Special Collection Article

International Journal of Distributed Sensor Networks

International Journal of Distributed Sensor Networks
2018, Vol. 14(1)

DOI: 10.1177/1550147717750374
journals.sagepub.com/home/dsn
SAGE

A knowledge-embedded lossless image compressing method for high-throughput corrosion experiment

Peng Shi[1], Bin Li[1], Phyu Hnin Thike[1] and Lianhong Ding[2]

Abstract

High-throughput experiment refers to carry out a large number of tests and attain various characterizations in one experiment with highly integrated sample or facility, widely adopted in biology, medicine, and materials areas. Consequently, the storing and treating of data bring new challenges because of large amount of real-time data, especially high-resolution images. To improve the storing and treating efficiency of high-throughput image, a knowledge-embedded lossless image compressing method is proposed. Based on the similarity of a series of high-throughput images, it accomplishes the high compression ratio according to the difference between the target images and one reference image. Meanwhile, the knowledge extracted from the image, such as edge information and differences from the reference image, is recorded into the compressed file. The key steps include similarity comparison, edge detection, coordinate transformation, and dictionary encoding. The method has been successfully applied into high-throughput corrosion experiment facility, a typical intelligent cyber-physical system. To evaluate the performance, corrosion metal, face, and flower images are compressed by our method and other lossless image compression methods. The results show that our method has fairly high compression ratio. Moreover, the embedded knowledge can be read directly from the compressed file to support further study.

Keywords

Coordinate transformation, edge detection, high-throughput experiment, knowledge embedded, intelligent cyber-physical system, lossless image compression

Date received: 31 May 2017; accepted: 23 November 2017

Handling Editor: Xiuzhen Cheng

Introduction

High-throughput experiment refers to carry out a large number of tests in parallel.[1] The history of high-throughput idea can be traced back to the work of Kennedy in 1965. At that time, he was tired of the repeated process of synthesis and characterization one by one in his new material screening work. Consequently, he put forward the idea of "multiple sample" (i.e. simultaneous synthesis and characterization of multiple specimens). Because of the limitation of technical conditions at that time, his idea did not get much attention in the field of materials, but have been promoted and applied in bioinformatics and pharmacology. With Xiang et al.'s[2] creatively realizing combinatorial material chip technology based on "combinatorial chemistry principle" in 1995, the method and idea of high-throughput test return to the material science area. Now the high-throughput

[1]National Center for Materials Service Safety, University of Science and Technology Beijing, Beijing, China
[2]School of Information, Beijing Wuzi University, Beijing, China

Corresponding author:
Lianhong Ding, School of Information, Beijing Wuzi University, No. 321 Fuhe Street, Tongzhou District, Beijng 101149, China.
Email: lhdingbwu@sina.com

technology is widely adopted in the field of biology, medicine, and materials.

Besides the integration of large number of tested objects into one sample, high-throughput experiment usually collects various characterization data during the experiment for quantitative comparison study. For example, material corrosion experiment collects surface morphology, chemical-electronic signal, and solution concentration during the corrosion process. To accomplish an automatic high-throughput experiment, a platform with data collecting, storing, and treating must be established, which is a typical cyber-physical system (CPS).[3] Some researchers have designed several kinds of experimental facility to accomplish automatic high-throughput material corrosion experiment.[4,5] However, they did not pay much attention to the storage and treatment of large amount of real-time data from high-throughput experiment. Since many specimens are being tested and monitored in parallel, one high-throughput experiment produces large amount of data. Consequently, the storage and treatment of data bring new challenges, especially for real-time recording high-resolution images.

In this article, we design a high-throughput image compression algorithm and accomplish the fast extraction of the key information from the compressed high-throughput image, which provides a feasible scheme for data management and knowledge retrieval in CPS. The rest of the article is arranged as follows. Section "Related works" introduces the related works of high-throughput experiment, image compression, and corrosion evaluation methods. Section "Knowledge-embedded lossless image compression method" describes the proposed knowledge-embedded image compression method. Section "Application in high-throughput corrosion testing" introduces the application of the proposed method in high-throughput corrosion experiment. In section "Experimental results and discussion," we compare our method with other compression methods by different kinds of images. Finally, section "Conclusion and future works" concludes our work and points out the future works.

Related works

High-throughput experiment and image

The idea of high-throughput has been applied in different fields, especially in scientific experiments. Although the objects and processes in the experiments are different, the basic objective is to improve the efficiency and decrease the cost by high-throughput way. Commonly used high-throughput experiments are as follows:

1. *High-throughput screening*. The system, based on molecular and cellular levels of experimental methods, uses microplate as experimental tool carrier and performs test procedures with automated operating systems.[6] It collects data with a sensitive and rapid testing instrument, analyzes experimental data with a computer, and detects tens of millions of specimens at the same time.
2. *High-throughput sequencing*. High-throughput sequencing technology also known as the next generation sequencing technology, which is able to sequence hundreds of thousands of DNA molecules in parallel.[7]
3. *High-throughput characterization of materials*. The material composition and structure are characterized by high-throughput method, which usually uses the electromagnetic spectroscopy method of X-ray, ultraviolet, and infrared.[8] The rate of the characterization is affected by the luminous flux density and the spatial resolution of the beam focusing.
4. *High-throughput material corrosion*. Material corrosion is a time cost procedure. High-throughput technology can obviously decrease the time cost. In high-throughput material corrosion experiment, corrosion degree of several materials in different environments can be characterized. The corrosion environment can be controlled by the combination of different materials, electrolyte concentration, surface roughness, and other factors.

