Structure in On-line Documents

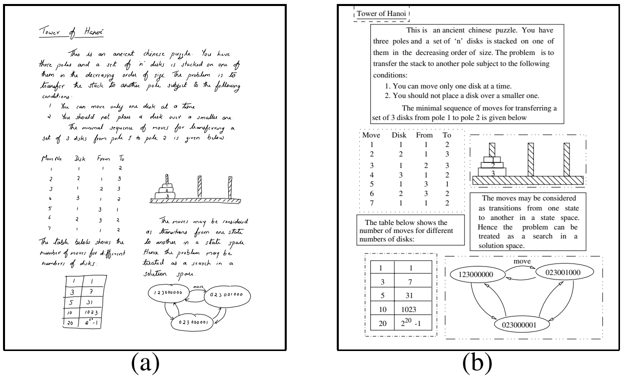
Anil K. Jain and Anoop M. Namboodiri
Department of Comp. Sci. and Engg.
Michigan State University
East Lansing, MI 48824
{jain, anoop}@cse.msu.edu

Jayashree Subrahmonia
IBM T.J. Watson Research Center
Yorktown Heights
NY 10598
jays@watson.ibm.com

Abstract

We present a hierarchical approach for extracting homogeneous regions in on-line documents. The problem of identifying and processing ruled and unruled tables, text and drawings is addressed. The on-line document is first segmented into regions with only text strokes [1] *and regions with both text and non-text strokes. The text region is further classified as unruled table or plain text. Stroke clustering is used to segment the non-text regions. Each non-text segment is then classified as drawing, ruled table or underlined keyword using stroke properties. The individual regions are processed and the results are assembled to identify the structure of the on-line document.*

Keywords: *On-line documents, text strokes, non-text strokes, table identification.*



Tower of Hanoi

This is an ancient chinese puzzle. You have three poles and a set of 'n' disks is stacked on one of them in the decreasing order of size. The problem is to transfer the stack to another pole subject to the following conditions:

1. You can move only one disk at a time.
2. You should not place a disk over a smaller one.

The minimal sequence of moves for transferring a set of 3 disks from pole 1 to pole 2 is given below

Move	Disk	From	To
1	1	1	2
2	2	1	3
3	1	2	3
4	3	1	2
5	1	3	1
6	2	3	2
7	1	1	2

The table below shows the number of moves for different numbers of disks:

1	1
3	7
5	31
10	1023
20	2^{20} -1

The moves may be considered as transitions from one state to another in a state space. Hence the problem can be treated as a search in a solution space.

Figure 1. Segmentation of on-line documents. (a) Input. (b) Desired output.

1. Introduction

The problem of segmenting document pages into homogeneous regions containing unique semantic entities is of prime importance in automatic document understanding systems [9]. Several algorithms exist that recognize printed or handwritten text. However, most of these algorithms assume that the input is a plain text and the text lines and words in the text have been properly identified and segmented by a preprocessor. A typical handwritten document page may contain several regions of interest such as underlined keywords, different types of tables, diagrams, sketches and text (see Figure 1(a)). The main task of a segmentation algorithm is to identify contiguous regions of text, graphics and tables in such a document (see Figure 1(b)) for document retrieval based on semantic entities and full transcription of the handwritten document.

Most of the research in document analysis has focused on off-line (scanned) documents. Examples of this work include page decomposition [6], locating embedded text in color images [15], skew detection [14] and table identification [7, 4, 13, 3, 8]. With the introduction of devices like *IBM ThinkPad TransNote* © [12] and Electronic Whiteboards it is now possible to store and process the entire on-line document. The digitizing tablet captures temporal information of the strokes, which is recorded as a sequence of (x, y) coordinates of the locus of the stylus/pen. This additional temporal information can be used for both text recognition as well as segmentation of documents. The work in on-line document analysis till date is limited to segmentation of text lines [11, 1, 10]. In this work, we extend it to understanding more of the page layout.

