Application of Improved Lzc Algorithm in the Discrimination of Photo and Text

ChengJing Ye\textsuperscript{1,a}, Donghai Zeng\textsuperscript{2,b}

\textsuperscript{1}Guangdong Vocational Institute of Science and Technology, Guangzhou 510640, Guangdong, China
\textsuperscript{2}Guangdong Vocational Institute of Science and Technology, Guangzhou 510640, Guangdong, China
\textsuperscript{a}yecj2000@163.com, \textsuperscript{b}zdhwjy2000@163.com

Keywords: Lempel-Ziv complexity  Multi-scale  LZC  Photo and text

Abstract. For the commonly used image two value method, there is a lack of precision and accuracy of the processing image containing noise or uneven illumination and so on. A new method for two valued difference image of complexity at different scales. Divide the signal sequence into several areas and do multi-scale binaryzation process to the sequence. Whith no increasing the number of symbols, get more meticulous and more accuracy complexity by the traditional LZC algorithm. The experiments show that the new method has advantages of certain anti-noise abilities, can be regarded as the feature of photo and text. Therefore, We can apply the improved LZC algorithm to the distinction between photo and text.

Introduction

It is not an easy thing to extract effective information from nonlinear complex signals. Therefore, we have to develop various methods to extract characteristic quantity of signals, of which the LZC algorithm developed by Lempel and Ziv is a simple and fast algorithm to measure the complexity of signals. Today the LZC algorithm has been applied in various aspects to extract the features of complex signals. To automatically find the content we need from a lot of files and images is an example of extracting features of complex signals. And of course the LZC algorithm can be applied in this aspect. However, the coarse graining method of the LZC algorithm has flaws, and if it is directly used, it may cause serious consequences. In this article, the author will improve the LZC algorithm with respect to the flaws and apply the improved LZC algorithm to distinguish photo and text.

Lempel-Ziv Complexity Algorithm

In 1976, Lempel and Ziv proposed a complexity algorithm to measure the increase of sequence and the increasing speed of new models with time. We call this complexity “Lempel-Ziv Complexity”. Its algorithm is as follows:

(1) Reconstruction of time sequence: first, obtain the average value of the subject sequence; then reconstruct a symbol sequence with the subject sequence. Let the x which is bigger than the average value be 1, and the x smaller than the average value be ‘0’.

(2) Behind an existing substring in the (‘0’, ‘1’) sequence, add a character Q or a character string Q, and we obtain a new symbol sequence S. Let Sv be the sequence after removing the last character of the character string S. Then check whether Q belongs to the substring that already shows in Sv. If it does, adding Q calls “copy”. Then we can extend Q, which is to increase k, and then repeat the steps above until Q no longer shows in Sv. If Q does not show, it is called “insert”; when it is “insert”, place a “.” behind Q; then consider all characters before “.” a target sequence, and repeat the steps above until the sequence ends.

(3) After (2) is completed, we can obtain a character string sectioned by “.”. The total number of the sections is defined as “complexity” c(n);
According to Lempel and Ziv’s study, the \( c(n) \) of almost all \( s \in \{0, 1\} \) tends to a constant value:

\[
\lim_{n \to \infty} c(n) = b(n) = \frac{n}{\log_2 n}
\]

(1)

Therefore \( b(n) \) is an asymptotic behavior of the sequence, which can be used to normalize \( c(n) \), becoming relative “complexity measure”:

\[
C_{LZ} = \frac{c(n)}{b(n)} \quad 0 \leq C_{LZ} \leq 1
\]

(2)

From the description above we can see that: the LeMpel-Ziv Complexity (LZC) algorithm can reflect the degree of disorder of sequences. Therefore, the dynamics-based LZC algorithm is often used as a measurement. For example, (1) the LZC algorithm can be used to calculate the EEG (electroencephalogram) signals under different states and then distinguish different states according to relative difference, and based on that, we may conduct further studies on some mechanisms of the brain; (2) the LZC algorithm can be used to distinguish text and image, which will be described in detail below.