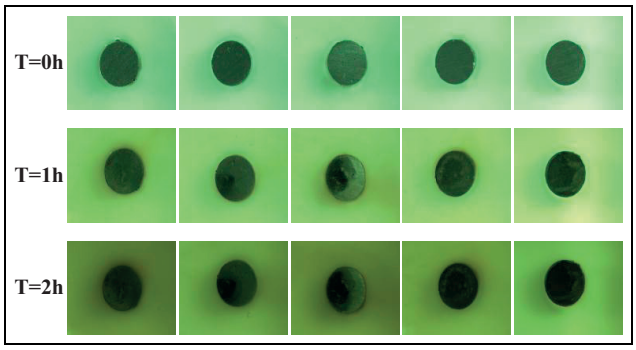


Figure 1. A group of high-throughput corrosion image.

Most high-throughput experiments require observing the surface morphology of the specimens, so the image of the specimens' surface must be monitored and recorded during the experiment, called high-throughput image.

Figure 1 shows a group of high-throughput image during one metal corrosion experiment. The first row is the initial morphology of the specimens. The middle and the last rows are the morphology after 1- and 2-h reactions, respectively. Since high-throughput experiment is usually designed with a series of similar parameters and samples for easily quantitative comparison

study, high-throughput images have similarity both among specimens (horizontal comparison of multiple specimens) and with time series (longitudinal self-comparison of one specimen).

For material corrosion study, the analysis of the surface morphology is important to evaluate the corrosion degree and investigate the corrosion mechanism. With the assistant of IT technology, the surface image can be automatically treated by program, and correct decision can be made.

Image compressing methods

Image compression is considered to deal with the images in order to decrease the space of storage. Depending on whether it can completely restore the original data, traditional image compression methods are divided into three categories, including lossless compression, lossy compression, and mixed compression. So far, many papers have improved the traditional methods. For example, Wu and Zheng[9] solved the blocking artifacts problem of JPEG algorithm. Stabno and Wrembel[10] improved the efficiency of Huffman coding with object-oriented refactoring. In order to better serve scientific research, lossless compression or targeted hybrid compression should be applied to compress high-throughput image. However, current lossy compression methods lack pertinence and discard important information for research. And, the compression ratio of lossless compression algorithm is not high. So, a new compression method is needed to compress the high-throughput image.

Corrosion evaluation based on image

During the study of material corrosion, the appearance of surface is important to evaluate the corrosion degree. The idea of image processing can be used in the grade evaluation of material corrosion. In 1981, Itzhak et al.[11] scanned the surface of AISI 304 material by digital scanner, soaked for 20 min in 10% $FeCl_3$ solution at 50°C. They also developed a computer program to count the corrosive pitting and calculated the corrosion rate according to the scanned images. EN Codaro analyzed the appearance of pitting of Al-Ti alloy and gave a quantitative description method for pitting. The method can be utilized to represent the evaluation procedure of corrosion.[12] Wang et al.[13] collected the morphology of sea water corrosion of carbon steel and established the relationship between the corrosion degree and the apparent color and edge by gray correlation method. Xu et al.[14] investigated the relationship between the apparent gray value of image and the depth of corrosive pitting in the specimen by fractal dimension method. They found out that the relationship was nearly linear and gave the equations.

Knowledge-embedded lossless image compression method

According to the demand of high-throughput experiment analysis, we propose a new compression method based on similarity of high-throughput images. The method uses the features of high similarity, both in color and contour, to achieve the complete preservation of image important research information and good image compression ratio. The other advantage is that the compression method records the knowledge of image analysis result, such as corrosion edges, difference of images, into the compressed file.

The principal idea of our method is to extract some key characteristics, like corrosion edge, from the target image and compress it according to its similarity with the reference image. The main process is shown in Figure 2. At first, one image is selected as a reference image from each group high-throughput images. Then, the key boundary points of the target image are extracted, according to which the target image is divided it into different regions. Then, each region is transformed based on the SIFT match with the reference image.[15] Finally, the processed pixel residuals and

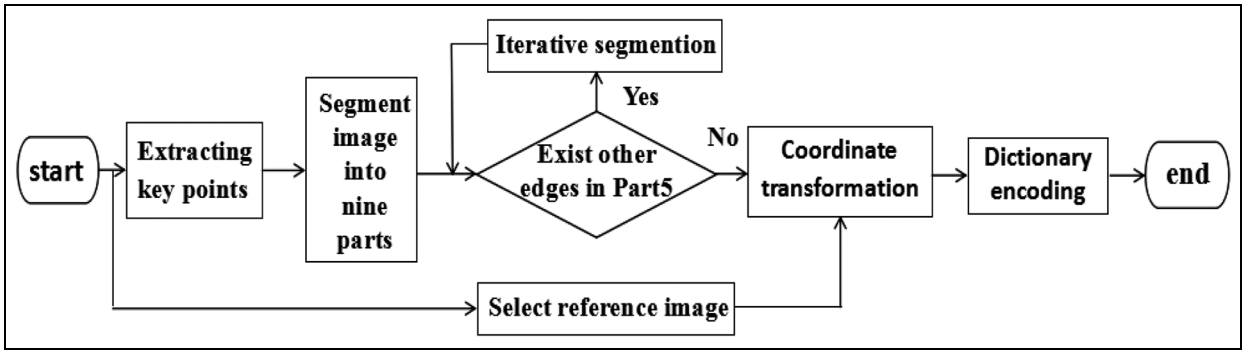


Figure 2. High-throughput image compression process.

the reference image are compressed by dictionary encoding.[16]

Extracting key points information on edge

Because the contour of adjacent high-throughput image is similar, we consider to finding the representative points on the edge of the image and applying them into the compression processing. These representative points are referred to as key points in this article.

