

Features for Neural Net Based Region Identification of Newspaper Documents

Tim Andersen, Wei Zhang
Computer Science Department, College of Engineering
Boise state University
Boise, ID 83725 USA
E-mail: tim@cs.boisestate.edu
wzhang@onyx.boisestate.edu

Abstract

Several features for Neural Network based document region identification are tested. Specifically, this paper examines features for non-text region identification. The Neural Network based region identification algorithm is a key component of a document recognition system that segments a document into regions, classifies them into text, graphic, photo, and other region types, and then uses this classification to guide the processing and analysis of the image. The input data are unusually challenging: low quality images of newspaper documents obtained from microfilmed archives. The results compare favorably with other results reported in the literature.

1. Introduction

Document image analysis involves recognizing text, graphics and pictures in images and extracting important information as a human would. Textual processing and non-textual processing are two categories of document image analysis dealing, respectively, with the text components and non-text components of a document image. For text components, the goal is to separate them into headline and paragraph components, link them together according to article and correct word ordering, and recognize the text through an OCR system. The goal for non-text components is to separate them into graphics/drawings (non-half-tone) and photo(half-tone) categories and apply different algorithms to these regions in order to improve the display quality. The identification and elimination of non-text regions from the document before presentation to the OCR system has the added benefit of increasing the OCR accuracy and speed. Identification of headline regions is the first step to link headlines and paragraphs together to create the article index. Therefore, region classification is an important step in document image analysis.

Many of the existing page segmentation algorithms have the region classification modules embedded in their systems. Numerous ways to segment and classify regions have been proposed in the literature, ranging from purely top-down approaches that recursively split a page into smaller components, to purely bottom-up approaches that attempt to cluster individual connected components into larger and larger entities, with many variations in between [3] [5].

Top-down approaches usually segment the page into homogenous regions before the classification stage. After homogenous regions are produced by the segmentation process various features are extracted from the region and used to classify it into an appropriate category. A number of different features have been proposed for region classification. Some of these include the area of the connected component of a block, number of black pixels in the region, the mean horizontal black run length, component width height ratio, component density, mean length of black intervals to the mean length of white intervals, number of black intervals over a certain length, feature-based interaction map [12], texture discrimination masks [3], periodicity measures [14], the measures of visual attention (legibility, complexity, attractiveness, etc.) [6][7][8], and many others. After the features are extracted a binary decision tree classifier or simple thresholding is often used to classify the region.

Bottom-up approaches usually classify each image pixel first and then group the pixels into regions. The texture Co-occurrence Spectrum technique [4], convolution [1], wavelet packets [9], multiscale texture segmentation [15], texture discrimination masks [13], mask based local textural characteristics extraction [11], etc. are some of the techniques which can be used to classify each image pixel. Some bottom-up approaches [2][10] that are based on geometric relations group the pixels according to various patterns and then use simple statistical features to classify them. Our approach, which falls into the top-down category, will be introduced in the next section.

2. Training and test set creation.

The segmentation algorithm used to generate training samples comes from a document processing and recognition system that is used to process newspaper documents scanned from microfilm. The document processing system has the following basic steps. First, a grayscale image is scanned from microfilm, and is then de-skewed and binarized using a local adaptive thresholding algorithm. Second, the segmentation algorithm, which is based on a modified recursive X-Y cut algorithm, over-segments the binarized image into regions using horizontal and vertical projections of the page to determine where to split regions. Because of the free format of newspaper documents, with overlapping and mixed text and non-text areas, the recursive X-Y cut projection algorithm cannot create homogenous regions if it does not over-segment the image. Third, a neural network is used to classify the over-segmented regions created by the segmentation step. Fourth, the regions are coalesced into larger regions, and a second pass classification is done to (attempt to) correct any misclassifications. At the end, an XML document is created to store the pertinent information for each region (bounding box location, region type, relationship of regions).

In the first classification pass the algorithm extracts features from the over-segmented regions and uses a neural network based classifier to identify each region. The neural networks used in the classification step are standard multilayer perceptrons trained with back-propagation. There are several reasons for choosing a neural network based classification scheme for the task of region identification. First, neural networks are very general and robust, and any feature that can be encoded as a number can be utilized by the neural network to perform its decision task. Second, as new features are designed, the neural network can be automatically retrained with these features to increase the accuracy of classification. Third, the system is much less dependent on user-defined parameters and does not require a programmer or expert to fine tune the system to new data (such as newspaper documents from different eras or publishers). If a different newspaper needs to be processed, then the only human effort required is to manually zone and label a representative sample of documents from that newspaper that can then be used to automate the production of a training set, and a new network can be trained to label the regions for that paper.