**Improvement of the LZC Algorithm**

According to its introduction, the features of the LZC algorithm are apparent. Merits: (1) the data needed to calculate Lempel-Ziv complexity are relatively short and the quantities of calculation are not large; and (2) without machine learning theory, relevant data features and laws can be extracted. However, the first step of calculating Lempel-Ziv complexity is to reconstruct the original time sequence into \((0,1)\) sequence. With such coarse graining processing method, it will cause loss of information and even change the dynamic characteristics of original signals completely. Therefore, the author has improved the LZC algorithm. The LZC algorithm can be improved from two aspects: (1) preprocessing of signals; and (2) improvement of the coarse graining method. This article mainly focuses on the research of the improvement of the coarse graining method.

With respect to the preprocessing of signals, use the wavelet decomposition algorithm to decompose the low frequency of signals, or use the wavelet packet decomposition algorithm to decompose the high frequency of signals, to filter singular value and remove noise. As the application of the wavelet decomposition algorithm and wavelet packet decomposition algorithm is mature, they will not be discussed further in the article.

**Traditional Coarse Graining Method**

The premise and key to the calculation of Lempel-Ziv complexity is to coarse-grain original signals, while the traditional coarse graining method of original signals is binaryzation. Suppose the known time sequence is \( \{x(i)|i=1,2,\ldots,n\} \). Reconstruct \( x(i) \) using the binaryzation process. Set

\[
x_{\text{ave}} = \left(\sum_{i=1}^{n} x(i) \right)/n
\]

(3)

where is the average value of the original sequence \( x(i) \). Let \( \{S(i)|i=1,2,\ldots,n\} \) denote the signal sequence after \( x(i) \) is reconstructed. When the element \( x(j) \) in the original sequence is \(<\), assign the symbol 0 to \( S(j) \); otherwise, assign 1 to \( S(j) \). When \( j = 1,2,\ldots,n \), a symbol sequence \( S(i) \) composed of 0 and 1 is thus established.

**Multi-scale Binaryzation Process**

The multi-scale binaryzation process is an improved method of the traditional coarse graining process. On the basis of the traditional coarse graining process, a more reliable sequence can be obtained with the multi-scale process. The multi-scale binaryzation process is described as follows: (1) Divide original signals into several sections, which is equivalent to a coarse graining process. The degree of the coarse graining is determined by the number of the sections. The more the number of the
sections is, the lower the degree of the coarse graining process is. In this article, original signals are equally divided into four sections, that is, calculate the average value of the whole signal sequence first, then divide it into two sections with the average value as the dividing line; and then obtain the average values of the two sections respectively and divide the two sections into another two sections based on their respective average value. Finally the signals are divided into four sections.

(2) If the value of the first point of the signal sequence is larger than the average value, indicate the point as 1; otherwise indicate it as 0. The value setting method of the first point is the same as that in the traditional binaryzation process.

(3) With respect to the second point of the signal sequence and the points behind, the binaryzation result is determined by comparing with the previous point. If it increases to another section, the indicate the binaryzation value of the point as 1; if it decreases to another section, indicate the binaryzation value of the point as 0; it is in the same section, the binaryzation value of the point is the same as that of the previous point.

![Figure 1. Comparison between two binaryzation methods.](image)

Take Figure 1 for example. Figure (a), The sequence is divided into two sections. The sequence (11110000) is obtained through reconstruction with the traditional binaryzation method, Figure (b), The sequence is divided into four sections. The sequence (11010110) is reconstructed with the multi-scale binaryzation method. According to the traditional binaryzation method, all of the four points A, B, C and D are larger than the average value, so they are indicated as 1; while E, F, G and H are smaller than the average value, so they are indicated as 0. According to the multi-scale binaryzation method, A for the first point, the value is larger than the average, so it is indicated as 1; B increases to another section, so it is indicated as 1; C decreases to another section, so it is indicated as 0; and so forth.

