Impact of Document Spectral Hue Intensity on Fax Compression Ratio

Student authors: Vladimir R. Ristić[1], Nemanja J. Mitić[1] and Dušan S. Marjanović[1]
Mentors: Radomir S. Stanković[2] and Dušan B. Gajić[2]

Abstract - **This paper presents the results of a student research done on analysis of impact of grey colour presence on the efficiency of the one-dimensional fax compression. The objective of analysis is to validate the implementation of advanced compression methods. Through performing a simulation of the modified Huffman compression, using a C++ software implementation, various experimental results were gathered. For example, after analysis of results a strong dependency of middle intensity spectral hue area coverage on compression ratio is detected. The implementation has solid real-world performance, but it was primarily developed with educational purposes in mind.**

Keywords - **One-dimension fax compression, modified Huffman compression, fax compression methods.**

I. INTRODUCTION

The main intention of the research leading to this paper was to obtain practical knowledge about the significance of advanced compression methods. The methodology rests upon simulation of real environment conditions using representative document patterns. Architecture of dialog based MFC application used for this research is simple. Input document image is a BMP file. Access to the bits that encode document image pixels is enabled using *dynamic_bitset* class [5] from the Boost C++ Libraries [4]. Application is developed in Microsoft Visual Studio 2010 IDE. Results are gathered manually and processed by Microsoft Excel 2007. More complex pattern sets require further aplication version to be featured with automatically gathering and graphically presenting the results through its own GUI.

A. Fax standards

ITU-T, former Comité Consultatif International Téléphonique et Télégraphique is a part of the International Telecommunications Union. Although this body declares its acts as "Recommandations", international standards are mandatory in a form of national laws of countries which adopted them.

Standard T.3 (Group 3 fax) introduced digital devices using analog lines for data transmission and was first to consider the image data compression standardization. Standard T.4 involves significant compression algorithm improvement and T.6 standard (Group 4 fax) is implying digital telecommunication lines (for example ISDN) [2].

B. Data compression

Data compression algorithms can be lossy or lossless [1]. If we consider data as a coded entity from real world, and information as useful data then even lossy compression algorithms save the actually desired information. Lossless compression saves all the data, useful at the present moment but also data that might be useful later. Fax data compression is saving all data, so the lossless compression is implemented.

C. Facsimile compression

Result of scanning process is a monochrome bit mapped file, which is to be send to destination fax. Compression is performed to remove data redundancy, thus decreasing the file size to reduce the file transferring time. **Modified Huffman (MH) compression** method is used, a run-length codebook based statistical algorithm [2]. Statistically, there will be more white areas on document than black ones, so codebook contains shorter words for white pixel sequences. MH algorithm takes into count most common pixel sequences which are 2-4 of black and 2-7 of white, so they are coded with shortest codes. Here is the main compression problem. White area is represented with white pixels and black area is represented with black ones. When middle intensity spectral hue is scanned, optical scanning system generates alternating sequence of black and white pixels. As the single pixel sequence is rare it will be coded with word longer than one bit [1]. It means that those sequences will deteriorate the compression. Determining the rate of this impact is the primary goal of this work. Method described above is the so called **one-dimension fax compression** [1]. It is used for T.3 Recommendation standard implementation. Advanced T.4 standard is based on Modified Read (MR) which is known as two-dimension fax compression. This method initializes encoding in the manner of MH method for the first scanned line on page, but second line is then compared to the first one and differences are encoded. Every next line is compared to the previous one as reference to encode the further

Student authors:
[1]Vladimir R. Ristić, Nemanja J. Mitić and Dušan S. Marjanović are with the University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia, E-mails: nemanja.mitic@elfak.rs, studentristicvladimir@gmail.com, dusan.marjanovic@elfak.rs.
Mentors:
[2]Radomir S. Stanković and Dušan B. Gajić are with the University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš, Serbia, E-mails: radomir.stankovic@gmail.com, dusan.gajic@elfak.ni.ac.rs.

differences. This method is especially effective when differences are small or none. Bad side of MR method is that data transition error propagates through the whole page. T.4 standard does not provide correction of errors, but provides number limitation of lines encoded by MR between the lines encoded by MH method. T.6 Recommendation standard allows more MR encoded lines between MH encoded lines due to improvement of reliability of digital lines [1].

D. ITU-T document patterns

In developing code for T.3 standard, ITU-T took eight representative documents as statistical patterns. Optimal codebook was efficiently derived through counting all black and white pixel sequences. As those documents are copyrighted by ITU-T they cannot be used as experimental material [1]. So, new reference document set is created.