In order to ensure the compression efficiency, we must choose an accurate and fast edge extraction method to find out the key points. Because the Canny algorithm has the characteristics of precision and strong anti-interference ability,[17] it is adopted to extract the contour points of the image. In the final step, the edge points selected by the method of double threshold connection are marked.[18]

Suppose there are a set of high-throughput images, denoted as I_0, I_1, I_2,..., I_n (ordered by the similarity of experimental parameters). I_{i-1} is conducted as the reference image of I_i ($0 < i < n + 1$). In image I_i, we first filter closed edges with the number of edge points greater than 100 (to avoid the influence by small areas) and select four points as the key points from them: the leftmost edge point $a_1(x_1, y_1)$, the top edge point $a_2(x_2, y_2)$, the rightmost edge point $a_3(x_3, y_3)$, and the lowest edge point $a_4(x_4, y_4)$.

Iterative image segment

After extracting the four key points of each image, we divide the image into nine regions and number in the sequence for these areas, as shown in Figure 3.

Formulae (1)–(3) are coordinate representation of region division, in which h and w are the width and height of the image, respectively

$$\text{Part } 1,4,7: 0 \le x < x_1, m \le y < n \quad (m = 0, y_2, y_4; n = y_2, y_4, h) \tag{1}$$

$$\text{Part } 2,5,8: x_1 \le x \le x_3, m \le y < n \quad (m = 0, y_2, y_4; n = y_2, y_4, h) \tag{2}$$

$$\text{Part } 3,6,9: x_3 < x < w, m \le y < n \quad (m = 0, y_2, y_4; n = y_2, y_4, h) \tag{3}$$

Because of the rules of key point extraction and region division, Part 5 contains all eligible closed edges.

In many cases, an image often has several edges. Nested loop segmentation is adopted to further divide the image until all the closed edges are concerned, shown in Figure 4. We construct a two-dimensional array by find contours method in OpenCV, where each closed edge is a one-dimensional (1D) array. The element of a 1D array is the edge point on the closed edge.

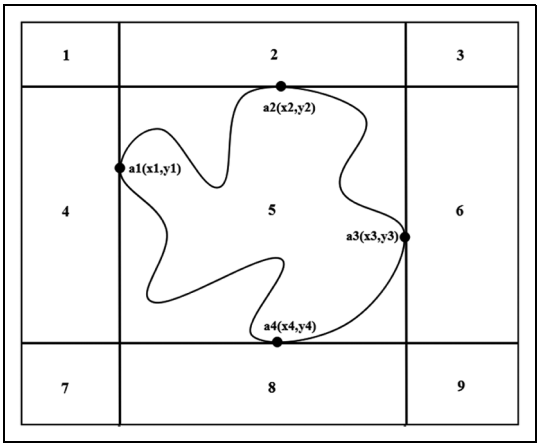


Figure 3. Segment image into nine parts by four key points.

When there are multiple edges in an image, iterative segmentation method first eliminates arrays with existing edge point on the edge of Part 5. And then, extract key points from this two-dimensional (2D) array and make regional division according to the above method if Part 5 have closed edges inside.

Coordinate transformation and pixel difference calculation

First, the Part i ($0 < i < 10 + 8t$, t is the number of iterations) and the reference image are matched with the SIFT feature points. SIFT is a robust and effective identifying targets method which is adaptable to different illumination and different positions. It proposes some essential features of an image. It can maintain invariance to rotation, scale scaling, brightness changes and a certain degree of stability to the change of view, affine transformation and noise. Typical applications of SIFT include object recognition, robot location and navigation, 3D modeling and video tracking. The SIFT method takes the feature point as the center and the neighborhood of 16×16 as the sampling window. The relative direction of the sampling points and the feature points is classified into eight directional histograms by Gauss weighting, and finally, 128 dimensional vectors are obtained. When the SIFT feature vectors of the two images are generated, the Euclidean distance of the feature vectors of the key points can be used as the similarity measure of the key points in the two images. A pair of feature points with minimum distance of feature points is found, and the difference between them is $(\Delta x_j, \Delta y_j)$. Set the two SIFT feature point vectors are M $(m_1, m_2, \ldots, m_{128})$ and N $(n_1, n_2, \ldots, n_{128})$.[19] Their distance is calculated by Formula (4)

$$D = \sum_{i=1}^{128} (m_i - n_i)^2 \tag{4}$$

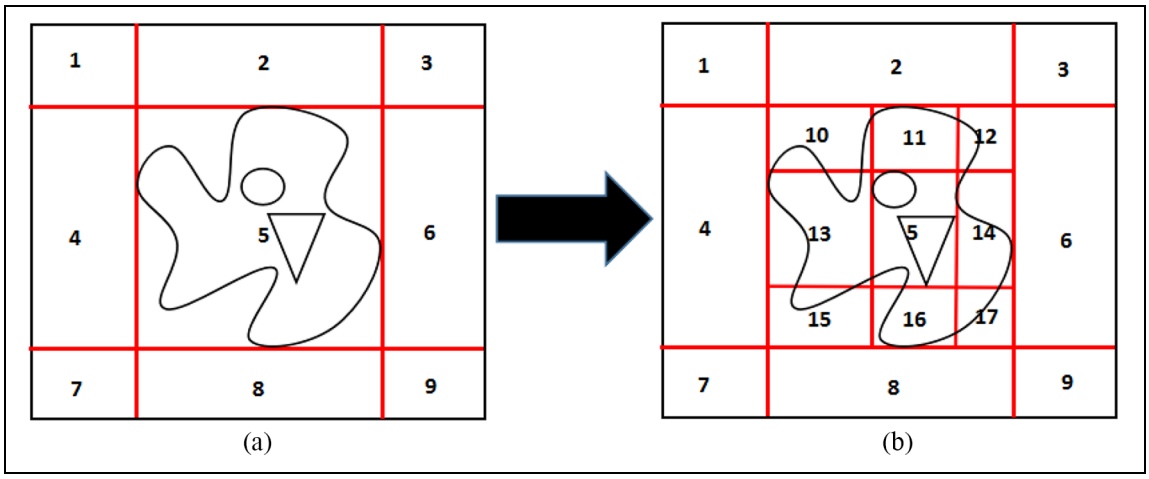


Figure 4. The nested high-throughput image segmentation method: (a) non-iterative segmentation and (b) iterative segmentation.

