Abramowitz and Stegun — A Resource for Mathematical Document Analysis

Alan P. Sexton

School of Computer Science
University of Birmingham, UK
a.p.sexton@cs.bham.ac.uk

Abstract. In spite of advances in the state of the art of analysis of mathematical and scientific documents, the field is significantly hampered by the lack of large open and copyright free resources for research on and cross evaluation of different algorithms, tools and systems.
To address this deficiency, we have produced a new, high quality scan of Abramowitz and Stegun's Handbook of Mathematical Functions and made it available on our web site. This text is fully copyright free and hence publicly and freely available for all purposes, including document analysis research. Its history and the respect in which scientists have held the book make it an authoritative source for many types of mathematical expressions, diagrams and tables.
The difficulty of building an initial working document analysis system is a significant barrier to entry to this research field. To reduce that barrier, we have added intermediate results of such a system to the web site, so that research groups can proceed on research challenges of interest to them without having to implement the full tool chain themselves. These intermediate results include the full collection of connected components, with location information, from the text, a set of geometric moments and invariants for each connected component, and segmented images for all plots.

1 Introduction

Reliable, high quality tools for optical analysis and recognition of mathematical documents would be of great value in mathematical knowledge management. Such tools would enable large scale, low cost capturing of mathematical knowledge both from scanned images of documents and from electronic versions such as the usual PDF formatted output from LaTeX, which does not include the semantic enrichments necessary for a knowledge-oriented representation of mathematics.

In spite of advances in the state of the art of analysis of mathematical and scientific documents [21], the field is significantly hampered by the lack of suitable large, open and copyright free document sets to serve as ground truth sets and as data sets for testing and cross evaluation of different algorithms, tools and systems.

A *data set*, in this context, is usually a collection of page images from mathematical documents, although they can be image clips of formulae from those pages, or collections of character, symbol or diagram images.

A *ground truth set* is an input data set together with validated correct recognition results for the input data set. These correct recognition results are normally produced manually or by automatic recognition followed by manual correction. In particular they include full character and formula identification in spite of problems such as touching or broken characters — even if these problems are beyond the current state of the art of document analysis to resolve. As such it provides an ideal recognition result by which document analysis systems can be trained and against which they can be evaluated.

The most significant data sets for optical analysis of printed mathematical documents is the series of data sets from the Infty project [8,16,18]:

- InftyCDB-1: 476 pages of ground truth text from 30 mathematics articles in English with 688,580 character samples, 108,914 words and 21,056 mathematical expressions.
- InftyCDB-2: has the same structure as InftyCDB-1. It has 662,142 characters from English articles, 37,439 from French, and 77,812 from German.
- InftyCDB-3: is divided into two data sets. InftyCDB-3-A contains 188,752 characters. InftyCDB-3-B contains 70,637 characters. Word and mathematical expression structure is not included.
- InftyMDB-1: contains 4,400 ground truth mathematical formulae from 32 articles, which are mostly the same articles as in InftyCDB-1.

While an excellent resource for the mathematical OCR community, and especially useful for optical character and mathematical formula recognition training purposes, the one drawback of these data sets is that the original articles are not copyright free. The articles can not be distributed with the data sets and the data sets are not allowed to contain sufficient information about the location of characters so that the articles could be reconstructed from the data. Hence one can not easily test and compare systems on the original articles while using the ground truth for evaluation purposes.

Another data set is UW-III [14], the technical document ground truth data set from the University of Washington, containing 25 pages of ground truth mathematics. This data set is not free, currently costing $200 plus shipping.

Garain and Chaudhuri [9,10] discusses their proposed ground truth corpus of 400 real and synthetic mathematical document images. However their available data set [9] is of 102 isolated expression images and 5 clips from mathematical document pages, all with ground truth information.

Ashida et al. [2] produced a ground truth data set of symbols from 1400 pages and formulae from 700 pages of mathematics (taken from *Archiv der Mathematik* and *Commentarii Mathematici Helvetici*). Unfortunately, this data set is not available for copyright reasons.

Building a ground truth for a significant number of mathematical document pages is a very expensive and labour intensive project. It is particularly unfor-

tunate that, to date, no entirely copyright and cost free data set, for which the original documents are also copyright and cost free, has yet been made available.

In the remainder of this paper, we discuss a project to make available a very large, high quality, copyright and cost free data set for optical analysis of mathematical documents which, although not yet ground truthed, has been prepared to simplify processing and support future community based ground truthing.

