Genuine Strikethrough Dataset

This dataset was created for the paper "**Strikethrough Removal From Handwritten Words Using CycleGANs**" by Raphaela Heil, Ekta Vats and Anders Hast, to appear in: 2021 International Conference on Document Analysis and Recognition (ICDAR).

For any questions regarding the content or creation, please feel free to contact: raphaela.heil@it.uu.se

For details regarding the license of this dataset, please see the included `LICENSE` file.

Content

This dataset contains registered pairs of clean and struck-through handwritten words which have been used in the context of the aforementioned paper for the task of unpaired strikethrough removal. The text, a passage from Bram Stoker's Dracula, was written by a single writer, using a blue ballpoint pen on regular white paper. The 756 word images have been systematically struck through, using one of the following stroke types: single horizontal line, double horizontal lines, diagonal, cross, wave, zigzag, scratch.

The dataset has been split into three subsets, each balanced with regard to the number of samples per stroke type. Each split contains csv-files that indicate the stroke type that has been applied to a particular image.

Train

Directory	Image Count	Description
struck	126	struck-through words
struck_gt	126	clean ground truth for the struck-through words from `struck`
clean	126	clean words, unrelated to `struck` and `struck_gt` (i.e. no overlap)

The images in `clean` and `struck` can e.g. be used to train an unpaired image-to-image translation algorithm.

Validation

Directory	Image Count	Description
struck	126	struck-through words
struck_gt	126	clean ground truth for the struck-through words from `struck`

Test

Directory	Image Count	Description
struck	378	struck-through words
struck_gt	378	clean ground truth for the struck-through words from `struck`