Page Layout Classification Technique for Biomedical Documents

Daniel X. Le and George R. Thoma
National Library of Medicine
8600 Rockville Pike, Bethesda, MD 20894

ABSTRACT

The structural layout information of scanned document pages is valuable for a wide range of document processing applications such as automatic document searching, document delivery and automated data entry. This paper describes the classification of scanned document pages into different classes of physical layout structures. The page layout classification technique proposed in this paper uses a combination of geometry-based and content-based zone features calculated from optical character recognition (OCR) output.

Geometry-based and content-based features are derived from geometric zone information and zone contents respectively. A new feature called “single and multiple column zone vertical area string pattern” is also proposed to normalize document image pages.

After normalizing document pages, a template matching algorithm calculates similarity classification features by matching vertical area string patterns of document pages to those of predefined layout document structures. Similarity classification features and both geometry-based and content-based zone features are then input into a rule-based learning system for the final decision on the page layout classification structure.

The performance of our document page layout classification scheme has been evaluated using a sample size of several hundred images of biomedical journal pages. Preliminary evaluation results show that our approach is capable of classifying journal pages into different classes of physical layout structures at an accuracy of more than 96 %.

Keywords: Document labeling, Page layout classification, Automated d ocument data entry, MEDLINE database, National Library of Medicine.

1. INTRODUCTION AND BACKGROUND

Automated document conversion systems are being developed for a variety of document related applications to convert paper-based document information to electronic format. Paper documents usually consist of text zones, or a mixture of text and non-text zones. In order to support automated document applications, techniques are required to identify the contents of each text zone (“labeling”). At the National Library of Medicine (NLM) we are developing an automated system [7, 8], the *Medical Article Record System* (MARS), to identify and convert bibliographic information from paper-based biomedical journals to electronic format for inclusion in the MEDLINE database used by biomedical professionals worldwide. This paper describes one component of this system: the classification of scanned document pages into different classes of physical layout structures.

Most document labeling techniques proposed so far in the literature [1, 2, 3, and 4] are based on the layout (geometric) structure and/or the logical structure of a document. Hones et al. [1] described an algorithm for layout extraction of mixed-mode documents. Taylor et al. [2] described a prototype system using 'feature extraction and model-based' approach. Tsujimoto et al. [3] presented a technique based on the transformation from a geometric structure to a logical structure. Tateisi et al. [4] proposed a method based on stochastic syntactic analysis to extract the logical structure of a printed document. Le [6] recommended a rule-based system for document labeling in which rules are derived from generic typesetting knowledge for English text. Other techniques [5, 7, 8 and 9] have used the outputs of OCR to further improve labeling accuracy. In this paper, we propose an automated technique to handle page layout classification by normalizing document image pages and applying a combination of a template string pattern matching algorithm and a rule-based system algorithm on features derived from OCR outputs. Preliminary evaluation results show that the system is capable of labeling text zones at a classification accuracy of more than 96.0%.

The rest of this paper is divided into eight sections. Section 2 describes the features used in our classification scheme. Section 3 provides a system overview. Section 4 presents classification features. Sections 5, 6, 7 and 8 describe in detail the page layout classification technique and experimental results. Section 9 contains a summary and conclusion.

2. STRING PATTERNS

Basic definitions necessary to understand our algorithms are given here.

2.1 Single and multiple column zone vertical areas

A single column zone vertical area of a binary image is defined as a vertical area in which only one text zone exists. A multiple column zone vertical area of a binary image is a vertical area in which more than one zone exists, and they are "vertically overlapped". Two zones are considered to be vertically overlapped if the top and/or the bottom coordinates of one zone are within the top and the bottom coordinates of another zone. Figure 1 shows an example of the single and multiple column zone vertical areas.

2.2 Vertical area string patterns

Let "M", "S", and "*" be the vertical areas of multiple column zone, single column zone, and empty line spaces. Let "C", "L", "R", "l", and "r" be the zone location features: Center, Left, Right, Left of Center, and Right of Center. Let "N", "Y" and "+" be No, Yes, and "Don't Care" respectively. There are two types of patterns: geometry-based and content-based patterns that are defined as follows.

2.2.1 Geometry-based "single and multiple column zone" vertical area string pattern

The "single and multiple column zone" vertical area string pattern is the combination of characters "M", "S", and "*" that represent the top-to-bottom vertical areas of a binary image. An example of this type of string pattern is shown in Figure 1(1) as "M*S*S*S*S*S*S*M*S".