2 Abramowitz and Stegun

Abramowitz and Stegun [1], or A&S, is an encyclopedia of mathematical functions. As it was published by the United States Government Printing Office, it is copyright free and hence fully available for researchers (and anyone else) to scan, report on and make their results freely available. Its history [3] and the respect in which scientists have held the book make it an authoritative source for many types of expressions, diagrams and tables, as witnessed by the article on the book in Wikipedia:

> "Abramowitz and Stegun is the informal name of a mathematical reference work edited by Milton Abramowitz and Irene Stegun of the U.S. National Bureau of Standards (now the National Institute of Standards and Technology). Its full title is Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.
> Since it was first published in 1964, the 1046 page Handbook has been one of the most comprehensive sources of information on special functions, containing definitions, identities, approximations, plots, and tables of values of numerous functions used in virtually all fields of applied mathematics. The notation used in the Handbook is the de facto standard for much of applied mathematics today."[1]

The fact that A&S was printed pre-TEX means that it can help researchers to avoid over-tuning their systems to the much more readily available TEX/LATEX sourced documents. For TEX documents, the availability of the ArXiv [5] with its very large collection of TEX based documents with full sources, makes alternative methods of obtaining data sets for OCR much more productive. Indeed, the ArXiv is already being data mined for the purposes of analysing semantic content [15]

One could argue that the ArXiv is such a large and rich resource that there is no need for a resource such as the one discussed here based on A&S. We respond to this argument as follows:

1. Fonts and typesetting are different enough between TEX and non-TEX based documents to significantly effect recognition. There is an enormous amount of mathematics in non-TEX based material still available only as bitmap images. Using only TEX based documents to train and test document analysis

[1] Taken from: `http://en.wikipedia.org/wiki/Abramowitz_and_Stegun`

systems could only handicap those systems in just the same way that not using any TEX based document data sets would: both TEX based and non-TEX based data sets are necessary.
2. It is significantly easier to produce a ground truth automatically when TEX sources, or PDF produced from TEX, are available. However, this does not cover the significant range of problems that earlier printing technologies introduce, such as the number and type of broken and connected characters, the level of ink bleed, problems caused by manual, rather than automatic typesetting etc. Modern printing followed by re-scanning produces artificial data sets that suffer from a much restricted and biased set of problems relative to those that occur in the wild. Hence it is valuable to have a data set such as this one that results from a true, non-digital printing process in addition to the data sets that can be built from collections such as the ArXiv.
3. A&S is of interest in itself. There is interest in matching the equations in A&S with those in the DLMF [12], it contains sections on probability that are not included in the DLMF and the tables in A&S, due to the care and rigour with which they were produced [3], are of interest in themselves if they could be accurately recognised.

The nature of A&S makes it a particular challenge for mathematical document analysis systems. Its 1060 pages contains a very high density of mathematical formulae (c.f. Figure 1). It has a complex layout that makes segmentation difficult — relatively few and short plain text passages, a two column default layout with lines usually, but not always, separating them and with frequent switches to single column layouts.

$$f(s) = \mathscr{L}\{F(t)\} = \int_0^\infty e^{-st} F(t)\,dt$$

(a) Equation 1

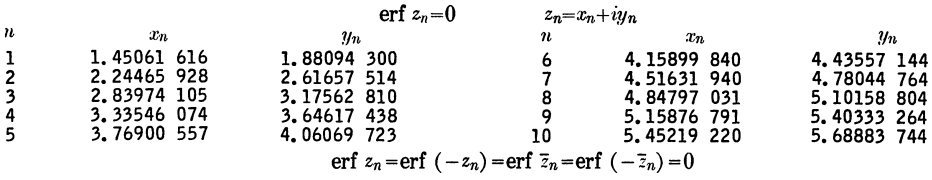
$$p(n) = \frac{1}{\pi\sqrt{2}} \sum_{k=1}^{\infty} \sqrt{k} A_k(n) \frac{d}{dn} \frac{\sinh\left\{\frac{\pi}{k}\sqrt{\frac{2}{3}}\sqrt{n-\frac{1}{24}}\right\}}{\sqrt{n-\frac{1}{24}}}$$

(b) Equation 2

Fig. 1. Example Equations from A&S

A&S also has a very large numbers of tables (c.f. Figure 2), both of numbers and of mathematical formulae. Many of these tables span multiple pages, most without any inter-cell lines but many with a complex line structure and a signif-

icant number rotated to landscape format. Some of the tables have been printed with approximately 5pt font sizes. Often the layout is spatially squeezed with obvious manual adjustments.



erf $z_n = 0$ $\quad z_n = x_n + iy_n$

n	x_n	y_n	n	x_n	y_n
1	1.45061 616	1.88094 300	6	4.15899 840	4.43557 144
2	2.24465 928	2.61657 514	7	4.51631 940	4.78044 764
3	2.83974 105	3.17562 810	8	4.84797 031	5.10158 804
4	3.33546 074	3.64617 438	9	5.15876 791	5.40333 264
5	3.76900 557	4.06069 723	10	5.45219 220	5.68883 744

erf z_n = erf $(-z_n)$ = erf $\bar{z}_n$ = erf $(-\bar{z}_n)$ = 0

Fig. 2. Example Table from A&S

There are a large number of often complex plots (c.f. Figure 3).