Set $I_i(x, y)$ $(0 \leq i \leq n-1)$ as the pixel function of image I_i. In this article, according to different regions, a coordinate difference transformation is defined to get the corresponding pixel difference. Denote the pixel difference function as $\Delta_i(x, y)$, K is the number of segmented regions, and the pixel difference of each region is calculated by Formula (5)

$$\text{Part } j : \Delta_i(x,y) = I_i(x,y) I_{i-1}\left(x + \Delta x_j, y + \Delta y_j\right) \quad (0 < j < K+1) \tag{5}$$

In order to further improve the compression ratio, this article uses the correlation of image interior pixels to further process the pixel difference and obtain the total pixel difference $D_i(x, y)$, shown as Formula (6)

$$D_i(x,y) = \Delta_i(x,y)(\Delta_i(x+1, y+1) + \Delta_i(x-1, y-1))/2 \tag{6}$$

Dictionary encoding

The last step of compression uses dictionary compression method to encode the total pixel difference $D_i(x, y)$. The dictionary encoding is adopted to further compress the storage space of an image. LZ77 algorithm is a representative of the dictionary compression and also known as sliding window compression algorithm. The core idea is to use the data structure of the data to compress the data. An example of LZ77 encoding is shown in Figure 5.

The main steps of LZ77 algorithm are as follows:

1. Set the encoding position as the start of the input stream.
2. Find the maximum matching string in the search area of the sliding window.
3. If the string is found, output (offset, matching length), the window sliding forward matching length; otherwise, output (0, 0, the first character of the area to be encoded), the window sliding forward a unit.
4. If the encoding area is not empty, return to step 2.

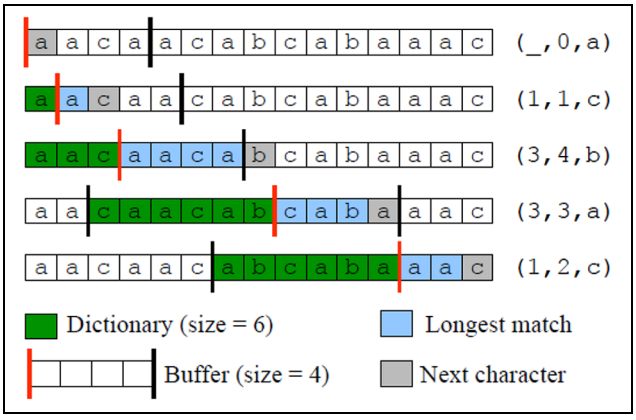


Figure 5. An example of LZ77 encoding.

LZ77 has a high compression ratio for the file with high repetition rate and large file size, especially when the repetition rate is very high. However, encoding repetitive continuous data will occupy more space encoding. It will affect the encoding efficiency. Thus, an improved lZ77 compression algorithm is used to do the last step of high-throughput image. It mainly applies the advantage of run length encoding to LZ77 algorithm, which effectively improve the encoding efficiency and compression ratio, shown in Figure 6.

At first, calculate the appearing time of each number in $D_i(x, y)$ and record the time, behind each number, denoted as P. Extract all the original numbers from P and put them into another sequence, denoted as P1. All the numbers of time are recorded into a new sequence, denoted as P2. Then, LZ77 compression algorithm is used to merge their dictionaries and encode P1 and P2,

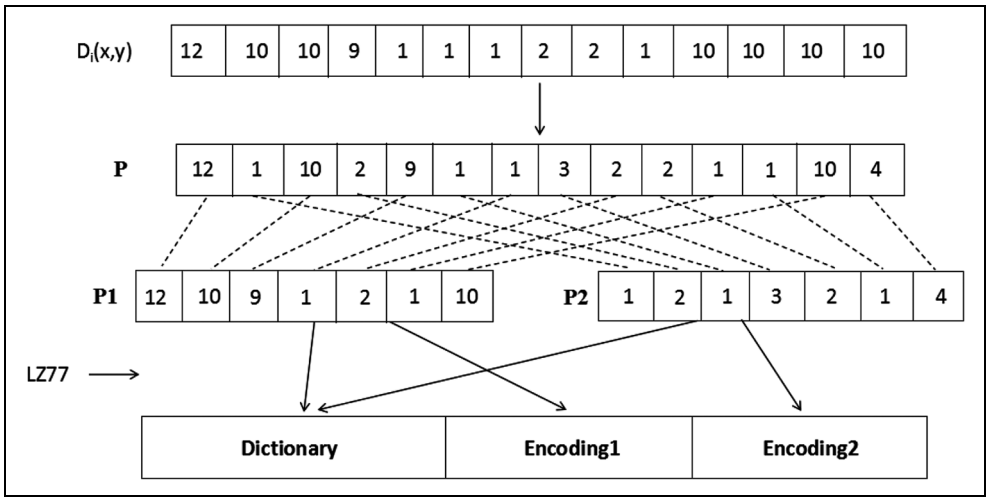


Figure 6. Improved LZ77 compression algorithm.

respectively.[20] Since LZ77 algorithm is lossless, $D_i(x, y)$ can be obtained by a reverse process.