2.2.2 Geometry-based "zone location" vertical area string pattern

The "zone location" vertical area string pattern is the combination of characters "C", "L", "R", "l", "r", and "+" that represent the relative location of a zone against the vertical middle line of a page. The following logic is used to determine the zone location in a page.

```
If | Zone Vertical Middle Line – Page Vertical Middle Line | is less than or equal to CENTER_THRESHOLD
Zone is center "C".
Else If Zone Vertical Middle Line is less than Page Vertical Middle Line and Zone Right Coordinate is greater than Page Vertical Middle Line
Zone is left "L".
Else If Zone Vertical Middle Line is greater than Page Vertical Middle Line and Zone Left Coordinate is less than Page Vertical Middle Line
Zone is right "R".
Else If Zone Vertical Middle Line is less than Page Vertical Middle Line and Zone Right Coordinate is less than or equal to Page Vertical Middle Line
Zone is left of center "l".
Else If Zone Vertical Middle Line is greater than Page Vertical Middle Line and Zone Left Coordinate is greater than or equal to Page Vertical Middle Line
Zone is right of center "r".
End If
```

The CENTER_THRESHOLD is selected to be about two 12-point characters and for 300 dot per inch document, its value is about 100 pixels. The "zone location" vertical area string pattern of an image shown in Figure 1(2) is "++C+C+C+C+C+L+++C".

2.2.3 Content-based "single and multiple text lines zone" vertical area string pattern

The "single and multiple text lines zone" vertical area string pattern is the combination of characters "Y", "N", and "+". "Y" characters are for zones having more than one text line and "N" characters are for one text line zones. The "single and multiple text lines zone" vertical area string pattern of an image shown in Figure 1(3) is "Y+Y+Y+Y+N+Y+N+Y+N".

2.2.4 Content-based "Nth order font size zone" vertical area string pattern

The "Nth order font size zone" vertical area string pattern is the combination of characters "Y", "N", and "+". *The smaller the order, the larger the font size*. "Y" characters are for zones of which font sizes are categorized as Nth order and "N" characters are for zones not having Nth order font size. Examples of the "1st and 2nd order font size zone" vertical area string patterns are shown in Figures 1(4) and 1(5) as "N+Y+N+N+N+N+N+N+N" and "N+N+Y+N+N+N+N+N+N" respectively.

2.2.5 Content-based "Nth order percentage of capital characters zone" vertical area string pattern

The definition of "Nth order percentage of capital characters zone" vertical area string patterns is similar to the definition presented in the previous section 2.2.4. The difference is that the percentage of capital characters compare to total characters of a zone is used instead of the font size. *The smaller the order, the larger the percentage.* Figures 1(6) and 1(7) show the "1st and 2nd order percentage of capital characters zone" vertical area string patterns as "N+Y+N+N+N+N+N+N+N" and "N+N+Y+N+N+N+N+N+N" respectively.

3. SYSTEM OVERVIEW

The page layout classification technique described in this paper is one component of our 2nd-generation MARS system under development at NLM [7, 8]. The classification process takes a scanned binary image as its input, performs OCR and calculates classification features for each text zone and for the entire image page, and then labels each text zone as title, author, affiliation, abstract, or others.

Classification features include both geometric (geometry-based) layout features and non-geometric (content-based) layout features. Geometry-based features calculated include page content coordinates, zone coordinates, zone dimensions, zone locations, zone order, number of columns, column dimensions, and column locations. Content-based features derived from zone contents obtained during the OCR operation include total characters, total capital characters, total punctuation marks, number of text lines, average font size, and average line spacing.

Using OCR output generated by a commercial 5-engine OCR system developed by Prime Recognition Inc. (PR) [10] for each zone of a page, geometry-based and content-based zone features are calculated and both will then be used to create several predefined types of vertical area string patterns for the entire page. Finally zone features and vertical area string patterns are input into a combination of template matching and rule-based learning system for label classification.

The purpose of creating vertical area string patterns, especially the "single and multiple column zones" pattern, is to normalize the document image page. Generally, the number of text lines in a labeled zone such as title, author, affiliation, or abstract is different from one article to another in a journal issue and therefore the labeled zone coordinates of one article may not be the same as those of another article. As a result, using the same document style guide, the geometric page layout of one article may not be the same as that of another article in the same journal issue. In order to overcome this problem of irregularity, we propose in this paper a new feature called "single and multiple column zone vertical area string pattern" that will be used to normalize the page images of a journal.

