International Journal of Advanced Science and Technology
Vol.124 (2019), pp.129-138
http://dx.doi.org/10.33832/ijast.2019.124.12

EFFICIENT EMBEDDED IMAGES IN PORTABLE DOCUMENT FORMAT (PDF)

Yair Wiseman

Computer Science Department
Bar-Ilan University
Ramat-Gan 52900, Israel
wiseman@cs.biu.ac.il

***Abstract*— The use of facsimile machines is gradually declined, because the email slowly but surely takes its function; however, the use of the facsimile format is still supported by the well-known format - Portable Document Format (PDF). In spite of that, PDF commonly prefers to use JPEG over using the facsimile format. This can be suitable for many images; however, when the image is a document, preferring JPEG over the facsimile format will not yield high-quality outcomes. In this paper we explain why the facsimile format can do better when an embedded image of document is enclosed in a PDF file.**

Keywords*— *PDF, Facsimile, JPEG

1. INTRODUCTION

Facsimile (AKA fax) is a technique to transmit documents over phone lines. Facsimile have been used as early as the 19th century; however, the use of Huffman codes to compress the transmitted data could be taken in the 19th century because Huffman codes was published just at 1952 [1].

Fax machines employ modified Huffman coding - CCITT Fax 3 & Fax 4 – protocol for transmission of fax documents over telephone lines [2]. Monochrome TIFF image compression format also make use of this technique [3]. This technique combines variable length codes of Huffman algorithm with the coding of repetitive data in run length encoding. Employing Huffman coding is very advantageous because Huffman coding has the ability to recover from a transmission errors [4,5] that sometimes occur in Fax machines.

The creation of the Multimedia Internet Mail Extension (MIME) format [6] for email attachments triggered off the acceptance of emails as a favorable and promising alternative for faxes. The MIME format facilitates transferring non-text files via the network by standard emails. This ability of MIME prevailed upon the fax main capability to transfer binary data [7].

Few years later Adobe Inc. developed a new format – Portable Document Format (PDF) [8]. PDF can contain Raster images (AKA Image XObjects) which can embed binary data within the PDF file [9]. Several types of binary data are allowed in PDF files. One of them is called "CCITTFaxDecode" which is actually a binary data compressed image exactly as fax compresses files [10].

Moreover, Adobe has been offering the PDF reader for free. So, as a result, emails have the capability to excellently transfer documents in the same format as

Received: November 26, 2018
Reviewed: January 28, 2019
Accepted: February 4, 2019

ISSN: 2207-6360 IJAST

fax transfers, nevertheless much faster. Therefore, the main function of the faxes has become outdated [11].

In spite of this, the programs that generate PDF files mostly does not use the CCITTFaxDecode type even when the PDF is a document and rather use the DCTDecode type that represents JPEG file, even though JPEG is not so suitable for document compression [12].

2. DOCUMENT COMPRESSION

JPEG is an effective compression technique for images with no sharp differences. When there are sharp differences within the image, JPEG will not compress these images well [13,14,15]. JPEG tends to smear the borderlines within the image instead of drawing them distinctly.

