International Journal of Technology (2015) 3: 345-354
ISSN 2086-9614

RECOGNIZING OFFLINE HANDWRITTEN MATHEMATICAL EXPRESSIONS (ME) BASED ON A PREDICTIVE APPROACH OF SEGMENTATION USING K-NN CLASSIFICATION

Sachin Naik[1], Pravin Metkewar[1*]

[1] *Symbiosis Institute of Computer Studies and Research(SICSR), A Constituent of Symbiosis International University (SIU), Atur Centre, Model Colony, PUNE-411016, Maharashtra State, India*

(Received: May 2015 / Revised: June 2015 / Accepted: July 2015)

ABSTRACT

Recognition of handwritten mathematical expressions has been an important topic for many researchers for decades. It remains one of the most challenging and exciting areas in pattern recognition. In the recognition process of offline handwritten mathematical expressions, segmentation is the most important process. Problems in ambiguities of identifying superscript and subscript in complex offline mathematical expressions remain one of the most important problem. To the best of our knowledge little work has been done in the segmentation of offline handwritten mathematical expressions with respect to superscript and subscript. In this paper an efficient segmentation technique for superscript, subscript and main characters within offline handwritten mathematical expressions has been proposed. This technique is based on the generation of predictions for superscript, subscript and main characters within handwritten mathematical expressions, which helps for the reconstruction of mathematical expressions during the recognition process with their spatial interrelationship. The proposed system was conducted as an experiment with a database of 300 samples of scanned mathematical expressions that comprised 2,000 symbols out of which there were 31 different types of Mathematical Symbols. The classification of the elements was carried out by the K-NN-classifier based on density features. This experiment shows remarkable results.

Keywords: Features extraction; K-NN classification; Mathematical Expressions (ME) Recognition; Segmentation

1. INTRODUCTION

Handwritten Mathematical Expression (ME) recognition has been a topic of intensive research in recent years due to its large number of applications A significant number of researchers have attempted to resolve the problem of ME recognition (Chan et al., 2000; Plamondon, 2000; Tapia & Rojas, 2007; Hu et al., 2014). The recognition of offline handwritten mathematical expressions differs from online handwritten recognition because it requires an explicit segmentation technique. The strategies for recognition of handwriting are based on segmentation techniques (Salim et al., 2007; Ha et al., 1998; Matan & Burges, 1991). The main challenge is to generalize segmentation techniques to accommodate a large set of mathematical expressions for recognition. Handwriting recognition is mainly related to optical character recognition.

* Corresponding author's email: sachin.naik@sicsr.ac.in, Tel. +91-20-25675601, Fax. +91-20-25675603
Permalink/DOI: http://dx.doi.org/10.14716/ijtech.v6i3.1069

A complete handwriting recognition system handles formatting, performs correct segmentation into characters, and recognizes words. Recognition of any handwritten characters with respect to any script is difficult since, the handwritten characters differ from person-to-person. In a mathematical expression, characters and symbols can be spatially arranged as a complex two-dimensional structure, possibly of different character and symbol sizes. These mathematical expressions contain different mathematical symbols in the following categories, like binding symbols, fence symbols and operator symbols.

Over the last two decades, researchers proposed many approaches to recognize mathematical expressions. In mathematical expressions, symbols are spatially arranged as a complex two-dimensional structure. Most of the proposed systems are for online mathematical recognition systems containing isolated symbols (Stefan et al., 1996). In very early work, an error-free symbol recognizer presented a coordinate grammar for 2D grammar recognition (Anderson, 1968). A prototype system was proposed that translated noise-free, typeset mathematical expressions into Lisp expressions using template matching based on a Hausdorff distance algorithm for recognising symbols (Fateman et al., 1996). A system was proposed for segmenting and understanding mathematical expressions within document in which connected components were considered as symbols for the use of directional feature symbol recognition (Anderson, 1968). A predictive algorithm for printed mathematical symbol segmentation is proposed based on multifactorial analysis that integrates several factors to determine the cut positions in touching characters (Garain & Chaudhuri, 2005). A robust and efficient system for recognizing typeset and handwritten mathematical notation using tree transformation is proposed, which presents a methodology and implementation of DRACULAE for rapid, robust recognition of typeset and handwritten mathematical expressions (Zanibbi et al., 2002). Many symbol recognition problems require the use of robust descriptors in order to obtain rich information about the data. However, the research for a good descriptor is still an open issue due to the high variability of the symbols' appearance. Rotation, partial occlusions, elastic deformations, intra-class and inter-class variations, or high variability among symbols due to different writing styles, are just a few problems (Escalera et al., 2009). In online ME recognition, conventional methods for symbol segmentation are used with respect to strokes, and the way of the handwriting style (Kenichi et al., 2004; Lehmberg, et al., 1996; Kanahori et al., 2000). However, these methods do not deal with tricky situations when incorrect stroke combination input is given as input (Kenichi et al., 2004). Another different approach in the recognition and retrieval of mathematical expressions based on four key problems includes query construction, normalization, and indexing and a relevant feedback method (Richard & Blostein, 2012).