We adopt a hierarchical approach to analyze on-line documents (see Figure 2). First, the individual strokes are classified as text or non-text strokes. The non-text strokes are then grouped into homogeneous regions based on their proximity to identify ruled tables and diagram regions. In the third stage, we focus on the (supervised) classification of tables, text, diagrams, and underlined keywords. In addi-

[1] A stroke is defined as the locus of the tip of the pen from pen-down to the next pen-up position.

Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR'01)
0-7695-1263-1/01 $10.00 © 2001 **IEEE**

tion to facilitating text recognition, understanding the structure of an on-line document opens up many applications. The spatial relationship of pictures and text in the document may be used to identify captions and labels. Page layout information allows the use of digitizing tablets as an interface for designing web pages. Search and retrieval of documents can be done based on an underlined keyword or a diagram specified by the user as a query.

2. Text versus Non-text Strokes

We use the *Stroke Length* and *Stroke Curvature* as features for classifying individual strokes as text or non-text. Before extracting the features, the strokes are re-sampled to make consecutive sample points to be equidistant and then low-pass filtered using a Gaussian kernel to reduce errors due to quantization and noise [2]. The number of points, n, in a stroke is used as a measure of its length. To measure the stroke curvature, the angular deviation from linearity is computed at each point of the stroke using its two neighboring points [2].

A two-dimensional feature space representation of the text and non-text strokes of a typical document page containing text, figures and tables is shown in Figure 3. A linear decision boundary easily separates the two classes. Figure 4 gives some examples of stroke classification.

3. Grouping of Non-text Strokes

The non-text strokes are clustered based on inter stroke distance to identify regions of tables or diagrams [5]. The minimum spanning tree (MST) of non-text strokes is constructed with the strokes as nodes and the shortest distance between them as edge weights. The MST is partitioned by removing inconsistent edges. An edge is defined to be inconsistent if its length is more than α times the average length of the edges incident on its two nodes.

The value of α has been empirically determined and is currently set at 3. A further restriction that, the inter-region distance should be greater than 20, is used to avoid splitting of uniform regions. In addition, an upper threshold of 200 was determined empirically for intra-region distances. Figure 5 shows the MST corresponding to the non-text strokes in Figure 4 (a). Three inconsistent edges were identified in this tree resulting in four connected components. The components identified are a ruled table, a state diagram sketch, a drawing illustrating the 'Tower of Hanoi' and an underlined keyword.

4. Table Identification and Processing

Handwritten tables can be classified into two categories: ruled and unruled. The ruled tables have lines separating

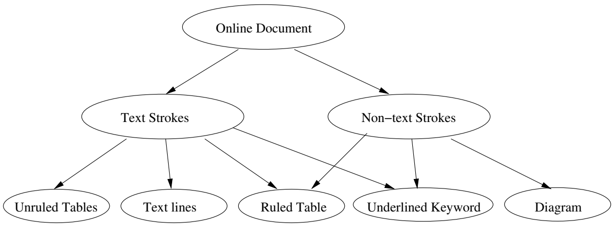


Figure 2. Classification of on-line documents.

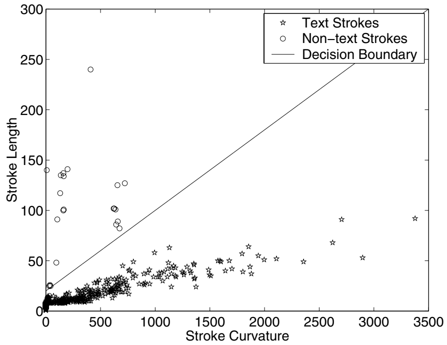
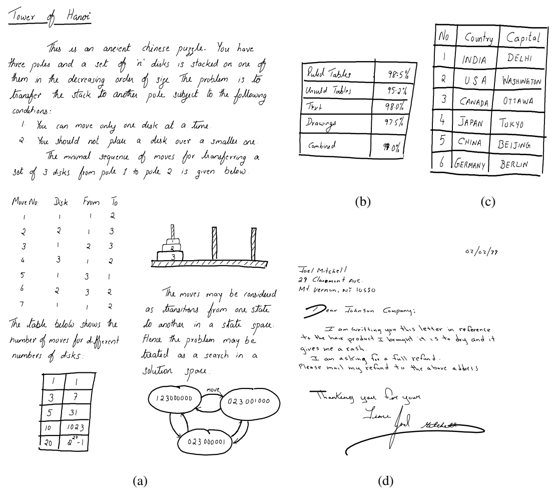


Figure 3. Text vs. non-text strokes.