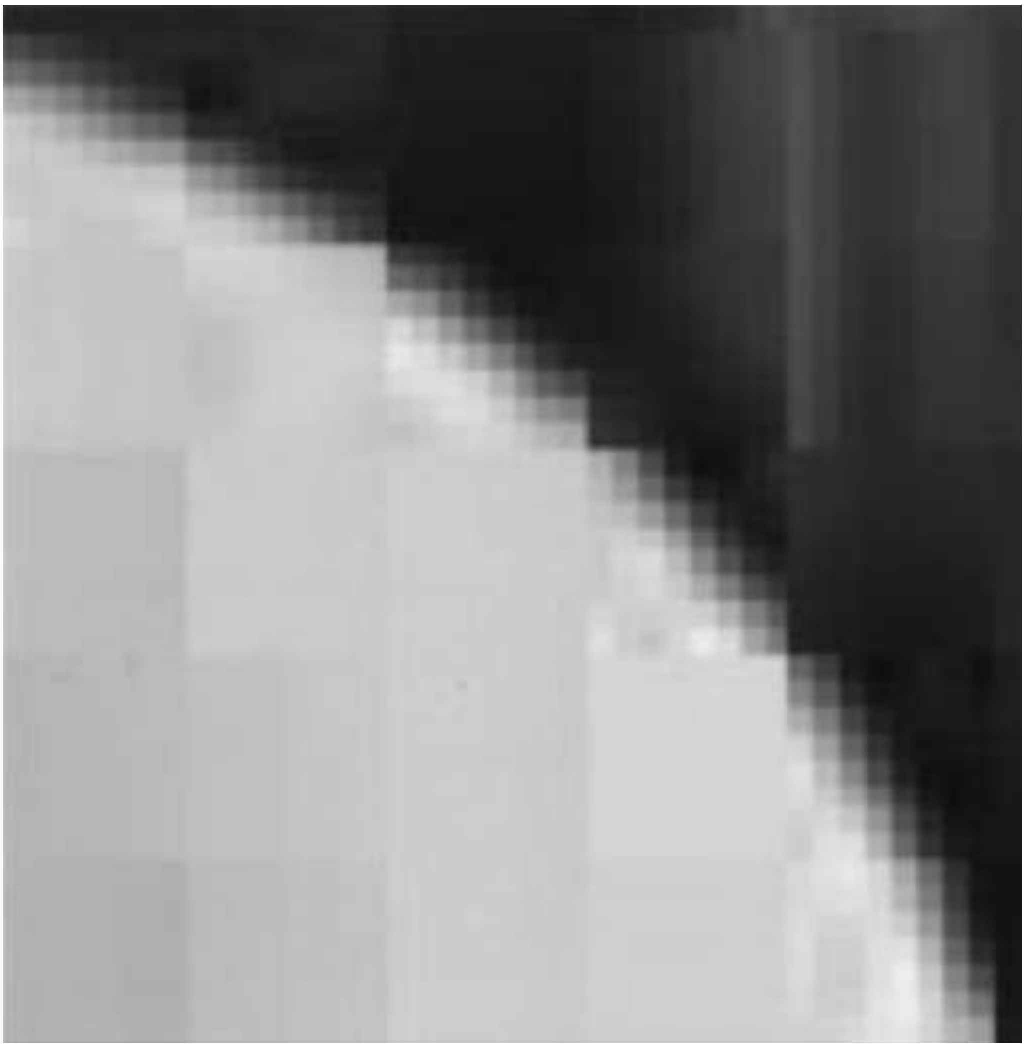
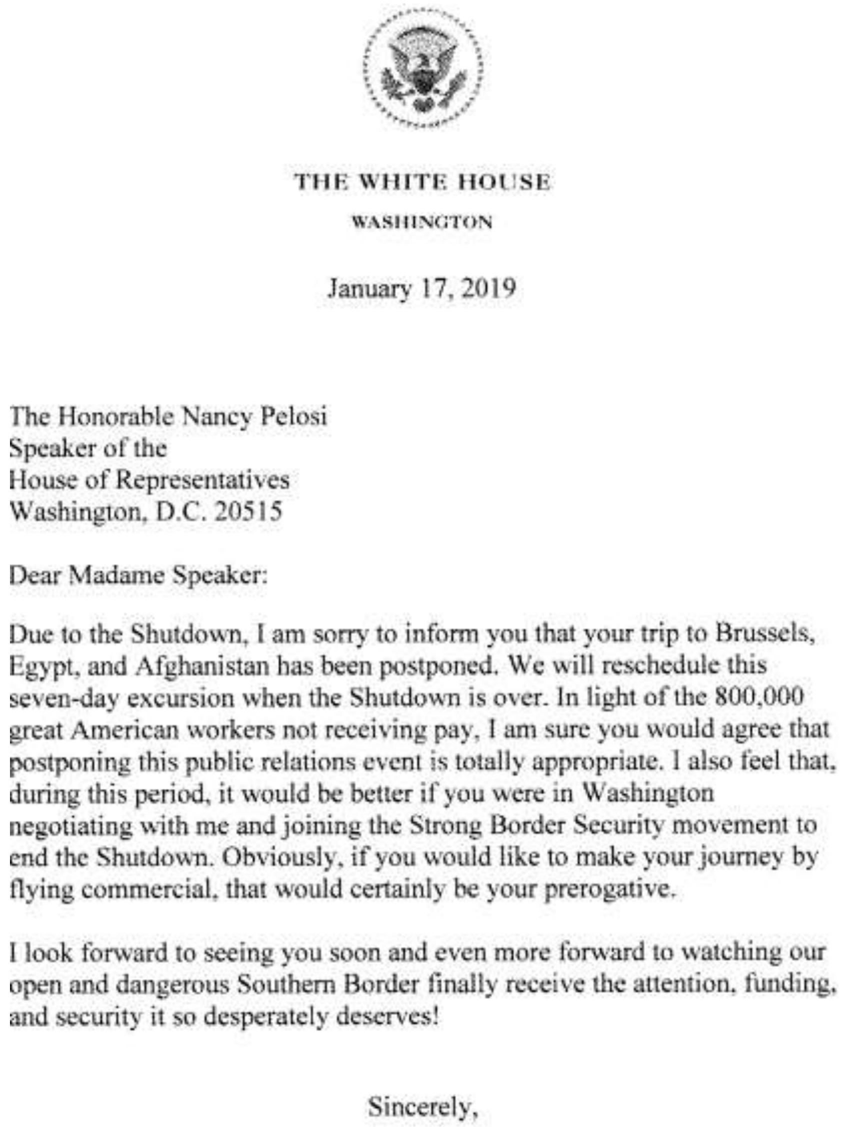


