Document Zone Classification Using Sizes of Connected-components

Jisheng Liang†, Ihsin T. Phillips†, Jaekyu Ha†, and Robert M. Haralick†

†Department of Electrical Engineering, University of Washington
Seattle, Washington 98195

†Department of Computer Science, Seattle University
Seattle, Washington 98122

ABSTRACT

In this paper, we describe a feature based supervised zone classifier using only the knowledge of the widths and the heights of the connected-components within a given zone. The distribution of the widths and the heights of the connected-components is encoded into an $n \times m$ dimensional vector in the decision making. Thus, the computational complexity is in the order of the number of connected-components within the given zone.

A binary decision tree is used to assign a zone class on the basis of its feature vector. The training and testing data sets for the algorithm are drawn from the scientific document pages in the UW-I database. The classifier is able to classify each given scientific and technical document zone into one of the eight labels: text of font size 8-12, text of font size 13-18, text of font size 19-36, display math, table, halftone, line drawing, and ruling, in real time. The classifier is able to discriminate text from non-text with an accuracy greater than 97%.

Keyword: Document image layout analysis, zone classification, connected components, decision tree classifier.

1 INTRODUCTION

Understanding scientific and technical documents involves estimating the rotation skew of each document page, determining the geometric page layout, labeling blocks as text, math, drawing, table, halftone, etc., determining the text of text blocks through an OCR system, determining the logical page layout, and formatting the data and information of the document in a suitable way for use by a word processing system or by an information retrieval system. However, the correctness of block labeling plays the key role in the success of such a document understanding system.

Many page layout analysis systems have incorporated the zone labeling module within their systems. Wahl et. al. extract features of the zones including the area of the connected component of the zone, the number of black pixels in the zone on the original document image, the mean horizontal black run lengths of the original image within the zones, and the height and width of the bounding rectangle of the zone. Text areas are classified into
extract connected component features such as component height, width, aspect ratio, density, perimeter, and area for classifying each block as text or non-text. Saitoh and Pavlidis\textsuperscript{5} classify each component into text, text or noise, diagram or table, halftone image, horizontal separator, or vertical separator, using block attributes such as block height, height to width ratio, and connectivity features of the line adjacency graph, and whether there are vertical or horizontal rulings. Pavlidis and Zhou\textsuperscript{6} label each zone as text or non-text using features such as ratio of the mean length of black intervals to the mean length of white intervals, the number of black intervals over a certain length, and the total number of intervals. Amamoto et. al.\textsuperscript{7} decide that a block is a text block if the length of the longest black run length in the vertical and horizontal directions is smaller than a given threshold. Each block is then assigned a class label from the set: text, figure, image, table, and separation line. Sivaramakrishnan et. al.\textsuperscript{2} extract features for each zone such as run length mean and variance, spatial mean and variance, fraction of the total number of black pixels in the zone, and the zone width ratio, and use a decision tree classifier to assign a zone class on the basis of its feature vector.

However, except for Sivaramakrishnan\textsuperscript{2} which involves extensive computation, there is not a systematic evaluation of the performance of the zone classification modules. In this paper, we describe a feature based supervised zone classifier using only the knowledge of the widths and the heights of the connected-components within the given zone. The distribution of the widths and the heights of the connected-components is encoded into a \( n \times m \) dimensional vector in the decision making. Thus, the computational complexity is in the order of the number of connected-components within the given zone.

A binary decision tree is used to assign a zone class on the basis of its feature vector. The classifier is able to classify each given scientific and technical document zone into one of the eight labels: text of font size 8-12, text of font size 13-18, text of font size 19-36, display math, table, halftone, line drawing, and ruling, in real time.

The training and testing data sets for the algorithm are drawn from the scientific document pages in the UW-I database. The performance evaluation of the algorithm for each zone type are reported.

2 ZONE FEATURES BASED ON THE SIZES OF CONNECTED COMPONENTS

The bounding box of a connected component is defined to be the smallest rectangle which circumscribes the connected component. The height and width of connected component are defined as the height and width of its bounding box. It is the common observation that different kinds of zones usually have different distributions of connected component heights and widths. For example, a text zone has many connected components with the size of individual characters, while a line drawing zone usually includes connected components with much larger sizes. The histogram of connected component bounding box widths and heights for different zone classes are given in Figure 1 and Figure 2.