Tower of Hanoi

This is an ancient chinese puzzle. You have three poles and a set of 'n' disks is stacked on one of them in the decreasing order of size. The problem is to transfer the stack to another pole subject to the following conditions:

1. You can move only one disk at a time.
2. You should not place a disk over a smaller one.

The minimal sequence of moves for transferring a set of 3 disks from pole 1 to pole 2 is given below.

Move No.	Disk	From	To
1	1	1	2
2	2	1	3
3	1	2	3
4	3	1	2
5	1	3	1
6	2	3	2
7	1	1	2

The table below shows the number of moves for different numbers of disks:

1	1
3	7
5	31
10	1023
20	$2^{20}-1$

The moves may be considered as transitions from one state to another in a state space. Hence the problem may be treated as a search in a solution space.

(a)

Ruled Tables	98.5%
Unruled Tables	95.2%
Text	98.0%
Drawings	97.5%
Combined	98.0%

(b)

No	Country	Capital
1	INDIA	DELHI
2	U.S.A	WASHINGTON
3	CANADA	OTTAWA
4	JAPAN	TOKYO
5	CHINA	BEIJING
6	GERMANY	BERLIN

(c)

02/02/99

Joel Mitchell
29 Claremont Ave.
Mt. Vernon, NY 10550

Dear Johnson Company:

I am writting you this letter in reference to the hair product I brought it is to dry and it gives me a rash.

I am asking for a full refund.
Please mail my refund to the above address

Thanking you for your
Time
Joel Mitchell

(d)

Figure 4. Examples of stroke classification. Non-text strokes are shown in bold. In (d) three text strokes are misclassified as non-text strokes because they have large stroke length and/or small curvature.

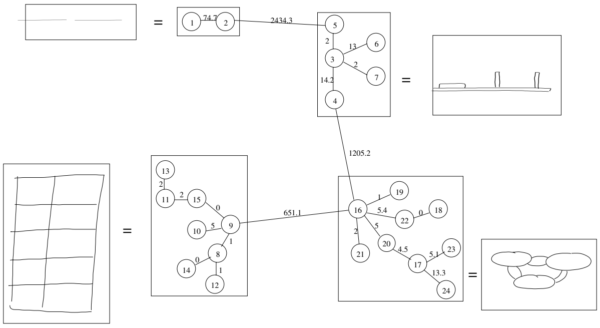


Figure 5. MST (with edge lengths) of the 24 non-text strokes in Figure 4(a).

the rows and columns of the table whereas an unruled table is an arrangement of text into different columns, without any explicit lines separating the columns (see Figure 4(a)). These two types of tables are treated separately as the identification of ruling lines provides an additional cue for ruled tables.

4.1. Ruled Tables

A ruled table is identified based on the presence of horizontal and vertical lines which can be detected using the Hough transform. The ruled tables in the 2-dimensional Hough space (r, θ) have significant peaks around angles of $\theta = 0^o$ and $\theta = 90^o$. A region is classified as a table or a diagram based on the lengths of the lines (size of the cluster in the transform space) and their orientations. A further restriction that tables contain at least 5 lines (four borders and at least one partition) helps to distinguish ruled tables from underlined key words. Figure 6 shows a typical ruled table and its (r, θ) representation. A total of 9 lines are detected which have been redrawn as straight lines in Figure 6 (c). An underlined keyword is detected when a region has a single horizontal line in the Hough transform space and has some text above it. Note that this method can detect broken underlines as can be seen in figure 8.