Application in high-throughput corrosion testing

In order to verify the efficiency of high-throughput image compression algorithm, we design and implement the corrosion experiments. In the experiment, the high-throughput image is collected in real time by the industrial camera. Then, we use this algorithm to compress and decompress the high-throughput image, recording the compression ratio, compression time and decompression time of the image, and test reduction effect. The high-throughput corrosion experimental platform designed in this article is divided into two parts: the experimental device and the image acquisition device. The experimental device is mainly composed of a multiple-solution electrolysis pool and samples. The image acquisition device is composed of industrial camera, the aperture, and the light shield and is responsible for real-time recording of high-throughput sample images. The main structure is shown in Figure 7.

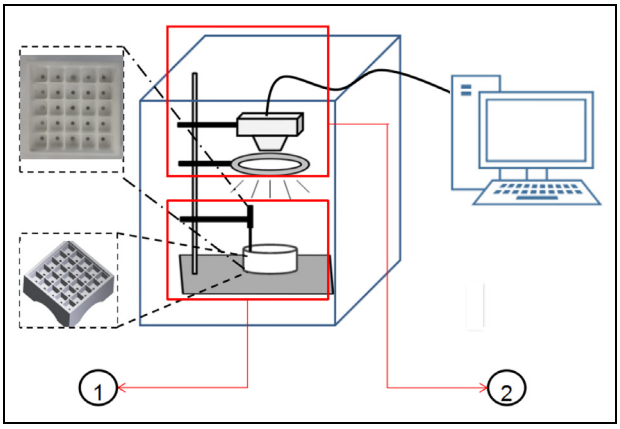


Figure 7. High-throughput corrosion experiment facility: ① experimental device and ② image acquisition device.

Table 1. The relationship between surface roughness and different types of sandpaper.

Sandpaper type (#)	200	400	800	1200	1500
Surface roughness (μm)	75.0	35.0	21.8	14.5	12.6

Experimental preparation

A multiple-solution electrolytic unit was designed and acted as the sample bearing platform in high-throughput corrosion experiment. Its appearance is shown in Figure 8. The multiple-solution electrolytic unit has five rows and five columns, consisting of 25-groove liquid pool which are parallel arranged. At the bottom of each liquid pool has a round hole with 5 mm diameter. The advantage of multiple-solution electrolytic unit is high flexibility. It can realize any combination of different metals and different solutions.

Because carbon steel has the characteristics of fast corrosion rate and obvious corrosion morphology, it is selected as the object of the basic corrosion characteristic study. This material was made into 25 bar samples, with 5 mm of diameter and 20 mm of length, whose intersecting surface was corrosion surface. Different solution concentrations are used between rows and rows, and different surface roughness are used between columns and columns. Parallel samples were polished, respectively, by 200#, 400#, 800#, 1200#, and 1500# sandpaper, so that distinguished the effect of surface roughness on material corrosion. The relationship between sandpaper model and surface roughness has been shown in Table 1.

According to 200#, 400#, 800#, 1200#, and 1500# sequence, the whole matrix was polished to the

Figure 8. A multiple-solution electrolytic unit.

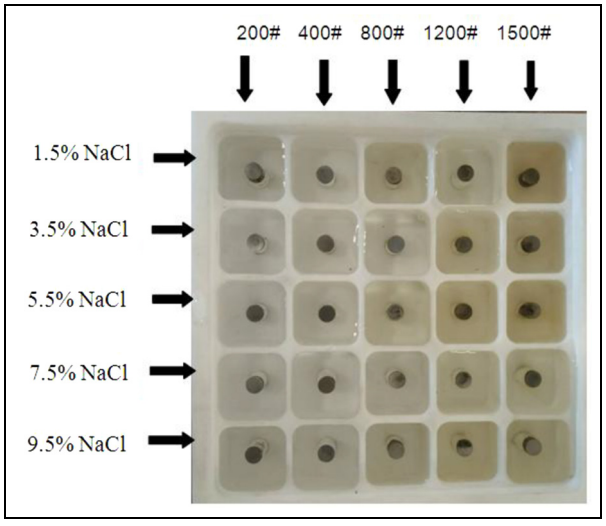


Figure 9. Fixed 25 specimens with different surface roughness in a multiple-solution electrolytic unit.

Table 2. Parameters of Daheng camera.

Name	Parameter
Model	MER-500-7UC
Interface	Mini USB 2.0
Resolution	2592 (H) × 1944 (V)
Frame rate	7 fps
Sensor	1/2.5″ CMOS
Pixel size	2.2 μm × 2.2 μm
Spectrum	Black-and-white/color

CMOS: complementary metal-oxide-semiconductor.

roughness degree of 1500# sandpaper first. Then, we selected the most edge of a column to protect, the remaining four columns were polished vertically by 1200# sandpaper. Then, we selected one column of four columns to protect and used 800# sandpaper to polish the remaining three columns. And so on, there would be 5 kinds of superficial roughness degree in the same matrix.

Different kinds of NaCl solution, with concentration of 1.5%, 3.5%, 5.5%, 7.5%, and 9.5%, were used in each line, respectively. The carbon steel rod wrapped by raw tape is inserted into the hole in the bottom of liquid pool, which ensures that the solution in the liquid pool does not leak.

Figure 9 shows the overall picture of samples, carbon steel in different concentration solutions was polished by different surface roughness of sandpaper, which made 25 specimens different with each other. It does reflect the idea of high throughput.

Image acquisition and processing

The image acquisition device was used to capture the variation characteristics of corrosion morphology with time on the surface of each sample. General image acquisition device can achieve real-time capture of all objects within the scope of the viewfinder.