Fig.1 JPEG image with sharp differences

Figure 1 shows a JPEG image with a sharp difference. The image has been enlarged with the intention of showing clearly each pixel. It can be easily seen that when a sharp change occurs, JPEG smears the pixels and pixels with intermediate levels of intensity are added [16,17].

As an example, the same document was sent by the CCITTFaxDecode type and by the DCTDecode type.



THE WHITE HOUSE
WASHINGTON

January 17, 2019

The Honorable Nancy Pelosi
Speaker of the
House of Representatives
Washington, D.C. 20515

Dear Madame Speaker:

Due to the Shutdown, I am sorry to inform you that your trip to Brussels, Egypt, and Afghanistan has been postponed. We will reschedule this seven-day excursion when the Shutdown is over. In light of the 800,000 great American workers not receiving pay, I am sure you would agree that postponing this public relations event is totally appropriate. I also feel that, during this period, it would be better if you were in Washington negotiating with me and joining the Strong Border Security movement to end the Shutdown. Obviously, if you would like to make your journey by flying commercial, that would certainly be your prerogative.

I look forward to seeing you soon and even more forward to watching our open and dangerous Southern Border finally receive the attention, funding, and security it so desperately deserves!

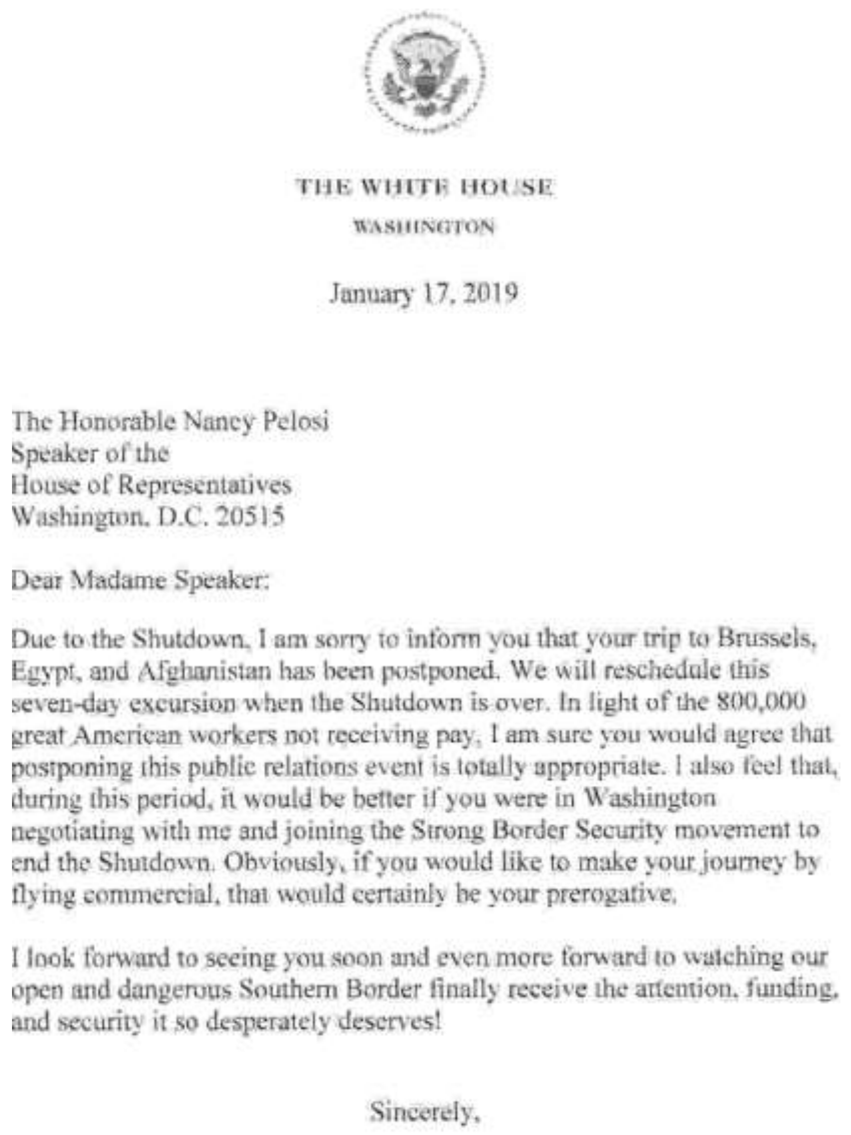
Sincerely,

Fig 2 Document sent by CCITTFaxDecode

Figure 2 contains the document sent by CCITTFaxDecode and Figure 3 contains the document sent by DCTDecode. It can be clearly seen that the CCITTFaxDecode loses less information and the document is more readable than the document compressed by DCTDecode.

While JPEG is better in pictures without sharp differences, faxes are dedicated to compress documents and documents by definition have many sharp differences. Therefore, faxes can do a better job than JPEG when there is a need to transmit documents.

Faxes can also obtain better compression efficiency, because the algorithm is dedicated to compress document with sharp changes. Accordingly, the image in Figure 2 was compressed into 57KB, whereas the image in figure 3 was compressed into 757KB. That is to say, the fax format has succeeded to compress the image into 7.5% of the same image compressed by JPEG.



THE WHITE HOUSE
WASHINGTON

January 17, 2019