The distribution of the widths and heights of the connected components is encoded into a \( n \times m \) dimensional feature vector (Figure 3), where \( n \) and \( m \) are the number of intervals for width and height respectively. In our experiment, we choose the value of \( n \) and \( m \) as 10 and 11. The width intervals are 10, 20, 30, 40, 80, 160, 320, 640, 1280, 2560. The height intervals are 10, 20, 30, 40, 50, 100, 200, 400, 800, 1600, 3200. For each zone, the number of connected components whose heights and weights fall in each bucket is counted and normalized. We use two different normalization methods, one is the number of connected components in each bucket divided by the total number of connected components within each zone, the other one is the number of connected components in each bucket divided by the area of each zone. So the feature values are the normalized connected components numbers corresponding to different height and width intervals.
Figure 1: illustrates the distribution of connected components heights and widths for each zone class. (a) Text; (b) Display math; (c) Table; (d) Halftone;
Figure 2: illustrates the distribution of connected components heights and widths for each zone class. (e) Line drawing; (f) Ruling.
3 DECISION TREE CLASSIFIER

A decision tree classifier makes the assignment through a hierarchical decision procedure. The classification process can be described by means of a tree, in which at least one terminal node is associated with each class and nonterminal nodes represent various collections of mixed classes.

For the construction of a decision tree, we need a training set of feature vectors with true class labels. Let $U = \{u_k : k = 1, \ldots, N\}$ be the unit-training set to be used to design a binary tree classifier. Each unit $u_k$ has an associated measurement $X_k$ with known true class. At any non-terminal node, let $\Omega^n$ be the set of $M^n$ classes still possible for a unit at node $n$. Let $U^n = \{u_k^n : k = 1, \ldots, N^n\}$ be the subset of $N^n$ training units associated with node $n$. If the number of units for class $c$ in node $n$ is denoted by $N^n_c$, we must have $N^n = \sum_{c=1}^{M^n} N^n_c$.

Now we describe how the decision rule works at node $n$. Consider unit $u_k^n$ which has measurement vector $x_k^n$. If the discriminant function $f(x_k^n)$ is less than or equal to a threshold, then $u_k^n$ is assigned to class $\Omega^n_{left}$, otherwise it is assigned to class $\Omega^n_{right}$. An assignment to $\Omega^n_{left}$ means that a unit descends to the left child node and an assignment to $\Omega^n_{right}$ can be understood in a similar way. Given a discriminant function $f$, the units in $U^n$ are sorted in such a way that $f(x_k^n) \leq f(x_{k+1}^n)$ for $k = 1, \ldots, N^n - 1$. Let $w_k^n$ be the true classes associated with the measurement vectors $x_k^n$. Then a set of candidate thresholds $T^n$ for the decision rules is defined by

$$ T^n = \left\{ \frac{f(x_{k+1}^n) - f(x_k^n)}{2} \mid w_k^n \neq w_{k+1}^n \right\} $$

(1)

For each threshold value, unit $u_k^n$ is classified by using the decision rule specified above. We count the number of samples $n_{Lc}^t$ assigned to $\Omega_{left}^n$ whose true class is $c$ and we count the number of samples $n_{Rc}^t$ assigned to $\Omega_{right}^n$ whose true class is $c$, that is,

$$ n_{Lc}^t = \# \{ u_k^n \mid f(x_k^n) \leq t \text{ and } w_k^n = c \} $$

$$ n_{Rc}^t = \# \{ u_k^n \mid f(x_k^n) > t \text{ and } w_k^n = c \} $$

Let $n_L^t$ be the total number of samples assigned to $\Omega_{left}^n$ and $n_R^t$ be the total number of samples assigned to
\[ n^t_L = \sum_{c=1}^{M^n} n^t_{Lc} \quad \text{and} \quad n^t_R = \sum_{c=1}^{M^n} n^t_{Rc} \]

We define the impurity \( IP_n^t \) of the assignment made by node \( n \) to be

\[
IP_n^t = \sum_{c=1}^{M^n} \left( -n^t_{Lc} \log \frac{n^t_{Lc}}{n^t_L} - n^t_{Rc} \log \frac{n^t_{Rc}}{n^t_R} \right)
\]  

The discriminant threshold \( t \) is chosen such that it minimizes the impurity value \( IP_n^t \). The impurity is such that it gives a minimum value when the training samples are completely separable.

The learned discriminant function splits the training subset into two subsets, generating two child nodes. The process is repeated at each newly generated child node until a stopping condition is satisfied, and the node is declared as a terminal node based on a majority vote.