The printing suffered from some problems in that there are a large number of characters that are clearly broken, and other characters that are touching, as well as reliably reproduced dirt or marks on the pages. These faults do not diminish the readability of the text to a human, but cause issues for OCR software.

3 Rescan and Analysis

A digital capture of a new copy of A&S was carried out. It was scanned at 600dpi to 8 bit grey scale TIFF images. The files contain all pages in the book, including the front and back matter and blank pages separating sections. The latter were included so that the relationship between TIFF/PDF file page numbers and book page numbers would remain as simple as possible.

The original printing process for the book was such that there are a significant number of printing flaws in the original book. Many flaws in the scanned images are faithful reproductions of these printing flaws rather than artifacts of the scanning process. In particular, most pages of the book have some slight skew — up to 1.35° in the worst cases. While the scanning process undoubtedly introduced some level of skew, most of the skew appears in the original book.

Post-scanning processing of the images was carried out to deskew, binarise and remove any connected components smaller than the size of the smallest correct dot.

The deskewing was carried out automatically based on a projection profile approach and, although it is by no means perfect, it has reduced the skew in all cases. The resulting processed images, at 600dpi and reduced to 300dpi, are available at the project web site.

3.1 Connected Components and Geometric Moments

A connected component in a monochrome image is a maximal set of connected foreground pixels. Extracting them from an image is typically the first step of

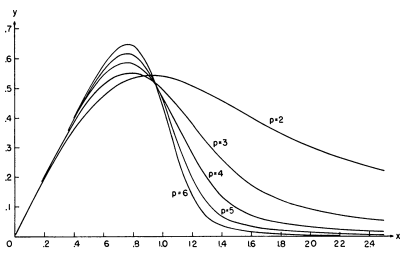


FIGURE 7.2. $y = e^{-z^p} \int_0^z e^{t^p} dt.$

$p = 2(1)6$

(a) Line Plot 1

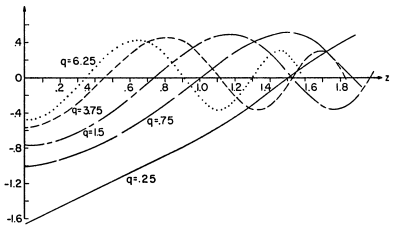


FIGURE 20.13. *Radial Mathieu Function of the Second Kind.*

(b) Line Plot 2

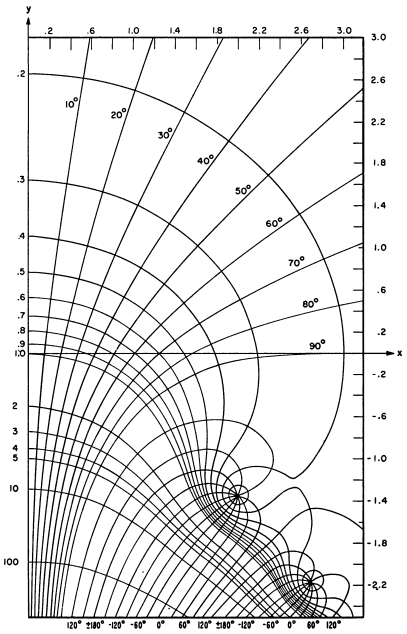


FIGURE 7.3. *Altitude Chart of $w(z)$.*

(c) Chart

Fig. 3. Example Plots from A&S

an OCR system after image pre-processing (binarisation, noise removal etc.). All connected components from the 600dpi monochrome images of A&S were extracted and have been made available. Although each connected component could be stored as a g4 compressed monochrome TIFF image, we have stored them as 8 bit grey scale TIFF images with an alpha, or transparency, channel and deflate compression. The foreground colour for these images is black, as expected. However the background colour has been set to fully transparent white. The result is that the original page image can be easily reconstructed for test purposes by drawing all the connected component images of a page in the correct location on a white page, The transparent background of the images ensures that no individual image masks any part of another image just because the bounding boxes of the images overlap. The resulting image files are not significantly larger than the monochrome versions and can easily be flattened to true monochrome if so desired.