The Honorable Nancy Pelosi
Speaker of the
House of Representatives
Washington, D.C. 20515

Dear Madame Speaker:

Due to the Shutdown, I am sorry to inform you that your trip to Brussels, Egypt, and Afghanistan has been postponed. We will reschedule this seven-day excursion when the Shutdown is over. In light of the 800,000 great American workers not receiving pay, I am sure you would agree that postponing this public relations event is totally appropriate. I also feel that, during this period, it would be better if you were in Washington negotiating with me and joining the Strong Border Security movement to end the Shutdown. Obviously, if you would like to make your journey by flying commercial, that would certainly be your prerogative.

I look forward to seeing you soon and even more forward to watching our open and dangerous Southern Border finally receive the attention, funding, and security it so desperately deserves!

Sincerely,

Fig. 3 Document sent by DCTDecode

3. WHY FAX COMPRESSES DOCUMENTS BETTER

The compression algorithm of faxes assumes that only white pixels and black pixels are present. If a pixel has another color, the fax will convert it to either white or black according to its brightness.

Faxes use a version of run length encoding [18]. Each line of pixels in the document is encoded as a series of alternating sequences of white and black pixels. One sequences of white pixels followed by a sequence of black pixels and then again a sequences of white pixels followed by a sequence of black pixels and so forth.

Sequences of 63 pixels or less are encoded by a terminating code, which means the fax expects to subsequently encode a sequence of the other color, whereas sequences of 64 pixels or greater will require a makeup code which means the subsequent sequence of pixels will be of terminating code of the same color. E. g. a sequence of 200 pixels of the same color will be encoded as a makeup code of 192 pixels followed by a terminating code of 8 pixels. The makeup codes are employed to denote sequences in multiples of 64 from 64 to 2560.

For example A4 usually has 2480 pixels in one line. Figure 4 shows an example of pixel line of 2480 pixels in this order:

- 0 white pixel.
- 2 black pixels.
- 3 white pixels.
- 5 black pixels.
- 2 white pixels.
- 4 black pixels.
- 11 white pixels.
- 1 black pixel.
- 2434 White pixels.
- 1 black pixel.
- 2 white pixels.
- 15 black pixels.