II. Algorithm of Choice for Compression

Crucial moment during the development process is **fax compression method** choice. Properly selected algorithm must meet both the requirements for high rate lossless compression and for high efficiency, resulting in short image processing time. As those requirements are in contradiction, the primary goal is finding an optimal solution.

Total fax device costs, C_t consists of initial costs C_i (device price) and operating costs C_o (including phone bills):

$$C_t = C_i + C_o \quad (1)$$

$$C_t = f'_{ci}(A_c) + f''_{tt}(A_c) \quad (2)$$

Initial costs (blue curve in Figure 1.) directly depend on compression algorithm complexity and operating costs (red curve) directly depend on compression ratio, i.e. operating costs indirectly depend on compression algorithm complexity. In this case, sum of two opposite depending functions of the same argument (green curve) has a local minimum. Projection of the local minimum onto the axe of compression algorithm complexity indicates an optimal algorithm.

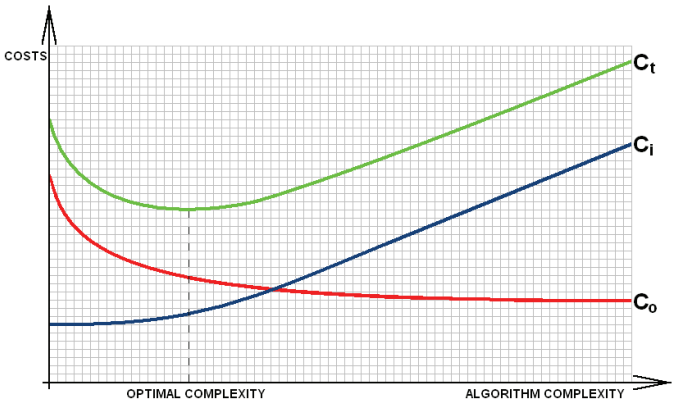

Fig. 1. Fax device costs dependencies on algorithm complexity.

Factors, additionally influencing this optimum, can be classified into two groups: as environmental and as implementational. A fax device development doesn't set the environmental factors (quality of telecommunication infrastructure, payment rates etc.), but must consider them. What development sets are implementational factors (device resolution, modulation type, data transmission rate etc.). From the standpoint of economy, key feature of cost effective device exploitation, is implemented compression algorithm.

If some document property has a huge negative impact on fax compression rate using conventional methods, advanced compression methods should be implemented to solve this problem.

III. Reference Document Set

For this research, the original reference was unavailable, but the document description was available. As the focus of experimental work was on compression rate deterioration, caused by incensement of middle intensity spectral hue coverage, reference document set was modified. Facsimile optical system converts any document image into monochromatic file. More intense hue document is converted into a bit map that contains more dense black pixels among white pixels, and vice versa. Half intensity hue is converted in alternating black and white pixel sequence.

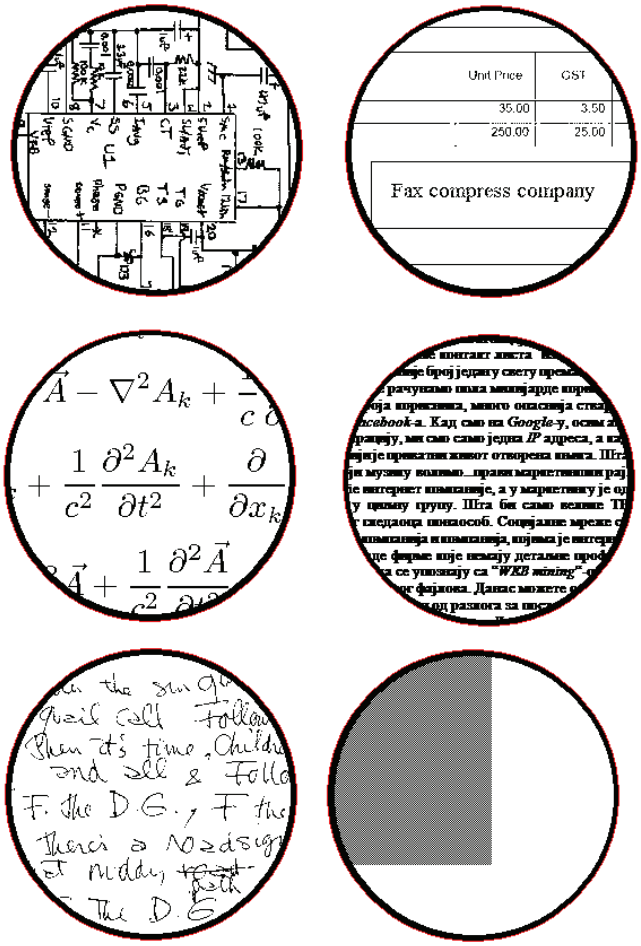

Fig. 2. Reference document parts.

