Performance Evaluation of Mathematical Formula Identification

Xiaoyan Lin[1], Liangcai Gao[1], Zhi Tang[1,2]
[1]Institute of Computer Science and Technology, Peking University, Beijing, China
[2]State Key Laboratory of Digital Publishing Technology, Beijing, China
{linxiaoyan, glc, tangzhi}@ pku.edu.cn

Xiaofan Lin
Vobile Inc
Santa Clara, California, USA
xiaofan@vobileinc.com

Xuan Hu
College of Software
Beihang University
Beijing, China
huxuan@sse.buaa.edu.cn

***Abstract*—This paper presents a performance evaluation system for mathematical formula identification. First, a ground-truth dataset is constructed to facilitate the performance comparison of different mathematical formula identification algorithms. Statistics analysis of the dataset shows the diversities of the dataset to reflect the real-world documents. Second, a performance evaluation metric for mathematical formula identification is proposed, including the error type definitions and the scenario-adjustable scoring. The proposed metric enables in-depth analysis of mathematical formula identification systems in different scenarios. Finally, based on the proposed evaluation metric, a tool is developed to automatically evaluate mathematical formula identification results. It is worth noting that the ground-truth dataset and the evaluation tool are freely available for academic purpose.**

Keywords-Mathematical formula identification; performance evaluation; ground truth; evaluation metric

I. INTRODUCTION

Mathematical formula identification is a critical step in mathematical notation recognition and scientific document management. It aims at detecting and segmenting mathematical formula regions from the document pages. In the past decade, a number of mathematical formula identification algorithms have been reported [1-3, 10-13]. However, there are few direct performance comparisons of the different methods due to the following obstacles in ground-truth dataset, evaluation metric, and automatic evaluation methods:

1) Ground-truth dataset.

Most of the reported mathematical formula identification approaches are evaluated on their own datasets, which are application specific and unavailable for other researchers. To solve this problem, some datasets were proposed [5-6]. As an early effort in this regard, UW-III [5] was proposed to evaluate layout analysis in document images. There are only 25 document pages containing 100 mathematical formulas in this dataset. Obviously, the scale is too small to be representative of the mathematical documents in the real world. In recent years, Suzuki et al. [6] presented Infty dataset for performance evaluation of mathematical formula recognition. However, the areas of mathematical formula are not given directly. Garain et al. [7] presented a corpus for mathematical expression recognition with in-depth statistical analysis. Also, Ashida et al. [8] presented a large-scale dataset for mathematical formula recognition. Unfortunately, both of these two datasets are not available to the public.

All the reported datasets for mathematical formula recognition target scanned document images. In addition, most documents in the existing datasets were published too early to obtain the corresponding source PDF documents. As a result, for mathematical formula recognition methods focused on PDF documents [9-11], it is difficult to compare the performance directly with image-based methods.

2) Performance evaluation metric.

Currently, the most widely used metric for mathematic formula identification is *precision* and *recall*. Although this evaluation metric is straightforward, there are some drawbacks: *a)* It treats different types of errors equally while the real-world penalties of different error types vary significantly. In general, a missed symbol in an equation is much better than the whole equation being misrecognized. *b)* The developers of the algorithms can get very few clues to improve the algorithms using such simplistic benchmark. *c)* In reality, mathematical formula identification system would be applied in specific contexts. The existing performance metric is not capable of adapting to different scenarios. In other words, it cannot tell the strengths and weaknesses of applying the algorithm to a particular context.

In [3, 12-13], evaluation metrics are proposed to evaluate different types of formula identification results, including *perfect*, *partial*, *wrong* and *missing*. Actually, mathematical formula identification results are more complex than those four cases. For instance, the *partial* result is possibly caused by erroneously splitting or merging the true formula regions. And this metric is not able to tell exactly which case occurs.

3) Automatic evaluation tool.