In reviewing the above research work, it is summarized that most of work is for online simple handwritten expressions. In this area, the offline features can be adopted for online recognition of handwritten mathematical symbol recognition (Davila et al., 2014). The ambiguities and results in offline handwritten mathematical expressions can be improved by applying better segmentation techniques. This work is concerned with the complex offline, handwritten, as well as mathematical expressions containing superscripts, subscripts and special symbols. The recognition rate of ME is verified using proposed segmentation techniques along with K-NN classification. K-NN is a classification technique which is used to classify the components within mathematical expression based on nearest training set using feature space. In the next few sections of this paper, different stages of recognition process i.e., pre-processing, segmentation, feature extraction, K-NN classification and recognition are discussed thoroughly. In this paper a novel segmentation technique has been proposed that helps in improving the recognition rate for ME.

2. OVERVIEW OF THE PROPOSED SYSTEM

Figure 1 below diagrams the architecture of proposed system, which consists of processes like Input, Pre-processing, Segmentation, Feature Extraction, Classification, Symbol Recognition and Reconstruction of ME involved during each stage of the recognition process.

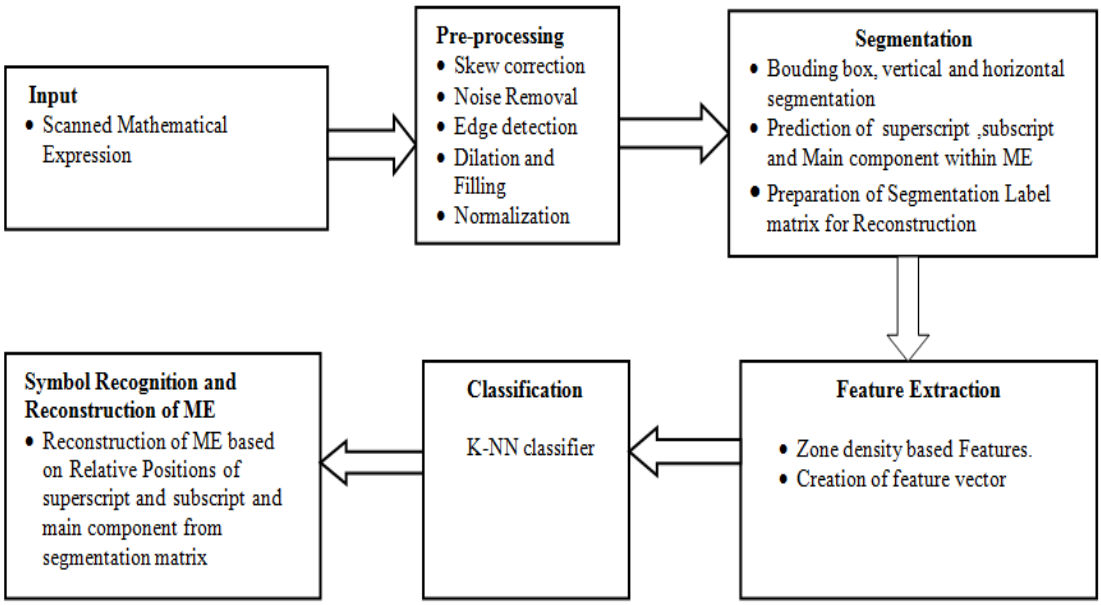

Figure 1 Proposed System for Handwritten ME

2.1. Pre-Processing

The key functionality of pre-processing is to improve the quality of image such that it increases the chances of success for further processes. In this experiment the pre-processing techniques are used for enhancing the contrast of the image, removal of noise and isolating the components of a mathematical expression which is of interest for further processing. A handwritten ME was scanned using a scanner at 600dpi and any one of the file formats like jpg, bmp etc. A scanned ME typically in a grey scale is given as an input for the pre-processing step. The pre-processing of image is carried out to reduce some undesirable variability that can have an effect on the recognition process. The different operations, like binarization, noise removal, edge detection, dilation and filling required to connect disconnected components of characters within images, smoothing, and normalization of randomly scanned images to standard size were performed. The binarization process converts the grey scale image into binary form using a global thresholding method by assuming two classes of pixels as foreground and background. In this procedure, the