An Online Repository of Mathematical Samples

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

School of Computer Science
University of Birmingham

Motivation

- Growing community working on recognition, parsing and digital exploitation of mathematical formulas
- Difficult to obtain data to reliably compare systems
- No collection of sample images of mathematical formulas readily available (exception: Suzuki's Ground Truth Set)
- We have currently plenty of images and are wondering what to do with them
- Categorise them and make them available
- Hoping for input from others
- Inspired by similar sets in other areas: TPTP, SAT-benchmarks

Basic Idea

- Build a repository of math formula images
- Support the compilation of sample sets for testing and benchmarking
- Recognise scanned math, electronically born documents, some handwritten math
- Possibly distinguish with respect to different mathematical subjects
- Collect different result files for samples
- Support a community effort

Data Overview

- Primary content are image files of formulas.
- Many will be clipped from single, larger documents.
- Each image will have several administrative files, some of them optional.

Data Overview (ctd)

- Sample File (Required)
  TIFF or InkML file of a single formula or a page with formulae.

Data Overview (ctd)

- Sample File (Required)
  TIFF or InkML file of a single formula or a page with formulae.
- Provenance (Required)
  The provenance and copyright information for the sample.

Data Overview (ctd)

- Sample File (Required)
  TIFF or InkML file of a single formula or a page with formulae.
- Provenance (Required)
  The provenance and copyright information for the sample.
- Source (Optional)
  The original document file from which the formula has been taken, e.g., PDF, Postscript, or multi-page TIFF

Data Overview (ctd)

- Sample File (Required)
  TIFF or InkML file of a single formula or a page with formulae.
- Provenance (Required)
  The provenance and copyright information for the sample.
- Source (Optional)
  The original document file from which the formula has been taken, e.g., PDF, Postscript, or multi-page TIFF
- Clip file (Optional)
  Bounding box and position of glyphs in sample in JSON format. We have a tool to easily generate this.

Data Overview (ctd)

- Sample File (Required)
  TIFF or InkML file of a single formula or a page with formulae.
- Provenance (Required)
  The provenance and copyright information for the sample.
- Source (Optional)
  The original document file from which the formula has been taken, e.g., PDF, Postscript, or multi-page TIFF
- Clip file (Optional)
  Bounding box and position of glyphs in sample in JSON format.
- Attribute File (Optional)
  Tagging information for the sample necessary for retrieval.

Data Overview (ctd)

- Sample File (Required)
  TIFF or InkML file of a single formula or a page with formulae.
- Provenance (Required)
  The provenance and copyright information for the sample.
- Source (Optional)
  The original document file from which the formula has been taken, e.g., PDF, Postscript, or multi-page TIFF
- Clip file (Optional)
  Bounding box and position of glyphs in sample in JSON format.
- Attribute File (Optional)
  Tagging information for the sample necessary for retrieval.
- Annotation Files (Optional)
  User generated result files in different formats, e.g. LaTeX, MathML, etc.

Categorisation

- Sample files are categorised by attributes that enable their goal-directed retrieval.
- Recall our goal:
  Faciliate the compilation of samples for testing and benchmarking
- Software in question might have different usage
  - Recognise scanned math or electronically born documents
  - Aimed at different mathematical subjects
- Retrieved set of images should be customisable
- Samples are assigned attributes with respect to
  1. image quality,
  2. semantic origin of the formula, and
  3. syntactic formula structure.

Quality Attributes

Each sample has exactly one attribute defining how much information is available for the recognition process.

- Perfect Information
- Rendered Image
- Scanned Image
- InkML

Quality Attributes

Each sample has exactly one attribute defining how much information is available for the recognition process.

- Perfect Information
  - Contains information about the actual components of that formula, such as characters, fonts, etc.
  - Examples are formulae in Postscript or PDF format.
- Rendered Image
- Scanned Image
- InkML

Quality Attributes

Each sample has exactly one attribute defining how much information is available for the recognition process.

- Perfect Information
- Rendered Image
  - Some bitmapped image format.
  - No noise, skewing problems or other artifacts associated with optical scanning.
  - Typically electronically generated from a perfect information format.
- Scanned Image
- InkML

Quality Attributes

Each sample has exactly one attribute defining how much information is available for the recognition process.

- Perfect Information
- Rendered Image
- Scanned Image
  - Bitmapped image originating from an optically scanned sample.
  - Can contain noise, skew, etc.
- InkML

Quality Attributes

Each sample has exactly one attribute defining how much information is available for the recognition process.

- Perfect Information
- Rendered Image
- Scanned Image
- InkML
  - InkML file, containing the data (stroke path, pressure, etc.)
  - obtained from online handwriting input device, e.g. graphic tablet, electronic pen or pad computer.