The test and training sets are created from the manually zoned regions in the following manner. First, each document is manually zoned and classified. Each region is bounded by a single rectangle and is labeled with one

region type. All of this information is stored in one XML file (one XML file per newspaper page).

Second, a modified recursive X-Y cut algorithm is used to segment the newspaper pages and to create the regions with rectangle boundary and various extracted features. The segmentation algorithm used in this step is the same as that used to segment images by the production system, and tends to (drastically) over-segment the image. This segmentation algorithm works independently of the classification algorithm.

Third, a region type is assigned to each over-segmented region by examining the ground truth and making the assignment based on region boundary geometry relationships. Currently, text regions that are contained in advertisements are discarded from the training set. Finally, all of the extracted machine generated regions are randomized and separated into training and test sets according to a user chosen split rate.

Section 3 introduces the features used for text/non-text separation. The experiments and result of the tests are presented in section 4. Discussion and conclusion are given in sections 5 and 6.

3. Feature set

The features we use are based on the region and connected components in the region or touching the region. In the experiments, many different features beyond the features that are listed below, almost 110 features in all, were tested. Some of these features were obtained from the literature, while some of the features were designed by us. The set of features listed below were found, after many experiments on a variety of newspaper documents, to produce the best generalization performance. In order to ease the selection of which features to use for the neural network, we developed a visual tool that incorporates a fast decision tree learning technique to assist in the selection of good combinations of features.

The following definitions are used in the feature definitions. Here x and y refer to connected components which are bounded by rectangular bounding boxes.

$$overlap(x, y) = \begin{cases} 1 & \text{if } x \text{ overlaps with } y \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$numOverlap(x) = \sum_y overlap(x, y) \quad (2)$$

$$area(x) = \text{the size of the bounding box for } x \quad (3)$$

$$sumOverlapArea(x) = \sum_y area(x \cap y) \quad (4)$$

$$surface(x) = \text{the surface area of } x \quad (5)$$

$weight(x)$ = the number of pixels for x (6)

$height(x)$ = the height of x (7)

$bigComp(x, \theta) = \begin{cases} 1 & \text{if } weight(x) \geq \theta \\ 0 & \text{otherwise} \end{cases}$ (8)

$numBigComp(rect) = \sum_{x \in rect} bigComp(x, \theta)$ (9)

$numComp(rect) = \text{sum of components}$ (10)

The most obvious characteristics of a text region are related to its regularity. For example, connected components in text regions generally do not overlap with each other (here overlap refers to overlap between the rectangular bounding boxes of the components, not between the pixels of the components), since each component in a text region is of relatively the same size and surrounded by a buffer of white space. With non-text regions, especially graphics and pictures, the connected components that are part of the graphic will often have significant overlap.

Because the regions generated by the initial segmentation step are often small (over-segmented), some regions may entirely contain only one or two (or even none) entire components. In order to handle the case where there are only a few components in the region, the amount of overlap in a region is also determined from components that “touch” the region, rather than just from components which are strictly inside the region. Therefore, we gather two sets of overlap statistics for one region, one based on the connected components which touch the region, and the other is based on those connected components which are strictly contained within the region. These two sets of overlap features complement each other. The overlap related features input into the neural network are calculated as follows,

Sum of the number of overlaps for all components in the region:

$$sumNumOverlap(rect) = \sum_{x \in rect} numOverlap(x) \quad (11)$$

Maximum number of overlap for a connected component in the region:

$$\max_{x \in rect} (numOverlap(x)) \quad (12)$$

Total size of overlap area for the region:

$$totalOverlapArea(rect) = \sum_{x \in rect} sumOverlapArea(x) \quad (13)$$

Average value of percent of overlap area based on one connected component:

$$avgPerOverlap(rect) = \frac{\left(\sum_{x \in rect} sumOverlapArea(x) / area(x) \right)}{|\{x | x \in rect\}|} \quad (14)$$