There were 2,672,788 connected components extracted in total and the number of connected components per page ranges from 270 to 12824, with an average of 2572.

A data file, in comma separated value (CSV) format, was prepared that identifies the correct page, location and bounding box size of all extracted connected components.

One of the most historically popular approaches to optical printed character recognition is based on geometric moments and moment invariants [6,13,20]. In order to lower the barrier to entry for students and groups to mathematical document analysis research, we have pre-calculated and provided a set of features based on these moments for each connected component and included them in the CSV data file. The features included are

- An aspect ratio feature:
$$\frac{1}{2} + \frac{h - w}{2\max(h, w)}$$
where h, w are the height and width respectively of the bounding box of the connected component. This returns a number between 0 and 1.
- m_{00}: the $(0, 0)$ geometric moment, which corresponds to the number of foreground pixels in the connected component.
- $\eta_{20}, \eta_{11}, \eta_{02}, \eta_{30}, \eta_{21}, \eta_{12}, \eta_{03}$: all the second and third order normalised central geometric moments (by definition, $\eta_{00} = 1$ and $\eta_{01} = \eta_{10} = 0$). These are translation and scale invariant, but not rotation or general affine transform invariant.
- $I_1, I_2, \ldots, I_8$: the extended set of Hu invariants [11]. Hu defined 7 rotation invariant moments, $I_1, \ldots, I_7$. However, this set was shown by Flusser [6] to be neither independent (in particular, I_3 can be derived from I_4, I_5 and I_7 and therefore can be omitted without loss of recognition power) nor complete, hence the addition of the I_8 moment.

In the form that these features are provided, they can easily be used as the basis for any number of pattern classification approaches such as a metric space based k-nearest neighbour or cosine similarity [4,19]. However, the ranges of the values are such that some features will overwhelm others without further scaling. For this reason we include the ranges for all features in Table 1.

4 Special Clips

A particular asset of A&S is its wealth of plots. There is a large range of plots encompassing simple line plots, contour maps and branch cut plots, as shown in Figure 3. There is interest in work on analysing and recognising such plots [7,17]. To support that work we have manually clipped and extracted all 399 plots from A&S and provided them, with location and bounding box information, with the resources available from our web site. Together with the feature information about the connected components in A&S, this provides a low barrier to entry for research in this area.

Feature	Min	Max
aspect	0·0006785795	0·9989733
m_{00}	6	2187349
η_{20}	$5{\cdot}212947e-05$	64·39165
η_{11}	−10·16923	12·31475
η_{02}	0·0001275819	132·2743
η_{30}	−135·2704	158·1855
η_{21}	−27·66605	28·00989
η_{12}	−39·19975	15·09384
η_{03}	−347·5878	135·2028
I_1	0·15625	132·2746
I_2	0	17496·43
I_3	0	120749
I_4	0	120845
I_5	−524301100	14597710000
I_6	−28445·19	15781260
I_7	−172762300	548055900
I_8	−148817·6	281045·5

Table 1. Numeric Feature Ranges

5 Conclusions

The main aim of the work has been to provide a much needed resource of high quality for the mathematical document analysis community. While other data sets provide the very important aspect of ground truthing, none of them are, to date, fully open and free. The data set reported on here, while not provided with ground truth data, is fully open and free, and future work to develop a ground truth based on automatic recognition and community based crowd-sourced correction is planned.

A further aim has been to lower the barriers to entry to research for students interested in this area. Without such a data set, and partially processed initial analyses, an undergraduate or MSc student has too little time, after completing the software for initial processing, to pursue the more interesting goals of symbol recognition, formula recognition and layout analysis. It is hoped that with this data, this will no longer be the case.

We plan on adding full character recognition information to this data set in the near future and invite contributions from other research groups to enhance the data set.

The project web site, and all data sets and images from the project, is available at `http://www.cs.bham.ac.uk/~aps/research/projects/as/`.

Because of their size (approximately 19GB), the full, *losslessly compressed* set of original, 600 dots per inch, grey scale scanned images without any image processing applied are not accessible directly from the web site, although they are available from the author on request.

A version of both the original grey scale images and the deskewed grey scale images in TIFF format with lossy JPEG compression at a compression level of 60% is available from the web site. Each of these sets comes to 2.5GB and they are actually quite usable for document analysis purposes.

It is expected that these data sets may be useful for research on binarisation, noise reduction, deskewing, or grey scale optical character recognition.