"MER-500-7UC" camera from Daheng Graphics Company is adopted as image acquisition device, whose specific parameters are shown in Table 2. The camera is connected with computer by USB 2.0 port and provides programming interface to control its action. Its size is only 29 mm × 29 mm × 29 mm, and it is easy to be used in various environments.

For enough light on the specimens, we take light-emitting diode (LED) ring light as a light source. It can supply an even light for the measured object. To avoid the influence from the environmental light, a light shielding cover was used. The stable light condition makes the direct comparison of images at different stages possible. Light shield selection was relatively simple, as long as it can cover the whole system and block the environmental light in.

Image compression

High-throughput images collected by this experiment can be divided into different groups according to row, column, or different time of the same specimen. The adjacent high-throughput images in a group have high similarity, so that we can use the method designed in this article to compress them better.

We take high-throughput images collected from each line in a multiple-solution electrolysis cell as a group. At the same time, five groups of high-throughput images were collected and there are five high-throughput images in each group.

Each group images is compressed together with the lossless compression method described above. The related parameters of compression and decompression are shown in Table 3.

Corrosion evaluation

Material corrosion evaluation is an important part of material corrosion experiment. It is also the knowledge

Table 3. High-throughput image compression and decompression parameters.

Time (h)	Group number	Image number	Group size before compression (MB)	Group size after compression (MB)	Compression ratio	Compression time (s)	Decompression time (s)
T = 0	1	5	72.05	17.2	0.24	16.8	14.3
	2	5	72.05	12.8	0.18	18.6	17.3
	3	5	72.05	13.8	0.19	15.7	13.2
	4	5	72.05	18.6	0.26	17.9	17.2
	5	5	72.05	19.1	0.27	20.3	18.5
T = 1	1	5	72.05	13.4	0.19	14.9	14.3
	2	5	72.05	20.2	0.28	16.7	16.0
	3	5	72.05	13.2	0.18	15.9	16.4
	4	5	72.05	15.6	0.22	15.6	14.7
	5	5	72.05	14.4	0.20	18.9	16.5
T = 2	1	5	72.05	15.7	0.22	17.5	18.4
	2	5	72.05	17.3	0.24	15.7	13.6
	3	5	72.05	11.8	0.16	18.0	16.5
	4	5	72.05	13.2	0.18	18.7	17.6
	5	5	72.05	12.6	0.17	19.2	15.9
T = 3	1	5	72.05	13.5	0.19	17.6	18.6
	2	5	72.05	14.2	0.20	14.3	15.2
	3	5	72.05	15.1	0.21	17.4	17.8
	4	5	72.05	18.4	0.26	16.8	15.9
	5	5	72.05	14.5	0.20	18.0	18.4
T = 4	1	5	72.05	14.8	0.21	17.4	17.5
	2	5	72.05	13.2	0.18	16.0	18.4
	3	5	72.05	16.3	0.23	16.8	15.5
	4	5	72.05	15.0	0.21	20.7	17.5
	5	5	72.05	12.9	0.18	19.3	20.2

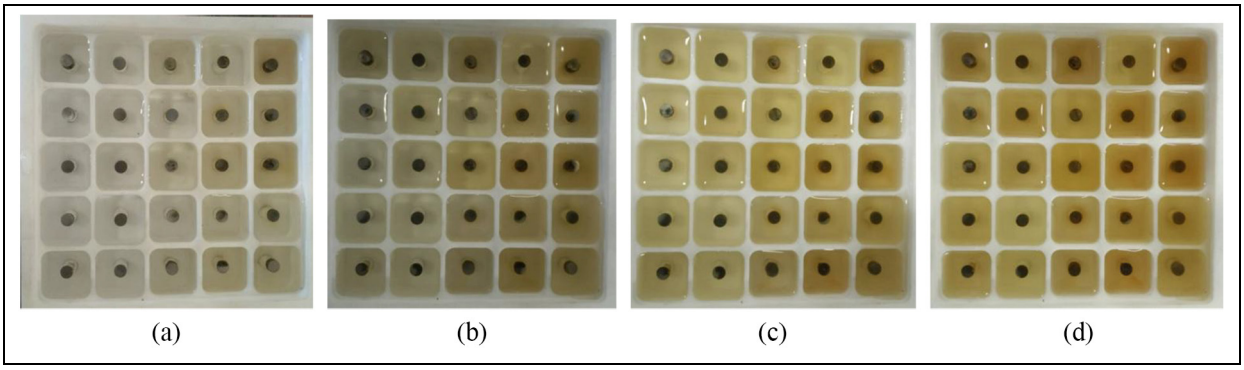


Figure 10. High-throughput corrosion image after a certain time: (a) 1 h, (b) 2 h, (c) 3 h, and (d) 4 h.

that researchers want to get from the CPS. If the traditional image compression methods are adopted, the image will be compressed and stored at first. Then, the compressed image is decompressed and the edge information is extracted one by one. It is a time-consuming and laborious process for a large number of high-throughput image data (Figure 10).

Because the compression algorithm designed in this article can extract the coordinates of four key points directly in each round, we only need to extract the eight key points of the first two rounds from the compressed file to construct the region of Part 5 and Part 5′ as shown in Figure 11.

