DML 2009

Josef B. Baker; Alan P. Sexton; Volker Sorge
An Online Repository of Mathematical Samples

In: Petr Sojka (ed.): Towards a Digital Mathematics Library. Grand Bend, Ontario, Canada, July 8-9th, 2009. Masaryk University Press, Brno, 2009. pp. 49--57.

Persistent URL: http://dml.cz/dmlcz/702559

Terms of use:

© Masaryk University, 2009

Institute of Mathematics of the Academy of Sciences of the Czech Republic provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.

This paper has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* http://project.dml.cz

An Online Repository of Mathematical Samples

Josef B. Baker, Alan P. Sexton, and Volker Sorge

School of Computer Science, University of Birmingham
Email: J.B.Baker|A.P.Sexton|V.Sorge@cs.bham.ac.uk
URL: www.cs.bham.ac.uk/~{jbb|aps|vxs}

Abstract. With a growing community of researchers working on the recognition, parsing and digital exploitation of mathematical formulae, a need has arisen for a set of samples or benchmarks which can be used to compare, evaluate and help to develop different implementations and algorithms. The benchmark set would have to cover a wide range of mathematics, contain enough information to be able to search for specific samples and be accessible to the whole community.
In this paper, we propose an on-line system and repository where researchers may upload samples of mathematics in various formats such as scanned images, images directly rendered from born-digital documents, or born-digital document extracts. The system will support community tagging of these samples with attributes about their syntactic structure, semantic origin, image quality and source. Each sample in the database may then be searched for by any of its associated attributes, and users could download sets of sorted or random formulae to meet their own requirements. Associated with the system will be freely downloadable tools to assist in extracting and clipping mathematical samples from various kinds of documents to prepare them for uploading.
Additionally, the system will allow users to annotate each sample with their own files, in LaTeX, MathML, OpenMath and other formats. The intention here is that these annotation files will correspond either to the recognition results of the users' own systems on the samples, or manually constructed results. We believe that this facility will help to build a community verified ground truth set, available to anyone accessing the system.

1 Introduction

When conducting research into mathematical formula recognition, it is necessary to have a high quality collection of images of mathematical formulae for testing and evaluation purposes. If such databases are made publicly available, then it can make a significant contribution to the whole field, as researchers around the world can make realistic and scientific comparisons of their systems based on published performance results on common data sets. Further, it can accelerate research in the area as having authoritative test sets reduces barriers to entry for new researchers in the field. This has already proved successful in other areas of mathematical software

Petr Sojka (editor): DML 2009, Towards a Digital Mathematics Library, pp. 49–57.
© Masaryk University, 2009 ISBN 978-80-210-4781-5

systems [1,2]. For the area of mathematical formula recognition, Suzuki, Uchida and Nomura recognised this need and developed their *"Ground-Truthed Mathematical Character and Symbol Image Database"* [3]. In this work they captured all symbols from 30 articles (467 pages) from a number of mathematical journals and manually tagged every symbol with critical information that a formula recogniser would need to identify; e.g. type, font, category etc. The dataset is an excellent resource for anyone working on mathematical document analysis from optically scanned images.

Research in formula recognition is broader than recognising optically scanned images from mathematical journals. There is significant interest in recognition of mathematics from online input devices such as graphic tablets, pad computers or electronic whiteboards. Recognition of offline handwritten documents is also of interest. If the document to be analysed is in document formats such as PDF or PostScript, then a somewhat different recognition problem is posed if one wishes to take the opportunity they offer to obtain a better recognition result than can be obtained by simply rendering the document to a bitmapped image format and applying the usual optical formula recognition to the result. Each of these related research areas would benefit enormously from appropriate high quality common data sets.

Unfortunately, producing such data sets is a very labour intensive, and hence expensive, undertaking. In this paper, we propose an approach to providing such data sets that, we believe, will spread the large effort required over a relatively large community of interested parties rather than imposing the workload on a small dedicated team. Our approach, based on the provision of a web database application together with free ancillary desktop applications, has further benefits, such as a continually growing set of data, ease of access and, as a side-effect, applicability to other areas of librarianship and mathematical knowledge management via the provision of data sets suitable for research on mathematical search, mathematical data mining and meta-data extraction.