In addition to the overlap features, we also use what we like to call “cheesiness” features. With the local adaptive thresholding algorithm we use, text areas often differ from graphics and pictures in the binarized image in terms of the surface area to mass ratio of connected components. When viewed at the pixel level, a connected component from a picture region will often look similar to “Swiss cheese”, while a text component has a more solid appearance and smoother edges. The “cheesiness” features capture this notion and are calculated as follows:

$$avgBigCompCheesiness(rect) = \frac{\left(\sum_{x \in rect} bigComp(x, \theta) surface(x) / weight(x) \right)}{numBigComp(rect)} \quad (15)$$

$$avgCompCheesiness(rect) = \frac{\left(\sum_{x \in rect} surface(x) / weight(x) \right)}{numComp(rect)} \quad (16)$$

Cheesiness of whole region

$$regionCheesiness(rect) = \frac{\left(\sum_{x \in rect} surface(x) \right)}{\left(\sum_{x \in rect} weight(x) \right)} \quad (17)$$

In general, the largest component in a photo region is much larger than the largest component that occurs in a text region. In addition, graphic regions tend to contain more small components than text regions. So we also include features designed to capture this information.

$$maxCompWeight(rect) = \max_{x \in rect} (weight(x)) \quad (18)$$

$$avgBigCompHeight(rect) = \frac{\sum_{x \in rect} bigComp(x, \theta) height(x)}{numBigComp(rect)} \quad (19)$$

$$numSmallComp(rect) = \sum_{x \in rect} (1 - bigComp(x, \theta)) \quad (20)$$

And finally, text components tend to have fairly consistent ratios of black to white pixels in the rectangular bounding box that bounds the text component. The average of these ratios for each region is used, and the standard deviation of the ratios for each region is also used since the degree to which this statistic varies within a region is often more indicative of region ID than the average.

$$avgBWRatio(rect) = \frac{\sum_{x \in rect} \frac{weight(x)}{area(x) - weight(x)}}{numComp(rect)} \quad (21)$$

$$stdevBWRatio(rect) = \text{stdev of b/w ratios} \quad (22)$$

4. Experiments

In these experiments, a networks composed of 20 nodes in one hidden layer, and one output node were used. Because the number of non-text regions is far less than the number of text regions, the positive and negative samples are separated into two distinct pools, and every training cycle a sample is randomly chosen from one of these pools based on the balance rate. The balance rate used for these experiments forces an equal number of positive samples and negative samples to be presented to the network.

The document images used in the training and test sets were provided by iArchives, which is a document processing company located in Orem, Utah. These pages included 220 pages from the "Chicago Daily Tribune", 1928, 124 pages from "The Dallas Morning News", 1928, 55 pages from "The Washington Post", 1997, and 13 pages from "The Austin American Chronicle", 1941. All of these pages were scanned from microfilm and as such have a great deal of noise and are of relatively poor quality in general.

In order to generate training and test sets, first small regions are created by the modified recursive X-Y cut projection algorithm and the regions are labeled (text and non-text) based on the ground truth, then the regions are randomized and split into a training (70%) and test set (30%). The first results that we give are restricted to pages from the "Chicago Daily Tribune", 1928. The

accuracy is given in table 1. The accuracy for all of the pages in the dataset is given in table 2.

Table 1. classification results of text and non-text region on Chicago newspaper

	Text (detected)	Non-text (detected)	Total (detected)	Accuracy (%)
Text (actual)	21638	361	22022	98.26
Non-text(actual)	70	1723	1793	96.10
Total			23815	98.09

Table 2. classification results of text and non-text region on all newspaper

	Text (detected)	Non-text (detected)	Total (detected)	Accuracy (%)
Text (actual)	43204	916	44120	97.92
Non-text(actual)	121	3033	3154	96.16
Total			47274	97.81

5. Discussion

Although the training set that contains all of the data has more training samples the accuracy for this data is a little bit lower than the single source, since it is more difficult for a single net to classify many different kinds of newspaper documents.

In [2] the pattern is extracted using a fast scan method, and classification uses pattern characteristics such as spread and pattern context to segment the pattern, the total accuracy is 98%. [3] uses a simple mask that makes use of the different correlation properties to classify the region. The test documents are from the University of Washington Document Image Database. The region location is provided by the UWDB, which is from ground truth. The best total accuracy is 98.3%. [12] uses texture features derived from a feature based interaction map. The region location for the classification algorithm is obtained from ground truth. The text identification accuracy is 97--98%, the non-text identification accuracy is 84%--89%, and the total accuracy is 96%. [14] uses multi-scale analysis and top-down approach to segment, and uses a periodicity measure to classify. The test documents are from UWDB. The region location accuracy is 97.7%, the text identification accuracy is 99.7%, the image identification accuracy is 97.1%. Our results are better than [12], approximately equal to [2][3], and slightly worse than those reported in [14]. But when taking into account the performance of their recursive X-Y cut method the region location accuracy is 91%, the text identification accuracy just 94%, and image identification accuracy just 87.5%.

