Mathematical Symbol Extraction From Document Images - A Comprehensive Review

A. SAKILA, Dr. S. VIJAYARANI

Abstract: The global effects of high speed internet access as hundreds of millions browse for information/multimedia, look up map directions, interact through email/social networks/ video chat, etc. Nowadays document images play a vital role in digitized organization and digitized libraries. Digitized means paper documents are converted into image format by using digitized equipment's. Optical Character Recognition (OCR) is a one of the document image analysis technique, which is used to convert document image into editable text format. Mathematical document identification is a unique challenge in document image analysis that deals with identifying mathematical symbols in a document and then classifying the document as math's and non-math's regions based on density of the mathematical symbols. Formulas are involved in mathematical documents, either as isolated formulas, or embedded directly into a text line. They have a number of features, which distinguish them from conventional text. This paper provides the basic concepts of the mathematical symbol recognition and its essential characteristics.

Index Terms : Document Images, Document Image Analysis, Optical Character Recognition (OCR), Mathematical Symbol Extraction.

1 INTRODUCTION

Document Image Analysis (DIA) is an essential and challenging research domain in the field of computer science. Information retrieval from the document images is a difficult to perform, hence number of techniques and procedures are used for document image analysis [16]. Extracting information from the document images is challenging problem as it compared with digital texts [12] [13]. The present scenario, makes the world's knowledge and information more accessible than before. In the classical OCR system [3], a mechanical apparatus used to evaluate the incidence of light reflected back from a printed character when illuminated through a set of character templates. A character detection would occur when the light emitted from the template overlapped the character (assumed to be in dark print) sufficiently to prevent light from being reflected upon the medium. To identify and recognize the text from document images, it is essential to first carry out document layout analysis techniques which will determine how the document is partitioned. The text will be recognized with an understanding of where the columns of text are, which portions of text indicate headings or quotes, and which segments correspond to images, tables, captions, etc. If the text is not partitioned appropriately prior to recognition then the textual output will become unpredictable. There are different DIA techniques are available for improving the efficiency of OCR system. Mathematical Symbol Recognition is an interesting and tedious research area in the field of DIA. In math recognition, the key problems are difficult to detecting the symbols and equation, because there are number of symbols are used to construct the equation with different shapes. However, most of the techniques focus on mathematical formulas themselves and do not recognize the whole mathematical document [4].

- *Author A. Sakila is currently pursuing Ph.D in Computer Science at Bharathiar University, Coimbatore, Tamilnadu, India. E-mail: sakivani27@gmail.com*
- *Co-Author Dr. S.Vijayarani is currently working as Assistant Professor in Department of Computer Science at Bharathiar University, Coimbatore, Tamilnadu, India. E-mail: vijimohan_2000@yahoo.com*

It is a very mature area of research, development in this area continues in order to increase recognition support for the broad spectrum of languages, formats, and subject matter of printed documents. The autonomous recognition of all printed documents would not only expedite the global advancement of knowledge and wisdom, but would also have tremendous implications toward every individual in society.The remaining portion of the paper is organized as follows. Section II presents the related works of the mathematical symbol recognition. Section III gives the classification techniques mathematical symbol recognition. Research issues and challenges are discussed in Section IV. Section V gives the conclusion.

2 RESEARCH ISSUES AND CHALLENGES IN MATHEMATICAL SYMBOL RECOGNITION

Mathematical Symbol Recognition (MSR) is more complicated than normal OCR, because difficult to identify the mathematical expressions from document image [7] [8]. Mathematical formulas themselves and do not recognize the whole mathematical document. Issues and challenges in mathematical symbol recognition are described below.

- The mathematical symbols recognition is not an easy task because it contains both characters from the English alphabets and letters from Latin and Greek. In addition, it also contains the numerals, as well as other symbols. Hence document mathematical symbols recognition is a major area of concern [6].
- Generally characters from document images are consecutively recognize from left side to right side. Instead of Math's expressions all the symbols are not horizontally written. This issue is due to the problem of the relationship between symbols.
- Some symbols for building formulae have different shapes in different situation but keep same meaning [9]. In these cases the established method of dictionary databases has been considered for recognizing mathematics symbols.
- In math's formulae the meaning of the different elements depends on their shape and spatial occupied position [7]. These attributes create difficulties in the building of good classifiers for mathematical symbols.

