Statistical Analysis of Compression Methods for Storing Binary Image for Low-Memory Systems

Roman SLABY, Jana NOWAKOVA, Radim HERCIK

Department of Cybernetics and Biomedical Engineering, Faculty of Electrical Engineering and Computer Science, VSB–Technical University of Ostrava, 17. listopadu 15, 708 33 Ostrava-Poruba, Czech Republic

roman.slaby@vsb.cz, jana.nowakova@vsb.cz, radim.hercik@vsb.cz

Abstract. *The paper is focused on the statistical comparison of the selected compression methods which are used for compression of the binary images. The aim is to asses, which of presented compression method for low-memory system requires less number of bytes of memory. For assessment of the success rates of the input image to binary image the correlation functions are used. Correlation function is one of the methods of OCR algorithm used for the digitization of printed symbols. Using of compression methods is necessary for systems based on low-power micro-controllers. The data stream saving is very important for such systems with limited memory as well as the time required for decoding the compressed data. The success rate of the selected compression algorithms is evaluated using the basic characteristics of the exploratory analysis. The searched samples represent the amount of bytes needed to compress the test images, representing alphanumeric characters.*

Keywords

Data compression, exploratory analysis, image recognition, testing of statistical hypotheses.

1. Introduction

Digital image processing is a fast growing discipline, which is caused by the rapidly increasing development of computer technology and using of this discipline in various fields of human activity. Recently, the usage of systems using digital image processing in industrial applications is frequently observed trend. Such systems are often only the single-purpose systems for evaluating a specific industrial process with minimal costs. Minimal costs of the realized systems can be achieved for example by using low-power microcontrollers, where the greatest problem can be in the insufficient size of RAM. The deficiency in many cases can be solved by the addition of Flash memory because the cost of this type of memory is lower than using the external RAM.

Recognition systems use correlation functions for digitizing printed alpha-numeric characters using a set of patterns. Individual patterns represented by the binary values are compared by correlation function with the input image. The highest level of agreement expresses the digital printed form of alphanumeric character. For accurate digitization of printed character using the correlation function it is necessary to use a large set of characters, containing all possible characters that may occur in the input image. For this reason it is required corresponding memory space for all patterns. In such systems, the patterns are stored in flash memory; the RAM is reserved for purposes of correlation functions.

Image processing speed depends on the effectiveness of the implemented detection algorithm, but also on the speed of the microprocessor itself. Because, the time required for writing or reading data from the flash memory can affect the speed of image processing. Therefore, it is necessary to ensure the amount of accesses to the data that is stored in Flash memory.

One of the possible solution is to use a compression algorithm to compress binary patterns representing alphanumeric digits and their subsequent storage on the mini-mum number of pages used Flash memory [1], [2], [3].

2. Compression Methods

Data compression is a special procedure for data storing or data transporting, whose task is to reduce the data stream. All compression methods are based on the fact that commonly used methods of data representation are designed to allow easy manipulation of data, regardless of the fact that some parts of the data are often repeated. The consequence is that data for storing often needs more space than is absolutely nec-

essary. Lossless compression methods remove the redundant information without losing any data. In the paper two compression methods are presented and it is assessed the difference in amount of bytes which are needed for encoding the images [4], [5].

2.1. Compression Method 1

The compression method 1 is lossless compression method used for compressing binary patterns. It is based on RLE lossless compression method with the difference that for encoding only one byte is used.

The principle of the method is based on encoding information using a single byte, where the upper 7th bit represents the number of repeated occurrences. The principle of the compression method 1 is depicted in Fig. 1 [4], [5].

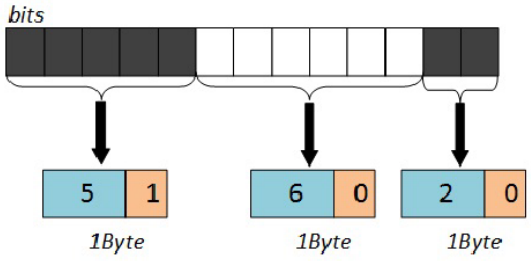

Fig. 1: The principle of compression method 1 [1].