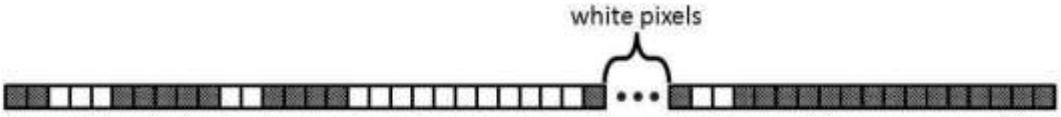


Fig. 4 Example of a pixel line in a document

Studies have shown that in average images sent by fax machines are 85 percent white and only 15 percent are black [19]. Accordingly, fax machines assume that any pixel line begins by a sequence of white pixels and if the pixel line does not begin by a sequence of white pixels, a sequence of white pixels in length of zero, will begin the encoded pixel line.

Accordingly, the code words that will be sent in order to represent the pixel line of Figure 4 are:

00110101, 11, 1000, 0011, 0111, 011, 01000, 010

After that, the 2434 cannot be represented by only one terminating code, because 2434 is larger than 63, so the fax will break it into a makeup code followed by a terminating code. The multiple of 64 that is the closest to 2434 is 2432, so the fax will take the makeup code of 2432 white pixel which is:

000000011101

Next to this makeup code, the fax will send the rest of the pixels in this sequence, which are two white pixels and their terminating code is:

0111

Finally, the fax will send the rest of the pixel line as it is described in Figure 4 and their codes will be:

010, 0111, 000011000

Unlike JPEG format, faxes assume that there will be sharp differences and in view of that faxes encode the pixel lines as a series of alternating sequences of white and black pixels. Documents usually have sharp changes because they contain black letters on white background; therefore, faxes can suitably handle such documents with black letters.

In addition, faxes are more robust to errors because each line of pixels ends with a special codeword named EOL (end of line). EOL consists of eleven zero bits and one bit with the value one (000000000001). There was a suggestion to make the addition of such an EOL codeword to the JPEG format [20,21]; however, this suggestion has not been accepted.

Fax machines assume that there might be an error in the transmission like missing bits or a bit that changed its value from one to zero or zero to one. In such cases it could happen that the fax machine will not be able to know where the codewords for the sequences of bits begin and end. However, this error will not last more than one pixel line.

In order to make sure that all EOL codewords are decoded correctly, fax codewords' table stipulates that there will be no codeword consists of more than seven leading zeroes and also there will be no codeword consists of more than three ending zeros, so even if the receiver fax misplaces the codewords boundaries and two codewords might be conjoined, there will be no more than ten continuous zeros in any erroneous codeword.

If the receiver fax gets eleven zeros, it will know that this should be a part of the end of line codeword and will continue scanning until it gets a bit with the value one. Just upon receiving the bit with the value one, the fax will start to print a new line. If one bit in any scanned line gets corrupted, the worst damage will be a loss of the rest of the line. If the EOL codeword itself gets corrupted, the worst damage will be a loss of the next line.

4. PHOTOGRAPH COMPRESSION

Unlike documents, photographs have much fewer sharp differences [22,23,24]; therefore, the advantage of fax compression ratio for photographs will be smaller and also the in many cases the fax compress will produce a considerably less clear image with many noticeable flaws.