Another difficulty in formula identification performance evaluation is the lack of automatic evaluation tools. It is too costly to evaluate manually after the algorithm or the thresholds are modified. As a result, the developers cannot get objective evaluation after improving the system. Automatic evaluation ties closely to the construction of ground truth and evaluation metric. It would be easier to develop a tool to evaluate performance automatically if the ground-truth dataset is established, the ground truth format is well defined to be automatic parsing, and an appropriate evaluation metric is accepted by the community. Unfortunately, to our best knowledge, no automatic evaluation tool is available in formula identification performance evaluation up to now.

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.68

To address the aforementioned obstacles, this paper presents a performance evaluation system for mathematical formula identification. The contributions of this paper are as follows: *a)* A ground-truth dataset for mathematical formula identification is proposed. Statistics analysis shows the wide coverage of the dataset. Besides, digitally originated PDF files and the corresponding scanned images are included in the dataset to facilitate the comparison among both image-based and PDF-based approaches. *b)* A detailed evaluation metric for mathematical formula identification is proposed, which not only quantizes the significance of different categories of errors, but also can be adapted for different application scenarios. The evaluation metric is calculated from different perspectives. Both area-based and symbol-based evaluation metrics are defined. *c)* An automatic evaluation tool for mathematical formula identification is developed, based on the proposed ground truth format and the evaluation metric. It is worth noting that the ground-truth dataset and the evaluation tool presented in this paper are freely available for academic purpose.[1]

The proposed performance evaluation system comprises three components, which would be discussed in the following three sections. First, a ground-truth dataset is constructed for evaluating mathematical formula identification methods. Second, the evaluation metric is defined to classify and quantify different errors, and an overall performance scoring scheme is proposed to adapt to different application scenarios. Third, a tool is developed to evaluate the performance of mathematical formula identification methods automatically.

The rest of the paper is organized as follows. Section II presents the ground-truth dataset. Performance evaluation metric and automatic evaluation tool are introduced in Section III and Section IV respectively. We draw conclusions and discuss future work in Section V.

II. GROUND-TRUTH DATASET

A. Organization

To build a dataset which can reflect documents in the real world, we collect documents through crawling PDF documents from CiteSeerX. More than 1,000 PDF documents are crawled. The dataset is intended for non-commercial academic usage, and for any redistribution or commercial usage permission should be obtained from the copyright owners of the individual documents. Among these documents 194 digitally originated PDF documents are manually selected to construct the dataset. When selecting the documents, several factors are considered, such as the publication year and the page layout. Further details about the selection will be discussed in the second part of the section. For each document, at least one and at most eight pages containing mathematical formulas are selected. In total, the dataset contains 400 document pages with 1,575 isolated formulas, and 7,907 embedded formulas.

To facilitate comparison of approaches targeting both scanned document images and PDF documents, the dataset includes not only digitally originated PDF files, but also their corresponding document images. The document images are rendered from the corresponding PDF documents at 500 dpi.

B. Statistical Analysis

To construct a representative dataset with wide coverage, the following aspects are considered:

1) Source. Document pages are collected from different types of documents, including journals, conference proceedings, books, and technical reports, etc. The distribution of each source type is shown in Table I.

2) Publication year. The page layout of documents and the mathematical formula typesetting have evovled a lot over time. Taking this into account, documents with publication years ranging from 1977 to 2010 are selected.

3) Domain. Documents in the dataset are selected from different domains, including computer science, mathematics, biology, and physics.

4) Page layout. The page layouts of documents in the proposed dataset are diverse. 65 percentages of documents are single-column and the remaining are multi-column.

5) Producer and PDF version. For approaches targeting at digitally originated PDF documents, the mathematical symbol parsing process would vary significantly with different PDF producers and versions. To evaluate the robustness of a method dealing with different PDF producers and versions, PDF documents generated by different producers and in different versions are included in the dataset. Distribution of the producers and versions of the PDF documents are illustrated in Table II and Table III.

