

A Unified Methodology for Document Structure Analysis

Jisheng Liang[†] Ihsin T. Phillips[‡] Robert M. Haralick[†]

[†] University of Washington, Seattle, USA

[‡] Seattle University, Seattle, USA

{jliang, yun, haralick}@george.ee.washington.edu

1 Introduction

In this paper, we formulate document image structure analysis as a partitioning problem. The goal of the problem is to find an optimal solution to partition the set of glyphs on a given document to a hierarchical tree structure where entities within the hierarchy at each level have similar properties and compatible semantic labels. This paper describes a document image structure extraction algorithm that is probability based, where the probabilities are estimated from an extensive training set of various kinds of measurements of distances between the terminal and non-terminal entities with which the algorithm works. The off-line probabilities estimated in the training then drive all decisions in the on-line segmentation module. An iterative, relaxation-like method is used to find the partitioning solution that maximizes the joint probability.

We have implemented algorithms that extract text-lines and text-blocks using this framework. The algorithms were evaluated on the UW-III database [1] of some 1600 scanned document image pages. The text-line extraction algorithm identifies and segments 99.76% of text-lines correctly, while the preliminary result of the text-block extraction shows 91% accuracy.

2 The Methodology

Let \mathcal{A} be the set of entities at the source_level within the hierarchy. Let Π be a partition of \mathcal{A} and each element τ of the partition is an entity on the target_level. Let L be a set of labels that can be assigned to elements of the partition. Function $f : \Pi \rightarrow L$ associates each element of Π with a label. $V : \wp(\mathcal{A}) \rightarrow \Lambda$ specifies measurement made on subset of \mathcal{A} , where Λ is the measurement space.

The problem can be formulated as follows: given initial set \mathcal{A} , find a partition Π of \mathcal{A} , and a labeling function $f : \Pi \rightarrow L$, that maximize the probability

$$\begin{aligned} P(V(\tau) : \tau \in \Pi, f, \Pi | \mathcal{A}) &= P(V(\tau) : \tau \in \Pi | \mathcal{A}, \Pi, f) P(\Pi, f | \mathcal{A}) \\ &= P(V(\tau) : \tau \in \Pi | \mathcal{A}, \Pi, f) P(f | \Pi, \mathcal{A}) P(\Pi | \mathcal{A}). \end{aligned} \quad (1)$$

By making the assumption of conditional independence, that when the label $f(\tau)$ is known then no knowledge of other labels will alter the probability of $V(\tau)$, we can decompose the probability (1) into

$$P(V(\tau) : \tau \in \Pi, f, \Pi | \mathcal{A}) = \prod_{\tau \in \Pi} P(V(\tau) | f(\tau)) P(f | \Pi, \mathcal{A}) P(\Pi | \mathcal{A}). \quad (2)$$

The possible labels in set L is dependent on the target_level and on the specific application. For example, $l \in L$ could be text content, functional content type, style attribute, etc.

The above proposed formulation can be uniformly applied to construction of the document hierarchy at any level, e.g., text-word, text-line, and text-block extractions, just to name a few. For example, as for the text-line extraction, given a set of glyphs, the goal of the text-line extraction is to partition glyphs into a set of text-lines, each text-line having homogeneous properties, and the text-lines' properties within the same region being similar. The text-lines' properties include, deviation of glyphs from the baseline, direction of the baseline, text-line's height and width, etc. As for the text-block segmentation, for example, given a set of text-lines, text-block segmentation groups text-lines into a set of text-blocks, each block having homogeneous formatting attributes, e.g. homogeneous leading, justification, and the attributes between neighboring blocks being similar.