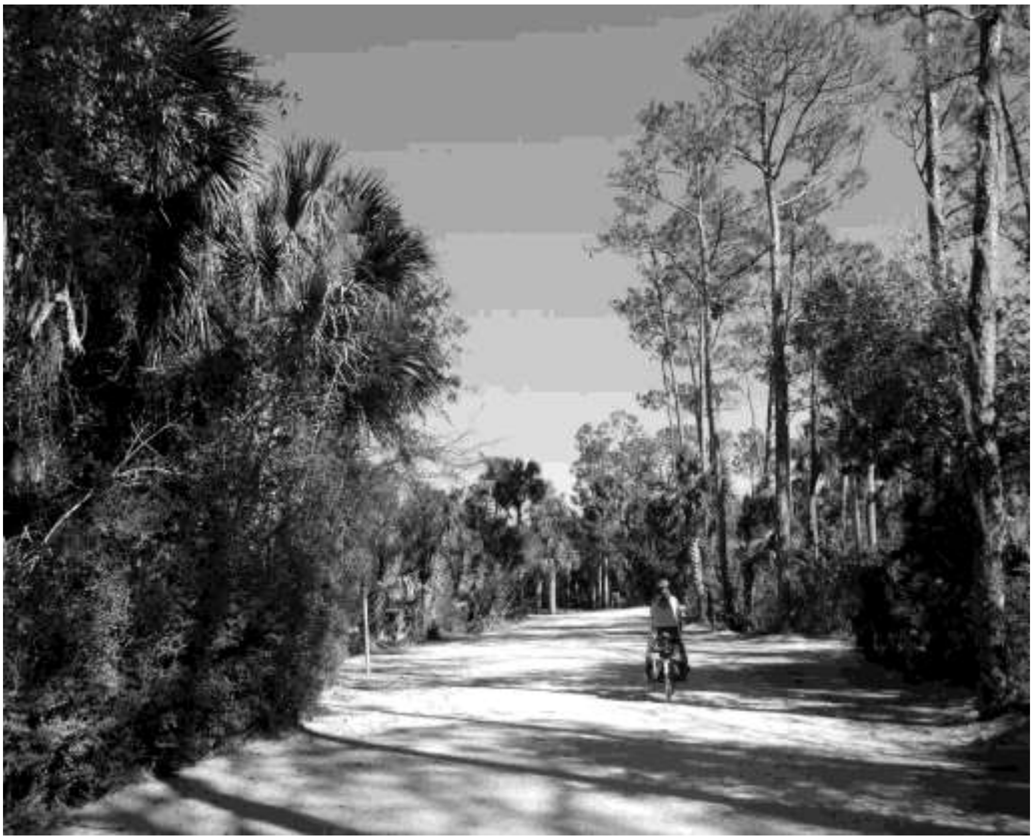


Fig. 5 Grayscale image compressed by JPEG

Figure 5 contains an image compressed by grayscale JPEG. The image is clear and almost unblemished. Just in areas in the image where a sharp difference occurs, the blocks containing this sharp difference will be compressed into a larger sequence of bits [25,26,27,28]. Also, when there is a significant difference between the average hues of the last block in a block line and the average hues of first block in the successive block line, like the lines in the upper part of the image, the DC will be compressed into a large sequence of bits [29,30].

The same image is shown in Figure 6; however, in this case the image has been compressed by the algorithm of fax machines. Fax machines have only white and black pixels; no gray pixels are available in faxes. The gray sky was in a gray hue that sometimes is closer to white and sometimes is closer to black. Consequently, the sky looks like it is made of white and black pixels and for that reason, the quality of the image is much lower.

The compression ratio of this image is also not as good as the compression ratio of documents. The compression ratio achieved by the fax was 125KB and the compression ratio achieved by JPEG was 198KB, which means the JPEG file size was 63.1% of the fax file size. This reduction is much smaller than the reduction in document file sizes, where a fax can reduce the file size into less than 10% of the JPEG file size in most cases.