The system we are proposing will allow users to upload mathematical formula sample to a central repository. The samples can be in any of a number of types and formats, e.g. handwritten bitmapped image, InkML online recording [4], clips from PDF documents or others. Related data, for example provenance and copyright information can be uploaded. Once the samples are available online, they will be made available for the community to tag with attribute information, describing various properties of the formulae such as, among others, whether they contain sub- or superscripts and what area of mathematics they pertain to. The system will provide selection facilities, so that users can select subsets of the full data set based on tagging information and format or type of the formula samples. These selections can then be downloaded for the users to conduct their own research or tests. Finally, users can annotate formula samples in the database with their own data. Thus one user might add a manually written Content MathML description of a formula, while another might add the LaTeX representation of the formula generated by their research formula recogniser tool.

In this system, we intend to collect and tag only mathematical formula samples. Thus, we do not collect diagrams, tables, or bibliographies. While it is necessary to allow uploading of entire documents, e.g. PDF or PostScript document sources for formula samples clipped from such a document, the focus of our repository is the formula sample itself, and not the containing document.

2 Data Overview

The primary aim of our database is to collect samples of single formulae that can be used for testing and developing mathematical formula recognition software. We anticipate that the majority of these formulae will originate from larger documents, e.g., taken from a scanned books or other form of electronic document. The actual formula image can then be viewed as a clip from the larger document and information relating the clipped image to the original source can be used to record the formula's provenance but could also prove helpful for the recognition process and should therefore be kept. In addition, information on the basic components of a formula sample, i.e., the actual glyphs occurring in the image, has to be computed and stored as well as related to the origin. Compiling this information is a laborious task and is therefore ideally done automatically. We will briefly describe a tool for clipping formulae that can achieve this in Sec. 4.

Our primary data are images containing single formulae. In addition each image has several mandatory and optional administrative files. The former contain necessary information about the image and its origin, while the latter can be files for recognition results and the original source of an image. In detail the database contains the following component files for each formula sample:

Sample File: (Required) Either an InkML file or A TIFF file of a single formula or a page containing formulae. The quality of the TIFF file may vary depending on the origin of the sample.
Provenance: (Required) The provenance and copyright information for the sample. Without this information we can not be confident that we can legally make the corresponding samples available online.
Source: (Optional) The original file from which the formula has been taken, e.g., usually a PDF or PostScript document or a multi-page TIFF file. This will not always be available, for example when the formula comes from a scanned image.
A sample may be a single standalone item, e.g. a single image of a single formula, or it may be one of a set of samples, all of which have been obtained from the same source, e.g. a set of formulae clipped from a paper or book. In the latter case, it is possible to upload the source document only once, but indicate the samples from the source by reference (e.g. source reference, page number and clip rectangle on the page).
Clip file: (Optional) A file associated with an image sample containing the location and size of a clipped formula, along with the bounding box and

position of every glyph within the clipped area. This file is in the JSON [5] format, which provides a simple and compact formalism that is usable by the majority of the contemporary programming languages. A clip file is automatically generated by the clipping software described in Sec. 4.
If a clip file is not provided with a sample, then it is assumed that the entire sample contains only a single formula and all glyphs in the sample are part of that formula. In this case, the system will automatically generate and store a clip file for the sample that covers the entire image. InkML samples do not require clip files.
Attribute File: (Optional) A file identifying the tags associated to a particular clip. We describe the components of this file in the next section.
We expect that most users will use the web interface to interactively set tags for individual sample. However, we provide this feature to support interactive desktop tools, or automatic analysis and tagging tools, by allowing their output tagging decisions to be uploaded to the repository.
Annotation Files: (Optional) These are user generated and will be associated with individual samples. Multiple files with different formats may be associated with each sample. This allows users to add and release documents such as MathML or OpenMath descriptions, LaTeX source or other items. These files may either be manually written or may be the output of analysis tools applied to the sample.

3 Categorisation

The primary aim of the database is to provide a means to quickly compile a set of samples for testing and benchmarking of mathematical formula recognition software. The retrieved set of images should be customisable with respect to the particular aim of the software in question, e.g., recognition systems might be optimised for paper scanned mathematics or electronically born document, or different systems might aim at different mathematical subjects.

To enable this we assign each sample particular attributes in different categories. In detail, we assign formulae attributes from three different types of categories:

1. image quality,
2. semantic origin of the formula, and
3. syntactic formula structure.

