Image Processing by Compression: An Overview

A. Roman-Gonzalez, K. Asalde-Alvarez

***Abstract*—This article aims to present the various applications of data compression in image processing. Since some time ago, several research groups have been developing methods based on different data compression techniques to classify, segment, filter and detect digital images fakery. It is necessary to analyze the relationship between different methods and put them into a framework to better understand and better exploit the possibilities that compression provides us respect to image processing; compression becomes very easy techniques to apply without much technical requirement. In this article we will also see the types of compression and specific image compressors.**

***Keywords*—Data compression, image processing, NCD, Kolmogorov complexity, JPEG, ZIP.**

I. Introduction

ACTUALLY and increasingly more often, we can see through the various research works, that data compression is being used for applications in image processing. In fact, very surprising we can see that many authors use data compression techniques for classification and / or segmentation images, filter or denoising image, artifacts detection in images, detecting altered images, etc. The analysis of digital images is of great importance in many fields for which there are many methods as shown in [1], and if the compression can help more easily to this purpose, it is a breakthrough.

In [2] and [3] the authors present methods of image classification based on data compression, in the first case using a video compressor such as MPEG4 and in the second case is used as general purpose compressor as ZIP compressor and images compressor as JPEG. Generally in [4] is a summary of the application of compression techniques in the classification of different data types. The authors of [5] present a new general method for detecting forged images also using data compression. In [6] the authors apply lossy compression to filter the noise in the images. Also the data compression is used as a technique for the detection of artifacts in satellite images as shown in [7], [8] and [9].

Using compression techniques in different areas described above, it becomes possible on the basis of information theory, complexity theory and the various applications around these notions.

The information theory was developed by Claude E. Shannon in 1948 to find fundamental limits in compression and reliable storage of data communication.

A. Roman-Gonzalez, Universidad Nacional San Antonio Abad del Cusco, Universidad Andina del Cusco, Universidad Alas Peruanas; Perú. & TELECOM ParisTech – France.
avid.roman-gonzalez@ieee.org

K. Asalde-Alvarez, Universidad Nacional San Antonio Abad del Cusco, kristaasalde@gmail.com

Considering the probability approach, we can establish a first principle of the measurement information. This establishes that the message that has more probability, it provide less information. This can be expressed as follows:

$$I(x_i) > I(x_k) \Leftrightarrow P(x_i) < P(x_k) \qquad (1)$$

Where:

$I(x_i)$: Amount of information provided by x_i

$P(x_i)$: Probability of x_i

According to this principle, is the probability of a message to be sending and not its content, which determines their informational value.

For a message x_1, with an occurrence probability $P(x_1)$, the information content can be expressed by:

$$I(x_i) = \log_2 \frac{1}{P(x_1)} \qquad (2)$$

Where: $I(x_i)$ will have as unit, the bit.

Within information theory, we have the algorithmic approach that talk about the Kolmogorov complexity $K(x)$ which is defined as the length of the shortest program capable to producing x on a universal machine. Intuitively, $K(x)$ is the minimum amount of information necessary to generate x through an algorithm.

$$K(x) = \min_{q \in Q_x} |q| \qquad (3)$$

Q_x is the set of codes that instantly generate x. The Kolmogorov complexity $K(x|y)$ from x to y is defined as the length of the shortest program that computes x when y is given as an auxiliary input for the program. The function $K(x|y)$ is the length of the shortest program that produces the concatenation of x and y. But $K(x)$ is a non-calculable function.

Both, the probabilistic approach of Shannon and the algorithmic approach of Kolmogorov for information theory, have a relationship with data compression.

The compression serves to transport the same information, but using the least amount of space. Data compression is fundamentally based on search data series repetitions. We have 2 methods of compression: lossless compression and lossy compression. Compressor ZIP is a general purpose lossless compressor based on the combination of LZW code and the Huffman code. The compressor JPEG-LS is a lossless compressor, the algorithm begins with a prediction process, each image pixel value is predicted based on the adjacent pixels, after that, it is necessary to apply an entropy encoder, and get the compressed image. JPEG -DCT Compressor is a lossy compressor, the first step is to divide the

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

image into blocks of 8x8 pixels, to each block, the second step is to apply a Discrete Cosine Transform (DCT), after that, to apply a quantifier and finally an entropy encoder for to get the image compressed.

The structure of this paper is as follows: In Section II we present the image classification based on compression methods. In Section III we show the segmentation of images. Section IV presents a method for detecting forged images based on data compression (lossy compression). In Section V is shown a method where data compression also serves for artifacts detection. Finally in Section VI we present the conclusions.

II. Image Classification

Based on the Normalized Compression Distance (NCD) presented in [10] and [11] which is one applications of Kolmogorov complexity, in [3] the authors perform image classification.

The Normalized Information Distance (NID) is a similarity measure proportional to the length of the shortest program that represents *x* given *y*, as far as the shortest program that represents *y* given *x*.

$$NID(x,y)=\frac{\max\{K(x\mid y),K(y\mid x)\}}{\max\{K(x),K(y)\}}=\frac{K(x,y)-\min\{K(x),K(y)\}}{\max\{K(x),K(y)\}} \quad (4)$$

Since the Kolmogorov complexity *K(x)* is a non-calculable function, in [10] the authors approximate *K(x)* with *C(x)* where *C(x)* is the compression factor of *x*. Based on this approach we obtain the Normalized Compression Distance NCD (Normalized Compression Distance).

$$NCD(x,y)=\frac{C(x,y)-\min\{C(x),C(y)\}}{\max\{C(x),C(y)\}} \quad (5)$$

Where *C(x,y) i*s an approximation of the Kolmogorov complexity *K(x,y)* and represents the file size by compressing the concatenation of *x* and *y*.

