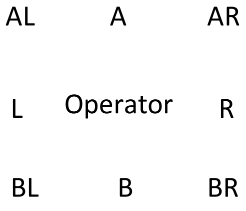

Figure 1: *Relative placement of symbols in an expression*

2.The symbol/expression have geometric complexity which is determined by the horizontal width occupied by the expression. The component symbols in the expression are arranged in a number of horizontal lines that increased the complexity as seen from a typical example shown in Figure 2.

$$\lim_{n\to\propto}\frac{\sum_{i=1}^{n} i^2}{\sum_{i=1}^{n} i^3}$$

Figure 2: *Illustration of geometry complexity in an expression*

3. Nested structure of the expression as seen in Figure 3a, in the figure, the argument of arc tan, that contains the square root operator and further involves fraction operator is a nested structure. Similarly a matrix image consists of element that may have fractions, square roots etc., as in Figure 3b.

$$\text{Y}=\tan^{-1}\sqrt{\frac{1-\cos x}{1+\cos x}}$$

(a)

$$\begin{pmatrix} 2x & \sqrt{\frac{x}{2}} \\ \frac{x+1}{x^2+1} & 4 \end{pmatrix}$$

(b)

Figure3: *Illustration of nested structure in an expression*

4.Links representing the structure of ME. Consider the expression $\bar{x_t} = 5x_{t-k} + p^2$.

The horizontal/base line together with the spatial relationships of the expressions are shown in the Figure 4. Understanding these is a part of structural analysis.

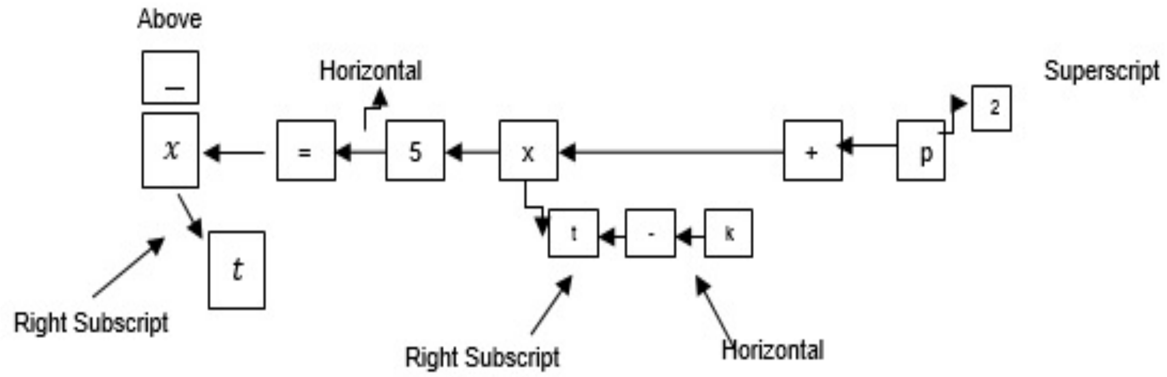

Figure 4: *Spatial relationships in an expression*

.

III. Proposed methodology

The first stage of the proposed method consists in symbol segmentation using connected component labelling method. For a given input image containing a mathematical expression, binarization is carried out. The binary image is scanned pixel by pixel and its pixels are grouped into components based on their connectivity. Once the connected components are extracted, their bounding boxes are obtained.

The bounding box of connected component (symbol) is defined to be the smallest rectangle which circumscribes the connected component (CC) [18]. Labels are then assigned to the CCs. The area, aspect ratio and centroid of the CCs are calculated. The centroid is a point attribute that helps determine the symbols location. Most of the single CC symbols can be segmented in this process but symbols like : , i , % , = , ≥, ≤ , ÷ , etc., which are multi CC are not segmented.

For example division (÷), a multi CC consists of three CC that consist of one horizontal line (-) and two dots (.). to resolve the problem of multi CC symbols the minimum spanning tree is utilized.

Considering the centroid of the CCs as vertices and distances between the centroids as weights a minimum spanning tree is constructed. The minimum distance between the components is used to resolve the ambiguity in the multi CC symbols like | | , i , % , = , ≥, ≤ , ÷ etc., .

The resolved multi CCs are labelled as composite symbols and the spanning tree is updated to group the similar components of remaining symbols if any in the expression. The mathematical functions like min, max, SIN, COS are also resolved through this process and symbol segmentation of ME is thus carried out.

