Artificial Intelligence and its Scope for Natural Language Processing in Compressed Document Images

Artificial intelligence and NLP provide a wide scope for accomplishing Intelligent document image processing and comprehension. Understanding a document image involves the tasks such as document cleaning [3][6], layout analysis, text and non-text segmentation, extraction of paragraphs, text lines, words and characters. In the context of document images, AI and NLP support developing different intelligent applications like document/text classification, sentiment analysis, text similarity, text summarization. In the current digital era, huge volumes of document images are being generated in the compressed form for the efficiency of storage and transmission. In such a scenario accomplishing direct and intelligent processing of compressed documents without decompression is a challenging research issue. The research papers [2] and [1] accomplish the task of text line and word segmentation directly in JPEG compressed Printed and Handwritten text document images.

Segmentation in handwritten document images is very challenging, due to the presence of uneven spacing, variable font sizes, overlapping and touching components, and it becomes much more challenging if it is to be done directly in the compressed images. Intelligent algorithms have been developed to accomplish text line segmentation in compressed document images such as tunneling algorithm in case of Run Length Encoding and space penetration algorithm in case of JPEG. Subsequently intelligent strategies for word and character segmentation have also been proposed. On the other hand, printed documents like newspapers, magazines, research articles have their own segmentation challenges like different layout, presence of text and non-text components, variable and complex background, different types of font size and font style. Intelligent algorithms have been developed to tackle these challenges both in compressed domain and conventional pixel domain.

Searching for a keyword in a document image or simply word-spotting is a very critical problem in case of document images. Addressing these problems in case of compressed document image will be a really difficult task, specifically in document images that exist with different resolutions, layout, font size, font style, spacing etc. Intelligence driven algorithms like OCR based partial decompression method and completely OCR-less and decompression-less methods have been investigated to address these issues in compressed document images.

In all, the research ideas proposed here are really groundbreaking, since they open up new vista in the field of AI and NLP driven image analysis, addressing many interdisciplinary research issues related to Storage, Transmission and efficient Bandwidth management.

Recent Research Papers and Ph.D. Theses:

1. Bulla Rajesh, Mohammed Javed, P. Nagabhushan, “Automatic Tracing and Extraction of Text-Line and Word Segments Directly in JPEG Compressed Document Images”, IET Image Processing, April 02, 2020.
2. Bulla Rajesh, Mohammed Javed, P. Nagabhushan,"Segmentation of Text-Lines and Words from JPEG Compressed Printed Text Documents Using DCT Coefficients”, Published in IEEE Data Compression Conference (DCC2020), Page 389, March 24-27, 2020
3. Mohammed Javed, Tryambak Bhattacharjee, P. Nagabhushan, “ Enhancement of Variably Illuminated Document Images Through Noise Induced Stochastic Resonance”, Published in IET Image Processing, Volume 13 (13), Pages 2562-2571, November 2019.
4. Mohammed Javed, “On the Possibility of Processing Document Images in Compressed Domain”, Ph.D. thesis, University of Mysore, 2016
5. R. Amarnath, “ Segmentation in compressed Document Images”, Ph.D. thesis, University of Mysore, 2019
6. Nirmala S, “Enhancing the readability of text in document images”, Ph.D. Thesis, University of Mysore, 2010