6) Frequency of formulas. Document pages with different mathematical formula frequencies are selected in the proposed dataset. The number of isolated and embedded formulas in each document page is counted separately. The statistics on the frequecies is shown in Table IV.

TABLE I. DISTRIBUTION OF THE SOURCES

Source	Conference	Journal	Book	Report	Others	Total
Count	93	73	7	16	5	194

TABLE II. DISTRIBUTION OF THE PRODUCERS

PDF Producer	Count	PDF Producer	Count
AFPL Ghostscript	15	Dvipdfm	18
Acrobat Distiller	66	Dvips	4
Acrobat PDFWriter	10	PdfTeX	27
ESP Ghostscript	23	Others	12
GNU Ghostscript	10		
MiKTeX pdfTeX	9	Total	194

TABLE III. DISTRIBUTION OF THE PDF VERSIONS

PDF Version	1.1	1.2	1.3	1.4	1.5	1.6	Total
Count	1	64	66	64	1	8	194

TABLE IV. FREQUENCY OF MATHEMAICAL FORMULAS

Number of isolated formulas per page						
Range	[0, 3)	[3,6)	[6, 10)	[10, 15)	[15, +∞)	Total
Count	154	152	71	18	5	400
Number of embedded formulas per page						
Range	[0, 10)	[10, 20)	[20, 40)	[40, 60)	[60, +∞)	Total
Count	115	122	117	35	11	400

[1] http://www.founderrd.com/marmot_data.htm

C. Representation

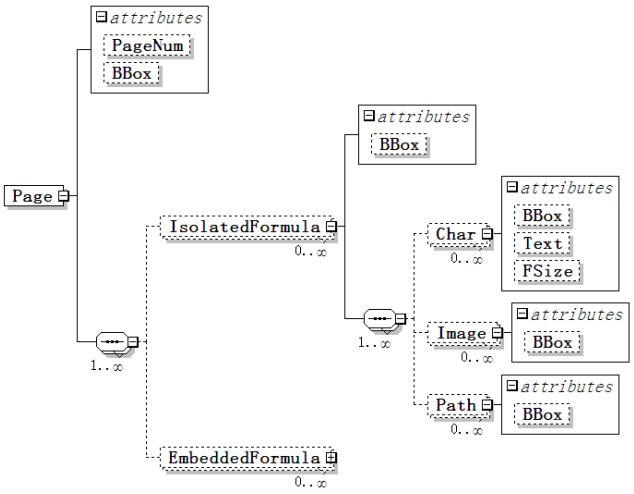

Figure 1. XML schema of ground truth (generated by *XMLSpy*)

The ground truth for each document page is an XML file, and the XML schema is shown in Fig. 1. The ground truth for each document page is stored within the <page> tag pair. The page number and the bounding box of the page are stored as the attributes named "PageNum" and "BBox".

Each isolated mathematical formula is stored within a pair of <IsolatedFormula>. The bounding box of the isolated formula is represented as the attributes named "BBox". The bounding box mentioned in this paper is the precise bounding box, which is represented by the coordinates of the top left corner and the bottom right corner.

The objects in the isolated formula are represented as the children of the <IsolatedFormula> tag pair. The objects parsed from the PDF documents include characters, graphics, and image objects. They are presented as <Char>, <Path>, and <Image> tag pairs respectively. For each object, its bounding box is stored as the attribute of the tag pair, named "BBox". For character objects, its Unicode and font size, named "Text" and "FSize", are stored as the attributes of <Char> tag pair.

The ground truth format of the embedded formula is similar to that of the isolated formula. Embedded formulas labeled in our ground truth dataset include the following three types of math notations: *a)* mathematical expressions in two-dimensional structure, such as subscript and fraction; *b)* named math functions and user-defined math functions; *c)* mathematical expressions with explicit math symbols, such as relation/operation operators and Greek letters.

The ground truth of our dataset is created through a semi-automatic ground-truthing tool named "*Marmot*", which will also be publicly available in near future.