optimum threshold is calculated for separating two classes, keeping the intra-class variance to a minimum. The input images are taken from different persons of varying age groups so there is a large possibility of 'visual noise' occuring in the images, such as extra pen dots, lines or blurs in scanning the image. The noise removal technique removes such noise from the image. Noise removal identifies all the connected components of the image and removes all of them under a given size. Slant correction simplifies recognition by eliminating some of the natural variations of people's handwriting. Size normalization adjusts the character size to a certain standard. The goal of character normalization is to reduce the within-class variation of the shapes of the characters in order to facilitate the feature extraction process and improve their classification accuracy. This series of operations enhance the quality of image suitable for image segmentation.

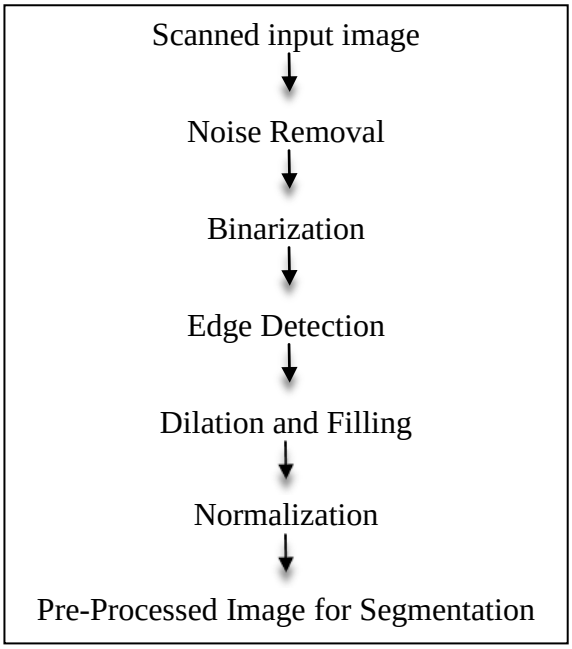

Figure 2 Preprocessing steps

2.2. Segmentation

This technique supports splitting of scanned mathematical expressions into sub-images, that is into individual characters so that individual characters within the mathematical expression are given as an input to further the recognition process (Lau & Mohan, 2013). Segmentation of a complex ME is a challenging task because of unconstrained handwritten expressions, overlapping and touching components, different character sizes, varied skew angles of characters and identification of spatial relations of symbols within mathematical expressions, which is one of the critical issues (Simistira et.al., 2014). The major activity involved in the segmentation process consists of separation of each component based on structural analysis. The technique proposed in this paper helps for predicting superscript components and subscript components of ME by measuring their spatial relationships. This technique combines bounding box and vertical and horizontal segmentation based on co-ordinates approximation for each component.

In this process the pre-processed input image is segmented into isolated characters and these isolated characters are labelled using a labelling process. The proposed algorithm for segmentation predicts the position of the components in the reconstruction phase. It also provides information about the number of components within the handwritten mathematical expression. A dynamic label matrix stores the labels for components within ME which consists of 3 rows and n columns, where the n value depends on the number of components within ME. This label matrix is useful for reconstruction of mathematical expressions.

