Table detection and ROI cell extraction using line detection and structure based

Bhaskar Menugula

Lead Engineer, HCL

Abstract The algorithm based on the inference that the tables have distinct columns, which implies that gaps between the fields are substantially larger than the gaps between the words in text lines.

Table detection steps:

The table detection algorithm depends primarily on:

- 1. Preprocessing
- 2. Formation of word blobs in text lines and
- 3. Finding out the set of consecutive text lines which would form a table
- 4. Vertical threshold between table lines.
- 5. Forming table cells
- 6. Processing table region of index

1. Preprocessing

A typical document image analysis contains image filtering techniques such as de-noising, skew-correction is applied on the document image. The input image is converted into gray scale and then it is converted into binary image.

2. Formation of word blobs in text lines.

Formation of text line is done by joining the word blobs in the line. A text-line would be

converted to a single rectangular block while a row of a table would consist of multiple

smaller blocks. Such a word grouping depends on measuring the average word-gaps

and finding out distance between consecutive connected components in a single text-line.

Connected components in a single text-line is obtained by using the formula:

$$f(Ca, Cb) = \begin{cases} 1 & \text{if } (T(Ca) \leq B(Cb) \text{ and } B(Ca) \geq T(Cb)) \\ 0, & \text{otherwise} \end{cases}$$

Where, Ca, Cb: two blobs.

T(Ca), B(Ca): Top and bottom of a rectangular blob.

Word blobs are formed by finding out by Histogram of inter-word gap between consecutive connected components in every text-line.

Distance between consecutive connected components Ca and Cb are obtained by using the formula:

$$D(Ca, Cb) = L(Cb) - R(Ca)$$

Where.

L(Ca), R(Ca): Left and Right of a rectangular blob.

The histogram of gap between two consecutive blobs mainly contains two humps, if the document image is non uniform font it may have more than two humps. The second peak of the histogram is word gap. Finally, morphological closing operation with a line-structuring element using second peak value forms the text-lines.

3. Finding out the set of consecutive text lines which would form a table.

This stage groups the candidate table lines from text-lines, which will be used for the identification of table 'region of index'. For a table row it is identified that word-gap between elements (column fields) are larger than word gap in the text-lines.

Morphological closing with second peak joins all the word-gap among text-lines except table lines, thus they are considered as candidate table lines, so based on this observation it is possible to separate table rows from text-lines.

Figure shows masking to text-lines from table rows

4. Vertical threshold between table lines.

It may be noted that the primary selection is not strong enough to detect all potential candidate lines; we may miss some of the rows of the tables. For this, gap between successive text-lines are compared and a maximum threshold value is formulated. If g(a) is the vertical gap between lines:

B(TELa): bottom of table-line

the upper-bound of the prominent peak of the histogram of g(a) gives the vertical threshold. There could be a possibility that consecutive text-lines are too far apart. This larger gap may affect the threshold calculation, so very high peaks in the vertical gap are eliminated.

Table-lines whose vertical gap is less than or equal to the threshold gap are grouped. Every group of text-lines form a table.

5. Forming table cells.

TELa

For forming cells, midpoint between successive rows as well as columns will be calculated from the text-line rectangles.

If Vr(a) is the vertical midpoint between every table rows and Hr(a) is the horizontal midpoint between each column.

$$gap = \{L(TELa1) - R(TELa)\} / 2$$

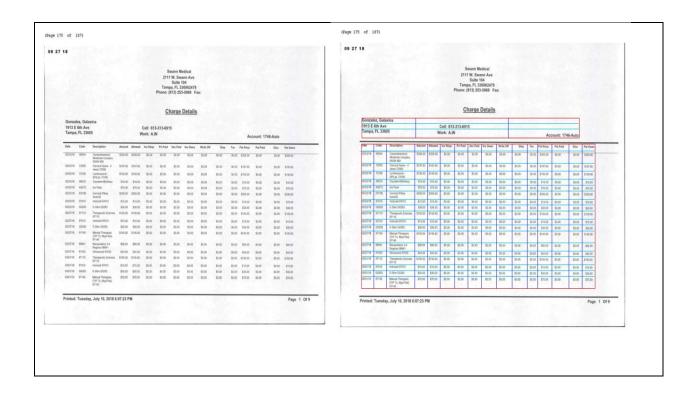
$$Vr(a) = R(TELa) + gap$$

$$Hgap = \{T(TELa1) - B(TELa)\} / 2$$

$$Hr(a) = B (TELa) + gap$$

Results:

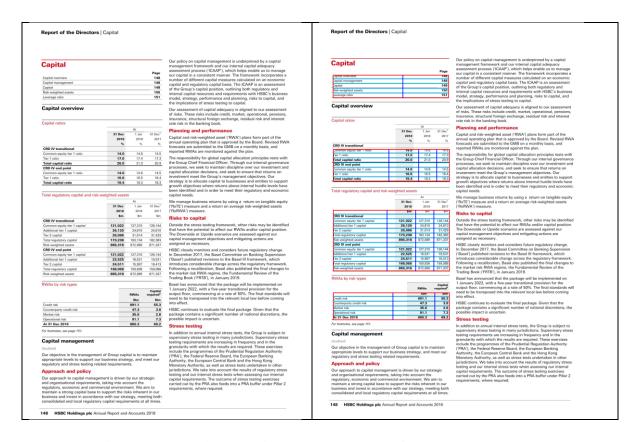
1. Table detection results without lines in image.



