

Improving and Implementing the Pattern Matching Technique for Compression of Farsi/Arabic and English Bi-Level Printed Textual Images Using Fuzzy Sets and Improved Correlation Coefficient

Seyed Ali HASANI*, Hadi GRAILU*, Mojtaba LOTFIZAD*
and Hosein SAHOOLIZADE**

* *Tarbiat Modares University of Technologie, Electronic Department, Tehran, Iran*

syedalihhasani@yahoo.com

lotfizad@modares.ac.ir

grailu@modares.ac.ir

** *Young Researchers club of Arak Islamic Azad University, Arak branch, Iran*

hosein_sahooli@yahoo.com

Abstract: Pattern matching is one of the most widely used methods for compression of bi-level printed textual images. The proposed PM-based lossy/lossless compression method of Farsi/Arabic and English textual images, works on the basis of improving the template matching technique as well as handling the residual patterns in the lossy compression. To this end, patterns are first compared using the proposed template matching technique, by improving the trend of calculating the correlation coefficient, and then, dividing them into fuzzy sets. Despite English script, in the Farsi and Arabic, there are numerous compounds of letters to create words. The proposed method has an efficient application in English script. But using fuzzy sets has increased its application in the Farsi/Arabic script well. Two prominent advantages of the proposed compression method to existing ones include higher compression ratio, and employing the human visual system in order to preserve the image quality as much as possible. The experimental results show the higher compression efficiency of the proposed method relative to the best existing ones such as the JBIG2 standard.

Key words: pattern matching, template matching, bi-Level textual image compression, fuzzy classification

INTRODUCTION

Today, many methods are used for the compression of natural images such as vector equalization [1], transform-based method [2] and fractals [3]. But, textual images have their redundancy at the symbol level instead of the pixel level. Therefore, the mentioned methods do not have enough efficiency for compression of such images. Another reason for the inefficiency of the mentioned compression methods for textual images is their low pass behavior. Methods such as JPEG and JPEG2000 smooth the high local variations of the image and act as a low pass filter. In bi-level textual images, contrary to the natural ones, almost all the information exist at the edges; thus, the mentioned low pass filter highly degrades the compressed image quality. Most of existing textual image compression methods are based on pattern

matching (PM) technique because this method can eliminate or decrease the redundancy at the pixel level. These methods are classified into two categories of lossy and lossless. The idea of using PM was first introduced in [4]. This method is lossy because does not code the residual patterns. Also in this method, no compression method is presented for prototypes. The combined symbol matching (CSM) method [5] was proposed for the improvement of the previous method in which for the compression of the residual image, the two dimensional run length (RL) coding have been used, similar to the G3 and G4 standards.

Then in the reference [6] a method was proposed which used a preloaded library by alphabet letters in several fonts, to improve the compression ratio. Also weighted AND-NOT method or WAN [7] was proposed to improve CSM. One of its author announced that under quantization noise, the

efficiency of CSM, PMS and WAN decrease. So, he proposed the combined size-independent strategy (CSIS) that operates rather independently from patterns size. Also in [9], a multi-stage method based on PM for bi-level textual image compression has been proposed.

The lossless compression methods based on PM have been modified and used in some standards such as G3, G4, JBIG1 and JBIG2. These methods use PM for the compression of the textual parts of image, but different in coding method and compression of graphic parts. JBIG2 uses the CSIS to form the library and the arithmetic coding to code the prototypes, indices and relative positions. Figure 1 shows the block diagram of basic PM method.

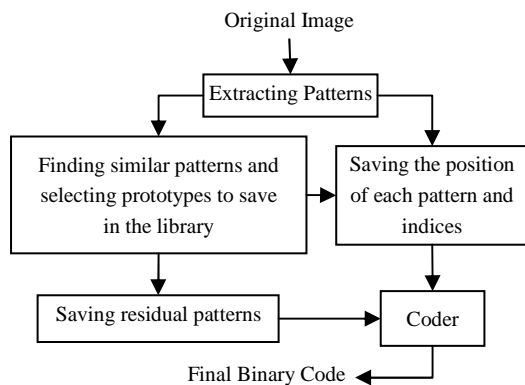


Figure 1. The block diagram of the basic PM method

Residual patterns contain differences between each pattern with the corresponding prototype, and are considered for better decompression.