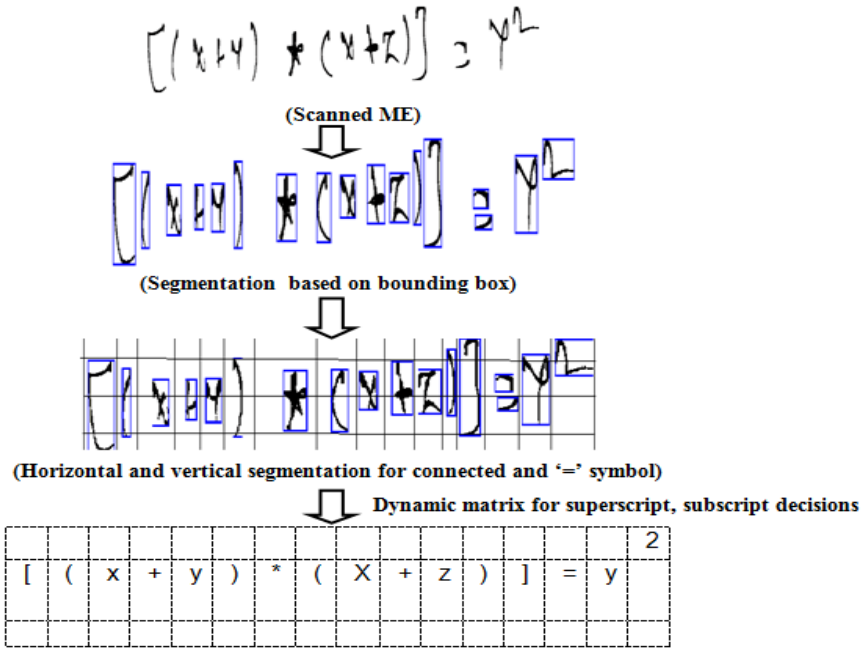

Figure 3 Experimental output of segmentation process

The predicted characters from the recognition process are mapped to the label matrix to identify the appropriate position of the components within a handwritten mathematical expression. The major steps involved in algorithm are as follows:

INPUT

Pre-Processed Handwritten Mathematical Expressions may consist of superscript and subscript components.

PROCESS

1. Apply the bounding box and vertical segmentation for connected components within ME
2. Compute the midpoint, top-left, top-right, Bottom-left, bottom-right location for each component.
3. Scan the first component, treat it as main character, go to step (5).
4. Scan second component,
   If (top_left (second_component) < top_right

   (first_component)) && (bottom_left (second_component) <=midpoint (first_component))

   Then first_component=superscript

   Go to Step (5)

   Else if top_left (second_component) >= mid_point (first_component) && bottom_left (second_component)>bottom_right (first_component)

   Then second_component=subscript

   Go to Step (5)

   Else If top-right (second_component) < mid_point (first_component) &&

Bottom_left (second_component) > mid_point (second_component)

Then second_component= Main character

Go to Step (5).

5. Locate the sequence of symbols in 3×n matrix at appropriate positions for the main character, superscript and subscript with label numbers.
6. Repeat the step (4) till end of expression.

OUTPUT

- 3×n matrix with appropriate location for superscript, subscript and main characters within HME.
- Segmented characters as sub-images within the Handwritten Mathematical Expressions

The above algorithm is applied for a variety of the 300 complex Mathematical Expressions which consist of 30 different types of mathematical symbols and numerals. The overall segmentation accuracy achieved using this approach is as follows:

Table 1 Segmentation result

No. Of Expressions	Total Number of ME Segmented	Segmentation Accuracy
300	294	98%

2.3. Feature Extraction

The selection of appropriate features is an important step in pattern recognition. In this proposed method, the segmented components of mathematical expressions are treated as a binary image, which is normalized to a nominal size of 48×48. The normalized segment is divided into 'n' equal zones where n=4, 9, 16 and 36, respectively, are considered for calculating the recognition rates. The density of the zone is computed by taking the ratio of total number of object pixels (i.e. pixels representing the numeral viz. binary 1) to the total number of pixels in the zone. This is carried out for all the zones in the image. A total of 65 features are extracted from each image and then for each image feature vectors are created. Based on the features vectors, then test sets and training sets are created.

In the equation below N is the number of object pixels in each zone Z, and T is total number of pixels in corresponding zone are computed as Equation (1).

$$\text{Density}(Z) = N / T \tag{1}$$

The steps involved in calculating the feature vector are as follows:

1. Segmented input image of size 48×48
2. Calculate density for n=4, 9, 16 and 36 which will provide 4, 9, 16 and 36 features.
3. Images for each zone are given below. For each image a feature vector is created, which is summation of all the features for zones 4, 9, 16 and 36. A total of 65 features are collected and a feature vector is formed.