Acknowledgements

We thank Bruce Miller of the National Institute of Standards and Technology, for providing a clean new copy of A&S for scanning.

We thank Bruno Voisin, of the Laboratory of Geophysical and Industrial Flows (LEGI), Grenoble, France, who allowed the use of his code to create PDF bookmarks for A&S as a basis for the bookmarks in the PDFs on the project web site.

References

1. Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. US Government Printing Office, Washington, 10th printing, with corrections. (December 1972)
2. Ashida, K., Okamoto, M., Imai, H., Nakatsuka, T.: Performance evaluation of a mathematical formula recognition system with a large scale of printed formula images. International Workshop on Document Image Analysis for Libraries pp. 320–331 (2006), http://doi.ieeecomputersociety.org/10.1109/DIAL.2006.30
3. Boisvert, R.F., Lozier, D.W.: Handbook of mathematical functions. In: Lide, D.R. (ed.) A Century of Excellence in Measurements Standards and Technology, pp. 135–139. CRC Press (2001), http://nvl.nist.gov/pub/nistpubs/sp958-lide/135-139.pdf
4. Cheriet, M., Kharma, N., Liu, C.L., Suen, C.Y.: Character Recognition Systems — A Guide for Students and Practitioners. Wiley & Sons Ltd., Hoboken, New Jersey (2007)
5. Cornell University Library: (2012), http://www.arxiv.org
6. Flusser, J., Suk, T., Zitová, B.: Moments and Moment Invariants in Pattern Recognition. Wiley & Sons Ltd., Chichester, UK (2009)
7. Fuda, T., Omachi, S., Aso, H.: Recognition of line graph images in documents by tracing connected components. Trans. IEICE J86-D-II(6), 825–835 (Jun 2003)
8. Fujiyoshi, A., Suzuki, M., Uchida, S.: Verification of mathematical formulae based on a combination of context-free grammar and tree grammar. In: Autexier, S., Campbell, J., Rubio, J., Sorge, V., Suzuki, M., Wiedijk, F. (eds.) Conferences in Intelligent Computer Mathematics (CICM 2008). LNCS, vol. 5144, pp. 415–429. Springer Berlin / Heidelberg (2008), http://dx.doi.org/10.1007/978-3-540-85110-3_35
9. Garain, U., Chaudhuri, B.B.: Ground truth datasets of mathematics, http://www.isical.ac.in/~utpal/resources.php
10. Garain, U., Chaudhuri, B.B.: A corpus for OCR research on mathematical expressions. IJDAR 7(4), 241–259 (2005), http://dx.doi.org/10.1007/s10032-004-0140-5

11. Hu, M.K.: Visual pattern recognition by moment invariants. IRE Transactions on Information Theory 8(2), 179–187 (February 1962)
12. Miller, B.: Personal communication (2011)
13. Mukundan, R., Ramakrishnan, K.: Moment Functions in Image Analysis. World Scientific, Singapore (1998)
14. Phillips, I., Chanda, B., Haralick, R.: UW-III english/technical document image database. University of Washington (2000), http://www.science.uva.nl/research/dlia/datasets/uwash3.html
15. Stamerjohanns, H., Kohlhase, M.: Transforming the arXiv to XML. In: Autexier, S., Campbell, J., Rubio, J., Sorge, V., Suzuki, M., Wiedijk, F. (eds.) Conferences in Intelligent Computer Mathematics (CICM 2008). pp. 574–582. LNCS, Springer Berlin / Heidelberg (2008), http://dx.doi.org/10.1007/978-3-540-85110-3_46
16. Suzuki, M., Uchida, S., Nomura, A.: A ground-truthed mathematical character and symbol image database. In: Eighth International Conference on Document Analysis and Recognition (ICDAR 2005). pp. 675–679 (2005), http://doi.ieeecomputersociety.org/10.1109/ICDAR.2005.14
17. Takagi, N.: On consideration of a pattern recognition method for mathematical graphs with broken lines. In: International Workshop on Digitization and E-Inclusion in Mathematics and Science (DEIMS12). pp. 43–51. Tokyo (2012)
18. The Infty Project: InftyCDB-1–3, InftyMDB-1 (2009), http://www.inftyproject.org/en/database.html
19. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Academic Press, 4th ed. (2009)
20. Yampolskiy, R.: Feature Extraction Approaches for Optical Character Recognition. Briviba Scientific Press, Rochester, NY (2007)
21. Zanibbi, R., Blostein, D.: Recognition and retrieval of mathematical expressions. International Journal on Document Analysis and Recognition pp. 1–27 (2012), http://dx.doi.org/10.1007/s10032-011-0174-4