Handwritten Mathematics Corpus

This handwritten mathematical corpus is described in the paper, Grammar-based techniques for creating ground-truthed sketch corpora, International Journal of Document Analysis and Recognition, Vol. 14 (2011) 65-74

To obtain a zip file containing the handwriting samples, please send the following information via email to mathbrush@gmail.com.

- Name and affiliation
- Email address to which the corpus will be sent
- Planned uses for the corpus

If you make use of the corpus in published work, please cite the paper in which it was published:

S. MacLean, G. Labahn, E. Lank, M. Marzouk, and D. Tausky, Grammar-based techniques for creating ground-truthed sketch corpora, International Journal of Document Analysis and Recognition, Vol. 14 (2011) 65-74

The contents of the corpus are described in the next section.

Corpus Contents

The corpus consists of samples collected from 20 study participants. Each participant's data is stored in a directory named `user`*n*, where *n* ranges from 1 to 20. Within each of these directories is an `auto-labels` subdirectory containing the automatically-generated ground-truth labels from our experiment using 50 symbol candidates per ink group without recognizer pre-training.

In the directories, each sample consists of 6 files with a common prefix. The files are as follows:

`*.spec`
: Contains a parenthesized parse-tree representation of the template expression.

`*.links`
: Contains the derivation string of the template expression (refer to the paper for details).

`*.tex`
: Contains a LaTeX representation of the template expression.

`*.gif`
: Contains a GIF image of the template expression.

`*.msink`
: Contains the expression as transcribed by the participant, in Microsoft's Tablet PC ink format.

`*.ink`
: Contains the expression as transcribed by the participant, in SCG ink format

SCG Ink format

The SCG ink format is a simple text-based representation of the (x,y)-coordinates drawn by the user. If you require pressure, tilt, etc. information, you should use the Microsoft format. However, the SCG format includes ground-truth labels, whereas Microsoft's format only includes ink data. Each SCG ink file is organized as follows:

1. The string "`SCG_INK`" followed by a newline.
2. A number indicating the number of strokes appearing in the ink, followed by a newline.
3. For each stroke:
    1. A number indicating the number of points appearing in the stroke, followed by a newline.
    2. For each point, two numbers indicating the x- and y-coordinates, separated by a space, and followed by a newline.
4. The string "`ANNOTATIONS`" followed by a newline.
5. A number of annotations, each of one of the following three forms:

    Symbol label

    A symbol label associates a symbol name with a group of ink strokes in the sample. These labels consist of the string "`SYMBOL`" followed by a stroke list, a symbol name, and a newline.

    e.g.

    ```
    SYMBOL <1, 2> x
    ```

    Symbol mapping

A symbol mapping associates a group of ink strokes in the sample with a terminal symbol in the associated `.links` file (i.e. the derivation string). These labels consist of the string "`SYMBOLMAP`" followed by a stroke list, a symbol index, and a newline.

e.g.

```
SYMBOLMAP <1, 2> 3
```

Relationship label

A relationship label indicates how two groups of ink are structurally connected in the mathematical expression. These labels consist of the string "`LINK`" followed by a stroke list, a relation name, a stroke list, and a newline. The relations indicated are as follows:

- `R`: "Right" - horizontal adjacency, first ink group leftmost
- `AR`: "Above-Right" - superscript position
- `BR`: "Below-Right" - subscript position
- `B`: "Below" - vertical adjacency, first ink group topmost
- `C`: "Contains" - containment, first ink group the container

e.g.

```
LINK <1, 2> R <3>
```

A stroke list is a comma-separated list of numbers enclosed in angle brackets (eg. <1, 2>) that specifies a group of ink strokes by their zero-based offsets in the ink section of the file.

The end of the annotation section is indicated by end-of-file.