The maximum impurity reduction is used as the stopping criterion for tree expansion. At a given node, the impurity before distinction \( IP_B \) and the impurity after distinction \( IP_A \) are computed. Let \( \theta \) be a predetermined value. If \( IP_B - IP_A < \theta \), partitioning is halted and this node is made a terminal node. The maximum depth of the tree and minimum number of samples to stop expanding node are also predetermined as stopping criterion.

In the simplest form of a linear decision rule (\( f \) is linear), one of the components of the measurement vector is taken and a set of candidate thresholds, \( T \), are computed for that component. The one that gives the maximum impurity reduction is chosen. This process is repeated for all the components in the measurement vector. Out of the thresholds computed for all the components in the measurement vector, the one that yields maximum impurity reduction is chosen.

At the testing stage, a feature vector is input to a decision tree, a decision is made at every non-terminal node as to what path the feature vector will take. This process is continued until the feature vector reaches a terminal node of the tree, where a class is assigned to it.

4 EXPERIMENTS AND RESULTS

We have applied the zone classification algorithm to a significant sized data set. The hold-out method is used for the error estimation. We divide the data set into \( N \) parts, train on the first \( N - 1 \) parts, and then test on the \( N \)th part. Then train on the \( N - 1 \) parts, omitting the \( N - 1 \)st part, and test on the \( N - 1 \)st part. Continue the training and testing, each time omitting one part from the decision tree construction procedure and then testing on the omitted part. Then combine the results of the \( N \) tests together to establish an estimate of the error rate.\(^8\) In this experiment, the value of \( N \) is chosen as 3.

The output of the decision tree is compared with the zone labels from the ground truth in order to evaluate the performance of the algorithm. A contingency table is computed to indicate the number of zones of a particular class label that are identified as members of another class.

The training and testing data set is drawn from the scientific document pages in the University of Washington document image database. It has 979 scientific and technical document pages with a total of 13726 zones. The class labels for each of the zones are obtained from the database. These zones belong to eight different classes: 3 text classes (of font size 8-12pt, font size 13-18pt and font size 19-32pt), math, table, halftone, line drawing, and ruling. These classes were assigned labels 1 – 8. The features were computed for every zone in the document.
Table 1: Normalized by the total number of connected components within each zone. Contingency table showing the number of zones of a particular class that are assigned as members of each possible zone class: In the table, $T_1$, $T_2$, $T_3$, $M$, $T$, $H$, $D$, $R$ represent text with font size $\leq 12\text{pt}$, text with font size $\geq 13\text{pt}$ and $\leq 18\text{pt}$, text with font size $\geq 19\text{pt}$, math, table, half tone, line drawing, and ruling, respectively.

<table>
<thead>
<tr>
<th></th>
<th>$T_1$</th>
<th>$T_2$</th>
<th>$T_3$</th>
<th>$M$</th>
<th>$T$</th>
<th>$H$</th>
<th>$D$</th>
<th>$R$</th>
<th>Total</th>
<th>Error %</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>11675</td>
<td>60</td>
<td>1</td>
<td>82</td>
<td>20</td>
<td>8</td>
<td>46</td>
<td>6</td>
<td>11504</td>
<td>1.87</td>
</tr>
<tr>
<td>$T_2$</td>
<td>73</td>
<td>115</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>16</td>
<td>4</td>
<td>204</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>$T_3$</td>
<td>1</td>
<td>84</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>17</td>
<td>595</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>$M$</td>
<td>100</td>
<td>2</td>
<td>0</td>
<td>385</td>
<td>1</td>
<td>0</td>
<td>24</td>
<td>0</td>
<td>114</td>
<td>2.15</td>
</tr>
<tr>
<td>$T$</td>
<td>42</td>
<td>1</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>7</td>
<td>27</td>
<td>0</td>
<td>133</td>
<td>0.87</td>
</tr>
<tr>
<td>$H$</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>121</td>
<td>27</td>
<td>0</td>
<td>156</td>
<td>1.19</td>
</tr>
<tr>
<td>$D$</td>
<td>80</td>
<td>5</td>
<td>7</td>
<td>18</td>
<td>35</td>
<td>294</td>
<td>5</td>
<td>459</td>
<td>3.06</td>
<td></td>
</tr>
<tr>
<td>$R$</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>745</td>
<td>244</td>
<td>0</td>
<td>257</td>
<td>1.57</td>
</tr>
</tbody>
</table>

Table 2: Normalized by the total area of zone. Contingency table showing the number of zones of a particular class that are assigned as members of each possible zone class: In the table, $T_1$, $T_2$, $T_3$, $M$, $T$, $H$, $D$, $R$ represent text with font size $\leq 12\text{pt}$, text with font size $\geq 13\text{pt}$ and $\leq 18\text{pt}$, text with font size $\geq 19\text{pt}$, math, table, halftone, line drawing, and ruling, respectively.