In the second stage, the structural analysis part is considered. Here the spatial relationship

between the symbols is identified. Horizontal profiling is used to split the math block containing the ME into five regions as shown in fig.5. Level three is the main base line of the ME and contains the parent/ dominant symbols. Level 1, level 2 and level 4, level 5 contain the super, super expressions, subscripts, sub expressions. The peak values in the profiles together with some heuristic rules are employed to get a coarse split of the region. The symbol component height and the y-coordinate of the centroid of each component is used to allot the symbol into the 5 regions.

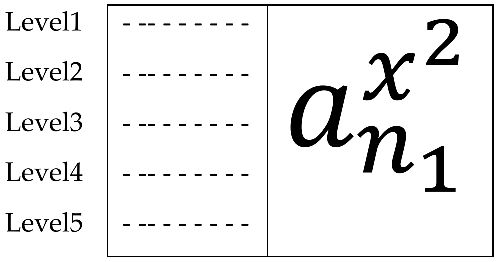

Figure 5: *A sample layout of the decomposition of a ME into five levels*

The statistical properties of the symbols together with the centroid values are utilized in finding the dominant base line. This is followed by the coarse identification of the relationship between two symbols such as above, below, in the same row, super script, sub script, prescript, nested using the relative geometric attributes of the bounding boxes of symbols. The horizontal and vertical projection profiles of CCs aid the process. The relationship tree based on the symbol dominance is then generated by constructing a minimum spanning tree that finds the different relationships among the components of the formulae.

IV. Results and discussion

The proposed methodology is tested on about 500 Mathematical Expressions taken from various mathematical documents collected from the internet including the database of Infty Project. The MEs considered cover various branches of mathematics. For handwritten samples we have considered only 20 expressions. The samples considered contain almost all mathematical expressions. The current algorithm is developed in MATLAB and Python. In the work, the connected components are sorted in the increasing order of x-coordinate and incase two or more (CC) have same x-coordinate, the y-coordinate are considered for sorting. The ordered CCs are scanned from left to right and basing on the spatial features the CCs are combined to form a single composite symbol. For e.g.; we consider some mathematical expressions and as a sample the components of the expressions are listed in Figure 6.The minimum spanning tree is used to find the spatial relations between the symbols of the expressions and the relationship obtained is shown in Figure 7. For the expression on printed documents the spatial relations could be identified correctly to the tune of 95%, however in handwritten cases the accuracy is around 60%. Further the element extraction could be accomplished with great ease.