III. Evaluation Metric

A. Identification Errors

To clearly reflect the strengths and weaknesses of an algorithm, it is necessary to distinguish different recognition result types. In the document page segmentation domain, errors are commonly categories as merging, splitting, and missing, and this categorization has been adopted in quite a few works on page segmentation evaluation. Formula identification is to detect and segment formula regions from the document pages, and it can be considered as a specialized page segmentation problem. After observing a large number of formula identification results, we refine the common result types of page segmentation into the following eight situations, which are also illustrated in Fig. 2:

For each page, let R_i denotes the i-th region in the identification formula region set of a page, and let G_j denotes the j-th region in the ground truth formula region set.

1) Correct. The detected region matches exactly one ground trouth formula region, as shown in Fig. 2-1. Let N_{cor}^{R} be the number of recogized regions belonging to this situation.

2) Missed. For a ground truth formula region, there exists no detected region to overlay it, as shown in Fig. 2-2. Let N_{mis}^{G} be the number of the regions in ground-truth dataset belonging to this situation.

3) False. The detected region does not overlay any region in ground-truth dataset at all, as shown in Fig. 2-3. Let N_{fal}^{R} be the number of such detected regions.

4) Partial. The detected region satisfies all of the following three conditions: *a)* It partitially overlays one of the ground truth regions; *b)* It does not overlay any other regions, such as non-formula regions or other formulas' regions; *c)* It does not split any ground truth formula regions (as defined in (8) in this section). An example of this situation is given in Fig. 2-4. Let N_{par}^{R} be the number of the recogized regions belonging to this situation.

5) Expanded. The detected region satisfies both of the following two conditions: *a)* It overlays at least one ground-truth region completely; *b)* It overlays non-formula regions or other formulas' regions. Examples of this situation are shown in Fig. 2-5. Let N_{exp}^{R} be the number of detected regions belonging to this situation.

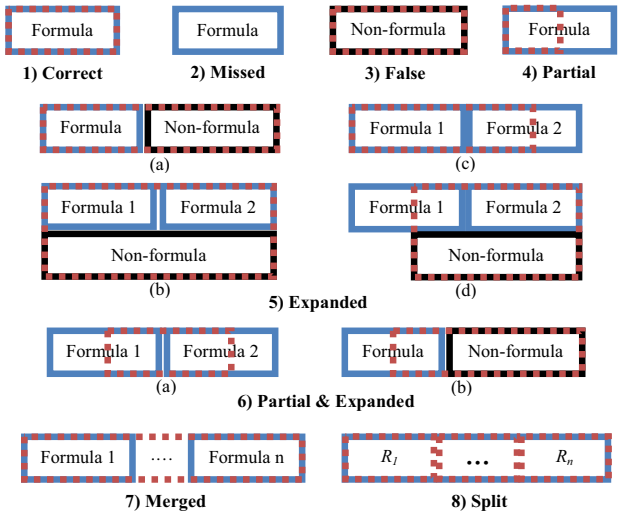

Figure 2. Example of formula identification errors. The blue boxes denote the ground truth formula regions. The red boxes denote the recognized regions. The black boxes denote the non-formula areas.

6) Partial & expanded. The detected region overlays not only part of the ground-truth regions, but also other regions (e.g. other formulas' regions, non-formula regions). Examples of this case are shown in Fig. 2-6. Let $N_{p\&e}^{R}$ be the number of the detected regions belonging to this situation.

7) Merged. The detected region overlays more than one ground-truth region completely, as shown in Fig. 2-7. Let N_{mer}^{R} be the number of such dectect regions.

8) Split. A ground-truth formula region is covered by more than one detected region completely, as shown in Fig. 2-8. Let N_{spl}^{G} be the number of the ground truth formula regions belonging to this situation.