The remainder of the paper is organized as follows: Section 1 explains the purposed method. Section 2 shows the experimental results of the proposed compression method and section 3 concludes this paper with a discussion on further improvement of the proposed compression method. Also the fourth section contains the references.

1. The Presented Method

The proposed compression algorithm uses the improved 2D correlation coefficient for estimation the similarity of each two patterns. After comparing the patterns two by two, different groups of patterns based on fuzzy sets are organized. To approach it, patterns are sorted in descending order by their area and each bigger pattern is compared with the smaller patterns. Also the arithmetic coding is used for coding the indices, relative positions, prototype library and residual patterns. Figure 2 shows the proposed compression method in this paper.

As shown in Figure 2, in the first stage, the presented method extracts and stors all existant patterns in the text.

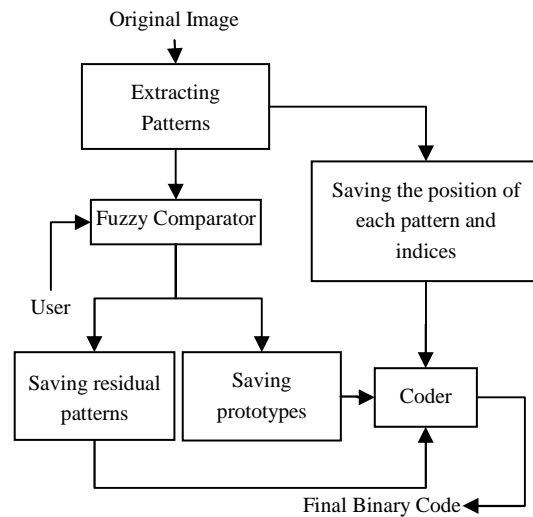


Figure 2. The block diagram of presented improved PM method

Figure 3 shows an image and its patterns for instance.

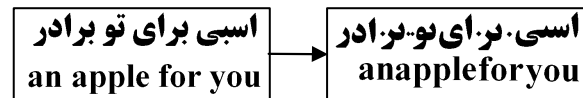


Figure 3. An instant of extracting patterns of an image

At the next step all, patterns are compared with each other by improved correlation coefficient and fuzzy sets to reduces the size of library. Figure 4 shows a corrected library and residual patterns resulted by the proposed algorithm.

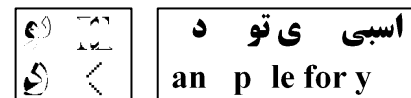


Figure 4. The prototype library of the image in Figure3 (Right) and its residual patters (Left)

To implement the PM technique, the image is scanned from up to down and left to right and the patterns are extracted. The upper left point of the bounding box of each pattern is considered as it's position.

1.1. The proposed comparison method using 2D correlation coefficient based on gravity center

As it was pointed out, so far, the correlation coefficient method has not been employed to compare the patterns. On the other hand, conventional pattern comparison methods contain long calculation and increase compression time. JBIG2 standard uses four continues local shifts and calculation of XOR, WXOR or WAN in each time to compare two patterns. This takes a long time and enlarges calculation. Also local shifts caused by undesirable noise on the edge of patterns, decrease the accuracy of comparison. Figure5 shows the local shift caused by a noise on the under edge.

Absolutely elimination of the effects of undesirable noise on the edges of patterns, increase the accuracy of comparison process.

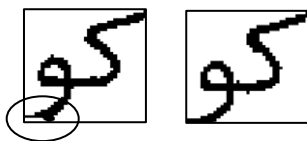


Figure 5. The resulted shift that caused by noise on the edge

The presented algorithm uses the gravity center to improve calculation of correlation coefficient. The main advantage of this method comparing with methods that use geometric center is elimination of the effects of noise on the edge. Because even just some affected pixels by noise cause much movement in geometric center. But whereas averaging is used to calculation of gravity center, affected pixels don't cause considerable movement in gravity center. The following steps at below describe improvement of 2D correlation coefficient that presented in this paper to compare two patterns, P₁ and P₂:

- A) Estimation of gravity center of patterns P₁ and P₂: This could be achieved by several methods like using index of black pixels:

$$P_0(x^0, y^0) = \{P_0 \in P(x, y) | P_0(x^0, y^0) = 1\}$$

$$X_c = \frac{\sum_{i=1}^A x_i^0}{A}, \quad Y_c = \frac{\sum_{i=1}^A y_i^0}{A} \quad (1)$$

A = Number of black pixels.

