smoothscan

Description

smoothscan is a tool to convert scanned text into a vectorized output form. Because printed text is assembled from fonts, each particular letter (like 'o') will have the same shape as every other 'o' in the document. We can take advantage of this, by building a table of such symbols, and represent each occurrence of a symbol with a reference to that symbol's table entry. This will save a lot of space, and a similar idea is used in djvu's jb2 mode and JBIG[a] for PDF.

smoothscan builds up this table, but instead of filling the table with the original raster images, it vectorizes each symbol. Vector images will look smoother than their raster equivalents, and can be scaled without introducing pixelation. These properties result in a smaller output file size, as well as making the scanned text images more readable.

smoothscan saves the vectorized images into a custom TrueType font and embeds the font into the output pdf file. Currently each symbol is mapped to an arbitrary letter in the font, but in future versions you could run OCR on each symbol, and ensure that the 'o' image is associate with the 'o' character encoding in the generated font.

To get good results, you must have good input. Higher resolution scans capture more detail about the shape of each symbol, so a higher quality vectorized version can be created. It's a good idea to process your scanned images using a toot like ScanTailor before running smoothscan.

Current smoothscan can onle process pure black and white 1bpp images, but in the future support will be added for other formats, especially ScanTailor's Mixed output mode.

smoothscan is currently targeted at GNU/Linux based systems, but Windows and OS X will be supported in future versions.