

The Component-Linking Tree Revisited

Bo Yuan¹, Leong Keong Kwoh¹, and Chew Lim Tan²

¹Centre for Remote Imaging, Sensing and Processing
National University of Singapore, Singapore 119260
{yuanbo, lkkwoh}@nus.edu.sg

²Department of Computer Science, School of Computing
National University of Singapore, Singapore 117543
tancl@comp.nus.edu.sg

Abstract. Component grouping has many uses in document image analysis. This paper introduces several important refinements for our component-grouping algorithm that is based on the component-linking tree [1]. The University of Washington English Document Image Database (UW-I) is used for the applicability evaluations.

1. Introduction

Texts in the printed documents are presented in blocks, such as paragraphs and columns, so that their logical relationships can be visually conveyed. In document image analysis (DIA), component grouping or page segmentation plays an important role in revealing such a relationship [2].

We have developed a component-grouping algorithm [1] that is fast and robust for DIA applications where the full-fledged page segmentation algorithms are either slow or overkill. In this grouping algorithm, a grouping function in Eq. (1) is evaluated for any two components with areas s_1 and s_2 . If the value is larger than the Euclidian distance between the two components, they are considered *directly linked*. Given a collection of components, all the unique combinations are evaluated for their linkages. The result is a collection of acyclic multi-trees as Fig. 1 shows.

$$f(s_1, s_2) = \sqrt{ks_1s_2/(s_1 + s_2)} \quad (1)$$

There are several choices regarding the use of Eq. (1):

- The type of components to choose: connected components, their upright/best-fit bounding-boxes, or their convex hulls?
- The fiducial-points [3] for the components: centroids, or bottom-centers?
- The setting of the adjustable parameter k : a single universal value, or a dynamic scheme that automatically adjusts its value for different samples?

This paper evaluates these choices and gives decisions based on the evaluations using the real world samples in the University of Washington English Document Image Database (UW-I) [4].

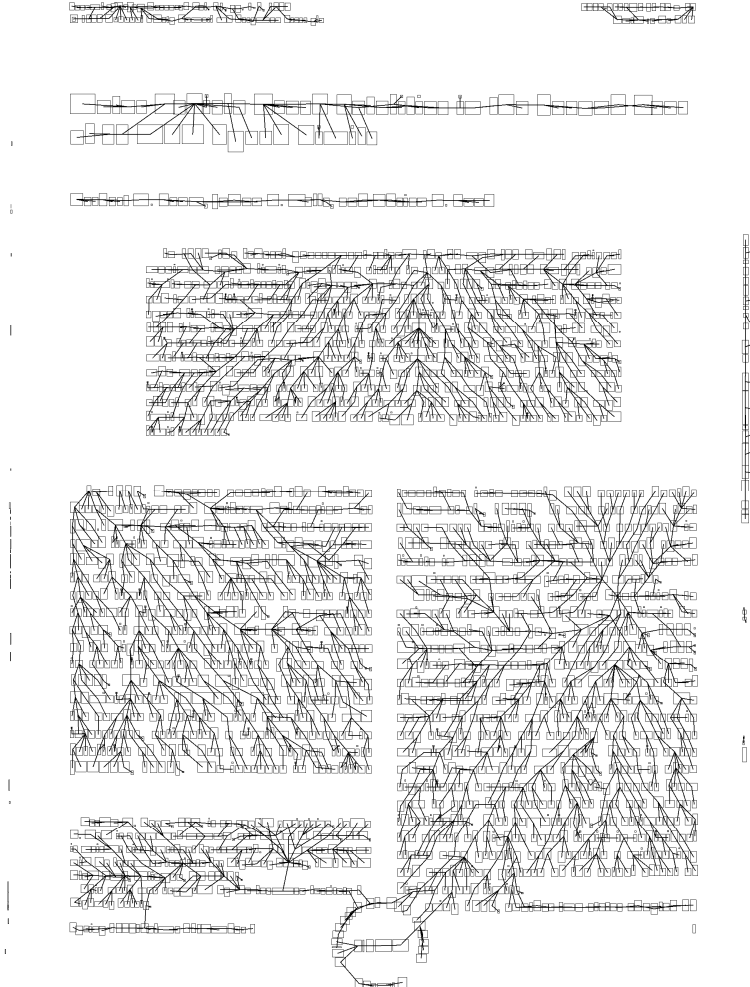


Fig. 1. The direct links among the connected components, which are established by the proposed grouping function in Eq. (1) on the image A04B in UW-I. The direct links are drawn among the centroids of the upright bounding-boxes of the connected components.

2. From Component Links to Text Blocks

The main goal of developing Eq. (1) is to reveal the block structures of the textual document images. It is called the grouping function because when the direct links are established, which can be seen in Fig. 1, the components that belong to the same tree form a group that closely represents a text block, be it a paragraph or a column.