3 RELATED WORKS

S. No	Author	Paper Title	Objective	Advantage	Disadvantage
1.	Jacob R. Bruce	Mathematical Expression Detection and Segmentation in Document Images	Proposed novel approach to mathematical expression detection and segmentation (MEDS) during the document layout analysis stage of OCR. The focus of this work is to enhancing the OCR quality.	Improved italics and bold detection used to make the detector more robust. The italics and bold detection implemented as part of Tesseract was used in feature extraction for this work but was found to not be very accurate and was thus not used to train the final classifier.	Another important future work item is to generate more data. The current amount of pages, 75, is a very small amount of data, and makes it difficult to get a truly objective understanding of the results.
2.	A. Kacem, A. Belaïd M. Ben Ahmed	Automatic extraction of printed mathematical formulas using fuzzy logic and propagation of context	Proposed a method that segmented the printed mathematical documents and automatically extract the formulas from the images	This method performed based on local and global segmentation. Some mathematical symbols are identified by existing models using fuzzy logic.	In this proposed method achieved acceptable recognize ratio for mathematical formula extraction.
3.	Xiaoyan Lin, LiangcaiGao, Zhi Tang, Xiaofan Lin, Xuan Hu	Mathematical Formula Identification in PDF Documents	Proposed hybrid method for formula identification in PDF documents.	The hybrid method is combined both rule-based approach and learning-based approach for improve the performance. The experimental results quite satisfactory than traditional rule-based and learning based methods.	This hybrid method only identifies the formulas from PDF document, but not converted into editable format.
4.	IffathFathima S, Ashoka K	Machine Learning Approach for Recognition of Mathematical Symbols	To analyze the efficiency of existing Support Vector Machines (SVM) and K-Nearest Neighbor (KNN) for symbol recognition.	KNN classifier gave the great accuracy compared to SVM classifier. The efficiency of KNN to increase the performance of symbol extraction.	They only recognize the document and symbol recognition from document images, not concentrated superscript and subscript from the formula.
5.	AzadehNazemi, Iain Murray	Mathematical Formula Recognition and Transformation to a Linear Format Suitable for Vocalization	Proposed a method for identified mathematical formulae then transforms into text and then LaTeX format.	The proposed method recognized math's formulae from document images and converts into LateX format.	The conversion results not give 100% accurate result.
6.	A. Kacem, A. Belaïd M. Ben Ahmed	EXTRAFOR: automatic EXTRAction of mathematical FORmulas	Proposed a method for automatic extraction of mathematical formulas from images of documents without character recognition.	They used fuzzy logic at CCXs labeling which has been useful to identify symbols and consequently to delimit formulas by a contextual analysis of their CCXs. They also used more complex alignment symbols of formulas and confirming the efficiency and the performance of our method for a large data base of mathematical formulas.	The CCXs labeling technique is only recognized the limited number of symbols.

4 MATHEMATICAL SYMBOL EXTRACTION

The mathematical symbol extraction system has required following steps. They are noise removal, binarization, segmentation, feature extraction and classification. Figure 1 shows the textual document image, it only text information. In Figure 2 displays the symbols and equations in the form of image.

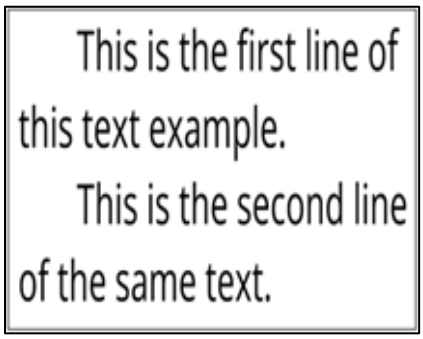

Figure 1 *Document Image with Text*

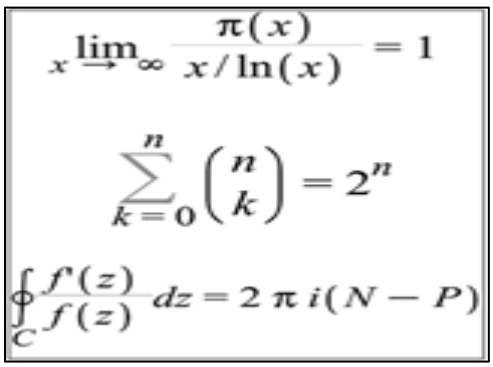

Figure 2 *Document Image with Symbols*

4.1 Preprocessing

This stage involves preparing/cleaning and enhancing the document images by resolving problems like illumination, low resolution aging effects, humidity, marks, fungus and stains. Hence there is a need for images enhancement techniques like de-noising and document images.