Set the coordinates of the four key points of the first round as (X_i, Y_i) (i = 1, 2, 3, 4). Set the coordinates of the four key points of the second round as (X_i, Y_i) (i = 5, 6, 7, 8). The order of the four key points is the same as the extraction order of section "Extracting key points information on edge." Set the area of Part 5 is SA, the area of Part 5′ is SA′. The corrosion ratio R and their calculation formulae are shown in Formulae (7)–(9). Here, we adopt the area of the rectangle instead of the area of closed edges for quick calculation

$$SA = (X_3 - X_1) \times (Y_4 - Y_2) \tag{7}$$

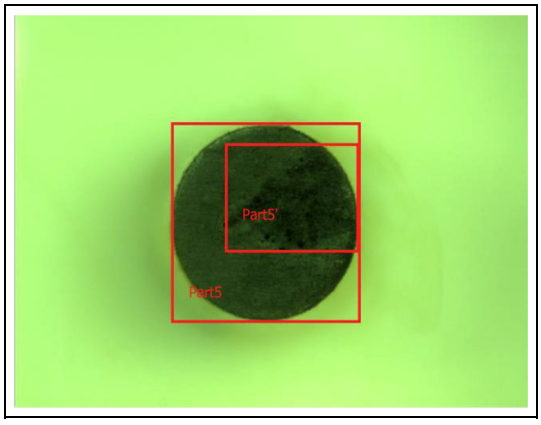


Figure 11. Region of Part 5 and Part 5′ in a high-throughput image.

Table 4. The corresponding relationship of grade value and corrosion area ratio.

Corrosion area ratio (%)	Corrosion grade value
No corrosion	10
$0 < R \leq 0.1$	9
$0.1 < R \leq 0.25$	8
$0.25 < R \leq 0.5$	7
$0.5 < R \leq 1.0$	6
$1.0 < R \leq 2.5$	5
$2.5 < R \leq 5.0$	4
$5.0 < R \leq 25$	3
$25 < R \leq 50$	2
$50 < R$	1

$$SA' = (X_7 - X_5) \times (Y_8 - Y_6) \tag{8}$$

$$R = SA' \div SA \tag{9}$$

Based on the standard GB/T 6461-2002, China standard of corrosion evaluation, the corrosion grade can be evaluated. The standard proposes two basic parameters: corrosion area ratio and corrosion grade, to evaluate the corrosion grade. The corresponding relationship between corrosion grade and corrosion area ratio is shown in Table 4. Corrosion area ratio is the percentage ratio between corrosion surface area and total surface area. The corrosion grade is marked from 1 to 10. The higher the grade is, the more serious the corrosion status is.

Lossless verification of image compression

Compared to the original image, whether the restored image is lossless cannot be judged by the eyes. So, we designed the parameter L to reflect the loss of image. Set I(x, y) as the original image pixel function and C(x, y) as reduced image pixel function. The length of the high and width of flux image is denoted as h and w, respectively. The specific expression of L is shown in Formula (10). If the value of L is equal to 0, it is proved that the compression method is lossless. The bigger the L, the greater the loss of the compression process is

$$L = \sum_{y=0}^{h-1} \sum_{x=0}^{w-1} (I(x, y) - C(x, y))^2 \tag{10}$$

Time cost analysis

The processing time of our method is mainly consumed in Canny edge extraction, SIFT feature matching, and dictionary encoding. The Canny algorithm needs to be executed twice to find the edge points and connect them, so the time complexity of Canny is approximately $O(n^2)$. At one loop of iteration, we need to traverse the edge points in Part 5 to find four new key points. Suppose the number of iterations is m, the time complexity of key point extraction and region segmentation is approximately $O(n^2) + O(n^m)$. The SIFT algorithm's time complexity lies in the establishment and vector matching. These two operations need to be recycled two times, so the time complexity of SIFT algorithm is approximately equal to $O(n^2)$. The dictionary encoding also requires two cycles to count the number of repeats and encode the data, so the time complexity of the algorithm is approximately $2O(n^2) + O(n^m)$. Therefore, the compression time is acceptable when the number of iterations is not high.

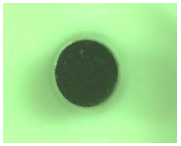
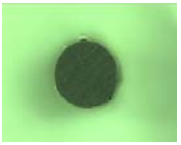
Experimental results and discussion

In order to evaluate the efficiency of the algorithm in this article, we select three groups of similar images and compress one image of them with the mainstream lossless compression algorithm and compare their compression ratio as shown in Table 5.

RAR and 7-Zip are the current mainstream universal compression algorithms. They are all dictionary compression algorithms which utilize the repetitive data in the image to compress the image. PNG adopts LZ77 lossless data compression algorithm. In order to achieve compression function, LZ77 algorithm uses the corresponding matching data appearing in the encoder or decoder to replace the current data. The essence of the Huffman algorithm is to re-encode the character itself for the statistical results, rather than for the repetition of characters or repeated substrings. In practice, the frequency of the symbol cannot be predicted, and it needs to be processed two times. So, it is slow to accomplish the compression and not practical.

RLE algorithm has the advantages of simple implementation, fast compression, and decompression. It can only be completed once the original data are scanned.

Table 5. Comparison of lossless compression methods.

Experimental images		Lossless compression algorithm	Our method	RAR	7-Zip	PNG	Huffman	Run length	LZW	JPEG 2000 (lossless mode)	Block matching (lossless mode)
Reference image 1	Compressed image 1	Before compression (M)	14.41	14.41	14.41	14.41	14.41	14.41	14.41	14.41	14.41
		After compression (M)	3.05	3.58	5.17	8.09	12.31	6.89	11.31	7.86	6.06
		Compression ratio	4.72	4.02	2.79	1.78	1.17	2.09	1.27	1.83	2.38
Reference image 2	Compressed image 2	Before compression (M)	14.41	14.41	14.41	14.41	14.41	14.41	14.41	14.4	14.41
		After compression (M)	3.87	3.67	5.76	8.61	11.62	8.45	12.95	5.52	4.15
		Compression ratio	3.72	3.92	2.50	1.67	1.24	1.71	1.11	2.61	3.47
Reference image 3	Compressed image 3	Before compression (M)	14.41	14.41	14.41	14.41	14.41	14.41	14.41	14.41	14.41
		After compression (M)	2.35	2.36	3.73	5.83	12.73	7.58	10.23	3.62	5.43
		Compression ratio	6.13	6.11	3.86	2.47	1.13	1.90	1.41	3.97	2.65

LZW: Lempel–Ziv–Welch.