3.1 Quality Attribute

Each sample can have exactly one type of quality attribute. This attribute defines primarily how much information is available that can be used for the recognition process and often depends on the origin of the formula. There are currently four different types:

Perfect Information: The sample is in a format that contains information about the actual components of that formula, such as characters, fonts, etc. Examples are formulae in PostScript or PDF format.
Rendered Image: The sample is given in some bitmapped image format, typically electronically generated from a perfect information format. The image therefore contains no noise, skewing problems or other artifacts associated with optical scanning.
Scanned Image: The sample is a bitmapped image, which comes from an optically scanned sample and can contain noise, skew, etc.
InkML: An InkML file, containing the data (stroke path, pressure, etc.) obtained by the use of an online handwriting input device, such as a graphic tablet, electronic pen or pad computer.

3.2 Semantic Attributes

Each image has a semantic attribute for its origin in some mathematical field. The semantic attribute corresponds to the two first digits of the 2000 Mathematics Subject Classification [6].

In general, few formula, viewed in isolation, can be reliably ascribed to a particular mathematical field. This attribute refers, not to the field the formula belongs to, but to the general mathematical field the *document* belongs to from which the formula was extracted. We cannot expect that this information will always be known, so we add a category of *unknown* or *unclassified* for images that are of unknown origin.

3.3 Structural Attributes

Each sample can be tagged with a set of attributes that express the structural composition of a mathematical formula. The basic attributes are given in the table below. Some attributes have a recursive flag, indicating that they can occur either in simple fashion or recursively. Thus a formula containing simple subscripts should be tagged the **Script** attribute. However, if it has a subscript which itself has a sub- or superscript, it should be tagged as **Script(Recursive)**.

Text: Formulae with interspersed text. e.g.,

$$a + b \text{ only when } x = 0$$

Script: Sub- or superscripts. e.g.,

$$a_3, \quad a^4, \quad a_3^4, \quad {}_2^1a_4^3, \quad a_{i_3}$$

Accents: Mathematical accents like vectors etc. e.g.,

$$\mathbf{a}, \quad \dot{a}, \quad \hat{a}$$

Fractions: Formulae containing division bars. e.g.,

$$\frac{a}{b}, \quad \cfrac{a}{1+\frac{b}{c+d}}$$

Containers: Containers are elements that fully contain another formula, i.e. their vertical and horizontal extent is large or equal to the contained formula. Examples are root symbols or boxes. e.g.,

$$\sqrt{a+b}, \quad \sqrt[i]{\sqrt{a+b}+c}$$

Limits: Elements with upper and/or lower limiting expressions. e.g.,

$$\sum_{i=1}^{n} n+i, \quad \lim_{n\to\infty} n$$

Fences: Formulae containing fencing or bracketing of some kind. Fences may be balanced (paired) or unbalanced (a single fence, or a 3 fence construct such as a set comprehension expression). They include vertical fencing such as under- or over-bracing or under- or over-lining as well as the more common horizontal fencing.

$$\left(A_y^x + B\right), \quad \left\{x \in X \middle| p(x) \wedge q(x)\right\}, \quad \underbrace{n(n-1)\dots(n-m+1)}_{m \text{ factors}}$$

Grids: Two dimensional array structures, usually representing matrices, tables or combinatorial expressions. e.g.,

$$\binom{x}{y}, \quad \begin{bmatrix} 1 & 2 & 3 \\ a & b & c \end{bmatrix}$$

Cases: Case statements. e.g.,

$$f(x) = \begin{cases} i & \text{if } x > 0 \\ j & \text{otherwise} \end{cases}$$

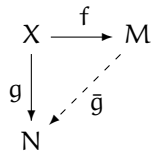
Ellipses: Formulae containing ellipses of various types (e.g. vertical, horizontal, diagonal or anti-diagonal, but also of different types, like vertically centred or on the baseline). e.g.,

$$a_1, \dots, a_n, \quad a_1 + \cdots + a_n, \quad \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ & \ddots & \vdots \\ \mathbf{0} & & a_{nn} \end{bmatrix}$$

Multiline: Equations or similar formulae that span multiple lines. e.g.,

$$\begin{aligned} x &= a + b + b \\ &= a + 2b \end{aligned}$$

Commutative Diagrams: Commutative diagrams as commonly found in algebra or category theory texts. e.g.,



All the above tags can be marked as occurring recursively except for **Text**, **Ellipses** and **Multiline**. An image that is not annotated is assumed to be a simple formula, for example, $a + b + 3 \cdot c$.

4 User Interface and Tools

We are currently developing a web-based user interface with two goals. On the one hand it should enable users to access the content of the repository by being able to download customised sample sets. On the other hand registered users should be able to add content to the repository as well as help tag and annotate existing content. For this purpose users can upload single samples together with their provenance information or sets of samples together with their source. A user created clip file can be uploaded or one with default parameters automatically generated. Quality and semantic information can be supplied for all files uploaded. Users, not necessarily those who uploaded the original sample, can use a simple web interface to tag individual samples or upload a tag file to bulk tag samples. Both semantic and quality information can be altered individually for each sample.

