	mismatched all	mismatched text	total
word seg.	227 (0.37%)	63	60936
groundtruth	318 (0.52%)	73	61027

Table 1: word segmentation result

Δx and Δy are difference of x- and y-coordinates of two consecutive bounding boxes listed in the character groundtruth file, and T_x and T_y are some positive integer values. Then we find correspondence of the symbols in the zone-based groundtruth file and the symbols associated with bounding boxes in the character groundtruth file. By grouping the bounding boxes which form a word, we can generate the word box groundtruth data for synthetic images in the database.

3.2 Evaluation of the Word Segmentation Algorithm

The output of the word segmentation algorithm is a set of word bounding boxes. To evaluate the performance, we need to compare the word box groundtruth data and the word bounding boxes produced by the word segmentation algorithm. Let $\mathcal{G} = \{G_1, G_2, \cdots, G_N\}$ denote the total of N groundtruth word bounding boxes and let $\mathcal{D} = \{D_1, D_2, \cdots, D_M\}$ denote the total of M detected word bounding boxes which are produced by the word segmentation algorithm. For our purpose, we simply compute $\mathcal{G} - \mathcal{D}$ and $\mathcal{D} - \mathcal{G}$.

Technical/scientific document images usually contain math expressions embedded in text lines. Our word segmentation algorithm will produce word bounding boxes based on the amount of space between consecutive symbols. Therefore, if a word bounding box produced by the algorithm contains a pure text word, it is the word bounding box in the true sense. However, if a word bounding box produced by the algorithm contains an inline math expression, we still call it a word bounding box because we are not concerned with the content of the box until symbol recognition is attempted.

Table 1 shows the number of elements in $\mathcal{G} - \mathcal{D}$ and $\mathcal{D} - \mathcal{G}$ versus the number of elements in $\mathcal{G}$ and $\mathcal{D}$. In the tables, "mismatched all" represents the number of mismatched word bounding boxes, and "mismatched text" represents the number of mismatched bounding boxes of pure text words. In the last column, the total number of bounding boxes are recorded.

4 Discussions

In this paper, we describe a new document page decomposition technique. The entire decomposition process is based on the analysis of the spatial configuration of bounding boxes of connected components. In our approach, connected components become the lowest level of the document hierarchy.

Correction of page skew is not of concern in this study. However, it is worth mentioning that performance of our decomposition method strongly depends on how much a document image is skewed. In fact, we may not be able to correctly extract text lines for about 0.5° skew of a letter-sized, single column, single spaced text document. Therefore, deskewing of the document image must precede the decomposition.

Our decomposition method has its own computational aspect: Once bounding boxes are obtained, the method does not refer to actual images. During the decomposition process, the method manipulates only bounding boxes. Hence, for a letter-sized document image at 300 dpi resolution, the number of computational units are reduced from $8.4 \times 10^6 (= 2550 \times 3300)$ pixels to at most a few thousands of bounding boxes. Our decomposition method [5][6] completes page decomposition within a few second and, thus, its superiority to the pixel-projection approach is obvious.

References

[1] Ihsin T. Phillips, Su Chen and R. M. Haralick, "English Document Database Standard," *Proc. ICDAR*, Japan, 1993.

[2] R.M. Haralick and L.G. Shapiro, *Computer and Robot Vision : Volume I*, Addison Wesley, 1992

[3] J. Kittler and J. Illingworth, "Minimum Error Thresholding," *Pattern Recognition*, Vol. 19, No. 1, pp.41–47, 1986

[4] Nobuyuki Otsu, "A Threshold Selection Method from Gray-level Histogramws," *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. SMC-9, No. 1, pp. 62–66, January, 1979

[5] Jaekyu Ha, Ihsin T. Phillips and R. M. Haralick, "Recursive X–Y Cut using Bounding Boxes of Connected Components," ISL Report, Dept. Electrical Eng., University of Washington, 1994.

[6] J. Ha, I.T. Phillips and R.M. Haralick, "Document Image Decomposition using Bounding Boxes of Connected Components," ISL Report, Dept. Electrical Eng., University of Washington, 1994.

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