The disadvantage is that it is inflexible, poor adaptability, and the average compression rate is low.

Compared with other algorithms, the LZW algorithm has the characteristics of self adaptation, that is, to say, different dictionaries can be built according to the compressed content. The compression effect of JPEG2000 using wavelet transform is worse than 7-Zip and RAR.

Because of the reasonable use of the similarity between the images, the average compression ratio of the algorithm designed in this article is greater than the above algorithm. The time cost of our method is less than that of 7-Zip and Block matching.[21]

Conclusion and future works

In this article, we design a new algorithm utilizing the similarity of high-throughput images to compress. It uses edge information to coordinate transformation, which will get more repeat data and then use the improved LZ77 algorithm to compress them. The knowledge, edge information and pixel differences, is embedded into the compressed file. It can be easily extract from the file to improve the treating efficiency for scientific study.

The method is successfully applied in the self-designed high-throughput corrosion experiment facility. The compression and decompression information of the high-throughput image is recorded and compared with other methods. The experimental results show that the compression ratio of the method is better than other mainstream methods. The time cost of compression and decompression is acceptable. There is no data loss after image compression and decompression by our method.

The future work will focus on the time cost optimization of image compression and decompression. Another work is to embed more knowledge for applications into the compressed file.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the 111 Project (grant no. B12012) and Breeding Project of BWU (no. 0541604654).

ORCID iD

Peng Shi http://orcid.org/0000-0002-5349-6383

References

1. Hughes JR, Roberts N, McGowan S, et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genet* 2014; 46(2): 205–205.
2. Xiang XD, Sun X, Briceño G, et al. A combinatorial approach to materials discovery. *Science* 1995; 268(5218): 1738–1740.
3. Jung H, Han H, Yeom HY, et al. CPS: operating system architecture for efficient network resource management with control-theoretic packet scheduler. *J Commun Netw* 2010; 12(3): 266–274.
4. White PA, Smith GB and Harvey TG. High-throughput corrosion quantification in varied microenvironments. *Corros Sci* 2014; 88: 481–486.
5. Shi P, Li B, Huo J, et al. A smart high-throughput experiment platform for materials corrosion study. *Sci Programming* 2016; 2016: 6876241.
6. Tseng YJ, Martin E, Bologa CG, et al. Cheminformatics aspects of high throughput screening: from robots to models: symposium summary. *J Comput Aided Mol Des* 2013; 27(5): 443–453.
7. Chen L, Wang L, Liu S, et al. Profiling of microbial community during in situ remediation of volatile sulfide compounds in river sediment with nitrate by high throughput sequencing. *Int Biodeter Biodegr* 2013; 85: 429–437.
8. Tsapatsaris N, Beesley AM, Weiher N, et al. High throughput in Situ EXAFS instrumentation for the automatic characterization of materials and catalysts. In: *Proceedings of the ninth international conference on synchrotron radiation instrumentation*, Daegu, Korea, 28 May–2 June 2007, pp.1739–1742. College Park, MD: American Institute of Physics.
9. Wu Z and Zheng N. Efficient rate-control system with three stages for JPEG2000 image coding. *IEEE T Circ Syst Vid* 2006; 16(9): 1063–1073.
10. Stabno M and Wrembel R. RLH: bitmap compression technique based on run-length and Huffman encoding. *Inform Syst* 2009; 34(4–5): 400–414.
11. Itzhak D, Dinstein I and Zilberberg T. Pitting corrosion evaluation by computer image processing. *Corros Sci* 1981; 21(1): 17–22.
12. Codaroa EN, Nakazatoa RZ, Horovistizb AL, et al. An image processing method for morphology characterization and pitting corrosion evaluation. *Mater Sci Eng: A* 2002; 334(1–2): 298–306.
13. Wang S, Kong D and Song S. Diagnosing corrosion modality system of metallic material in seawater based on fuzzy pattern recognition. *Acta Metall Sin* 2001; 37(5): 517–521.
14. Xu S, Weng Y and Li X. Characterization for corrosion pit distribution by using fractal dimension of image. *J Chinese Soc Corrosion Proc* 2007; 27(2): 109–113.
15. Zhou W, Li H, Lu Y, et al. SIFT match verification by geometric coding for large-scale partial-duplicate web image search. *ACM T Multim Comput* 2013; 9(1): 411–418.
16. Žalik B, Mongus D, Lukač N, et al. A universal chain code compression method. *J Vis Commun Image R* 2015; 29: 8–15.

17. Danos RJ, Frey AR, Wang Y, et al. Canny algorithm: a new estimator for primordial non-Gaussianities. *Phys Rev D* 2012; 86(4): 04352610435265.
18. Sidhu RK. Improved canny edge detector in various color spaces. In: *Proceedings of the 3rd international conference on reliability, infocom technologies and optimization*, Noida, India, 8–10 October 2014, pp.1–6. New York: IEEE.
19. Costanzo A, Amerini I, Caldelli R, et al. Forensic analysis of SIFT keypoint removal and injection. *IEEE T Inf Foren Sec* 2014; 9(9): 1450–1464.
20. Crochemore M, Langiu A and Mignosi F. Note on the greedy parsing optimality for dictionary-based text compression. *Theor Comput Sci* 2014; 525: 55–59.
21. Je C and Park HM. Optimized hierarchical block matching for fast and accurate image registration. *Signal Process: Image* 2013; 28(7): 779–791.