



# Segmentation of Text-Lines and Words from JPEG Compressed Printed Text Documents Using DCT Coefficients

Bulla Rajesh\*\*, Mohammed Javed\*\*, P. Nagabhushan\*\*, Watanabe Osamu\*  
 \*\*Department of IT, Indian Institute of Information Technology, Allahabad, U.P, India  
 \*Department of ECS, Takushoku University, Japan



## INTRODUCTION AND METHODS

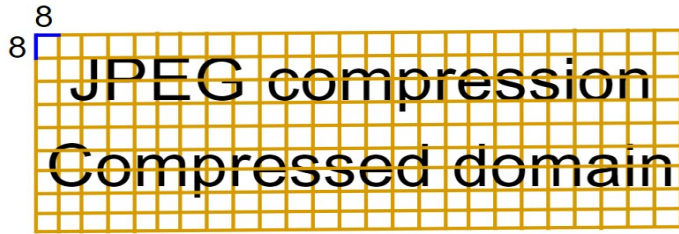


Figure 1. Illustration of how an 8x8 DCT block encodes printed text contents in pixel domain during JPEG compression

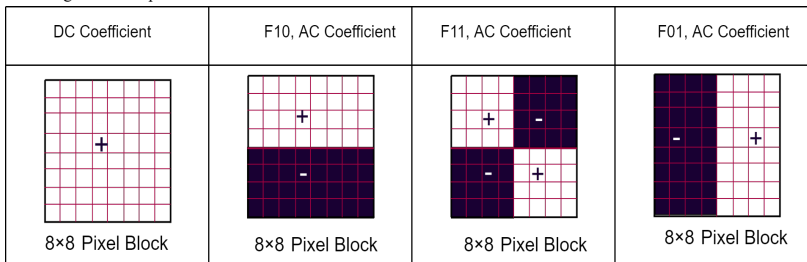


Figure 2. Pictorial visualization of F00 (DC) F10, F11 and F01 coefficients from a typical 8x8 DCT Block

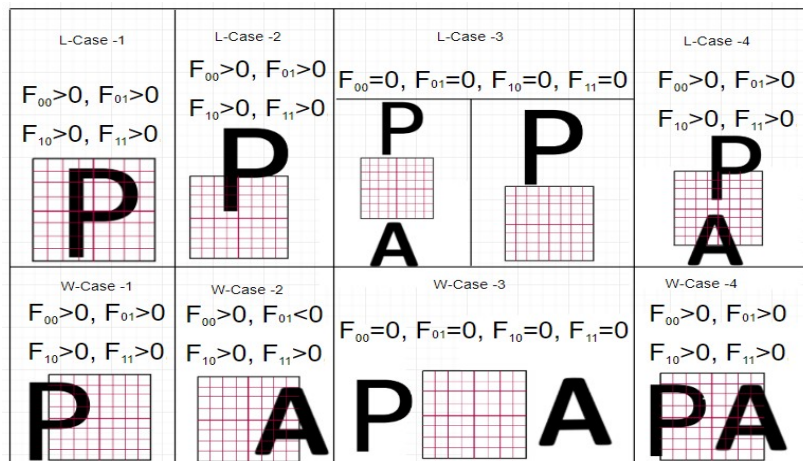


Figure 3. Depicting the various possible blocks shown in Fig. 1.

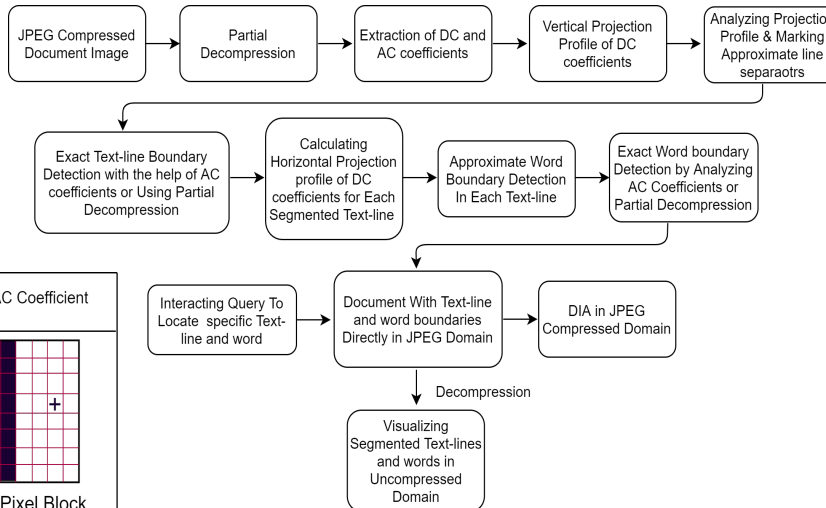


Figure 4. Proposed model for Text-line and Word segmentation in JPEG compressed documents.