Figure 6 shows three examples of pattern and their gravity centers.



Figure 6. The gravity center of patterns

- B) Translation of P₁ and P₂ to bigger and same size blank patterns R₁ and R₂: This translation must be such that the gravity centers of smaller patterns locate in the middle of bigger patterns (R₁ , R₂). Figure 7 shows this process.

- C) Calculation 2D correlation coefficient for R₁ and R₂ :

$$R_1 = f_{A \times B}(x, y), R_2 = g_{C \times D}(x, y) \quad (2)$$

$$M \geq A + C - 1, x = 0, 1, 2, 3, \dots, M - 1$$

$$N \geq B + D - 1, y = 0, 1, 2, 3, \dots, N - 1$$

$$Correlation = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} [f^*(m, n)g(x+m, y+n)] \quad (3)$$

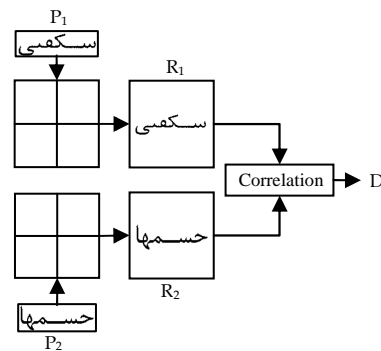


Figure 7. The improved correlation coefficient method using gravity center

Table 1 shows the similarity percentage of two patterns P₁ and P₂. As it is observed, this method can effectively determine the similarity among of patterns and eliminate the effects of local shifts as a result of noise.

P1	P2	D
سگما	بحسب	%7.96
نکمل	مسکلا	%22
مکس	محسو	%53.57
گر	گر	%85.38
بحسب	بحسب	%95

Table 1. Several Patterns and the magnitude of their similarity, that are calculated by the presented method

1.2. Using fuzzy sets

In all existing PM methods, two patterns are either similar or not similar. This paper considers the human visual criterion to improve the compression ratio and quality by employing fuzzy sets to determine the similarity of patterns.

For instance a person has classified patters by his visual comparison from the viewpoint of similarity rate to three groups:

- Similar patterns
- Semi-Similar patterns
- Not similar patterns

For the abovementioned three groups, the membership function of similarity could be considered as shown in Figure 8. According to the proposed membership function, each two patterns are a member of one of these three fuzzy sets: similar, semi-similar and not similar.

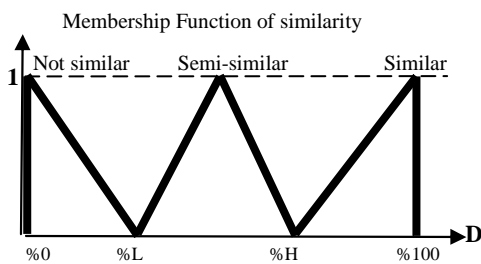


Figure 8. The membership function of similarity

So prototype library is organized as following:

- Saving one pattern from each two similar patterns as a prototype in the library.
- Saving one pattern from each two semi-similar patterns as a prototype and saving their residual patter.
- Saving each two not similar patterns separately in the library as a prototype.

Figure 9 shows organization of library and residual patterns by fuzzy sets.