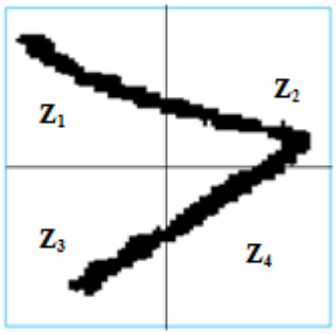

Figure 4(a) 4 Zone Feature space

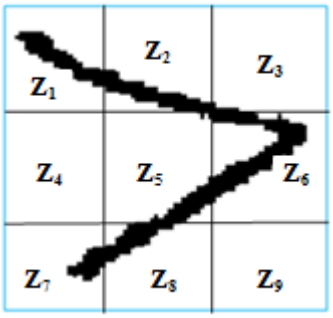

Figure 4(b) 9 Zone Feature space

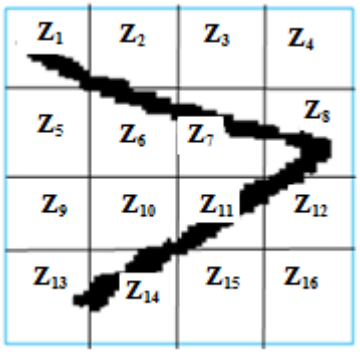

Figure 4 (c) 16 Zone Feature space

Figure 4 (d) 36 Zone Feature space

The feature vector constructed using 65 features from features space 4, 9, 16 and 36 are as shown in Equation (2).

$$\text{Feature Vector V [x]} = (\text{4 Zone Feature Space}) + (\text{9 Zone Feature Space}) + (\text{16 Zone Feature Space}) + (\text{36 Zone Feature Space}) \quad (2)$$

2.4. K-NN Classification

K-Nearest Neighbour (K-NN) algorithm method used to classify objects based on the closest training examples from feature space. Features extracted from each of the components from ME are sent to a classifier. The objective here is to interpret the sequence of components of ME taken from the test set. K-NN, which is a supervised learning algorithm where the result of a new instance-based object is classified based on majority of K-Nearest Neighbour category. In this way an object is classified by its distance from neighbours, with the object being assigned to the class most common among its *k* (a small positive integer) distance nearest to its neighbours using Euclidean distance. K-NN takes a test instance *x*, finds its *k*-nearest neighbours in the training data, and assigns *x* to the class occurring most often among those *k* neighbours.

2.5. Symbol Recognition and Reconstruction of ME

To evaluate the proposed approach, we have created a database consisting of complex mathematical expressions from 30 writers. Each writer was requested to write 10 different expressions which consist of superscript, subscript and special characters. Each expression contains 3 to 17 symbols. As a result, 300 MEs were collected and used in the experiments. These MEs consists of 28 different types of symbol components. In collecting handwritten ME samples, we observed and verified the writer's natural way of writing. Expressions are written in the writers' own style and pace freely. There are no restrictions imposed on writers about writing of MEs.

Table 2 Symbols used in ME

Type of Symbol	Samples used in ME
Digits	0,1,2,3,4,5,6,7,8,9
Alphabets	x,y,z
Special symbols and operators	- ± = * ≤ ≥ ≠ × . / > <
Parenthesis	[ ] ( ) { }

The overall recognition accuracy of complex MEs is 94.30%. As reconstruction is an important step for recognizing mathematical expressions, the appropriate position of label numbers from the segmentation matrix is used for identification of superscript, subscript and main components.

Table 3 Symbol recognition result

Total Number of Symbols and characters within HME	Recognition using K-NN Classification	Misinterpreted	Recognition Result
2000	1886	114	94.30%

The proposed steps for reconstruction of handwritten mathematical expressions are given below:

1. Scanned handwritten mathematical expressions as an input.
2. Apply pre-processing, segmentation, feature extraction and classification as proposed in sections 2.1, 2.2, 2.3, 2.4.
3. Use the predicted symbol and characters within handwritten mathematical expressions from classification and maps to the appropriate positions within the matrix created at the segmentation.
4. Display reconstructed handwritten mathematical expressions with the help of position numbers given in the 3×n segmentation matrix.

3. CONCLUSION

In this paper a novel segmentation technique for complex MEs based on their spatial interrelationships is presented. This method consists of a predictive approach for determining the superscript and subscript components of MEs. Based on the position of MEs a predictive matrix is prepared which is useful for reconstruction of mathematical expression. Classification was carried out using K-NN classification with density-based features. The proposed method is independent of font styles. The main recognition errors were due to ambiguity among similar shaped characters, operators and numerals. The proposed system is tested over a simple and complex ME which consists of superscript and subscript. Overall recognition rate for 31 different kinds of characters, numerals and operators is being achieved 94.30%.

Future work includes verifying the variety of mathematical expressions from categories like algebraic equations, logarithmic functions, hyperbolic functions, geometric formulas, limits and derivative formulas, which can be used for experimentation in conjunction with the proposed segmentation techniques and the results will be verified using a combination of different classifiers to improve the recognition accuracy.