4.1.1Noise removal

The raw document images are often degraded by noises and it is a random variation of image Intensity and visible as a part of grains in the image. Noise can be produced during scanning the documents, capturing the images, transmission, etc. Noise reduction is a significant aspect of image quality and it is also called as image smoothing. A color image will be converted to a gray image before proceeding with the noise removal procedure. The de-noised image is then transformed into a binary image with suitable threshold. Different noise removal techniques are available for removing the noise from document images. Some of the traditional noise removal techniques are mean (linear), median (non-linear), average, Gaussian, wiener and adaptive.

4.1.2Binarization

Binarization is used to improve the quality of the document image and it converts the document image text information (foreground) as black color and the background as white color. This method is very important step and challenging research problem in Document Image Analysis and document recognition system. Image thresholding is an essential for binarization [7], this task of thresholding is to extract the foreground (ink) from the background (paper). There are two different types of binarization, they are global threshold and local threshold. The global thresholding method is very popular for many document image analyses; it uses a single threshold for the entire image [8]. A local threshold method calculates a different threshold value for each pixel. Otsu, Isodata and Iterative Global thresholding are considered as a global threshold method. Bernsen's, Niblack and Sauvola / Pietikäinen's, Wolf's and Nick are local threshold method.

4.1.3 Skew detection and correction

Skew detection and correction is a pre-processing step for character recognition system. It is a difficult research problem in document image recognition system. Skew can be classified into two different types one is page skew another one is handwritten skew. Normally page skew is produced by scanning process. Some writers are cannot write straightly and some people put their signature crossly. Numerous techniques have been proposed as alternatives for skew angle detection of document images. Some of the skew detection techniques are Hough Transformation, Cross Correlation, K - Nearest Neighbour and Fast Fourier transformation.

4.2 Segmentation

Segmentation is necessary for character recognition system, it segments the document image into paragraphs, text lines, words and lastly characters. Normally document images only contains text information, hence it can be segmented easily. Figure 3 depicted the line segmentation. Instead of document images contains mathematical symbols has combination of subscript, super script, fraction, radical, integral, accent, matrix, operator, trigonometric function and large operators (summation, product, union, intersection), hence the equations are difficult to segment. Figure 4 depicted the difficulties in line segmentation, hence, symbols are not properly segmented.

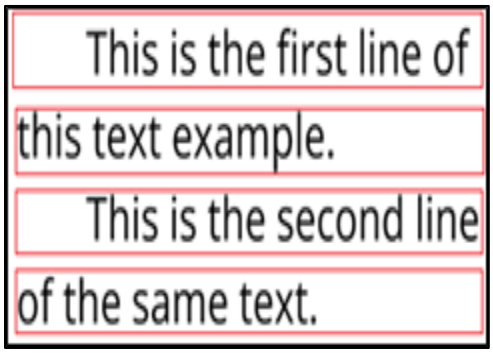

Figure 3 *Sample Text line segmentation*

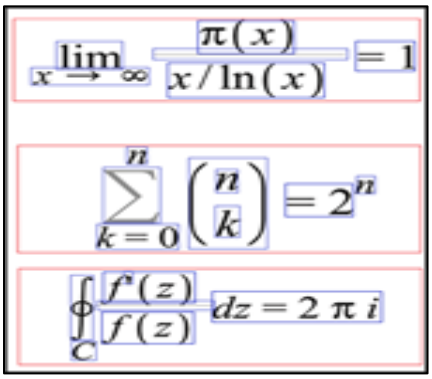

Figure 4 *Sample Document Image with symbols*

4.3 Classification

Classification is a well-established concept for organizing information and knowledge from document image. It is used to analyze the input images and identifying each character then translates the text images into character codes. Figure 1 shows the Sample Document Image which consist of both text and mathematical symbols / equations.

4.3.1 Hidden Markov Model (HMM)

The Hidden Markov Model (HMM) is finite set, these states are associated with (generally multidimensional) probability distribution. Transition probabilities name implies the states are ruled by set of probabilities. The observation and/or outcome of particular state can be generated rendering to

accomplice probability distribution. In HMM the outcome state is not visible to external observer, hence the outside states are "hidden". HMM is plays significant role in Mathematical symbol recognition and OCR system.

4.3.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a popular method for supervised learning method, which is used to document image classification. It is mainly involved in separating data into testing and training sets. SVM is to create the model based on training set, to predict the target value of the test data SVM given only the test data attribute. SVM classifier is used two important distinct classes for predict and classify the document image for recognition. SVM used kernel functions for classification, it may be differentand they are linear kernel, Gaussian Radial Basis Function (RBF), Polynomial kernel and Sigmoid (hyperbolic tangent).