In order to simulate documents with middle intensity hue, a series of BMP files was created containing various area percentage of alternating black and white pixels. Modified reference set consists of seven representative document images, one image file completely filled with black pixels, and four files with 6%, 13%, 25% and 50% of area filled with half-intense gray hue in form of alternating B/W pixels.

Figure 2. shows parts of image files used to simulate representative documents. Obviously, invoice document image (second circle part) contains longest sequences of black or white pixels, while dense document image (fourth circle part) consists of very short sequences of the same color pixels, not counting the last circle with alternating area. It is expected in experiments to get the result set that is in accordance with this note, i.e. best compression ratio for invoice image, and worst in case of dense document or document with alternating black and white.

IV. Software Simulation

Main dialog of graphic user interface is shown in Figure 3.

Fig. 3. Software simulation user interface.

Reason for choosing C++ rather than some other program languages was its high speed and performance of executive code, as well as its bit manipulation capabilities. Application is a simulation of one-dimension fax compression.

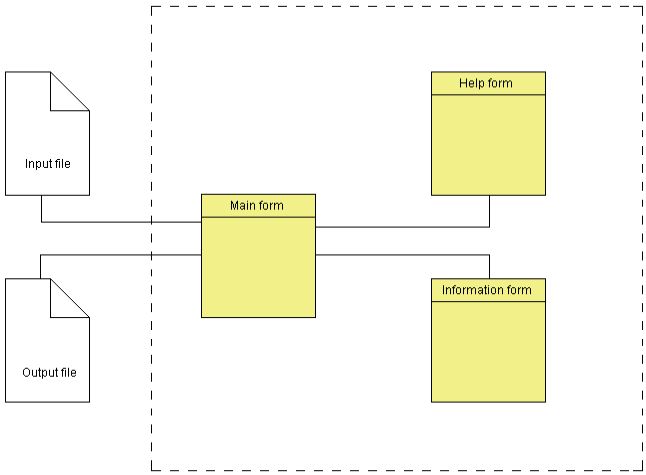

Fig. 4. Simulation program architecture.

Codebook is realized as a bit array vector. Input bitmap file is processed row by row, loading the bitset vector and counting the same color pixels in sequence within. Code words are written into a binary file. Decompression is not implemented, as it is not relevant for compression ratio data set.

V. Results and Analysis

As expected, the experimental results confirmed the theory. As Table I shows, there is a huge compression efficiency variation caused by document properties.

TABLE I
COMPRESSION RATIO AND TIME

No.	Document type	In (B)	Out (B)	Ratio (%)	Time (ms)
1	Business letter	109416	84819	77	31
2	Electric circuit	104960	95922	91	47
3	Invoice	83412	27722	33	16
4	Dense report	100032	108400	108	47
5	Equations	65604	44534	67	15
6	Dense document	108600	122288	112	46
7	Handwritings	54450	37594	69	16
8	Black area 100%	75000	2900	3	16
9	Gray area 50%	75000	327643	436	94
10	Gray area 25%	75000	170227	226	47
11	Gray area 13%	75000	86429	115	31
12	Gray area 6%	75000	48394	64	15

We can note that very dense documents have negative compression ratio. Average pixel sequence length is close to one. This effectively approaches to negative impact of gray