Given an initial set \mathcal{A} , we first construct a partial ordering of the elements of \mathcal{A} , according to their physical locations and reading direction. Let $A = (A_1, A_2, \dots, A_M)$ be a linearly ordered set (chain in \mathcal{A}) of input entities. Let $\mathcal{G} = \{Y, N\}$ be the set of grouping labels. Let A^P denote a set of element pairs, such that $A^P \subset A \times A$ and $A^P = \{(A_i, A_j) | A_i, A_j \in A \text{ and } j = i + 1\}$. Function $g : A^P \rightarrow \mathcal{G}$, associates each pair of adjacent elements of A with a grouping label, where $g(i) = g(A_i, A_{i+1})$. Then, the partition probability $P(\Pi|A)$ can be computed as follows,

$$\begin{aligned} P(\Pi|A) &= P(g|A) = P(g(1), \dots, g(N-1) | A_1, \dots, A_N) \\ &= P(g(1) | A_1, A_2) \times \dots \times P(g(N-1) | A_{N-1}, A_N) \\ &= \prod_{i=1}^{N-1} P(g(i) | A_i, A_{i+1}). \end{aligned} \quad (3)$$

Therefore, the joint probability is further decomposed as

$$P(V(\tau) : \tau \in \Pi, f, \Pi|A) = \prod_{\tau \in \Pi} P(V(\tau) | f(\tau)) \times P(f|\Pi, A) \prod_{i=1}^{N-1} P(g(i) | A_i, A_{i+1}). \quad (4)$$

An iterative search method is developed to find the consistent partition and labeling that maximizes the joint probability of Equation (4). First, the grouping probability between each pair of adjacent input entities is computed, by observing the spatial relationships between the pair. An initial partition is determined based on the initial grouping probabilities. Then, we start to adjust the partition and assign labels to the elements of the partition, by maximizing the labeling probability. At each iteration, the adjustment that produces the maximum improvement of the joint probability (4) is selected. The iteration stops when there is no improvement on the joint probability. A detailed description of our method is presented in [2].

3 Text-line Extraction [3]

Without loss of generality, we assume that the reading direction of the text-lines on the input page is left-to-right. The text-line segmentation algorithm starts with the set of the glyph bounding boxes of a given textual document.

Given observations on a pair of adjacent glyphs A_i and A_{i+1} , where the glyph is represented by a bounding box (x, y, w, h) , we compute the probability that A_i and A_{i+1} belong to the same text-line,

$$P(g(i) | h_i, h_{i+1}, w_i, w_{i+1}, d(i), o(i)),$$

where h_i is the height of A_i , w_i is the width, $d(i)$ is the distance between A_i and A_{i+1} , and the vertical edge overlap is $o(i)$.

Let $T = (A_i, A_{i+1}, \dots, A_k)$ be an extracted group of glyphs. Function f associates T with a label of homogeneous text-line,

$$P(\text{angle}(T), \text{dev}(T)|f(T)),$$

where $\text{angle}(T)$ is the direction of the baseline of T , and $\text{dev}(T)$ is the mean absolute deviation of glyphs from the baseline.

Let (T_1, \dots, T_q) be a sequence of text-lines within a region R . Each text-line $T_p \in R$ is represented by its height and width (w_p, h_p) . Therefore, $W_R = (w_1, \dots, w_p)$ is a sequence of text-line width and $H_R = (h_1, \dots, h_p)$ is a sequence of text-line height. The probabilities that the sequence of text-lines within R have homogeneous height and width are estimated as $P(\text{median}(W_R), \max(W_R))$ and $P(\text{median}(H_R), \max(H_R))$.

4 Text-block Extraction

Given observations on a pair of adjacent text-lines T_i and T_{i+1} , where a text-line is represented by its bounding box (x, y, w, h), we compute the probability that T_i and T_{i+1} belong to the same block,

$$P(g(i, i+1)|h_i, h_{i+1}, d(i), o(i), e_l(i), e_c(i), e_r(i)).$$

The measurements made on a pair of adjacent text-lines are: x-height h_i and h_{i+1} , inter-line spacing $d(i)$, horizontal overlap $o(i)$, left edge offset $e_l(i) = x_i - x_{i+1}$, center edge offset $e_c(i) = x_i - x_{i+1} + (w_i - w_{i+1})/2$, and right edge offset $e_r(i) = x_i - x_{i+1} + w_i - w_{i+1}$.