Semantic Attributes

- Each image has an attribute for its origin in some mathematical field
- Attribute refers to the field of the document the image belongs to
- We use the first two digits of the 2000 Mathematics Subject Classification
- For images of unknown origin we have a category *Unclassified*

Structural Attributes

- Each sample can be tagged with a set of attributes.
- Express the structural composition of a mathematical formula.
- Some attributes can be flagged as recursive.
- An image that is not annotated is assumed to be a simple formula. Example: $a + b + 3 \cdot c$.

Structural Attribute List

- Text
- Script
- Accents
- Fractions
- Containers
- Limits
- Fences
- Grids
- Cases
- Ellipses
- Multiline
- Commutative Diagrams

Structural Attribute List

- Text
- Script
- Accents
- Fractions
- Containers
- Limits

Structural Attribute List

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

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

- Script
- Accents
- Fractions
- Containers
- Limits

Structural Attribute List

- Text
- 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
- Fractions
- Containers
- Limits

Structural Attribute List

- Text
- Script
- Accents Mathematical accents like vectors etc. e.g.,

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

- Fractions
- Containers
- Limits

Structural Attribute List

- Text
- Script
- Accents
- Fractions Formulae containing division bars. e.g.,

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

- Containers
- Limits

Structural Attribute List

- Text
- Script
- Accents
- Fractions
- Containers
  - Elements that fully contain another formula
  - Their vertical and horizontal extent is larger or equal to the contained formula
  - E.g. root symbols or boxes

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

- Limits

Structural Attribute List

- Text
- Script
- Accents
- Fractions
- Containers
- Limits Elements with upper and/or lower limiting expressions. e.g.,

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

Structural Attribute List

- Fences
- Grids
- Cases
- Ellipses
- Multiline
- Commutative Diagrams

Structural Attribute List

- Fences
  - fencing or bracketing of some kind
  - balanced (paired) or unbalanced (a single fence, or a 3 fence construct such as a set comprehension expression)
  - also includes vertical fencing, e.g.under- or over-bracing or under- or over-lining

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

- Grids
- Cases
- Ellipses
- Multiline
- Commutative Diagrams

Structural Attribute List

- Fences
- Grids
  - two dimensional array structures
  - e.g., matrices, tables or combinatorial expressions

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

- Cases
- Ellipses
- Multiline
- Commutative Diagrams

Structural Attribute List

- Fences
- Grids
- Cases Case statements. e.g.,

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

- Ellipses
- Multiline
- Commutative Diagrams

Structural Attribute List

- Fences
- Grids
- Cases
- Ellipses
  - vertical, horizontal, diagonal or anti-diagonal
  - different types, e.g. vertically centred or on the baseline

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

- Multiline
- Commutative Diagrams

Structural Attribute List

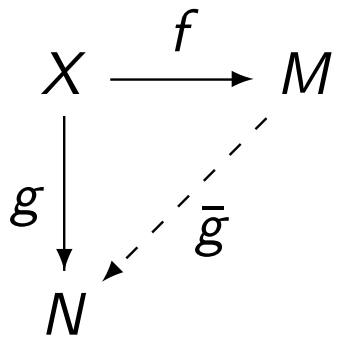
- Fences
- Grids
- Cases
- Ellipses
- Multiline Equations or similar formulae that span multiple lines.

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

- Commutative Diagrams

Structural Attribute List

- Fences
- Grids
- Cases
- Ellipses
- Multiline
- Commutative Diagrams as commonly found in algebra or category theory texts.



Software Tools

- Clipping program
  - PDF, Postscript or multipage TIFF input
  - GUI to clip formulas from an image file
  - Returns clip file in JSON format containing information like, glyph positions and bounding boxes
- We have about 1000 images, primarily clipped from PDF files
- We are putting together the web front-end to register and categorise samples
- We are developing an evaluation tool for rendered result files

Issues

- Quality assurance
  - How can we assure the quality of the structural annotations?
  - Restrict write access to the repository to registered users.
- Copyright issues
  - What about images from copyrighted material?
  - Using images for one's own experiments should be no problem.
  - But making them easily and freely available in large numbers might be a different matter.
  - Taking all formulas from a single book might exceed anything covered under fair use.
  - Intended solution: moderated submission to ensure we have valid free-use copyright agreements

Conclusions

- Repository of Mathematical Samples as a service to the community
- Webinterface is currently under construction and should be available soon
- Will enable to contribute and categorise samples
- Categorisation is important as it should aid better testing, analysing and comparing systems
- Copyright issues?
- We appreciate your input!