4.2. Unruled Tables

The text strokes contain both plain text and unruled tables. To identify the unruled tables, first individual text lines are identified and adjacent lines are incrementally grouped. The line detection algorithm used here is similar to the approach proposed in [11].

The inter-line distance, d_l, is estimated from the autocorrelation of the y-axis projection of the text. Line separations are decided by valleys in the histogram. To avoid local minima, we choose only those points as valleys which have the smallest magnitude within a window of width d_l. Figure 7 shows the text lines detected from the document in figure 1 (a). Identification of unruled tables is primarily based on the x-axis projection of the region to be classified. The table regions tend to have prominent peaks and valleys in this projection, while text regions have a more uniform projection (see Figures 9 (c) and (d)). Text also varies in appearance due to differences in inter-word and inter-line spacings. We assume that the inter-word distances are more than double the inter-stroke distances within a word.

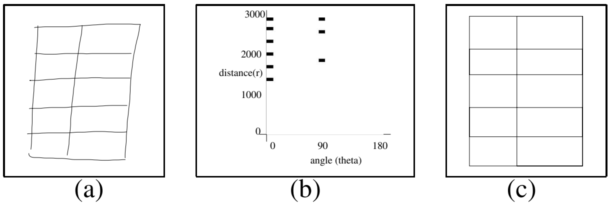


Figure 6. Finding Ruling lines. (a) A typical table. (b) Peaks in the 2-Dimensional Hough transform. (c) Table cleaned up by identifying the ruling lines.

4.3. Table Processing

In the case of a ruled table, the horizontal and vertical lines provide the cell boundaries. The cell boundaries in unruled tables are determined by dividing each line at vertical cell boundaries given by the valleys in the projection histogram. The cell contents are easily identified by collecting the text strokes within each cell boundary. The text in each cell is supplied to a text recognizer [2] and the results are written into an ASCII table for exporting it to *Microsoft Excel* ©. Figure 10 shows the result of exporting the ruled table in Figure 4 (b) into *Excel*. Note that the numerical data in individual cells were correctly identified although the text recognizer incorrectly recognized the word 'Unruled Tables' in the second row as 'Annual Tables'.

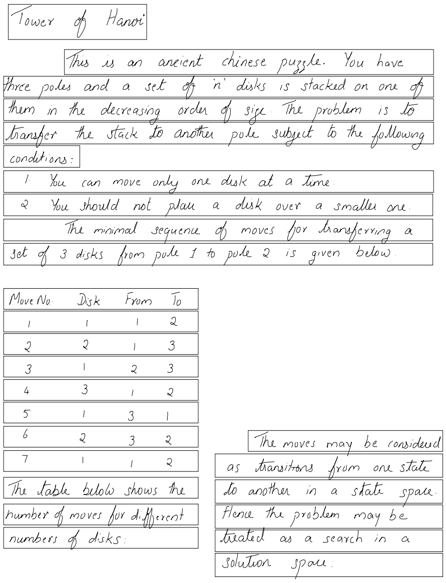
The final result of processing the on-line document in figure 1 (a) is shown in figure 8.

5. Experimental Results

The experimental data was collected from 123 different people [3] without any restriction on the style or content of data. Text vs. non-text classifier was trained on a set of

[2] *IBM Ink Manager* © software shipped with the *CrossPad*©.

[3] Most of this data was collected by the Pen Computing group, IBM T.J. Watson Research Center.



Tower of Hanoi

This is an ancient chinese puzzle. You have three poles and a set of 'n' disks is stacked on one of them in the decreasing order of size The problem is to transfer the stack to another pole subject to the following conditions:

1 You can move only one disk at a time

2 You should not place a disk over a smaller one.

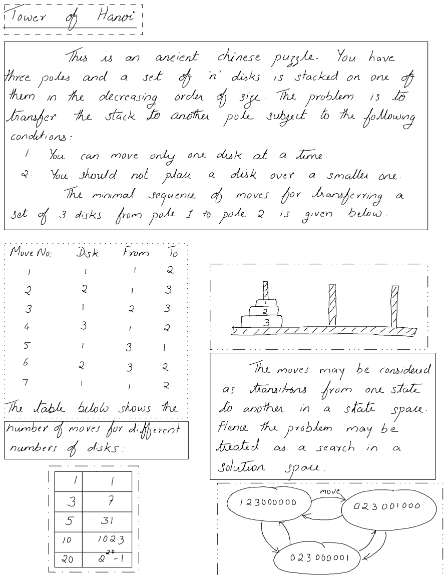
The minimal sequence of moves for transferring a set of 3 disks from pole 1 to pole 2 is given below.