In our previous work [1], the connected components are directly used, and their sizes are used as s_1 and s_2 . Further experimental results show that the upright

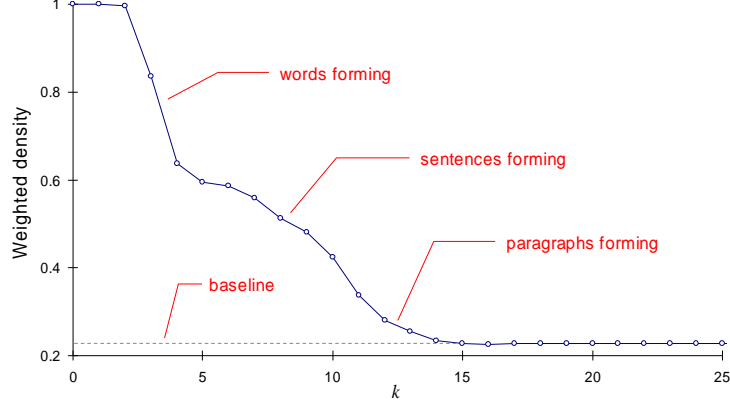


Fig. 2. The component-grouping stages determined by the value of k . This result is obtained from using the image A000 in UW-I.

bounding-boxes of the connected components are a better choice overall than that of the connected components for the same font size because the upright bounding-boxes have areas closer than that of the connected components. Furthermore, the detection speed of the upright bounding-boxes are 3 times faster than that of the convex hulls.

The parameter k plays a crucial role in Eq. (1). As shown in Fig. 2, when k is small, the characters in a document are all isolated (no links among the components). When k is very large, almost any two components can be linked. In between the two extremes of k , words and paragraphs can be formed by the trees of the component links. In Fig. 2, we choose the weighted density ρ_{wtd} of the trees (groups) in Eq. (2) as the criterion to evaluate the various component linking (grouping) stages.

$$r_{wtd} = \frac{\sum_{g=0}^{N-1} r_g S_g}{\sum_{g=0}^{N-1} S_g} = \frac{\sum_{g=0}^{N-1} \sum_{i=0}^{n_g} S_i}{\sum_{g=0}^{N-1} S_g} \quad (2)$$

In Eq. (3), N is the number of component groups in an image, n_g is the number of components in group g , s_i is the area of component i in group g , and s_g is the area of group g . Note that the upright bounding-boxes are used to calculate all the areas.

We used the full set of 979 real document images and the 168 synthesized images in UW-I [4] to estimate the proper range of k in forming paragraphs. Fig. 3 shows the distribution of the smallest k that reaches the baseline in Fig. 2 for the individual test images using the connected components and their bounding-boxes, respectively. The top chart in Fig. 3 clearly shows that k varies widely for different sample images, while the bottom chart has well-defined peaks.

3. Conclusions

The final choices obtained from this paper for the component-grouping function in Eq. (1) are: (i) use the upright bounding-boxes to represent the components, as real

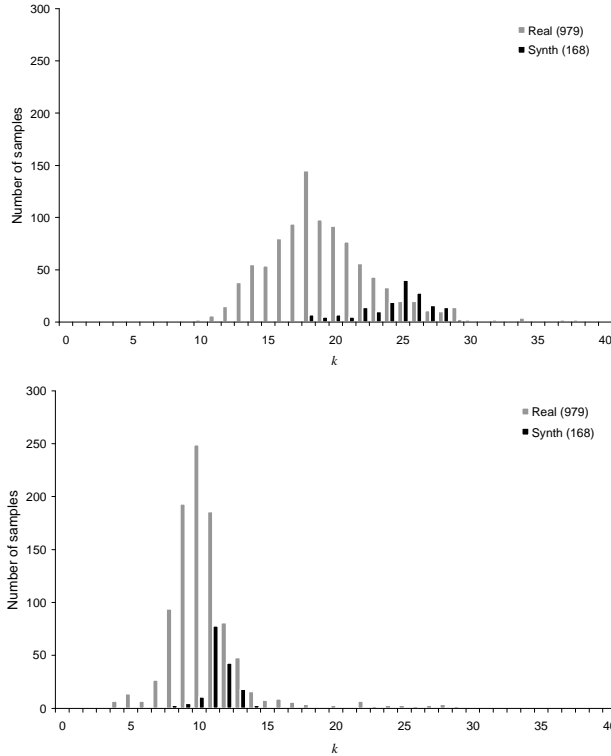


Fig. 3. The distribution of the weighted densities versus the smallest k that reaches the baseline in Fig. 2 for the test images in UW-I when the connected components are used (top) and when the bounding-boxes of the connected components are used (bottom).

world document images have skew angles within $\pm 3^\circ$ [4]; (ii) use the centroids of the upright bounding-boxes of the components as the fiducial points; (iii) use a single value of $k = 20$ for the whole batch of samples, such as the UW-I samples. Different batches may use different values. This choice is resolution independent because both Eq. (1) and the Euclidian distance are proportional to resolution.

References

1. B. Yuan, and C. L. Tan, "A Multi-Level Component Grouping Algorithm and Its Applications", Proceedings of the Eighth International Conference on Document Analysis and Recognition, pp. 1178-1181, Seoul Korea, 29 August - 1 September 2005.
2. G. Nagy, "Twenty Years of Document Image Analysis in PAMI", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22 (1), pp. 38-62, January 2000.
3. A. L. Spitz, "Determination of image skew angle from data including data in compressed form", United States Patent 5245676, September 1993.
4. S. Chen, M. Y. Jaisimha, J. Ha, I. T. Phillips, and R. M. Haralick, UW English Document Image Database I Reference Manual, 1993.