The traditional coarse graining method can only show whether the signal sequence is larger or smaller than the average value, while the multi-scale coarse graining method can not only show the change of the signal sequence more accurately with “0” and “1”, but also indicate a certain scale of rising trend and extension of the signal sequence with “1” and a certain scale of declining trend and extension of the signal sequence with “0”. By comparison we can see that the multi-scale coarse graining method describes a changing process in a certain extent and scope, and it is more elaborate and logic than the traditional coarse graining method.

### Discrimination Photo and Text with the LZC Algorithm

The LZC algorithm is used to analyze one dimensional sequence, while photo and text are two dimensional graphics. They seem to have no relation. However, photo and text can be scanned into one dimensional sequence by rows or lines, and then the LZC algorithm can be used to analyze them. Therefore, LZC complexity can also be used in two dimensional graphics. Based on past experiences, when calculating LZC complexity, the sequence with a length of 2000 is more suitable. Therefore, we have chosen the 50×50 square area in image and changed the image area into one dimensional sequence to analyze the difference in complexity of photo and text under difference scales. The experiment results show that the complexity of image and text increases with the decrease of the
computation scale. However, the increase of the complexity of photo is larger than that of text. Based on that difference, we can distinguish photo and text.

The specific method of using the LZC algorithm to distinguish photo and text is as follows: calculate the complexity of photo under four scales S1, S2, S3 and S4 (from large to small) and obtain LZ1, LZ2, LZ3 and LZ4; obtain the slope K that indicates the change of complexity with the change of scale using the least square method, and distinguish text and photo based on the difference in k. In specific calculation, regarding the photo with a relatively small size, choose 50×50 image area only; regarding the photo with a relatively large size, choose several areas according to the photo size and take the average value of the k values of the several areas as the feature of the image.

In the article, 50 photo and 50 texts have been used to do the experiment. Of the images, some were downloaded from the Internet and some were taken by digital camera, and the smallest image size is 50×50 while the largest is 874×585. All of the photo and texts are 8-digit grayscale images converted by scanning or being taken pictures of by digital camera, As shown in Figure 2, part of the experimental sample map. The red square area in the graph is a random selection of image complexity calculation area.
Table 1 shows the number of k values of the texts and photo in different sections.

<table>
<thead>
<tr>
<th></th>
<th>photo</th>
<th>texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0.0,0.02)</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>(0.02,0.04)</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>(0.04,0.06)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>(0.06,0.08)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(0.08,0.10)</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>(0.10,0.12)</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>(0.12,0.14)</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>(0.14,0.16)</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>(0.16,0.18)</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>(0.18,0.20)</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>(0.20,0.22)</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

The experiment results show that the k value of 98% (49) of the photo is larger than or equal to 0.08, and that of 86% (43) is larger than or equal to 0.12. The k value of 96% (48) of the texts is smaller than 0.08, and that of 88% (44) is smaller than 0.06. From the experiment results we can see that the method proposed in this article can be used to distinguish photo and text, and has the following characteristics:

For text, if the picture is serious noise, the computational complexity will be affected, so before the computational complexity of image denoising the appropriate.

For photo, if the photo in the background gray value is single, will also affect the computational complexity, as the scale decreases and little change.

Summary

Traditional LZC algorithm is a classic image of the two values of the algorithm, But it is the problem of the fine degree difference and the accuracy of the input image with noise or uneven illumination. So this article is aimed at these problems, the author improved it using the multi-scale binaryzation algorithm to make it more elaborate and logic; then the author discrimination photo and text with the improved LZC algorithm. The results show that the accuracy of the discrimination is very high (over 95%), There is no need to increase the number of symbols when using the method. Moreover, the method has higher accuracy and can be used to distinguish photo and text and expands the application scope of the LZC algorithm. Later we will conduct further studies on the following aspects: (1) how to use the method to discrimination image and photo, and image and text; (2) take the complexity method under different scales as a feature, combine the feature with other features of image, use existing classification tools to classify the image, compensate for the deficiencies of different classification features, improve classification accuracy and enhance its practicability.

References


