Method for Encoding the Message Source According to the Characteristics of the Structural Group and Their Quantitative Measure

Kseniya Kurin
National Aviation University
Kyiv, Ukraine
yak333@ukr.net

Oleksandr Yudin
State Service for Special Communication and Information Protection of Ukraine
Kyiv, Ukraine
yudin.ok8@gmail.com

Olha Suprun
National Aviation University
Kyiv, Ukraine
o.n.suprun@gmail.com

Viktoriia Ivannikova
Dublin City University
Dublin, Ireland
viktoriia.ivannikova@dcu.ie

Iryna Tregubenko
State Service for Special Communication and Information Protection of Ukraine
Kyiv, Ukraine
wr.irttri@gmail.com

Oleksii Yudin
State Service for Special Communication and Information Protection of Ukraine
Kyiv, Ukraine
yudinalex.ua@gmail.com

2022 IEEE 4th International Conference on Advanced Trends in Information Theory (ATIT) | 979-8-3503-3262-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ATIT58178.2022.10024239

***Abstract* — In presented article, the expedience of the use of structural code in the tasks of compression of data is proved. The structural signs of binary sequences are defined. Structural signs of binary sequences and limitations that are laid on them are defined. The rule of determination of the number of binary sequences which form structural groups is described. The rule of calculation of the sequence number of binary sequence based on the number of bit changes is described. The estimation of the expected degree of compression is done for the sequences of different length. Also the testing of proposed approaches are presenting, and comparissons to classic algorithms were made.**

***Keywords* — compression, structural coding, structural signs, bit changes, structural group, compression coefficient**

I. Introduction

The idea of data compression is a quite natural, it also appears in ordinary language in the form of various abbreviations. The main reason for using data compression in communication systems is the desire to transmit and store information with the greatest efficiency. In some sense, the task of data compression consists in extracting from the data stream the most significant and unique part, which will allow to restore all original information.

Due to the trend to the continuous intensive development of computer graphics, a certain field of compression theory, namely image compression algorithms, is rapidly developing.

The most compact representation of data according to the well-known work of Claude Shannon [1] on information theory is determined by the entropy. It is a numerical characteristic that determines the lower limit of possible compression. This value provides a quantitative assessment of the information contained in the message, based on its statistical characteristics.

However, recently, new improbable approaches to the estimation of the information quantity have been appeared in information theory. This is due to the fact that most of information processes are not strictly statistical ones and their information content cannot be evaluated only from the standpoint of the probabilistic theory of information [2]. This proves the necessity to use non-probability methods for determination of the quantity of information and their adequate usage.

Statistical compression methods are based on the assessment of the probability of appearance of the elementary symbols in the array of information, which determines the codes of variable length, that are compared with correspondent symbols.

The most common image compression technologies JPEG and JPEG-2000 currently provide high degrees of compression due to the reduction of psychovisual redundancy and further statistical coding of the transformant components of orthogonal transformations. Psychovisual redundancy is reduced due to the zeroing of high-frequency composite parts of the transformants components as a result of their subsequent quantization. Statistical compression methods such as Huffman coding (for the JPEG standard) and arithmetic coding (JPEG-2000) are used to encode quantized transformants. These coding methods are characterized by the following disadvantages [3,4]:

– number of machine operations allocated to the processing of transformed images can be up to 70% from the total number of operations of their compression procedure; this is due to the fact that it is necessary to take into account the number of operations spent to calculate statistics, to build code tables and to organize a double pass through the processed data, and during restoring the entire fragment will be restored only after recoding of all uneven code words;

– number of operations required to perform statistical coding for transformants formed on the basis of DKP and DVP reaches 80% of the total number of operations spent on obtaining a compressed image, and may exceed the number of operations for performing transformations (for modes ensuring high quality of image);

– due to the necessity of synchronization and marking of uneven code combinations on the boundaries of the processed fragments of the transformed image, the complexity of the software and hardware implementation of the coding procedure increases significantly;

– it is quite difficult to implement parallel processing of statistical codes;

979-8-3503-3262-9/22/$31.00 ©2022 IEEE

– during processing of highly coherent images, the compression degree is sharply reduced. It is caused by the increasing of the values of high-frequency components of the transformant, as well as by the necessity to use code tables and markers;

– statistical coding will not provide additional compression of the transformant if there is a series of zeros of small lengths.

Therefore, the goal of the research is to develop a coding method that uses the characteristics of an information message, which are fundamentally different from the statistical ones, taking into account the possibility of its further usage in image compression technology [5,6].

