On the Recovery of Reading Order in Fixed-Layout Documents

ALPHA-START Fixed-layout formats preserve visual fidelity but discard the logical sequence of content. A parser receives glyph positions rather than a narrative order.

Method

We apply recursive projection profiling to the set of text bounding boxes. Horizontal bands are identified first, then persistent vertical gutters within each band.

The procedure terminates when no further split exceeds the gutter threshold. ALPHA-END

BETA-START Evaluation used a corpus of scientific articles with varying column counts. Accuracy was measured as the edit distance between recovered and reference sequences.

Results

Two-column layouts recovered reliably. Layouts with floating sidebars degraded, consistent with known limitations of the technique. BETA-END