B. Overall Performance Scoring

In most of the reported work, *precision* and *recall* are used for evaluating mathematical formula identification. Using this benchmark, different types of errors are all treated the same way. However, the error types of the mathematical formula identification are complex and their implications vary a lot. For example, a totally misrecognized result is considered to be worse than a partially misrecognized one. Besides, the developers cannot learn the strengths and weaknesses of the method or get clues for future improvement from this simplistic evaluation metric.

Furthermore, a mathematical formula identification system would be applied to various application scenarios, which tend to have different degrees of error tolerance. A generic evaluation metric like *precision* and *recall* is not able to tell which algorithm is optimal for a specific application context. For instance, in the mathematical formula retrieval task, false identification might be OK compared with missing some formulas. Because the missed ones which might be wanted result can never be found back once it is missed identified. However, false identification of formulas is more damaging than a total miss in layout analysis, because it mistakes other components of the page as formulas, which would influence the recognition of other components, such as paragraph. In this paper, an overall performance ***Score*** is computed based on the significance of different types of identification results, as defined in (1). The range of ***Score*** is [-1, 1], and the larger value of ***Score*** shows the better formula identification performance. Weights of different types of errors are W_{cor}, W_{mis}, W_{fal}, W_{par}, W_{exp}, $W_{p\&e}$, W_{mer}, and W_{spl}. They can be set according to a specific application scenario. $\boldsymbol{W}$ is calculated by summing each weight when the corresponding situation exists. For example, W_{cor} is added to $\boldsymbol{W}$ if and only if N_{cor}^{R} is above 0. The terms in the scoring function are defined as follows:

1) For situations of ***correct***, ***missed***, and ***false***, numbers of such cases are counted as N_{cor}^{R}, N_{mis}^{G}, and N_{fal}^{R}.

2) For situations of ***partial***, ***expanded***, and ***partial & expanded***, S_{par}, S_{exp}, and $S_{p\&e}$ are defined to evaluate the validity degree of the recognized result. In this paper, the propotion of correctly recognized region is considered the validity degree. Values of S_{par}, S_{exp}, and $S_{p\&e}$ can be calculated at both the area level and the symbol level, as defined in (2), (3), and (4).

At the area level, S_{par}, S_{exp}, and $S_{p\&e}$ are defined as S_{par}^{A}, S_{exp}^{A}, and S_{exp}^{A}. Their values are calculated according to the proportion of the correctly recognized areas. At the symbol level, S_{par}, S_{exp}, and $S_{p\&e}$ are defined as S_{par}^{S}, S_{exp}^{S}, and S_{exp}^{S}. Their values are calculated according to the proportion of correctly detected symbols. If a symbol in the detected region satisfies the following two conditions, it is considered to be correctly identified: First, the Unicode of the recognized symbol should be the same as the ground truth symbol. Second, the bounding box of the recognized symbol overlaps the ground truth symbol by more than $Th_{overlap}$ percentages. In our system, $Th_{overlap}$ is set as 95.

3) For the ***merged*** situation, S_{mer} is calculated as the number of ground truth regions that are overlapped by one recognized region, as defined in (5). Let $N_{R_i \cap G}$ denote the number of ground-truth formula regions that are completely overlapped by the *i*-th recognized formula region.

4) For the ***split*** situation**,** S_{spl} is calculated as the number of recognized regions that split the specific ground-truth formula region, as defined in (6). Let $N_{G_j \cap R}$ denote the number of recognized regions that split the *j*-th ground-truth formula region.

$$Score = \begin{pmatrix} W_{cor}N_{cor}^{R} - W_{mis}N_{mis}^{G} - W_{fal}N_{fal}^{R} + W_{par}S_{par} \\ +W_{exp}S_{exp} + W_{p\&e}S_{p\&e} + W_{mer}S_{mer} + W_{spl}S_{spl} \end{pmatrix} / WN$$

$$N = N_{cor}^{R} + N_{mis}^{G} + N_{fal}^{R} + N_{par}^{R} + N_{exp}^{R} + N_{p\&e}^{R} + N_{mer}^{R} + N_{spl}^{G} \quad (1)$$