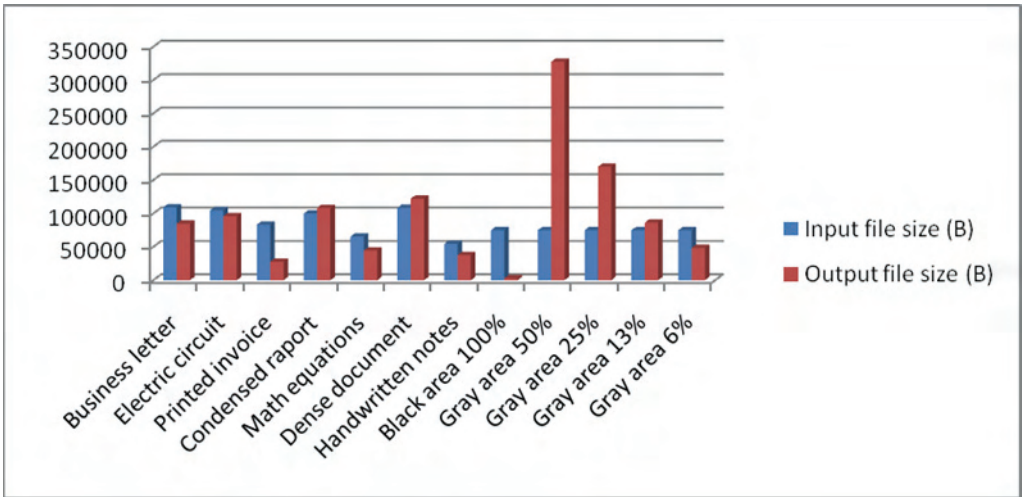

Fig. 5. Input file size variations.

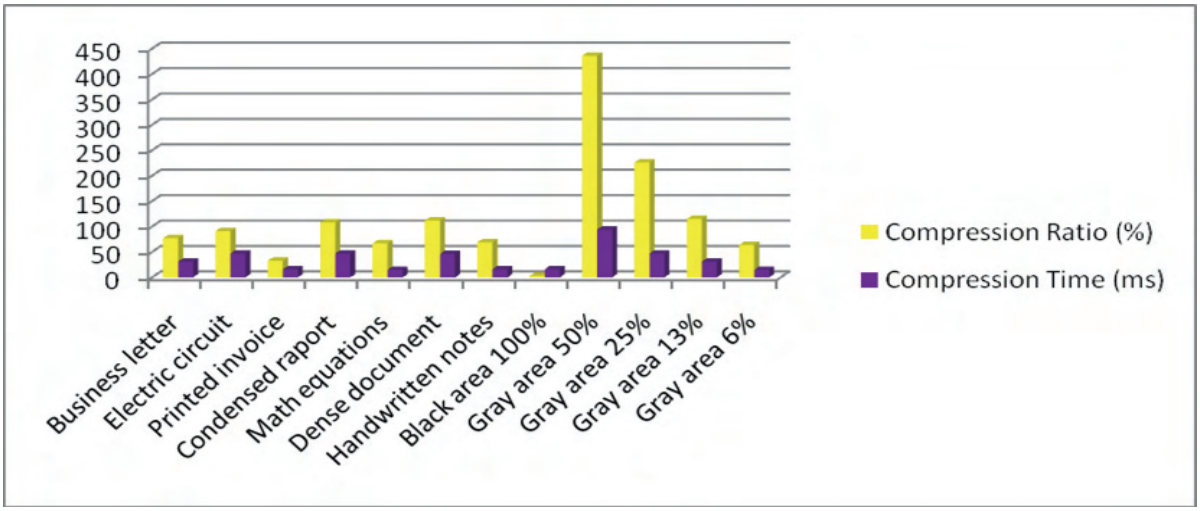

Fig. 6. Compression ratio and time variations.

hues on compression ratio. Invoice has huge white area coverage, resulting in high compression ratio.

The minimal rate of compression minimal is 3%. This is also theoretical minimum due to contextual data in coded file.

Strong compression ratio and compression time dependencies on percentage of middle spectral hue coverage on document are very notable facts. For turning compression rate into negative value, it is enough that the document contains more than 10% of this area. Still, facsimile is prevalent use for documents with high contrast degree rather than with middle intensity spectral hues.

VI. Conclusion

The programming implementation is able to execute both the task of image compression and of compression ratio measuring. It was found appearance of hues of middle intensity has strong negative impact on compression ratio. Experimental results in conjunction with the starting assumption lead us to conclusion that advanced compression methods application is conditionally sustainable. This stands under the assumption that the telecommunication infrastructure is of high quality which provides low data transmission error rate.

Device implementation environment is a key factor to be considered when implementing two-dimension facsimile compression method. In digital infrastructure environment this is method of choice.

References

[1] David Salomon, *Data Compression: The Complete Reference*, 3rd edition, Springer, 2004. (ISBN 0-387-40697-2)
[2] http://en.wikipedia.org/wiki/Fax, 23. 12. 2010, 18:00.
[3] http://en.wikipedia.org/wiki/Data_compression, 24. 12. 2010, 16:00.
[4] http://www.boost.org, 14. 11. 2010, 9:00.
[5] http://www.boost.org/doc/libs/1_45_0/libs/dynamic_bitset/dynamic_bitset.html, 14. 11. 2010, 13:00.
[6] http://www.britannica.com/EBchecked/topic/199972/fax, 27.12.2010, 20:00.