Groundtruthed Word/Formula Image Database

InftyCDB-1

The database InftyCDB-1 includes all the characters and symbols of 467 pages of 30 English articles on pure mathematics (published 1970∼ 2000), and is organized so that it can be used as word image database or as mathematical formula image database. The ground-truth of each character is composed of type, font, quality (touched/broken) and link (relative position in math formula), etc.

The image data are stored separated into word or math formula units and arranged in alphabetic order indemendent of the content of papers. No whole page image is included in the database to avoid copyright problems.

The database includes 688,570 characters/symbols in total, and 157,058 characters/symbols in math expressions. All pages were scanned in 600 dpi and binarized automatically by the same commercial scanner (RICOH Imagio Neo 450). The quality of the resulting page images varies with the quality of paper, etc. Several page images are noisy and include a lot of abnormal characters, such as touching characters and broken characters.

InftyCDB-1 is a public database and freely usable for research and development purposes. The database can be used in the following researches, for example:

- development and evaluation of character and scientific symbol recognition,
- development and evaluation of mathematical formula recognition,
- analysis of words in mathematical documents.

Since all the caracter images appeared in the page images are included in the database, users can get training data or test data for character/symbol recognition. To all the special mathematical symbols in the database, their own code and symbol name are attached carefully. Since each alpha-numeric character in the database has its font attribute such as italic/upright, bold or not, the database can be used in the evaluation of these font distinction ability of character recognition.