$$S_{par}^{A} = \sum_{i=1}^{N_{par}^{R}} \frac{Area(R_i \cap G_j)}{Area(G_j)} \qquad S_{par}^{S} = \sum_{i=1}^{N_{par}^{R}} \frac{Symbol(R_i \cap G_j)}{Symbol(G_j)} \quad (2)$$

$$S_{exp}^{A} = \sum_{i=1}^{N_{exp}^{R}} \frac{Area(R_i \cap G_j)}{Area(R_i)} \qquad S_{exp}^{S} = \sum_{i=1}^{N_{exp}^{R}} \frac{Symbol(R_i \cap G_j)}{Symbol(R_i)} \quad (3)$$

$$S_{p\&e}^{A} = \sum_{i=1}^{N_{p\&e}^{R}} \frac{Area(R_i \cap G_j)}{Area(R_i)} \qquad S_{p\&e}^{S} = \sum_{i=1}^{N_{p\&e}^{R}} \frac{Symbol(R_i \cap G_j)}{Symbol(R_i)} \quad (4)$$

$$S_{mer} = \sum_{i=1}^{N_{mer}^{R}} \frac{1}{N_{R_i \cap G}} \quad (5) \qquad S_{spl} = \sum_{j=1}^{N_{spl}^{G}} \frac{1}{N_{G_j \cap R}} \quad (6)$$

IV. Automatic Evaluation Tool

A. Tool

Based on the ground truth format and the proposed evaluation metric, a tool is developed to evaluate the mathematical formula identification results.The input of the automatic evaluation tool includes the identification result, the ground truth file, and the parameters.

The weights for each type of errors can be set according to a specific application scenario via setting parameters. These weights are set to 1 by default. For situations such as *partial*, *expanded*, and *partial & expanded*, performance score can be calculated based on overlapped area or symbol, as defined in (2), (3), and (4). The user can choose either way (area or symbol) to compute the performance score through setting parameters. By default, area-based method is adopted calculate the performance score.

The identification result represented in the XML format as defined in Fig. 1 can be evaluated by the evaluation tool. It is worth mentioning that if the user chooses to calculate the performance score based on ***area***, the identification result needs to include only the bounding box of the identified isolated and embedded mathematical formulas. Otherwise, if the user chooses to calculate the performance score based on ***symbol***, not only the bounding boxes of the formulas but also their content objects (characters, images, and graphics) should be included in the identification results.

The automatic evaluation tool will output the number of each type of identification result and the overall performance score. Because further details about the errors might be needed, the document identifier and the page number will be recorded as well.

B. *Experiment Result*

Based on the proposed dataset and metric, we evaluate the mathematical formula identification methods presented in our previous work [11], including rule-based, SVM-based, and hybrid methods. Evaluation on other existing formula identification methods is not conducted in this paper due to: *a)* The source code of the reported mathematical formula identification methods is unavailable; *b)* Published description is not always sufficient for implementation; *c)* As far as we know, the only accessible math formula recognition tool is *InftyReader* provided by Infty [1]. However, it outputs no coordinates of either the identified formulas or symbols. Therefore results generated by *InftyReader* cannot be evaluated by our tool.

The evaluation tool is implemented in *Python* and run on a 2.5GHz PC with 2GB RAM. On average, it takes 20 seconds to evaluate 400 document pages. Fig. 3 illustrates the distribution of the eight result types of each method. It can be seen that the most significant strength of the rule-based method is that it produces less false recognition results. But it identifies much less formulas correctly and it also expands more formulas than the other methods. With the default parameter setting, the overall performance scores of rule-based, SVM-based, and the hybrid methods are 0.0001, 0.0114, and 0.0111, respectively.