2.2. Compression Method 2

Method number two is based on the same principle as first presented compression method except that the compression method 2 contains the start byte and the sequence of data bytes. The start byte contains information about the value of the initial element of the original pattern and data bytes contain information about the frequency of individual occurrences. (Fig. 2) [4], [5].

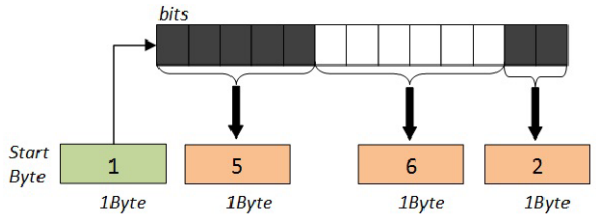

Fig. 2: The principle of compression method 2 [1].

3. Exploratory Data Analysis

The section describes a summary of used methods for exploratory analysis and definition of hypotheses.

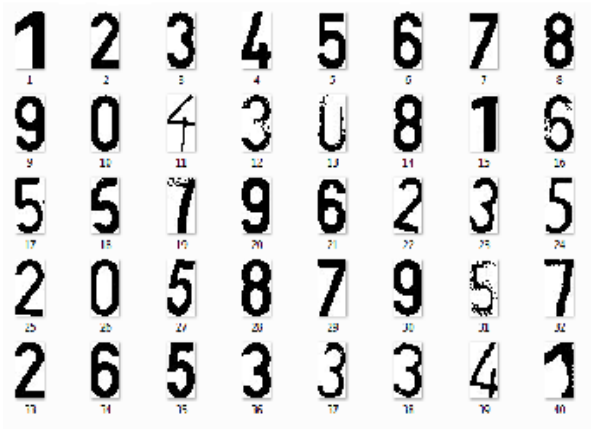

Fig. 3: The used patterns.

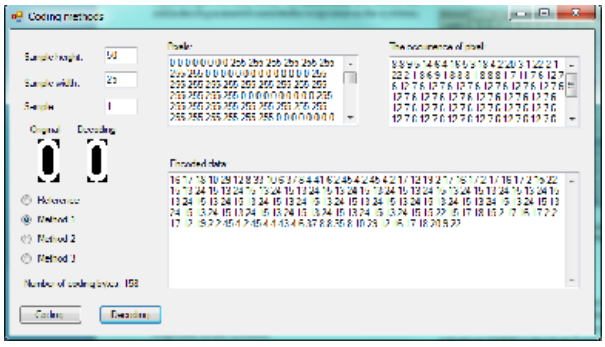

Fig. 4: User interface of the test application.

Samples represent the number of bytes required to encode each pattern. Digital representation of alphanumeric characters is depicted in Fig. 3.

Measurement was realized by specially developed user's application (Fig. 4). This application was developed in C#. Intuitive user interface allows you to set the basic parameters and to display the encoded data stream. The numbers of encoded bytes constituting the sampling, the individual patterns were then recorded in the table shown below.

The samples are formed by the numbers of encoded bytes, for every pattern the numbers of bytes are written in Tab. 1.

The statistical evaluation was realized using statistic tool SPSS version 18. Basic statistical parameters for both compression methods (Tab. 2) allow us easy comparison, and also the box chart is depicted in Fig. 5.

4. Hypotheses Testing

In the section the testing normality, homoscedasticity and assessment of statistical significance of the observed difference are shown.

Tab. 1: The numbers of encoded bytes for described compression methods for image digits patterns storing.

no.	comp. method 1	comp. method 2
1	109	97
2	141	101
3	154	119
4	148	101
5	153	115
6	158	135
7	132	89
8	174	129
9	150	123
10	164	119
11	154	109
12	207	169
13	243	201
14	182	147
15	148	103
16	213	183
17	145	115
18	154	123
19	200	157
20	182	149
21	165	143
22	152	115
23	158	127
24	158	109
25	150	113
26	178	151
27	174	137
28	165	119
29	135	89
30	179	149
31	289	251
32	172	127
33	133	92
34	152	121
35	166	125
36	151	119
37	172	129
38	153	123
39	168	129
40	128	107

Tab. 2: Basic statistical parameters.

	comp. method 1	comp. method 2
Count	40	40
Average	165,2	129,0
Median	158	123
Variance	995	970
Standart dev.	31,5	31,1
Minimum	109	89
Maximum	289	251
Range	180	162