As defined in section 2.2.1, the "single and multiple column zone" vertical area string pattern consisting of characters "M", "S", and "*" can be created by identifying vertical areas having single or multiple column zones from the top of a binary image to the bottom. Using this feature, we could have the same vertical area string patterns for document pages that use the same document style guide. Figures 1 and 2 show an example of two binary images having contents with different number of text lines but sharing the same "single and multiple column zones" vertical area string patterns.

CLASSIFICATION FEATURES

Features calculated for this page layout classification technique are based on an analysis of the page layout for each journal. Geometry-based features include zone dimensions, zone locations, umber of columns, column dimensions, and column locations. Content-based features derived from zone contents obtained during the OCR operation include total characters, total capital characters, total punctuation marks, number of text lines, average font size, and average line spacing.

A list of 16 features for each zone and 2 features for the entire page used in the page layout classification technique is presented as follows:

Geometry-based zone features:
Zone coordinates (left and top)
Zone dimensions (height and width)
Zone location (center, left, right, left and right of center)
Content-based zone features:
Zone content Total text lines Total characters
Total capital characters
Total punctuation marks Average font size Average line spacing
Geometry-based page features:
Page content frame coordinates (left and right)

The page content left/right frame coordinate is defined as the left-most/right-most coordinate of text zones in an image page.

5. PAGE LAYOUT CLASSIFICATION PROCESS

The page layout classification process consists of five steps: (1) scan journal images, (2) perform OCR, (3) calculate geometry-based and content-based zone features, (4) normalize by creating vertical area string patterns for the entire page, and (5) finally, submit geometry-based and content-based zone features and vertical area string patterns to a template matching and rule-based learning system for label classification. In the following subsection, each step will be discussed in detail.

5.1 Scan journal images

In this step, the first page of each article of a journal issue is scanned and saved as a binary document image. Image processing operations such as page orientation and skew detection are then applied to improve the quality of a scanned image. The images of pages in landscape mode are automatically rotated to be in portrait mode, and skewed page images are rotated to correct skew angle.

5.2 Perform OCR

Using a commercial 5-engine OCR system, each scanned binary document image is segmented into text and graphics zones. Each text zone is processed to deliver an OCR output (including zone coordinates, characters and their bounding boxes, confidence levels, font sizes and style attributes).

5.3 Calculate geometry-based and content-based features

In this step, using the OCR output generated for a scanned image page, sixteen geometry-based and content-based zone features and two page features as defined in Section 4 are calculated.

5.4 Normalize

Zone features are then used to create the “single and multiple column zones” vertical area pattern and the rest of vertical area patterns defined in Section 2.

5.5 Template matching and rule-based learning

In this step, the template matching algorithm matches vertical area string patterns of a binary image to those of predefined layout document structures of a given journal type to derive two types of similarity classification features: degrees of geometry-based similarity and degrees of content-based similarity. If both similarity degrees exceed a predefined weight threshold, the label classification of a predefined article page will be used to label zones of a binary image; otherwise the rule-based will be used to make the final decision on the page layout classification structure. The following algorithm summarizes the template matching and rule-based learning system procedure to label zones of a binary image of a particular document type using a predefined weight threshold of 170 points. Assume that the document type is using two orders of font size and two orders of the percentage of capital characters.

```
Set the weight matching to 0
If “single and multiple column zone” patterns are matched, add 100 points to weight matching.
If “zone location” patterns are matched, add 50 points to weight matching.
If 1st order font size zones is used and if its patterns are matched, add 10 points to weight matching.
If 2nd order font size zones is used and if its patterns are matched, add 10 points to weight matching.
If 1st order percentage of capital characters zones is used and its patterns are matched, add 10 points to weight matching.
If 2nd order percentage of capital characters zones is used and its patterns are matched, add 10 points to weight matching.
    If the weight matching is at least 170 points
    Label zones using the predefined labels vertical area string patters to handle page layout classification.
    Else
    Use rule-based learning system for classification.
    End if
Else
Use rule-based learning system for classification.
End if
```

6. PREDEFINED VERTICAL AREA STRING PATTERNS

For each journal type, a small set of article image pages are used to generate predefined vertical area string patterns and each pattern is labeled as title, author, affiliation, abstract, or other. Since all labeled zones consist of text only, it is reasonable to automate the generation of the predefined vertical area string patterns along with their label classifications by matching the content of zones labeled by the user against that of an image. Let “1”, “2”, “4”, “8”, “0” be title, author, affiliation, abstract, and others. An example of predefined vertical area string patterns of an image article shown in Figure 1 using up to two orders is as follows:

M*S*S*S*S*S*S*M*S	”Single and multiple columns zone” [see (1)]
++C+C+C+C+C+L+++C	“Zone location” [see (2)]
Y+Y+Y+Y+N+Y+N+Y+N	“Single and multiple text lines zone” [see (3)]
N+Y+N+N+N+N+N+N+N	“1st order font size zones” [see (4)]
N+N+Y+N+N+N+N+N+N	“2nd order font size zones” [see (5)]
N+Y+N+N+N+N+N+N+N	“1st order percentage of capital characters zones” [see (6)]
N+N+Y+N+N+N+N+N+N	“2nd order percentage of capital characters zones” [see (7)]
0+1+2+4+0+8+0+0+0	“Label”

7. RULE-BASED LEARNING SYSTEM

The rule-based learning system is used to classify zones of an image if the degrees of similarities do not pass the weight threshold test for the template matching algorithm. In order to support the rule-based learning system, the following predefined features are generated for each journal type:

1. Top-down sequences of document labels
2. Font sizes:

Total font size orders
What font size orders to be used?

3. Percentage capital characters:
Total percentage of capital characters orders.
What percentage of capital characters orders to be used?
4. Label having the maximum font size.
5. Number of text lines for each label.
6. Label locations (center, left, right, left and right of center)

The rule-based learning algorithm is implemented using the elimination and selection procedure. The algorithm begins with the *elimination* process by eliminating any zone for a particular label identification if the zone does not pass all rules for that label. The algorithm continues with the *selection* process in which a zone is assigned a particular label if it passes more rules for that label than any other zones.

It is important to note that rules and predefined vertical area string patterns can be generated and updated automatically for each document type using the article page image information and article zone labels identified and confirmed by users.

8. EXPERIMENTAL RESULTS

The page layout classification technique has been implemented and experiments have been conducted with binary document images selected from several different medical journals. All documents used in these experiments are 8.5 x 11 inches in size and were scanned at 300 dpi resolution. A test sample consisting of 524 article page images from four different journal types was used in the experiment, and the algorithm has correctly classified 503 image pages, giving an accuracy rate of 96 %. Most errors were due to the inaccurate segmentation generated by the autozoning feature of the commercial OCR system. For example, a zone of interest (such as title zone) would be split into multiple zones, or several different zones (such as author and affiliation zones) would be merged into a single zone.

9. SUMMARY AND CONCLUSIONS

A page layout classification technique based on a combination of template matching and a rule-based learning system has been presented. The technique provides meaningful labels for the contents of text blocks such as article titles, authors, affiliations, and abstract. The technique performed well on a set of 4 different biomedical journals and showed the possibility of extension to other journals. It is noted that our approach using the proposed feature called “single and multiple column zones” pattern is able to successfully handle pages having the same document style guide but different geometric page layouts.

10. REFERENCES

[1] F. Hones and J. Lichter, Layout Extraction of Mixed Mode Documents, Machine Vision and Applications 7, pp. 237-246, 1994.

[2] S. Taylor, R. Fritzson, and J. Pastor, Extraction of Data from Preprinted Forms, Machine Vision and Applications 5, pp. 211-222, 1992.

[3] S. Tsujimoto and H. Asada, Major Components of a Complete Text Reading System, Proc. IEEE, Vol. 80, No. 7, pp. 1133-1149, 1992.

[4] Y. Tateisi and N. Itoh, Using Stochastic Syntactic Analysis for Extracting a Logical Structure from a Document Image, Proc. IEEE Int. Conf. Neural Networks, Vol. 2, pp. 391-394, 1994.

[5] T. Hu et. al., A Prototype for Extracting Logical Elements from Tables of Contents of Journals, Int. Assoc. Patt. Recog. Workshop on Doc. Analysis System, Malvern, PA, 1997

[6] D. X. Le, "Document Analysis and Labelling System," *Ph.D. dissertation,* Computer Science Department, SITE, George Mason University, February 3, 1997.

[7] D. X. Le, J. Kim, G. Pearson, and G. R. Thoma, "Automated Labeling of Zones from Scanned Documents," *Proc. SDIUT,* at Annapolis, MD, pp.219-226, 1999.

[8] D. X. Le and G. R. Thoma, "Automated Labeling Using Integrated Image and Neural Processing," *Proc. SCI’99 and ISAS’99,* Orlando, FL, Vol. 6, pp.105-108, 1999.

[9] J. Liang et. al., The Prototype of a Complete Document Image Understanding System, Int. Assoc. Patt. Recog. Workshop on Doc. Analysis System, Malvern, PA, 1996.

[10] Prime Recognition Inc., Prime OCR Access Kit