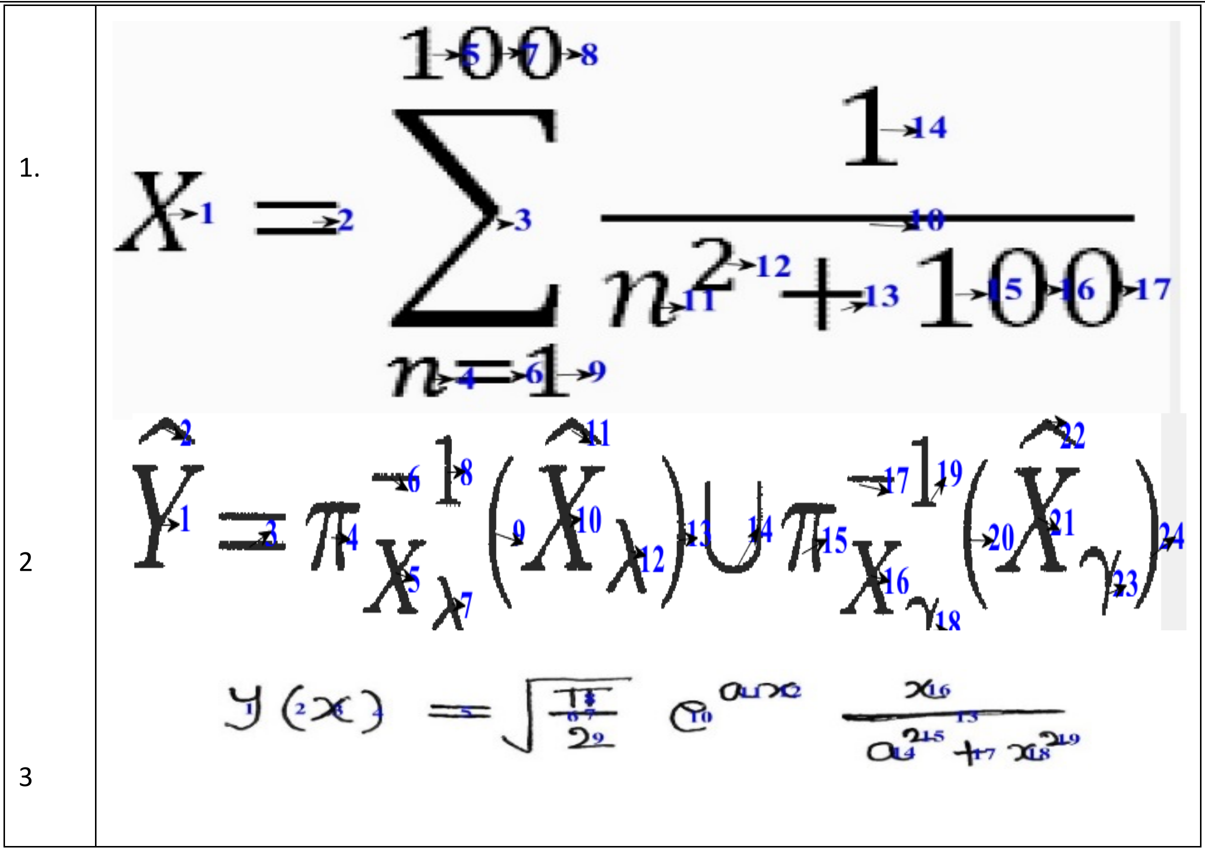

Figure 6: *Sample identification of different components of an expression*

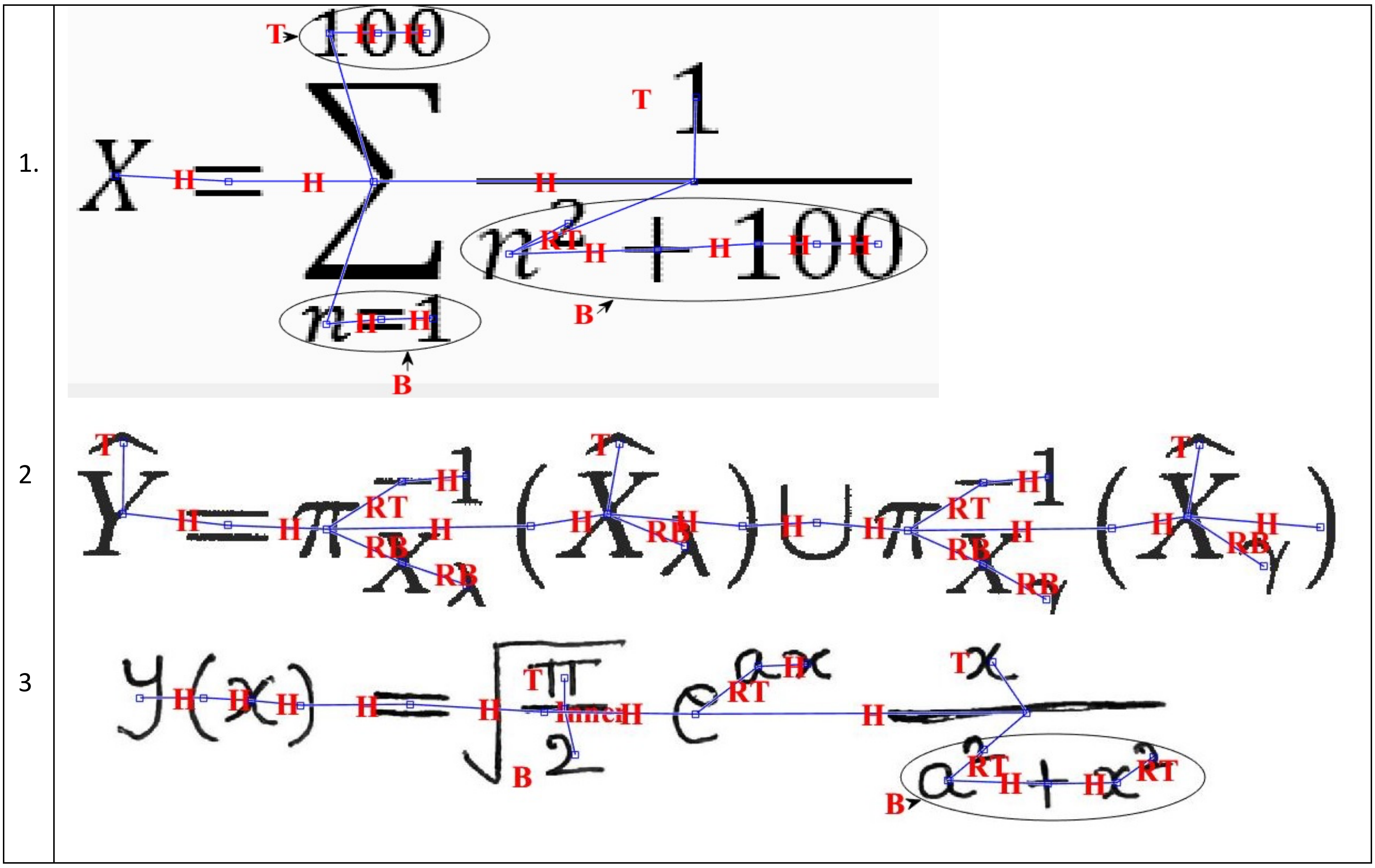

Figure 7: *The spatial relation between symbols in an expression*

V. Conclusion

In this paper, we focus on symbol segmentation and structural analysis of mathematical expressions in both printed and handwritten documents. The multiconnected components and context dependent symbols are resolved using the minimum spanning tree. The approach has been tested on around 500 MEs collected over the internet and the results are reported on them. We hope a good classifier can be added to the existing methodology to increase the accuracy on handwritten MEs.

References

[1] Zanibbi, R. and Blostein, D. (2012). Recognition and retrieval of mathematical expressions, International Journal on Document Analysis and Recognition. 15:331–357.

[2] Zanibbi, R. and Blostein, D. and Cordy, J.R. (2002). Recognizing mathematical expressions using tree transformation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(11):1455-1467.

[3] Chan, K. and Yeung, D. (2000). Mathematical expression recognition: a survey, International Journal on Document Analysis and Recognition,3:3–15.

[4] Namboodiri, A.M. and Jain, A.K. (2007). Document Structure and Layout Analysis In: Chaudhuri B.B. (eds) Digital Document Processing. Advances in Pattern Recognition. Springer, London, 29-48.

[5] Lin, X. Gao, L., Tang, Z. Baker, T. and Sorge, V. (2014). Mathematical formula identification and performance evaluation in PDF documents, International Journal on Document Analysis and Recognition. 17: 239–255.

[6] Aly, W. Uchida, S. and Suzuki, M. (2008). Identifying subscripts and superscripts in mathematical documents, Mathematics in Computer Science, 2:195-209.

[7] Garain, U. Chaudhuri, B.B. and Ghosh, R.P. (2004). A multiple-classifier system for recognition of printed mathematical symbols, Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004 1:380-383.