Move No.	Disk	From	To
1	1	1	2
2	2	1	3
3	1	2	3
4	3	1	2
5	1	3	1
6	2	3	2
7	1	1	2

The table below shows the number of moves for different numbers of disks.

The moves may be considered as transitions from one state to another in a state space. Hence the problem may be treated as a search in a solution space.

Figure 7. Text lines detected in Fig. 4 (a).



Tower of Hanoi

This is an ancient chinese puzzle. You have three poles and a set of 'n' disks is stacked on one of them in the decreasing order of size The problem is to transfer the stack to another pole subject to the following conditions:

1 You can move only one disk at a time

2 You should not place a disk over a smaller one.

The minimal sequence of moves for transferring a set of 3 disks from pole 1 to pole 2 is given below.

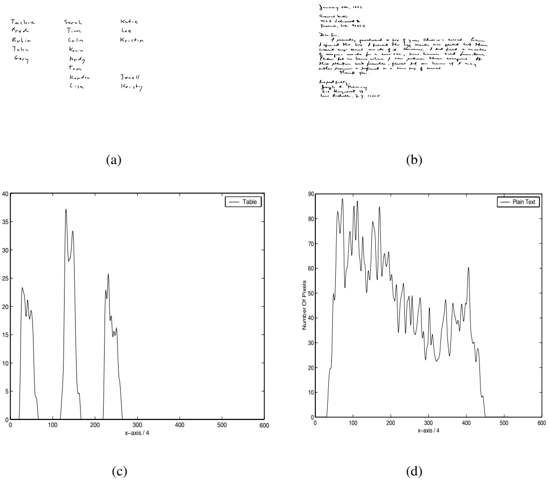
Move No.	Disk	From	To
1	1	1	2
2	2	1	3
3	1	2	3
4	3	1	2
5	1	3	1
6	2	3	2
7	1	1	2

The table below shows the number of moves for different numbers of disks.

The moves may be considered as transitions from one state to another in a state space. Hence the problem may be treated as a search in a solution space.

1	1
3	7
5	31
10	1023
20	$2^{20}-1$

Figure 8. Segmented document of Fig. 1 (a).



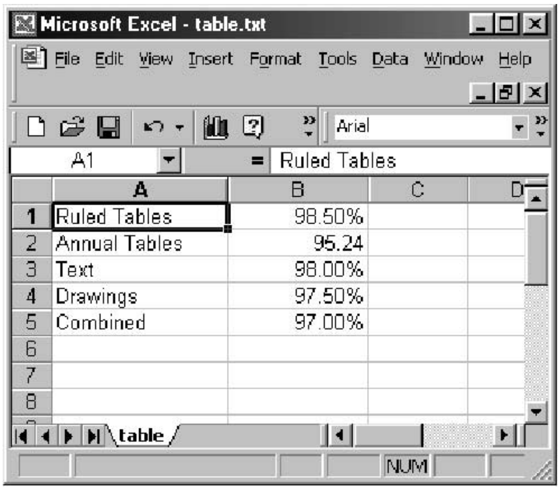
(a)

(b)

(c)

(d)

Figure 9. Projections of table (a) and text (b).



Microsoft Excel - table.txt
File Edit View Insert Format Tools Data Window Help
Arial
A1
=
NUM

Ruled Tables

	A	B	C	D
1	Ruled Tables	98.50%		
2	Annual Tables	95.24		
3	Text	98.00%		
4	Drawings	97.50%		
5	Combined	97.00%		
6				
7				
8				

Figure 10. A table imported into Excel.