Given a text-block $B \in \Pi$, we compute the probability that B has homogeneous leading, and certain type of text alignment. Let $B = (T_i, T_{i+1}, \dots, T_k)$ and $D_B = (d(i), d(i+1), \dots, d(k-1))$, where $d(j)$ is the inter-line space between A_j and A_{j+1} . Function f associates B with a label of homogeneous leading with probability, $P(\text{median}(D_B), \max(D_B)|f(B))$.

Given observations $V(B)$ on a text-block B , we determine the probability that B has certain text alignment type $f(B)$ (justification, indentation, and hanging), represented as $P(V(B)|f(B))$. Measurements $V(B)$ are the mean and maximum absolute deviation of text-lines' edges from corresponding edge of the text-block. The context constraint $P(f|\Pi, T)$ is modeled as a Markov chain,

$$P(f(B_t)|f(B_{t-1}), \dots, f(B_1)) = P(f(B_t)|f(B_{t-1}))$$

where $B_t \in \Pi$. Therefore, the alignment labeling probability

$$\prod_{\tau \in \Pi} P(V(\tau)|f(\tau))P(f|\Pi, T) \tag{5}$$

is actually a hidden Markov model. Given a partition, sequence of labels which maximizes the probability (5) can be computed using the Viterbi algorithm. The details of this algorithm can be found in [2].

5 Experimental Results

We applied our text-line extraction algorithm to the total of 1600 images from the UW-III Document Image Database [1]. The numbers and percentages of miss, false, correct, splitting, merging and spurious detections are shown in Table 1(a). Of the 105,020 ground truth text-lines, 99.76% of them

are correctly detected, and 0.08% and 0.07% of lines are split or merged, respectively. Examples of successful cases and failures are presented in [2] and [3]. Table 1(b) illustrates the numbers and percentages of miss, false, correct, splitting, merging and spurious detections of text-blocks.

The results of two existing techniques, rule-based and parametric model-based, are reported in [4] on the same 1600 images. The rule-based algorithm [5] computes the projection profile of connected component bounding boxes, and detects 94.78% of text-lines and 75.64% of text-blocks correctly. The parametric model-based algorithm [6] has an accuracy of 96.49% on text-line extraction and a performance of 72.96% on segmenting text-blocks.

Table 1: Performance of (a) text-line extraction algorithm and (b) text-block extraction algorithm.

	Total	Correct	Splitting	Merging	Mis-False	Spurious
Ground Truth	105020	104773 (99.76%)	80 (0.08%)	78 (0.07%)	79 (0.08%)	10 (0.01%)
Detected	105019	104773 (99.77%)	172 (0.16%)	37 (0.04%)	25 (0.02%)	12 (0.01%)

(a)

	Total	Correct	Splitting	Merging	Mis-False	Spurious
Ground Truth	21788	19828 (91.00%)	560 (2.57%)	1250 (5.74%)	1 (0.01%)	149 (0.68%)
Detected	21709	19828 (91.34%)	1219 (5.61%)	501 (2.31%)	0 (0.00%)	161 (0.74%)

(b)

References

- [1] I.T. Phillips. *Users' Reference Manual*, CD-ROM, UW-III Document Image Database-III, 1995.
- [2] J. Liang. *Document Structure Analysis and Performance Evaluation*, Ph.D. thesis, University of Washington, 1999.
- [3] J. Liang, I.T. Phillips, and R.M. Haralick. "A Statistically Based, Highly Accurate Text-line Segmentation Method," ICDAR '99.
- [4] J. Liang, I.T. Phillips, and R.M. Haralick. "Performance Evaluation of Document Layout Analysis Algorithms on the UW-III Data Set," *Document Recognition IV*, Proceedings of the SPIE '97.
- [5] J. Ha, R.M. Haralick, and I.T. Phillips. "Document page decomposition using bounding boxes of connected components of Black Pixels," *Document Recognition II*, Proceedings of the SPIE '95.
- [6] S. Chen, R.M. Haralick, and I.T. Phillips. "Extraction of Text Lines and Text Blocks on Document Images Based on Statistical Modeling," *International Journal of Imaging Systems and Technology*, Vol. 7, No. 4, pp. 343-356, Winter, 1996.