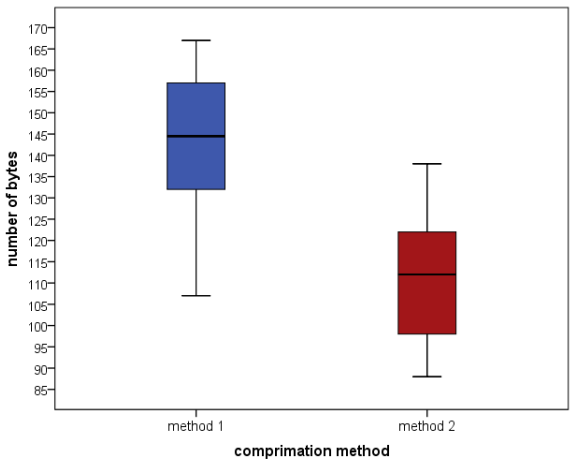

Fig. 5: Box charts for both compression methods.

4.1. Assessment of Normality

For assessment of normality the null (H0) and alternative hypothesis (HA) are defined

H0: Data comes from a normal distribution,

HA: Data do not come from a normal distribution.

$\chi 2$ test of goodness could not have been used, because of the low number of observations identified in the samples [7]. For assessment of normality Kolmogorov - Smirnov test [8] was used Fig. 6, and it was found that for the compression method 1 p-value was determined as 0,185 and for the compression method 2 p-value was determined as 0,082. It means that the null hypothesis is not rejected at 0,05 significance level for both compression method, and it therefore can be said that both samples come from normal distribution. Kolmogorov-Smirnov test statistics can be seen on the chart, which compares the actual and theoretical distribution function in Fig. 6.

4.2. Assessment of Homoscedasticity

Next step is an assessment of equality of standard deviation of samples [7]. First, it is necessary to establish null (H0) and alternative hypothesis (HA) again:

$$H0 : s_1 = s_2, \tag{1}$$

$$HA : s_1 > s_2. \tag{2}$$

The p-value was determined as 0,468. It means that, the null hypothesis is not rejected at significance level 0,05 and standard deviations of the samples can be considered as equal.

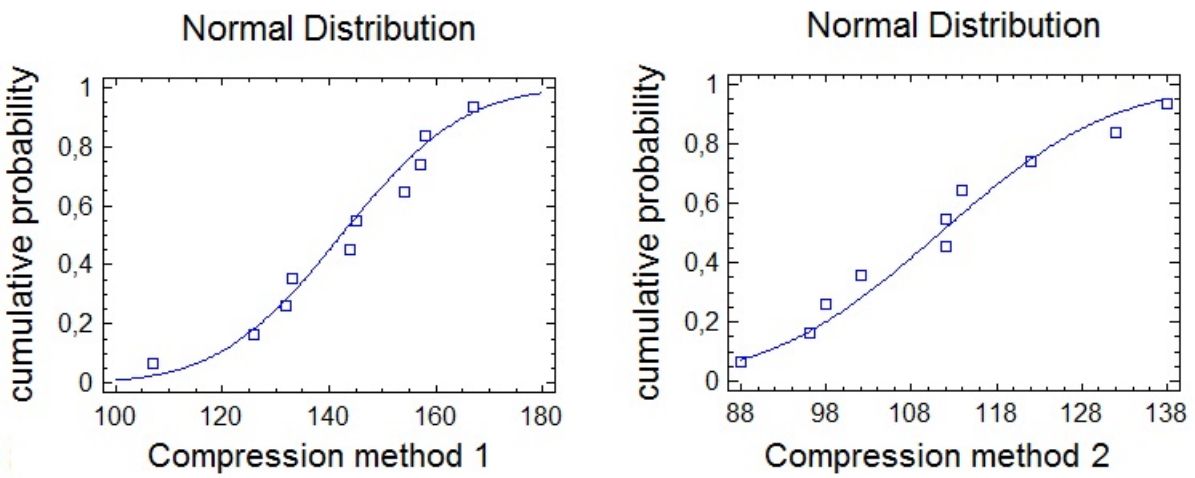

Fig. 6: Kolmogorov-Smirnov test statistics for compression compression method 1 and method 2.