## RESULTS AND ANALYSIS

Table 1. TEXT-LINE EXTRACTION RESULTS BY DECOMPRESSING THE SELECTED DCT BLOCKS

Image Type (DPI)	LINE		WORD		F-MEASURE	
	P1 (%)	R1 (%)	P1 (%)	R1 (%)	LINE (%)	WORD (%)
96	90	100	80	100	94.73	88.88
200	93	100	96	100	96.37	97.95
300	100	100	100	100	100	100

Table 2. Text-line and word extraction results through analyzing the F10, F01 and F11 coefficients.

Image Type (DPI)	LINE		WORD		F-MEASURE	
	P1 (%)	R1 (%)	P1 (%)	R1 (%)	LINE (%)	WORD (%)
96	90	92	80	90	90.98	84.75
200	93	100	96	100	96.37	97.95
300	100	100	100	100	100	100

Table 3. EXECUTION TIME ANALYSIS OF THE PROPOSED COMPRESSED DOMAIN ALGORITHMS WITH RESPECT TO PIXEL DOMAIN PROCESSING.

Approaches	Text-line & Word Seg Time (sec)	Speed Gain (%)
Pixel Domain (Decompression + processing)	0.2027	-
Partial Decompression	0.0144	92.89
Using AC coefficients	0.00132	99.34

## DISCUSSION

- o In a typical DCT 8x8 block, each coefficient value is the linear combination of the pixel values and the frequency level increase from top to bottom and left to right.
- o To solve all the cases in Fig. 3 except case 4, for locating the line boundaries in compressed domain, we need F10, F01 and F11 coefficients.
- o If  $F_{10} > 0$ , which means the text belongs to upper line and we mark the below block as the line boundary. Similarly for Word also. If  $F_{01} > 0$ , means word belongs to previous word.
- o Similarly, if  $F_{10} < 0$  then we mark the above block as the line boundary.
- o Likewise, we also observe the absolute value of F11 coefficient. If the  $|F_{11}| > 0$  then we can't divide that block but marks that block itself as boundary.
- o The another way of solution for F11 cases is to partially decompress the block at that position to locate the exact line separator in the pixel domain.
- o The flow diagram of the approach is shown in Fig. 4.
- o We tested our approach on different printed documents, results of both the methods have been shown in Table 1 and Table 2.
- o The observation is that the process in compressed domain has gain the improvement in the speed as shown in Table 3.

## CONCLUSIONS

- We proposed the possibility of working directly JPEG compressed domain for JPEG compressed document images.
- Two approaches have been proposed in this paper, first approach is using DC projection profile along with F10 and F11 AC coefficients.
- The second approach is using partial decompression at the expected line separators positions where the detection of line boundary is impossible in compressed domain.

## FUTURE DIRECTIONS

We want to continue our work in compressed domain to explore the techniques for building models to find solution to the problems like segmentation.

We want to incorporate the better representations and methods to further increase the performance that we achieved in this paper.

## Acknowledgements

This research was carried out in the Indian Institute of Information Technology, Allahabad and supported by the Ministry of Human Resource and Development, Government of India.

## Contact Information

Bulla Rajesh, Mohammed Javed  
 Department of IT, IIT Allahabad  
 Email: [rajesh091106@gmail.com](mailto:rajesh091106@gmail.com), [javed@iitaa.ac.in](mailto:javed@iitaa.ac.in)

## References

1. J. Mukhopadhyay, Image and video processing in the compressed domain. Chapman and Hall/CRC, 2011.
2. M. Javed, P. Nagabhushan, and B. B. Chaudhuri, "A review on document image analysis techniques directly in the compressed domain," Artificial Intelligence Review, vol. 50, no. 4, pp. 539-568, 2018.
3. C. Florea, M. Gordan, B. Orza, and A. Vlaicu, "Compressed domain computationally efficient processing scheme for jpeg image filtering," in Advanced Engineering Forum, vol. 8, pp. 480-489, Trans Tech Publ, 2013.
4. R. L. de Queiroz and R. Eschbach, "Fast segmentation of the jpeg compressed documents," Journal of Electronic Imaging, vol. 7, no. 2, pp. 367-378, 1998.