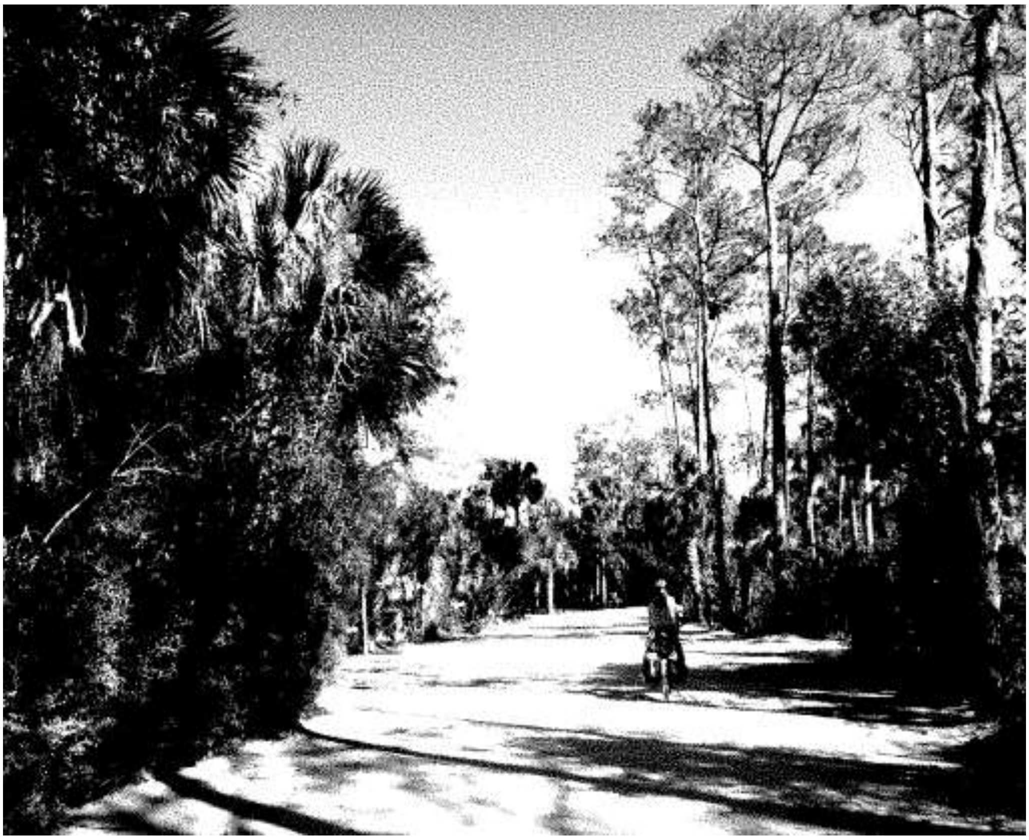


Fig. 6 Grayscale image compressed by a fax machine

6. CONCLUSIONS

Images are often embedded in PDF files [31]. Many application use JPEG as the default format for these images [32]; however, if the image is actually a document with black letters on white background, the fax machine format can be more effective in both compression efficiency and image quality.

In this paper we showed that checking the image compression efficiency by JPEG format and by the facsimile format before embedding the image into the PDF file, can be advantageous in both compression efficiency and visible quality of the embedded document.

REFERENCES

[1] D. A. Huffman, "A method for the construction of minimum-redundancy codes", Proceedings of the IRE 40, No. 9, pp. 1098-1101, (1952).

[2] International Telecommunication Union Standardization Sector, "Procedures for Document Facsimile Transmission in the General Switched Telephone Network", ITU-T (CCITT), Recommendation T.30, (1996).

[3] P. Aguilera, "Comparison of different image compression formats", Wisconsin College of Engineering, ECE 533, (2006).

[4] S. T. Klein and Y. Wiseman, "Parallel Huffman Decoding with Applications to JPEG Files", The Computer Journal, Oxford University Press, Swindon, United Kingdom, Vol. 46, No. 5, pp. 487-497, (2003).

[5] S. T. Klein and Y. Wiseman, "Parallel Huffman Decoding", Proc. Data Compression Conference DCC-2000, Snowbird, Utah, USA, pp. 383-392, (2000).

[6] N. S. Borenstein, "MIME: a portable and robust multimedia format for internet mail." Multimedia Systems, Vol. 1, No. 1, pp. 29-36, (1993).

[7] Y. Wiseman, K. Schwan and P. Widener, "Efficient End to End Data Exchange Using Configurable Compression", Proc. The 24th IEEE Conference on Distributed Computing Systems (ICDCS 2004), Tokyo, Japan, pp. 228-235, (2004).

[8] T. Bienz, R. Cohn and J. R. Meehan, "Portable Document Format (PDF) Reference Manual", Published by Adobe Systems Incorporated, ISBN 0–201–62628–4, Mountain View, California, USA, (1997).

[9] H. Déjean and J. L. Meunier, "A system for converting PDF documents into structured XML format", In International Workshop on Document Analysis Systems, pp. 129-140. Springer, Berlin, Heidelberg, (2006).

[10] G. Feng, M. G. Fuchs and C. A. Bouman, "Image rendering for digital fax", In International Society for Optics and Photonics, Color Imaging VIII: Processing, Hardcopy, and Applications, vol. 5008, pp. 504-513, (2003).

[11] L. Rosenthol, "Developing with PDF: Dive Into the Portable Document Format", O'Reilly Media, Inc., (2013).

[12] Y. Wiseman, "The Still Image Lossy Compression Standard – JPEG", Encyclopedia of Information and Science Technology, Third Edition, IGI Global, Vol. 1, Chapter 28, pp. 295-305, (2014).