4.3. Assessment of the Observed Difference - Student's T-test

Student's t-test is used in statistics to verify the statistical difference of means of the samples on selected significance level. Student's t-test can be used only for samples which come from a normal distribution and their standard deviation must be considered as equal on selected significance level [7], [8].

Determination of null and alternative hypotheses:

$$H0 : \bar{x_1} = \bar{x_2}, \tag{3}$$

$$HA : \bar{x_1} > \bar{x_2}. \tag{4}$$

P-value, as the result of the used Student's t-test, is $\ll 0{,}001$.

And it can be said that null hypothesis is rejected according to the result of Student's t-test on significance level 0,05 in favor of alternative hypothesis. So it means that the mean of sample of compression method 1 can be considered to be bigger than mean of sample of compression method 2.

5. Conclusion

For the recognition systems the selection of the appropriate compression algorithm is very important part, because of trend of low-memory system usage. In the paper the suitability of selected compression method was described using statistical analysis. Both presented samples of both used compression methods come from normal distribution and the standard deviation can also be considered as equal. According to it, the difference of means of samples was assessed using Student's t-test. Based on it, the determination of null and alternative hypotheses was defined and it can be said that compression method 2 is more efficient that compression method 1 (the mean of number of bytes is lower for compression method 2 in comparison with method 1) in data stream. So the presented compression method 2 can be preferable for systems with limited memory space.

Acknowledgment

This work was supported by SGS VSB–Technical University of Ostrava Grant No. SP2013/116, Project TACR TA01010632 and Grand–aided student R. Hercik, Municipality of Ostrava, Czech Republic.

References

[1] MACHACEK, Z., R. SLABY, R. HERCIK and J. KOZIOREK. Advanced System for Consumption Meters with Recognition of Video Camera Signal. *Electronics and electrical engineering*. 2012, vol. 18, no. 10, pp. 57–60. ISSN 1392-1215. DOI: 10.5755/J01.EEE.18.10.3062.

[2] LEIMER, J. Design factors in the development of an optical character recognition machine. *IEEE Transactions on Information Theory*. 1962, vol. 8, iss. 2, pp. 167–171. ISSN 0096-1000. DOI: 10.1109/TIT.1962.1057696.

[3] GASHNIKOV, M., N. GLUMOV and V. SERGEYEV. Compression method for real-time systems of remote sensing. In: *Proceedings 15th International Conference on Pattern Recognition*. Barcelona: IEEE, 2000, pp. 228–231. DOI: 10.1109/ICPR.2000.903527.

[4] SALOMON, David. *A concise introduction to data compression*. London: Springer, 2008. ISBN 978-1-84800-071-1.

[5] SALOMON, David. *A guide to data compression methods*. New York: Springer, 2002. ISBN 03-879-5260-8.

[6] MYATT, Glenn J. *Making sense of data: a practical guide to exploratory data analysis and data mining*. Hoboken: Wiley-Interscience, 2007. ISBN 978-047-0074-718.

[7] LEHMANN, E. L. and P. JOSEPH. *Testing statistical hypotheses*. 3rd ed. New York: Springer, 2010. ISBN 14-419-3178-3.

[8] SHI, Ning-Zhong and J. TAO. *Statistical hypothesis testing: theory and methods*. Singapore: World Scientific Publishing, 2008. ISBN 98-128-1436-1.

About Authors

Roman SLABY was born in Prerov, in 2006 he started his studies at the VSB–Technical University of Ostrava, he graduated in 2011 in Measurement and Control Engineering. Nowadays he continues to study in Technical Cybernetics. His areas of interest is image recognition and programming in C#.

Jana NOWAKOVA was born in 1987 in Trinec. She received her MSc. in Measurement and Control from VSB–Technical University of Ostrava, Faculty of Electrical Engineering and Computer Science in 2012. Nowadays she continues her studies in Technical Cybernetics at the same faculty. She is interested in addition to fuzzy modeling, also in statistical data processing in cooperation with University Hospital Ostrava.

Radim HERCIK was born in 1987 in Ostrava. He is a Ph.D. student and studies Technical Cybernetics on VSB–Technical University of Ostrava, Faculty of Electrical Engineering and Computer Science. His areas of interest are the image processing and signal analysis with focus on industrial use.