4. REFERENCES

Anderson, R.H., 1968. *Syntax-directed Recognition of Hand-printed Two-dimensional Mathematics*. Ph.D. Dissertation, Dept. Eng. Appl. Phys., Harvard Univ., Cambridge, MA

Chan, K-C., Yeung, D-Y., 2000. Mathematical Expression Recognition: A Survey. *IJDAR* 3: pp. 3–15, Springer-Verlag

Davila, K., Ludi, S., Zanibbi, R., 2014. Using Off-line Features and Synthetic Data for On-line Handwritten Math Symbol Recognition. *Proceeding The 14th International Conference on Frontiers in Handwriting Recognition* (ICFHR)

Escalera, Sergio, Fornés, Alicia, Sánchez, Gemma, Lladós, Joseph, 2009. Blurred Shape Model for Binary and Grey-level Symbol Recognition. *Pattern Recognition Letters*, Volume 30(15), pp. 1424–1433

Fateman, R., Tokuyasu, T., Berman, B., Mitchell, N., 1996. Optical Character Recognition and Parsing of Typeset Mathematics. *Journal of Visual Communication and Image Representation*, Volume 7(1), pp. 2–15

Garain, U., Chaudhuri, B.B., 2005. Segmentation of Touching Symbols for OCR of Printed Mathematical Expressions: An Approach based on Multifactorial Analysis. *Proceedings of the Eight International Conference on Document Analysis and Recognition* (ICDAR'05), IEEE.

Ha, T.M., Zimmermann, M., Bunke, H., 1998. Off-line Handwritten Numeral String Recognition by Combining Segmentation-based and Segmentation-free Methods. *Pattern Recognition*, Volume 31(3), pp. 257–272

Hu, Y., Peng, L., Tang, Y., 2014. Online-handwritten Mathematical Expression Recognition Method based on Statistical and Semantic Analysis, Document Analysis System (DAS), *The 11th IAPR International Workshop*, pp. 171–175, IEEE

Kanahori, T., Tabata, K., Cong, W., Tamari, F., Suzuki, M., 2000. *On-line Recognition of Mathematical Expressions using Automatic Rewriting Method*, ICMI, LNCS 1948, Springer, pp. 394-401

Kenichi, T., Naoya, Y., Kenji, M., Takayuki, K., Kensaku, M., Yasuhito, S., Tomoichi, T., 2004. A Study of Symbol Segmentation Method for Handwritten Mathematical Formula Recognition using Mathematical Structure Information. *Proceedings of the 17th International Conference on Pattern Recognition* (ICPR, 2004), pp. 630–633

Lee, H.J., Wang, J.S., 1995. Design of Mathematical Expression System. In: *Proceedings of ICDAR'95*, Canada, pp. 1084–1087

Lu, C., Mohan, K., 2013. Recognition of Online Handwritten Mathematical Expressions, *Project Final Report*, Stanford University

Matan, Burges, J.C., 1991. Recognizing Overlapping Hand-printed Characters by Centered-objects Integrated Segmentation and Recognition. In: *Proceedings of International Joint Conference on Neural Networks* (IJCNN), pp. 504–511, Seattle, USA

Plamondon, R.J., Sargur, N.S., 2000. On-line and Off-line Handwriting Recognition: A Comprehensive Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 22(1), pp. 63-84

Salim, O., Mouldi, B., Abderrazak, L., 2007. Segmentation and Recognition of Handwritten Numeric Chains. *Journal of Computer Science*, Volume 3(4), pp. 242–248

Simistira, Papavssiliou, Katsouros, Carayannis, 2014. Recognition of Spatial Relations in Mathematical Formulas. *Proceeding The 14th International Conference on Frontiers in Handwriting Recognition* (ICFHR),pp. 164–168, IEEE

Stefan, L., Winkler, H.J., Lang, M., 1996. A Soft-decision Approach for Symbol Segmentation within Handwritten Mathematical Expressions. In: *Proceeding International Conference on Acoustics, Speech, and Signal Processing* (ICASSP), pp. 3434–2437

Tapia, E., Rojas, R., 2007. *A Survey on Recognition of Online Handwritten Mathematical Notation Freie University at Berlin*. Institute fur Informatik,Germany

Zanibbi, R., Blostein, D., 2012. Recognition and Retrieval of Mathematical Expressions. *IJDAR* 2012, Springer-Verlag

Zanibbi, R., Blostein, D., James, R.C., 2002. Recognizing Mathematical Expressions using Tree Transformation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 24(11), pp. 1455-1467