II. MAIN PART

There is a class of compression methods among the set of compact data presentation technologies, which is based on the elimination of structural redundancy due to the identifying of structural regularities in binary sequences, based on some feature. They are called methods of structural coding [3].

Structural coding methods for increasing the degree of compression and saving the specified quality of the reproduced image at the same time are ensured by the compliance with the following requirements [3]:

1) binary data are processed, because binary representation is universal for presenting information from different sources;

2) the process of data compression is organized based on the removal of structural redundancy;

3) structural regularities are taken into account by several features simultaneously in the process of compression; at the same time in order to ensure the degree of compression, the features must be interdependent.

The choice of the structural nature features is determined by the following features [3]:

– they have a quantitative measure;

– they take into account the regularities that give possibility to exclude redundancy without loss of quality;

– only integer operations are used, which reduces processing time;

– knowledge of apriori information are not required for coding and is more resistant to non-stationary images.

Let's consider the following structural features of the binary sequences: number of transitions from a single element to a zero element $p_{1\to0}$ and the number of transitions from a zero element to a single element $p_{0\to1}$. Tables 1-3 show binary sequences with the lengths n equal to 4, 3, and 2 bits, correspondently. For each of the sequences, the values of structural features $p_{0\to1}$ and $p_{1\to0}$ are indicated. It should be mentioned that analysis of the values of structural features is carried out taking into account the fact that an additional zero bit is introduced into the binary sequence (highlighted in gray in the table). This sequence will be called extended one.

It is obviously that binary sequences, which are characterized by the same values of structural features can be grouped into structural groups. Structural groups (SG) are highlighted in different shades of gray in Tables 1-3.

TABLE I. VALUES OF SEQUENCE NUMBERS AND STRUCTURAL FEATURES FOR BINARY SEQUENCES WITH THE LENGTH $n = 4$ BITS

S/Number	Binary sequence	Structural features		Number of structural groups	Sequential number in the structural group
		$p_{0\to1}$	$p_{1\to0}$		
0	00000	0	0	0	0
1	00001	1	0	1	0
2	00010	1	1	2	0
3	00011	1	0	1	1
4	00100	1	1	2	1
5	00101	2	1	3	0
6	00110	1	1	2	2
7	00111	1	0	1	2
8	01000	1	1	2	3
9	01001	2	1	3	1
10	01010	2	2	4	0
11	01011	2	1	3	2
12	01100	1	1	2	4
13	01101	2	1	3	3
14	01110	1	1	2	5
15	01111	1	0	1	3

At the same time, coding will mean receiving sequence of its serial number in the given group. The sequence number value will be less than the decimal value of the binary sequence. Due to this fact, compression will be provided.

TABLE II. VALUES OF SEQUENCE NUMBERS AND STRUCTURAL FEATURES FOR BINARY SEQUENCES WITH THE LENGTH $n = 3$ BITS

S/Number	Binary sequence	Structural features		Number of structural group	Sequential number in the structural group
		$p_{0\to1}$	$p_{1\to0}$		
0	0000	0	0	0	0
1	0001	1	0	1	0
2	0010	1	1	2	0
3	0011	1	0	1	1
4	0100	1	1	2	1
5	0101	2	1	3	0
6	0110	1	1	2	2
7	0111	1	0	1	2

TABLE III. VALUES OF SEQUENCE NUMBERS AND STRUCTURAL FEATURES FOR BINARY SEQUENCES WITH THE LENGTH $n = 2$ BITS

S/Number	Binary sequence	Structural features		Number of structural group	Sequential number in the structural group
		$p_{0\to1}$	$p_{1\to0}$		
0	000	0	0	0	0
1	001	1	0	1	0
2	010	1	1	2	0
3	011	1	0	1	1

Tables 4-6 summarize the values of the binary sequences number in structural groups and the values of structural features, which determine them for sequences of different lengths.

TABLE IV. NUMBER OF BINARY SEQUENCES IN STRUCTURAL GROUPS ($n = 4$ BITS)

Structural features		Number of structural group	Number of sequences in a structural group
$p_{0\to 1}$	$p_{1\to 0}$		
0	0	0	1
1	0	1	4
1	1	2	6
2	1	3	4
2	2	4	1

TABLE V. NUMBER OF BINARY SEQUENCES IN STRUCTURAL GROUPS ($n = 3$ BITS)

Structural features		Number of structural group	Number of sequences in a structural group
$p_{0\to 1}$	$p_{1\to 0}$		
0	0	0	1
1	0	1	3
1	1	2	3
2	1	3	1