As mentioned in Sec. 2, we have already developed a tool for the clipping of formulae from existing documents given different input formats such as multi-page TIFF, PDF or PostScript. In the case of PDF or PostScript, a multi-page TIFF file is rendered from the given input document. The user can then interactively clip and save rectangular sections from each page of the document. The clipped sections are saved as the cut-out TIFF image together with a JSON file that contains the bounding box information of the glyphs contained in the image. In addition it records the provenance referring to the containing document, page number and coordinates of the clipped image.

It should be noted that PDF and PostScript documents provide a clipping facility. That is, various tools allow one to clip a page in these to a specific sub-rectangle. However, these are not true clips, in that the resulting PDF document still contains the drawing instructions for the whole original page. Instead, masking information is added to the existing page so the drawing instructions outside the clipping rectangle are suppressed and the page size information is updated. While, on rendering, this gives the visual result of clipping, the internal information does not appear clipped to software that analyses that information for formula recognition purposes (without fully interpreting the document description language). Hence directly clipped PDF or PostScript is

not a suitable format as a source in our repository. For this reason, if a PDF or PostScript sample is to be uploaded, the original document has to be uploaded together with the clip specification file (generated by our desktop Java tool) that identifies the exact details of the clip.

5 Issues

We briefly discuss two non-technical issues associated with the repository:

1. Quality assurance
2. Copyright issues

The concern with issue 1 is not related to the quality of the images, but the quality of the associated administrative files. In particular, the correct declaration of the provenance of formulae has to be of concern. Secondly, the provision of the structural annotations has to be controlled and protected from abuse. We currently plan to minimise these problems by restricting the write access to the repository to registered users.

Issue 2 is less simple to overcome and needs to be explored thoroughly in the future. The main question here is to what extent images taken from copyrighted material are subject to copyright themselves, in particular if formulae are systematically clipped from scanned books to put them into the repository. While the images can be used as material for experiments in ones own work under fair use, it needs to be investigated whether making all images readily available online will not violate copyright laws in certain countries. For the initial release of our software, we intend to moderate submissions to ensure that all submissions that we make available are covered by an appropriate valid free-use copyright agreement.

6 Conclusion

We have presented the design of a repository for images of mathematical formulae that should serve as a sample set and benchmark tool for mathematical formula recognition software. As has been mentioned before, the repository is currently work in progress, and our goal is to present its current design as a basis for further discussion to the community in order to gain feedback on further desirable features that should be incorporated.

The database is in its prototype development stage and currently contains around 1000 images taken from two freely available mathematics books [7,8], which can be freely downloaded as PDF files. The initial version of the system, with the limitation that we moderate submissions for valid free-use copyright, will be made publicly available. We cannot make the unlimited version publicly available until clarification of the copyright issues mentioned in Sec. 5. The clipping tool mentioned in Sec. 4 is implemented in Java and is currently available on request from the authors.

Acknowledgements

We thank the anonymous referees, whose comments helped us to improve this paper.

References

1. Hoos, H.H., Stutzle, T.: SATLIB: An online resource for research on SAT. In: Proceedings of the Third Workshop on Satisfiability (SAT 2000), IOS Press (2000) 283–292 http://www.satlib.org.
2. Sutcliffe, G., Suttner, C.: The TPTP Problem Library: CNF Release v1.2.1. Journal of Automated Reasoning **21**(2) (1998) 177–203
3. Suzuki, M., Uchida, S., Nomura, A.: A ground-truthed mathematical character and symbol image database. In: Proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR 2005), IEEE Society Press (2005) 675–679 http://www.inftyproject.org/en/database.html.
4. W3C: Ink markup language (InkML). (2006) http://www.w3.org/TR/InkML/.
5. Crockford, D.: JavaScript Object Notation. (2006) http://www.json.org/.
6. The American Mathematical Society: 2000 Mathematics Subject Classification (2000) http://www.ams.org/msc/.
7. Sternberg, S.: Semi-riemann geometry and general relativity (2003) http://www.math.harvard.edu/~shlomo/docs/semi_riemannian_geometry.pdf.
8. Judson, T.: Abstract algebra — theory and applications (2009) http://abstract.ups.edu/download.html.