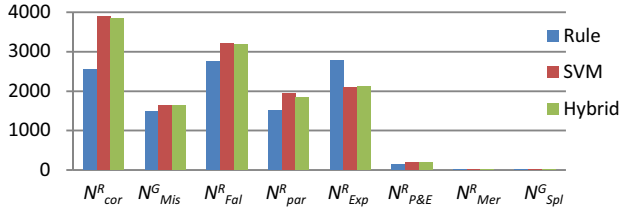

Figure 3. Formula identification results

V. CONCLUSIONS

A performance evaluation system for mathematical formula identification methods is presented in this paper. A ground-truth dataset with considerable number of document pages is constructed. Statistics analysis on the dataset shows that the dataset is representative of the documents in the real world. Based on this dataset, the performance of mathematical formula identification algorithms targeting document images or PDF documents can be compared directly. In addition, we propose a performance evaluation metric, which classifies the identification results in detail and quantifies the severities of different categories of errors. An overall performance score is defined and it can be adapted for different application scenarios. The proposed evaluation metric provides the information and means to gain insight into the identification results. Finally, an evaluation tool is developed to carry out the evaluation automatically. The ground truth dataset and the evaluation tool are freely available for academic purpose.

In the future, we plan to enlarge and refine the ground truth dataset and to develop evaluation tool for structure analysis of mathematical formulas.

ACKNOWLEDGMENT

This work is supported by National Basic Research Program of China, also named "973 Program" (No. 2012CB724108).

REFERENCES

[1] M. Suzuki, F. Tamari, R. Fukuda, S. Uchida and T. Kanahori, "Infty: an integrated OCR system for mathematical documents," Proc. ACM Symp. Document Engineering, pp. 95-104, 2003.

[2] D. M. Drake and H. S. Baird, "Distinguishing mathematics notation from English text using computational geometry," Proc. Int. Conf. Document Analysis and Recognition, pp. 1270-1274, 2005.

[3] U. Garain, "Identification of mathematical expressions in document images," Proc. *Int. Conf.* Document Analysis and Recognition, pp.1340-1344, 2009.

[4] A. Lapointe, and D. Blostein, "Issues in performance evaluation: a case study of math recognition," Proc. *Int. Conf.* Document Analysis and Recognition, pp. 1355-1359, 2009.

[5] University of Washington UW-III English/Technical Document Image Database. CD-ROM, 1996.

[6] M. Suzuki, S. Uchida, and A. Nomura, "A ground-truthed mathematical character and symbol image database," Proc. *Int. Conf.* Document Analysis and Recognition, pp. 675-679, 2005.

[7] U. Garain, and B. Chaudhuri, "A corpus for OCR research on mathematical expressions," *Int.* Journal on Document Analysis and Recognition, vol.7, pp. 241-259, 2005.

[8] K. Ashida, M. Okamoto, H. Imai, T. Nakatsuka, "Performance evaluation of a math formula recognition system with a large scale of printed formula images," Proc. Int. Conf. Document Image Analysis for Libraries, pp. 321-331, 2006.

[9] J. Baker, A. P. Sexton, and V. Sorge. "Comparing approaches to mathematical document analysis from PDF," Proc. *Int.* Conf. Document Analysis and Recognition, pp. 463-467, 2011.

[10] J. Baker, A. P. Sexton, and V. Sorge, "Towards Reverse Engineering of PDF Documents," Towards a Digital Mathematics Library, pp, 65-75, 2011.

[11] X.Y. Lin, L.C. Gao, Z. Tang, X.F. Lin, and X. Hu, "Mathematical formula identification in PDF documents," Proc. *Int.* Conf. Document Analysis and Recognition, pp. 1419-1423, 2011.

[12] A. Kacem, , A. Belaïd and M. Ben Ahmed, "Automatic extraction of printed mathematical formulas using fuzzy logic and propagation of context," *Int.* Journal on Document Analysis and Recognition, vol. 4, pp. 97-108, 2001.

[13] U. Garain, B.B. Chaudhuri and A.R. Chaudhuri, "Identification of embedded mathematical expressions in scanned documents," Proc. *Int.* Conf. Pattern Recognition, pp.384-387, 2004.