TABLE VI. NUMBER OF BINARY SEQUENCES IN STRUCTURAL GROUPS ($n = 2$ BITS)

Structural features		Number of structural group	Number of sequences in a structural group
$p_{0\to 1}$	$p_{1\to 0}$		
0	0	0	1
1	0	1	2
1	1	2	1

It is obviously that the value, which will be used for the binary sequence $A = \{a_1, a_2, ..., a_n\}$ coding in the structural group depends on the length of the binary sequence n and the values of the structural features $p_{0\to 1}$ and $p_{1\to 0}$. At the same time, the combination $p_{0\to 1}$ and $p_{1\to 0}$ simultaneously defines another feature that characterizes the binary sequence, i.e. number of the structural group S. Parameter S also determines the total number of transitions between zero and one in the sequence. Moreover, the values of S, $p_{0\to 1}$ and $p_{1\to 0}$ are connect by the following relations:

$$p_{0\to 1} = \left[\frac{S+1}{2}\right]; \quad (1)$$

$$p_{1\to 0} = \left[\frac{S}{2}\right]; \quad (2)$$

$$S = p_{1\to 0} + p_{0\to 1}. \quad (3)$$

The obtained data allows to identify the following relationships:

1. For a sequence with the length n bit the number of structural groups will be $n+1$.

This statement is easy to prove by considering the restrictions applied to the values of structural features:

- number of transitions from zero to one for a sequence with length n bit, taking into account the expansion, will vary in the range $0..\left[\frac{n+1}{2}\right]$;

- number of one-to-zero transitions $p_{1\to 0}$ for a sequence with length n bit taking into account the extension, will vary in the range $0..\left[\frac{n}{2}\right]$;

- $p_{1\to 0} + p_{0\to 1} \le n$;

- due to the fact that the extended binary sequence always starts from zero, the number of transitions from zero to one cannot exceed the number of transitions from one to zero;

- it is obviously that because of the mutual dependence $p_{0\to 1}$ and $p_{1\to 0}$ difference between the values of these features cannot exceed 1.

Therefore, the total number of structural groups for a sequence with the length n bit is determined by the number of possible combinations of the values $p_{0\to 1}$ and $p_{1\to 0}$. Taking into account the listed restrictions on the value of structural features, it can be stated that each possible value of a structural feature $p_{0\to 1}$ determines two combinations together with the value of the feature $p_{1\to 0}$, which is equal to it or is smaller by one. Then, the total number of possible combinations will be equal to:

$$2\times(\left[\frac{n+1}{2}\right]+1)-1 = n+1.$$

The unit that is subtracted takes into account one extra combination $p_{0\to 1} = 0$ and $p_{1\to 0} = 0$.

2. For each structural group S ($S \in 0..n$), the number of sequences in the group $N(S,n)$, and therefore the range of sequence numbers, which will be used for coding the sequences of the group, is defined as the number of ways with the help of which S transitions between zero and one can be placed at n positions in the extended sequence with the length $n+1$ bit. It is calculated as the corresponding coefficient of the Newton binomial of the n -th degree according to the formula:

$$N(S,n) = \binom{n}{S} = \frac{n!}{S!(n-S)!}. \quad (4)$$

This conclusion gives us possibility to make an assumption about the value of the biggest and the smallest degree of compression, which will be provided during usage of this coding method.

The biggest degree of compression will be achieved when a binary sequence with the length n is encoded by a sequence number 0 or 1, for which only one bit is needed. The numerical value of the maximum degree of compression will be equal to:

$k_{\max} = n.$

The smallest degree of compression will be obtained if the binary sequence is encoded by the value of the largest possible binomial coefficient for the given length of the binary sequence, reduced by 1 (it should be considered that the sequence numbering in the structural group starts from 0). At the same time, the value of the compression coefficient will be equal to:

$$k_{\max} = \binom{n}{\left[\frac{n+1}{2}\right]}.$$

Here, the operation $[\ ]$ means taking a whole part from a fractional number.

Numerically, the value of the smallest degree of compression will be equal to:

$$k_{\min} = \frac{n}{\left[\log_2\left(\binom{n}{\left[\frac{n+1}{2}\right]}-1\right)\right]+1}.$$

Table 7 shows the values of compression coefficients $k_{\max}$ and $k_{\min}$ for binary sequences of different lengths.

Therefore, the value of the sequential number of a binary sequence $A=\{a_1, a_2, ..., a_n\}$ in a structural group is a function of the length n of the binary sequence and the number S of the structural group.

We denote this value by $NUM(A,S,n)$.

