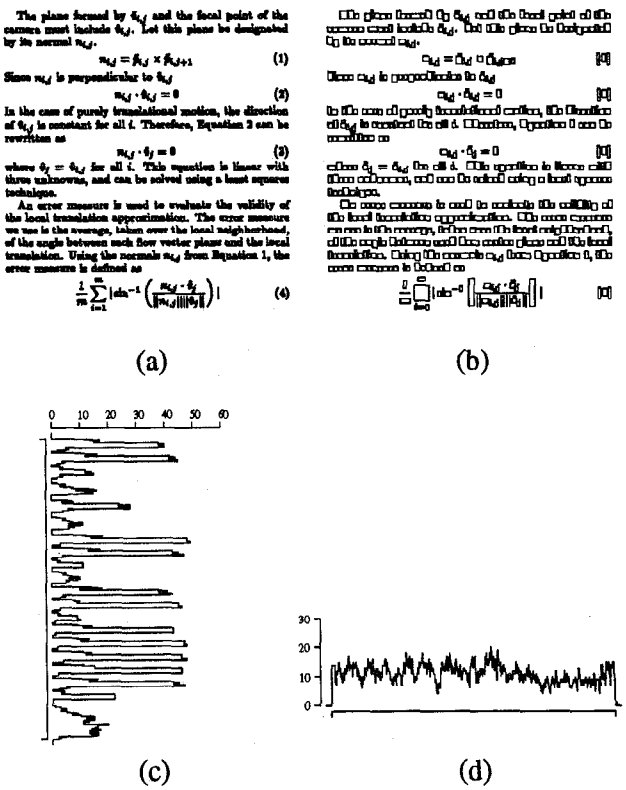

Figure 1: (a) an English document, (b) bounding boxes of connected components of black pixels, (c) horizontal projection profile, (d) vertical projection profile.

always larger than the number of symbols since multiple bounding boxes are produced for multi-component symbols. Our page decomposition scheme analyzes the spatial configuration of those bounding boxes of connected components to extract textlines, words, and paragraphs.

2.2 Projections of Bounding Boxes

Analysis of the spatial configuration of bounding boxes can be done by projecting those bounding boxes onto a straight line. Since paper documents are usually written in the horizontal or vertical direction, projections of bounding boxes onto the vertical and horizontal lines are of particular interest. While projecting bounding boxes onto the horizontal or vertical line, they will accumulate onto that line, which results in the projection profile. A projection profile is a frequency distribution of the projected bounding boxes on the projection line. The bounding box projection profiles provide important information about the number of bounding boxes aligned along the projection direction. Figure 1(c) and 1(d) shows the horizontal and vertical projection profiles of the bounding boxes in Figure 1(b).

2.3 Extraction of Textlines

In this step, the algorithm first determines the textline direction of the page by analyzing both horizontal and vertical projection profiles. Once the textline direction of the page is determined, the algorithm partitions the page bounding box into textline bounding boxes.

From Figure 1(c) and 1(d), it is easy to see that textlines are horizontally oriented: On the horizontal projection profile, there are distinct high peaks and deep valleys at somewhat regular intervals, whereas on the vertical projection profile, there is no such distinction. Since the bounding boxes are represented by the coordinates of two opposite end points, textlines are easily extracted and Figure 1(f) shows the result.

2.4 Extraction of Words

In this step, the algorithm groups the bounding boxes on each textline (produced from the last step) into bounding boxes of words.

The algorithm first computes the projection profiles within each of the textline bounding boxes. Figure 1(e) shows projection profiles within textlines. Next, the algorithm considers each of the projection profiles as a one-dimensional *gray-scale image*, and thresholds each of the images with threshold value 1 to produce a binary image. Note that, during the binarization, a symbol (or a broken symbol) with multiple bounding boxes may be merged into one, as well as, those adjacent symbols within the same textline whose bounding boxes are overlapping with each other. But this will not cause any problem in the result of our word extraction process, since our algorithm extracts words by merging bounding boxes based on the lateral proximity of neighboring boxes.

After such binarization, the algorithm performs a morphological closing operation on each of the binarized textline projection profiles with structuring element of appropriate size. The length of the structuring element is determined by analyzing the distribution of the run-lengths of 0's on the binarized textline projection profile. In general, such a run-length distribution is bi-modal. One mode corresponds to the inter-character spacings within words, and the other to the inter-word spacings. A threshold value can be chosen in the valley between the two dominant histogram modes. The two elegant techniques suggested

Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95)
0-8186-7128-9/95 $10.00 © 1995 IEEE