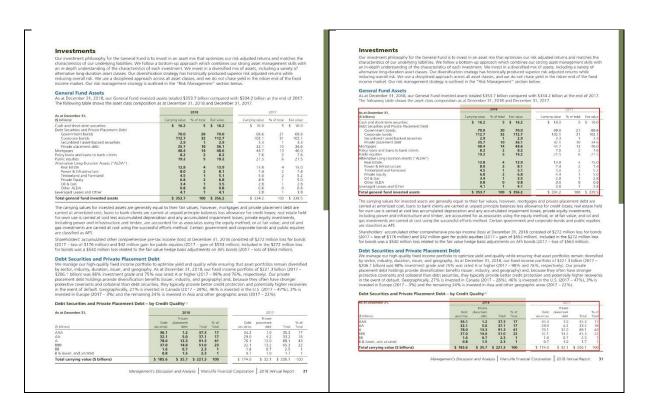
0 01 18		S. Marie			10 01	1 18					
Babinsk		Downgoing	Downgoing			Babinski:	Downgoing	Downgo	ing		
Reverse	e Radial:	Negative	Negative			Reverse Radial: Clonus:	Negative negative	Negative			
Clonus		negative	negative				negative	Inguitt			
Lympha	atic:	***	-		10.58	Lymphatic:	Right	Left			
Neck:		Right normal	Left		100000	Neck:	normal	normal			
Axillae		normal	normal			Axillae:	normal	normal			
Groin:		normal	normal			Groin:	normal	normal			
Cardiov	vascular				100.00	Cardiovascular					
		Right	Left				Right	Left			
Upper e Swelling	extremities:	Negative	Negative		10000	Upper extremities:					
Varicos		Negative	Negative Negative		100000	Swelling Varicosities	Negative Negative	Negative Negative			
						valicosities	(vegative	regetive			
Lower E Swelling	Extremities:	Negative	Negative		1 1 2 2 2	Lower Extremities:					
Varicos		Negative	Negative			Swelling Varicosities	Negative	Negative			
12.000						varicosities	Negative	Negative			
Palpatio	ion	Right	Left		1000	Palpation					
	extremities:						Right	Left			
Pulse		0	0			Upper extremities: Pulse	0	0			
	rature: Normal					Temperature: Normal	0	- u			
Lower E Pulse	Extremities:	0	0			Lower Extremities:					
	rature: Normal				1,000	Pulse	0	0			
						Temperature: Normal					
	ramination of Galt and Station: sine: tenderness					Examination of Gait ar	ad Stations				
Ribs: no					1000	Spine: tenderness	iu station.				
Pelvis: n					100.00	Ribs: normal					
RUE ter					120,030	Pelvis: normal RUE: tenderness					
	nderness					LUE: normal					
LLE nor	emal				10.000	RLE: tenderness					
Diagno	ostic Studies Order/I	Interpretations				LLE: normal					
Status	Ordered	Order	Completed Interpr	retation Report		Diagnostic Studies Oro	for Anterpretation				
result n	received 02/15/2018	MRI CERVICAL SPIN	NE W/O DYE	ADG on 12/06/17; CD and report reviewed. I	100	Status Ordere	d Order	Con	pleted Interpretation	Report	
				and report reviewed. I agree with the report.			2018 MRI CERVICAL SPIR			ADG on 12/06/17; CD	
				Broad Based disc	131324					and report reviewed. I	
				herniation C4-5 that abuts anterior cord	150200					agree with the report. Broad Based disc	
repult o	received 02/15/2018	MRI LUMBAR SPINE	E W/O DYE	ADG on 01/10/18: CD						herniation C4-5 that	
				and report reviewed. I	144					abuts anterior cord	
						result received 02/15/.	2018 MRI LUMBAR SPIN	IE W/O DYE		ADG on 01/10/18; CD	
Moya, A	Anyelis 0000003530	073 08/02/1973 02/15	5/2018 09:55 AM Page: 23/26		100000					and report reviewed. I	
			- Control of the Cont						Page: 23/26		

2. Table detection results with lines:





3. Table detection results having multi table structured ROI



References

S. Mandal, S. P. Chowdhury, A. K. Das, and B. Chanda. A simple and effective table detection system from document images. IJDAR, 8:172-182, 2006.