4.3.3 K-Nearest Neighborhood (KNN)

In machine learning, K-Nearest Neighbors (KNN) is a most popular and simple classification algorithms. It comes under the supervised learning domain and finds powerful application in pattern recognition, data mining and intrusion detection. KNN does not create any basic assumption about the data distribution, because it is a non-parametric. It performed based on two factors, which is suitable value for the parameter K and suitable similarity function. In the KNN takes less execution time and easy to interpretation. The KNN is a familiar for classify the characters from document images.

5 Conclusion

Information retrieval (IR) from document image is tedious task and it is plays major role in Document Image Analysis (DIA). Optical Character Recognition (OCR) is a one of the document image analysis technique, which is used to convert document image into editable text format. It only converts textual contents from document images, instead of symbols from document images, OCR does not produce the results. Mathematical Symbol Recognition and retrieval is a challenging and interesting research area in DIA. There is no tools are available for recognition the symbols from document images. This paper has reviewed basics concepts of mathematical symbol recognition. Different approaches in mathematical symbol recognition and it research issues and challenges are also discussed in this paper.

6 ACKNOWLEDGEMENTS

The authors thank the University Grants Commission (UGC), New Delhi (Official Memorandum No.F1-17.1/2016/RGNF-2016-17-SC-TAM-19477) for the financial support under Rajiv Gandhi National Fellowship for this research work.

7 REFERENCES

[1] Jacob R. Bruce, Mathematical Expression Detection and Segmentation in Document Images.

[2] H. F. Shantz, the History of OCR, Optical Character Recognition. Manchester Center: Recognition Technologies Users Association, 1982.

[3] P. W. Handel, "Statistical Machine," United States Patent Office. 1,915,993, Jun, 27, 1933.

[4] Kacem, A. Belaïd and M. Ben Ahmed "Automatic extraction of printed mathematical formulas using fuzzy logic and propagation of context"

[5] Xiaoyan Lin, LiangcaiGao, Zhi Tang, Xiaofan Lin, Xuan Hu, "Mathematical Formula Identification in PDF Documents", 2011 International Conference on Document Analysis and Recognition.

[6] IffathFathima S and Ashoka K, "Machine Learning Approach for Recognition of Mathematical Symbols", International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882, Volume 6, Issue 8, August 2017.

[7] AzadehNazemi and Iain Murray, "Mathematical Formula Recognition and Transformation to a Linear Format Suitable for Vocalization", International Journal on Computer Science and Engineering (IJCSE), ISSN: 0975-3397, Vol. 5 No. 09 Sep 2013.

[8] G,Erik . Miller .A. Paul, "Ambiguity and Constraint in Mathematical Expression recognition". Viola Massachusetts Institute of TechnologyArtificial Intelligence Laboratory545 Technology Square, Office 707Cambridge, MA 02139. In Proceedings of the 15'th National Conference on Artificial Intelligence (AAAI-98)

[9] P,Garcia. B, Couasnon, "Using a Generic Document Recognition Method for Mathematical Formulae Recognition". IRISA / INSA-D´epartement Informatique20, Avenue des buttes de Co¨esmes, CS 14315F-35043 Rennes Cedex, France, Graphics Recognition Algorithms and Applications Lecture Notes in Computer Science Volume 2390, 2002, pp 236-244.

[10] Xuedong Tian, RuihanBai, Fang Yang, JinyuanBai, Xinfu Li, "Mathematical Expression Extraction in Text Fields of Documents Based on HMM" Journal of Computer and Communications, 2017.

[11] Kacem, A. Belaïd and M. Ben Ahmed, "EXTRAFOR : automatic EXTRAction of mathematical FORmulas"

[12] Simone Marinai, "A Survey of Document Image Retrieval in Digital Libraries", 9th colloque International Francophone Sur l'Ecritet le Document (CIFED)-2006, pp. 193–198.

[13] Dr. S. Vijayarani and A. Sakila, "A Survey on Word Spotting Techniques for Document Image Retrieval" International Journal of Engineering Applied Sciences and Technology, Vol. 1, No. 1, Dec 2016.

[14] Francisco Álvaro, Joan Andreu Sanchez," Comparing Several Techniques for Offline Recognition of Printed Mathematical symbols", 2010 International Conference on Pattern Recognition.

[15] Mou-Yen Chen, AmlanKundu and SargurN. Srihari, "Variable Duration Hidden Markov Model and Morphological Segmentation for Document Word Recognition", IEEE transactions on image processing, vol. 4, no. 12, December 1995.

[16] Nawei Chen, Dorothea Blostein "A survey of document image classification: problem statement,classifier architecture and performance evaluation"1 June 2004.