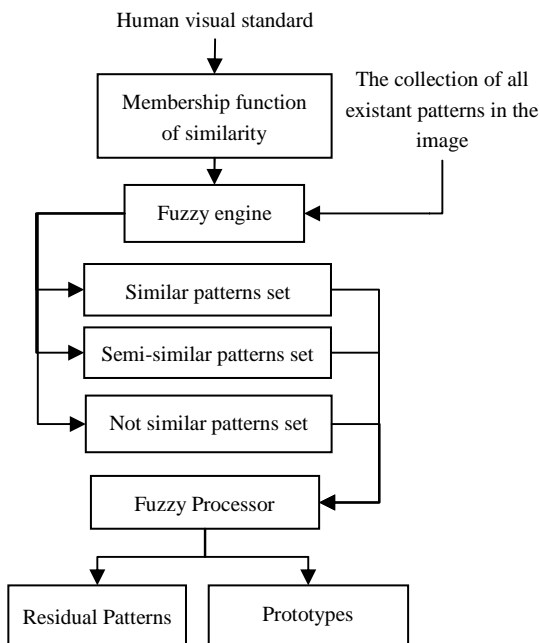


Figure 9. The organization of library and residual patterns by fuzzy engine

(4) Describes the function of system for comparison of two patterns A and B. As distinguished from (4) the compression will be lossless, if $L=H=100$.

$$2D - Corr(A, B) = D_{A,B}$$

$$\Rightarrow \begin{cases} \text{if } D_{A,B} \in [0, L] \Rightarrow \text{Not Similar} \\ \text{if } D_{A,B} \in [L, H] \Rightarrow \text{Semi-Similar} \\ \text{if } D_{A,B} \in (H, 100] \Rightarrow \text{Similar} \end{cases} \quad (4)$$

1.3. Improving the storage of residual patterns

Table 2 shows instances of semi-similar pairs P_1 and P_2 and the correspondent residual pattern that resulted during a real compression process.

P_1	P_2	R

Table 2. Members of semi-similar fuzzy set (P_1, P_2) and their residual pattern (R)

In many cases, residual patterns are similar to each other. In conventional methods and also in JBIG2 standard saving similar residual patterns decreases ratio and efficiency of compression. The presented algorithm in this paper compares residual patterns and eliminates similar patterns to achieve better efficiency. Experimental results show more than %5 improvement in compression ratio by employing this technique.

1.4. Presented new symbols for arithmetic coding

The presented algorithm in this paper uses a new context-based arithmetic coding that is relatively desirable and faster than the method in JBIG2. Because against used method in JBIG2, employed contexts in the new method don't overlap with each other and convert each four pixels to a symbol. So there are $2^4=16$ different symbols. Figure 10 (a) shows the resulted symbol collection.

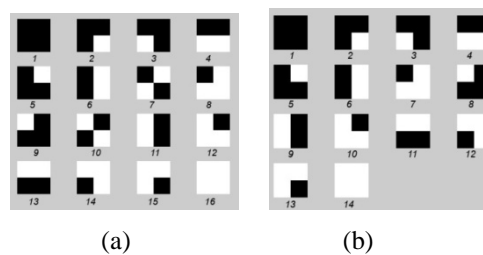


Figure 10. The new presented symbols

Using these symbols to convert patterns has a considerable result; symbols 7 and 10 in Figure 10 (a) appear rarely. Table 3 shows the probability of each symbol that are resulted from a compression process of 20 Farsi/Arabic textual images.

Symbol	1	2	3	4
Probability	%19.85	%1.2	%1.3	%1.4
Symbol	5	6	7	8
Probability	%1.27	%1.7	$\%2 \times 10^{-8}$	%0.99
Symbol	9	10	11	12
Probability	%2.02	$\%2.1 \times 10^{-8}$	%4.41	%0.847
Symbol	13	14	15	16
Probability	%4.06	%0.825	%1.197	%58.91

Table3. The probability of each symbol in Figure10 (a)

As shown in table 3, probability of symbols 7 and 10 are so weak. So, these two symbols could be ignored in a lossy compression. This connivance improves compression ratio and decreases compression time. Thus symbols decrease to 14 symbols as shown in Figure 10 (b). In continue we use these symbols to code all patterns.

2. Experimental Results

To experiment the presented method, a Farsi/Arabic and English image collection selected that contains 350 images with relatively extended spatial resolution (208,300 and 400 dpi). The selected images were not destroyed by deflection, rapture and adhesion. In the performed implementation, in order to increase the speed of algorithm, only the patterns with less than %15 difference in width and height have been compared. Also L and H selected respectively in the range of 70 to 75 and 90 to 95 dependent on image resolution. The selection of these ranges has a considerable effect on the final output. Figure 11 shows the division of compression ratio (bit per pixel) of JBIG2 standard to compression ratio of the presented method in this paper. This results show the efficiency of the presented method is more than JBIG2 standard.