TABLE VII. VALUES OF MAXIMAL AND MINIMAL COMPRESSION COEFFICIENTS FOR BINARY SEQUENCES OF DIFFERENT LENGTHS

Length of a binary sequence n, bit	Maximal compression coefficient $k_{\max}$	Minimal compression coefficient $k_{\min}$
4	4	1,333
8	8	1,143
16	16	1,143
32	32	1,067
64	64	1,049

It is undestandable that the coding procedure, i.e. determination of the $NUM(A,S,n)$ value can be carried out with the help of construction of coding tables, which are similar to the tables 1-3. Then, searching of the sequential numbers $NUM(A,S,n)$ in the tables, which will correspondent to the given binary sequences. On the receiving side, the decoding procedure will be realized in the same way. Basing on the received values n and S, code tables will be constructed, and according to the value $NUM(S,n)$ necessary binary sequence will be determined.

But intuitively, such coding principle seems to be impractical. Therefore, the next task will be to find a mathematical model that will describe the rule for calculating the value of the sequential number $NUM(A,S,n)$ for a binary sequence $A=\{a_1, a_2, ..., a_n\}$ according to the value of the structural feature S.

According to the value of structural feature S the sequential number $NUM(A,S,n)$ for a binary sequence $A=\{a_1, a_2, ..., a_n\}$ with the length n bit is calculated by the formula:

$$NUM(A,S,n) = \sum_{i=0}^{n-1}(a_{i+1}-a_i)\frac{(n-i)!}{l_i!(n-i-l_i)!}-a_n. \quad (5)$$

where, l_i is the number of residual transitions between zero and one elements of the binary sequence during processing the transition between a_{i+1} and a_i, which is numerically equal to:

$$l_i = l_{i-1} - |a_i - a_{i-1}|; \quad l_0 = s; \quad a_0 = 0.$$

III. CONCLUSIONS

The scientific novelty of the presented research lays in the following:

- the rationality of structural coding usage in data compression tasks is grounded;
- structural features of binary sequences and restrictions, applied to them are determined;
- the rule for determining the number of binary sequences, which form structural groups according to the common structural features is described and proved;
- the rule for calculating the sequential number of a binary sequence in a structural group according to the number of transitions between binary elements of the a sequence is described and proved;
- the expected degree of compression has been evaluated for sequences of different lengths.

The obtained results allow us to conclude that the proposed method for encoding binary data can be further used in compression technologies.

REFERENCES

[1] C. E. Shannon, "Probability of error for optimal codes in a Gaussian channel," in The Bell System Technical Journal, vol. 38, no. 3, pp. 611-656, May 1959, doi: 10.1002/j.1538-7305.1959.tb03905.x.

[2] Oleksandr Yudin, Olha Suprun, Ruslana Ziubina, Serhii Buchyk, Oleg Frolov, Natalia Barannik Efficiency Assessment of the Steganographic Coding Method with Indirect Integration of Critical Information: Proceeding of the International Conference on Advanced Trends in Information Theory (ATIT 2019), Kyiv, Ukraine, 18.12.19-20.12.19, 2019. IEEE Catalog Number: ISBN 978-1-72-81-61-44-0/19/$31.00©2019IEEE, pp.36-40. DOI: 10.1109/ATIT49449.2019.9030473

[3] N. Magotra, M. Kirstein, S. Sirivara and T. Hamill, "Quantitative and qualitative (subjective) perceptual measures for speech processing applications," Conference Record of The Thirtieth Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 1996, pp. 766-769 vol.1, doi: 10.1109/ACSSC.1996.601157.

[4] Yudin O., Ziubina R., Buchyk S., Suprun O., Matviichuk – Yudina O., Ivannikova V. Development of Metods for Identification of Informationcontrolling Signals of Unmanned Aircraft Complex Operator: Easten-European Journal of Enterprise Technologies, Vol 2, No 9(104) (2020), pp. 56-64. DOI: 10.15587/1729-4061.2020.195510.

[5] Yudin O., Barannik N., Ziubina R., Buchyk S., Frolov O., Suprun O. Efficiency Assessment of the Steganographic Coding Method with Indirect Integration of Critical Information: In.: 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT), pp. 36-40 (2019). DOI: 10.1109/ATIT49449.2019.9030473.

[6] Suprun O., Popereshnyak S., Suprun O., Wieckowski T. Intrusion Detection Method Based on the Sensory Traps System: Materials of International Conference Perspective Technologies and Methods in MEMS Design (MEMSTECH), Polyana, April 18-22, 2018, pp.122-126.