6. Conclusion

We have presented several critical features that we use in a neural network based region identification algorithm for newspaper documents. Using these features, the neural network based approach is able to achieve excellent identification accuracy. The high accuracy is especially surprising in light of the degree to which the regions are over-segmented. It is possible that the results reported in this paper could be improved if we use a soft decision. In other words, by using multiple output nodes (one node represents text, the other node represent non-text) in the neural network, it is possible to classify some regions as not homogenous - containing both text and non-text. With this information it should then be possible to re-partition this type of region into homogenous parts, which should make it possible for the classification pass to correct some of the mistakes from the initial segmentation pass.

Current work is also focusing on improving classification accuracy by using contextual information (information from adjacent regions) as well as features from the grayscale document. Initial work has shown that this can reduce errors by up to 50%.

Acknowledgments

This material is based on work supported by iArchives.

References

- [1] D. Patel, "Page segmentation for document image analysis using a neural network," *Opt. Eng.* 35(7) 1854-1861 (July 1996)
- [2] P. E. Mitchell, "Document page segmentation based on pattern spread analysis," *Opt. Eng.* 39(3) 724-734 (March 2000)
- [3] T. N. Pappas, S. H. Tseng and D. A. Kosiba, "A robust and efficient algorithm for bilevel document block classification," *International Conference on Image Processing 2001. Proceedings*, Vol. 1, pp. 1122-1125
- [4] J. S. Payne, T. J. Stonham, D. Patel, "Document segmentation using texture analysis," *12th Int. Conf. Pat. Rec.*, Jerusalem, Israel, Oct. 1994, 2, 380-382.
- [5] J. Wieser and A. Pinz, "Layout and analysis: finding text, titles, and photos in Digital Images of Newspaper pages,"

Proceedings of the 2nd International Conference on Document Analysis and Recognition, Oct 20-22 1993. pp. 774

[6] G. Maderlechner, A. Schreyer and P. Suda, "Extraction of relevant information from document images using measures of visual attention," *Proceedings of the 15th International Conference on Pattern Recognition*, Vol. 4, pp. 385-388

[7] W. Eglin, and A. Gangneux, "Visual exploration and functional document labeling," *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, 2001. pp. 816-820

[8] V. Eglin, S. Bres, and H. Emptoz, "Printed text featuring using the visual criteria of legibility and complexity," *Proceedings. Fourteenth International Conference on Pattern Recognition*, 1998, Vol. 1. pp. 942-944.

[9] K. Etemad, D. Doermann, and R. Chellappa, "Page segmentation using decision integration and wavelet packets," *Proceedings of the 12th IAPR*, Vol. 2 , pp. 345-349

[10] P. E. Mitchell, and H. Yan, "Newspaper document analysis featuring connected line segmentation," *Proceedings. Sixth International Conference on Document Analysis and Recognition*, 2001, pp. 1181-1185.

[11] P. S. Williams, and M. D. Alder, "Generic texture analysis applied to newspaper segmentation," *IEEE International Conference on Neural Networks*, 1996. Vol. 3, pp. 1664-1669

[12] D. Chetverikov, J. Liang, J. Komuves, R. M. Haralick, "Zone classification using texture features," *Proceedings of the 13th International Conference on Pattern Recognition*, 1996, Vol. 3, pp. 676-680

[13] A. K. Jain, and Y. Zhong, "Page segmentation using texture discrimination masks," *Proceedings of the International Conference on Image Processing*, 1995, Vol. 3, pp. 308-311

[14] D. Ryu, S. Kang, and S. Lee, "Parameter-independent geometric document layout analysis," *Proceedings. 15th International Conference on Pattern Recognition*, 2000, Vol. 4, pp. 397-400

[15] V. Wu, R. Manmatha, and E. M. Riseman, "TextFinder: An automatic System to Detect and Recognize text in images," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 11, November 1999 pp. 1224-1229