$1,304$ strokes ($1,202$ text strokes and 102 non-text strokes). A classification accuracy of 99.1% was achieved on an independent test set of $36,812$ strokes ($35,882$ text strokes and 930 non-text strokes). Most of the misclassifications were due to very short strokes in diagrams which were incorrectly identified as text strokes. About 99.9% of the text strokes were correctly classified (Figure 4).

The identification of tables is, in general, more difficult than identification of text due to the large variability in structure of unruled tables. A classification accuracy of 85.0% was achieved on a data set containing 105 unruled tables and 35 text regions. While most of the text samples were correctly identified, the misclassifications in the case of tables were mainly due to skew of columns and inconsistent column separation.

6. Conclusions

A hierarchical approach for extracting structural information from an on-line handwritten document is presented here. Stroke characteristics have been used to extract drawings and ruled tables. Structural information and temporal sequence of strokes were used to identify unruled tables. The system also allows user to modify the results at any stage of processing. The performance of the algorithm on a set of 150 pages of text, tables and drawings is reported. Currently we are working on extending the approach to identify different languages to facilitate text recognition and to clean up diagrams and sketches by identifying different drawing primitives.

References

[1] E. Bruzzone and M. Coffetti. An algorithm for extracting cursive text lines. In *Proceedings of the* 5^{th} *International Conference on Document Analysis and Recognition (ICDAR'99)*, pages 749–752, Bangalore, India, September 1999.

[2] S. D. Connell and A. Jain. Learning prototypes for on-line handwritten digits. In *Proceedings of the* 14^{th} *International Conference on Pattern Recognition*, pages 182–184, Brisbane, Australia, August 1998.

[3] Y. Hirayama. A method for table structure analysis using DP matching. In *Proceedings of the* 3^{rd} *International Conference on Document Analysis and Recognition (ICDAR'95)*, pages 583–586, Montreal, Canada, August 1995.

[4] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong. Table detection across multiple media. In *Proceedings of the Workshop on Document Layout Interpretation and its Applications*, Bangalore, India, September 1999.

[5] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[6] A. K. Jain and B. Yu. Document representation and its application to page decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):294–308, March 1998.

[7] T. G. Keininger and A. Dengel. The T-RECS approach for table structure recognition and table border determination. In *Proceedings of the Workshop on Document Layout Interpretation and its Applications*, Bangalore, India, September 1999.

[8] W. Kornfeld and J. Wattecamps. Automatically locating, extracting and analyzing tabular data. In *Proceedings of* $21st$ *Annual International ACM SIGIR Conference*, pages 347–349, Melbourne, Australia, August 1998.

[9] G. Nagy. Twenty years of document image analysis in PAMI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):38–62, January 2000.

[10] Y. Pu and Z. Shi. A natural learning algorithm based on Hough transform for text lines extraction in handwritten documents. In *Proceedings of the* 6^{th} *International Workshop on Frontiers in Handwriting Recognition*, pages 637–646, Taejon, Korea, August 1998.

[11] E. H. Ratzlaff. Inter-line distance estimation and text line extraction for unconstrained online handwriting. In *Proceedings of the* 7^{th} *International Workshop on Frontiers in Handwriting Recognition*, Nijmegen, Netherlands, September 2000.

[12] J. Subrahmonia. Pen computing: Challenges and applications. In *Proceedings of the* 15^{th} *International Conference on Pattern Recognition*, pages 60–66, Barcelona, Spain, September 2000.

[13] T. A. Tokuyasu and P. A. Chou. An iterative decoding approach to document image analysis. In *Proceedings of the Workshop on Document Layout Interpretation and its Applications*, Bangalore, India, September 1999.

[14] B. Yu and A. K. Jain. A robust and fast skew detection algorithm for generic documents. *Pattern Recognition*, 29(10):1599–1630, October 1996.

[15] Y. Zhong, K. Karu, and A. K. Jain. Locating text in complex color images. *Pattern Recognition*, 28(10):1523–1535, 1995.