سومین کنفرانس فناوری اطلاعات و دانش
۶ تا ۸ آذرماه ۱۳۸۶

Lossy/Lossless Farsi/Arabic Printed Binary Text Image Compression Based on Enhanced Pattern Matching

Hadi Grailu
Engineering Department,
Tarbiat Modares University,
Tehran, Iran
grailu@modares.ac.ir

Mojtaba Lotfizad
Engineering Department,
Tarbiat Modares University,
Tehran, Iran
lotfizad@modares.ac.ir

Hadi Sadoghi-Yazdi
Engineering Department,
Sabzevar Tarbiat Moalem
University of Sabzevar,
Sabzevar, Iran
sadoghi@sttu.ac.ir

Abstract —In this paper a lossless/lossy Enhanced Pattern Matching (EPM) method is proposed for Farsi/Arabic printed binary text image compression that uses the touching nature of characters in theses languages.
Using this method the library size will decrease and higher compression rate can be achieved. The PM method has been improved to include Farsi/Arabic texts. For comparison of patterns a generalized distance function is introduced which calculates the amount of similarity of two patterns. Also an encoding/transmission order is introduced for low bitrate progressive compression and low visual distortion.
The results show that the compression performance of proposed method is up to 6 times better than the conventional PM [7] for Farsi/Arabic text images.

I. INTRODUCTION

There are many methods for image compression such as vector quantization [1], transform-based [2] and fractal-based [3]. These methods remove or reduce the redundancy in pixel value level, but in text image the redundancy is in symbol level. So the above mentioned methods are not suitable for effective text image compression [4,5] and we should process the text images in symbol level.

There are few methods for binary text image compression and existing effective methods are mostly on the basis of Pattern matching (PM) or symbol matching idea.

PM first introduced in [6]. Its performance on printed text images is considerably better than general image compression methods such as JPEG.

Before employing PM some preprocessing such as skew correction could/should be done for higher performance.

The stages of PM method are as follows [6-7]:

a-) Extraction of all patterns. A pattern is a connected component which also is named blob or mark.

b-) Find similar patterns somehow and assign a prototype pattern for each similar patterns group. Generally a prototype is the most similar pattern to its corresponding group members.

Include all different prototypes in a library. In this stage, processing tasks such as noise removal can be done for example by removing all blobs having smaller size than a predefined threshold.

c-) For each pattern in original text image save index of its corresponding prototype in library and its relative position to previous processed pattern. Therefore at the end of this stage we have two sequences of integers.

d-) Compress and transmit the prototypes bitmaps and two integer sequences somehow.

e-) If we desire lossy compression the task could be finished here, but in the case of lossless compression the difference of each image pattern and corresponding prototype, named residual pattern, should also be compressed and transmitted.

The differences of PM-based methods lie in:

a-) How to compare the patterns.

b-) How to compress and encode the prototypes bitmaps, residual patterns and the integer sequences.

c-) How effectively the library is produced.

One of the most important parameters in determining the compression rate is the library size. Generally in lossy case, smaller the library size, larger the compression rate. But in lossless residual patterns compression has noticeable effect on total performance.