[13] Y. Wiseman, "Take a Picture of Your Tire!", In Proceedings of The 12th IEEE Conference on Vehicular Electronics and Safety (IEEE ICVES-2010), Qingdao, ShanDong, China, pp. 151-156, (2010).

[14] Y. Wiseman, "The Effectiveness of JPEG Images Produced By a Standard Digital Camera to Detect Damaged Tyres", World Review of Intermodal Transportation Research, Vol. 4, No. 1, pp. 23-36, (2013).

[15] Y. Wiseman, "Camera That Takes Pictures of Aircraft and Ground Vehicle Tires Can Save Lives", Journal of Electronic Imaging, Vol. 22, No. 4, 041104, (2013).

[16] Y. Wiseman and E. Fredj, "Contour Extraction of Compressed JPEG Images", ACM - Journal of Graphic Tools, Vol. 6, No. 3, pp. 37-43, (2001).

[17] E. Fredj and Y. Wiseman, "An O(n) Algorithm for Edge Detection in Photos Compressed by JPEG Format", Proc. IASTED International Conference on Signal and Image Processing SIP-2001, Honolulu, Hawaii, pp. 304-308, (2001).

[18] G. Feng and C. A. Bouman, "Efficient document rendering with enhanced run length encoding", In International Society for Optics and Photonics , Color Imaging XI: Processing, Hardcopy, and Applications, vol. 6058, p. 60580R, (2006).

[19] S. J. Urban, "Review of standards for electronic imaging for facsimile systems", Journal of lectronic Imaging, Vol. 1, No. 1 pp. 5-22, (1992).

[20] Y. Wiseman, "Enhancement of JPEG Compression for GPS Images", International Journal of Multimedia and Ubiquitous Engineering, Science and Engineering Research Support Society (SERSC), Vol. 10, No. 7, pp. 255-264, (2015).

[21] Y. Wiseman, "Alleviation of JPEG Inaccuracy Appearance", International Journal of Multimedia and Ubiquitous Engineering, Science and Engineering Research Support Society (SERSC), Vol. 11, No. 3, pp. 133-142, (2016).

[22] Y. Wiseman, "Real-Time Monitoring of Traffic Congestions", in proceedings of IEEE International Conference on Electro Information Technology (EIT 2017), Lincoln, Nebraska, USA, pp. 501-505, (2017).

[23] Y. Wiseman, "Tool for Online Observing of Traffic Congestions", International Journal of Control and Automation, Vol. 10, No. 6, pp. 27-34, (2017).

[24] Y. Wiseman, "Computerized Traffic Congestion Detection System", International Journal of Transportation and Logistics Management (IJTLM), Global Vision Press, Vol. 1, No. 1, pp. 1-8, (2017).

[25] Y. Wiseman, "Device for Detection of Fuselage Defective Parts", Information Journal, Tokyo, Japan, Vol. 17, No. 9(A), pp. 4189-4194, (2014).

[26] Y. Wiseman, "Fuselage Damage Locator System", Advanced Science and Technology Letters, Vol. 37, pp. 1-4, (2013).

[27] Y. Wiseman, "Automatic Alert System for Worn Out Pipes in Autonomous Vehicles", International Journal of Advanced Science and Technology, Science and Engineering Research Support Society (SERSC), Vol. 107, pp. 73-84, (2017).

[28] Y. Wiseman, "Safety Mechanism for SkyTran Tracks", International Journal of Control and Automation, Vol. 10, No. 7, pp. 51-60, (2017).

[29] Y. Wiseman, "Enhancement of JPEG Compression for GPS Images", International Journal of Multimedia and Ubiquitous Engineering, Science and Engineering Research Support Society (SERSC), Vol. 10, No. 7, pp. 255-264, (2015).

[30] Y. Wiseman, "Improved JPEG based GPS picture compression", Advanced Science and Technology Letters, Vol. 85, pp. 59-63, (2015).

[31] D. G. Barnes, M. Vidiassov, B. Ruthensteiner, C. J. Fluke, M. R. Quayle and C. R. McHenry. "Embedding and publishing interactive, 3-dimensional, scientific figures in Portable Document Format (PDF) files" PloS one, Vol. 8, No. 9, Paper no, e69446, (2013).

[32] W. L. Cheng. "Portable Document Format (PDF) - Finally, a Universal Document Exchange Technology", The journal of Technology Studies, Epsilon Pi Tau Publication, Vol. 28, No. 1, pp. 59-63, (2002).