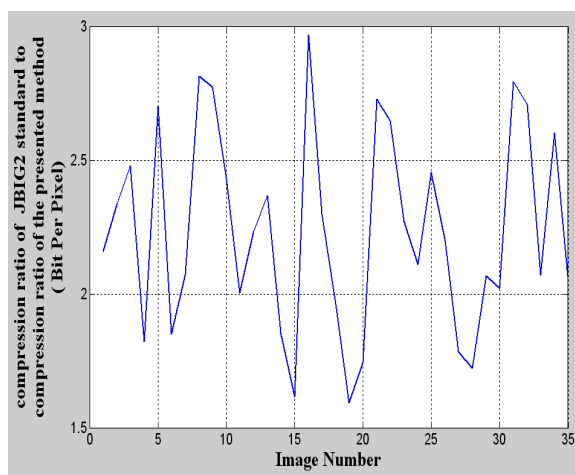


Figure 11 . The division of compression ratio (bit per pixel) of JBIG2 standard to compression ratio of the presented method in this paper

3. Conclusion

This article presents a PM-based method for compression of printed Farsi/Arabic and English printed textual images with relatively extended spatial resolution and simple structure. In the existing PM-based methods, the fuzzy sets have not been used. Whereas applying fuzzy sets in this paper to classify the patterns according to human visual comparison, causes improvement in quality and ratio of compression.

Experimental results show applying proposed method for long texts cause better compression ratio. Because more patterns have correspondent prototype in similar fuzzy set library. Also higher resolution

causes better compression. The presented method in this paper has compared with one of the most conventional standards and referring to experimental results, the presented method always gains better compression. The presented method is more proper for images with same size and font type. Invreasing the diversity of fonts ans sizes decreases the compression ration. But in the worst conditions, it will not be less than reference method.

Improvement of the presented method for images with different fonts and sizes is decided for continue. Also increasing the speed of compression process is considered as future works.

REFERENCES

- [1] A. Gersho and R. Gray, "Vector quantization and signal compression", Norwell, MA: Kluwer, 1992.
- [2] D. Solomon, "Data Compression, The Complete Reference," Fourth Edition, Springer-Verlag, London, 2007.
- [3] Y. Fisher *Editor*, "Fractal Image Compression," Springer-Verlag, New York, 1995.
- [4] R. N. Ascher and G. Nagy, "A Means for Achieving a High Degree of Compaction on Scan-Digitized Printed Text," *IEEE Trans. Comput.*, vol. 23, pp.1174-1179,1974.
- [5] W. K. Pratt, P. J. Capitant, W. H. Chen, E. R. Hamilton, and R. H. Wallis, "Combined Symbol Matching Facsimile Data Compression System," *Proc. IEEE*, vol. 68, no. 7, pp. 789-796,1980.
- [6] N.F. Brickman and W. S. Rosenbaum, "Word Autocorrelation Redundancy Match (WARM) technology," *IBM J. Res. Devel.*,vol. 26, pp. 681-686, 1982.
- [7] M. J. Holt and C. S. Xydeas, "Recent Developments in Image Data Compression for Digital Facsimile," *ICL Tech. J.*, pp. 123-146, May 1986.
- [8] M. J. Holt, "A Fast Binary Template Matching Algorithm for Document Image Data Compression," *Pattern Recognition*, J. Kittler Ed. Berlin, Germany, Springer Verlag, 230-239, 1988.
- [9] M. B. Carvalho, E. A. B. Silva and W. A. Finamore, "Multidimensional Signal Compression Using Multiscale Recurrent Patterns," *Signal Processing*, vol. 82, pp. 1559-1580, 2002.
- [10] I. M. Pu, "Fundamental Data Compression," Butterworth-Heinemann, 2006.
- [11] D. Salomon, " A Concise Introduction to Data Compression," *Springer-Verlag*, London, 2008.
- [12] K. Sayood *Editor*, "Lossless Compression Handbook," *Academic Press*, 2003.