<table>
<thead>
<tr>
<th></th>
<th>$T_1$</th>
<th>$T_2$</th>
<th>$T_3$</th>
<th>$M$</th>
<th>$T$</th>
<th>$H$</th>
<th>$D$</th>
<th>$R$</th>
<th>Total</th>
<th>Error %</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>11675</td>
<td>46</td>
<td>0</td>
<td>107</td>
<td>23</td>
<td>8</td>
<td>41</td>
<td>5</td>
<td>1164</td>
<td>0.92</td>
</tr>
<tr>
<td>$T_2$</td>
<td>70</td>
<td>115</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>8</td>
<td>2</td>
<td>204</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>$T_3$</td>
<td>2</td>
<td>2</td>
<td>82</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>7</td>
<td>95</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>$M$</td>
<td>138</td>
<td>3</td>
<td>0</td>
<td>334</td>
<td>4</td>
<td>1</td>
<td>32</td>
<td>0</td>
<td>312</td>
<td>3.5</td>
</tr>
<tr>
<td>$T$</td>
<td>60</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>43</td>
<td>2</td>
<td>23</td>
<td>0</td>
<td>104</td>
<td>0.67</td>
</tr>
<tr>
<td>$H$</td>
<td>15</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>87</td>
<td>4</td>
<td>46</td>
<td>2</td>
<td>196</td>
<td>4.3</td>
</tr>
<tr>
<td>$D$</td>
<td>84</td>
<td>8</td>
<td>5</td>
<td>32</td>
<td>18</td>
<td>31</td>
<td>275</td>
<td>2</td>
<td>480</td>
<td>4.0</td>
</tr>
<tr>
<td>$R$</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>3</td>
<td>244</td>
<td>0</td>
<td>257</td>
<td>1.06</td>
</tr>
</tbody>
</table>

For the purposes of creating a decision tree, the set of feature vectors was divided into 2 parts, one part with two thirds of the vectors, used for creating the decision tree and the other part with the remaining one third of the vectors, used for testing the tree. The testing is done three times each with a different third of the tree data set as test and two thirds of the data set for training. The contingency tables for the result of the classification algorithm with two different normalization methods are shown in Table 1 and Table 2. The table shows the number of zones of a particular class that are identified as members of a different class. The performance of this algorithm for all zone types, on average, is around 94 % to 95 %. However, if we only consider classifying zones into text or non-text, the classification accuracy is above 97 %. The contingency table is shown in Table 3.

The largest classification error occurs in confusing math zones and text zones. This is due to the fact that some display math zones have the similar connected componentswidths and heights distributions as text zones. The error of text zones confused with drawing or table zones is a result of underlined text in text zones and the appearances of significant number of text in drawing and table zones. Noting that our algorithm is connected

<table>
<thead>
<tr>
<th></th>
<th>Text</th>
<th>Non-text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>98.46 %</td>
<td>1.54 %</td>
</tr>
<tr>
<td>Non-text</td>
<td>14 %</td>
<td>86 %</td>
</tr>
</tbody>
</table>

Table 3: Contingency table showing the classification results of text and non-text zones. (a) Normalized by the total number of connected components within each zone; (b) Normalized by the total area of zone.
5 CONCLUSIONS AND DISCUSSIONS

After the page decomposition which partitions an document image into zones, correctly determining the zone class is very important for the further processes. We have presented a new method which classifies each given zone into one of the eight labels: text of font size 8-12, text of font size 13-18, text of font size 19-36, display math, table, halftone, line drawing, and ruling. In our approach, we only use the heights and widths of the connected components within each zone. Thus, the computational complexity is in the order of the number of connected-components within the given zone. We have tested our method on a significant sized data set. The accuracy for text and non-text distinction is greater than 97%.

Our future work will include the statistically analysis of the heights and widths distributions of connected components. For example, we are trying to take into the locations of distribution cluster of connected components heights and widths, and improve the classification accuracy between text, drawing and table zones.

6 REFERENCES