[8] Garain, U. and Chaudhuri, B.B. (2005). A corpus for OCR research on mathematical expressions, International Journal on Document Analysis and Recognition. 7:241–259.

[9] Pavan Kumar, P. Agarwal, A. and Bhagvati, C. (2018). Isolated structural error analysis of printed mathematical expressions, Pattern Analysis and Applications. 21:1097-1107.

[10] Huang, J. Tan, J. and Bi, N.(2020). Overview of Mathematical Expression Recognition, In book: Pattern Recognition and Artificial Intelligence, Proceedings International Conference on Pattern Recognition and Artificial Intelligence ICPRAI 2020, Y.Lu et al(Eds),LNCS 12068, 41-54.

[11] Zhelezniakov, D. Zaytsev, V. and Radyvonenko, O. (2021). Online Handwritten Mathematical Expression Recognition and Applications: A Survey. IEEE Access, 9:38352-38373.

[12] Fujiyoshi, A. and Suzuki, M. (2011). Minimum Spanning Tree Problem with Label Selection. IEICE Transactions on Information and Systems .E94-D(2).

[13] Yin, F., Liu, C.L. (2009). Handwritten Chinese text line segmentation by clustering with distance metric learning, Pattern Recognition,42: 3146-3157.

[14] Tapia, E., Rojas, R. (2004). Recognition of On-line Handwritten Mathematical Expressions Using a Minimum Spanning Tree Construction and Symbol Dominance. In: Llados, J., Kwon YB. (eds). Graphics Recognition Recent Advances and Perspectives. GREC 2003. LNCS, 3088,329-340.

[15] Fujiyoshi, A. and Suzuki, M. (2010). A variation of the minimum spanning tree problem for the application to mathematical OCR. Journal of Math-for-Industry,2:183-197.

[16] Blostei, D. and Grbavec, A. (1997). Recognition of Mathematical notation, In Handbook of Character Recognition and Document Image Analysis, 557-582.

[17] Rhee, T.H. and Kim, J.H. (2009). Efficient search strategy in structural analysis for handwritten mathematical expression recognition,Pattern Recognition, 42(12):3192-3201,

[18] Ha, J. Haralick, R.M. and I. T. Phillips. (1995). Recursive X-Y cut using bounding boxes of connected components. Proc of 3rd Int Conference on Document Analysis and Recognition, IEEE, 2:952-955.