For image classification using the approach based on Normalized compression Distance (*NCD*); the first step is calculate the distance matrix between the images based on *NCD* using the equation 5, thus $d_{ij} = NCD(I_i,I_j)$. Thus, we can calculate the distance matrix *D* between all images I_i as:

$$D=\begin{bmatrix} d_{11} & d_{12} & \dots & d_{1j} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2j} & \dots & d_{2n} \\ . & . & \dots & . & \dots & . \\ d_{i1} & d_{i2} & \dots & d_{ij} & \dots & d_{in} \\ . & . & \dots & . & \dots & . \\ d_{n1} & d_{n2} & \dots & d_{nj} & \dots & d_{nn} \end{bmatrix}$$

The matrix D is a square matrix of size *n* x *n*, where *n* is the number of images to classify. Finally we apply a hierarchical classification method as a dendrogram.

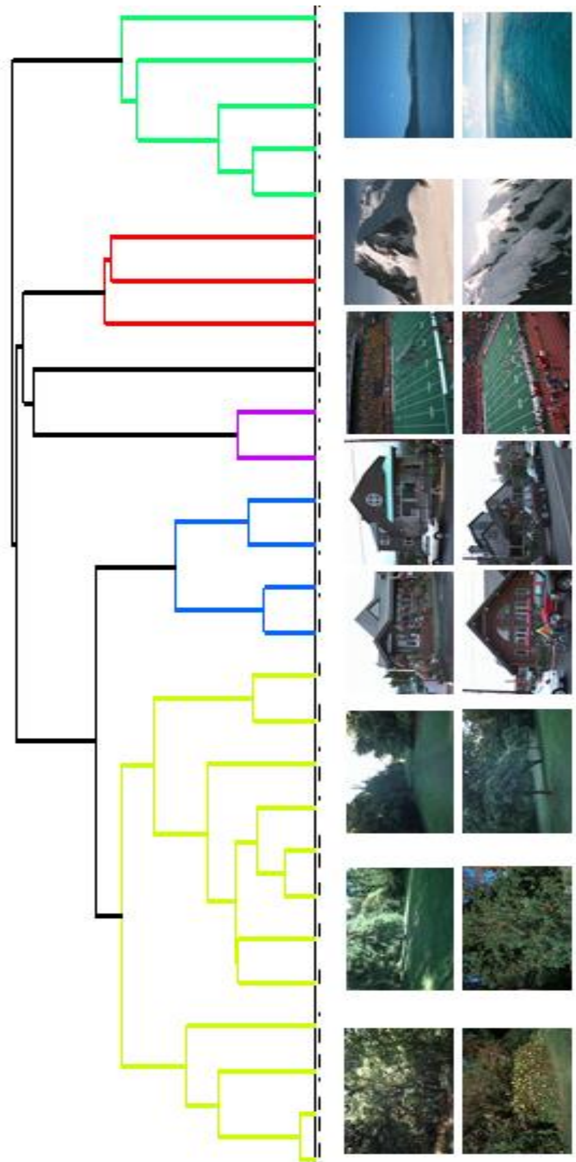


Fig. 1. Image Classification

There are also applications in remote sensing such as those presented in [12].

III. Image Segmentation

As for classification, we can use the same method described above for image segmentation. If we take an image and divide it into small pieces, we can calculate the distance matrix *D* between these different pieces, then we can classify the pieces that will give us as results a segmentation of the original image, this mainly in remote sensing applications.

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

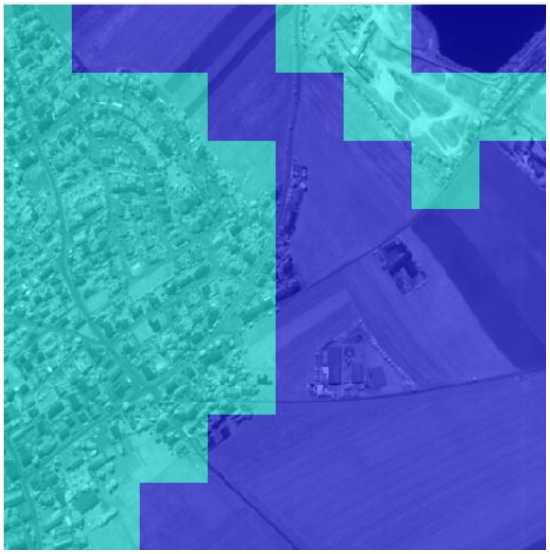


Fig. 2. Image Segmentation

IV. Image Fakery Detection

In [5] the authors use the Rate-Distortion experimental curve (RD) using lossy compression. The RD function indirectly measured the visual complexity of the images, for example, plotting the experimental RD curve where the horizontal axis represents the compression factor (size of the compressed image file / size of the original image file) and the vertical axis represents the distortion calculated using the Mean Square Error (MSE), we can make an analysis of the image.

We can say that the falsification of an image can alter the experimental RD curve of the image, for example we can see from Figure 3 that shown in (a) an original image and (b) the same image but with an alteration which has been removed the person.

(a) Original Image [5]

(b) Manipulated Image [5]

Fig. 3: (a) Original photograph of a Car Show. (b) is the same image (a) but the person was erased and replaced by duplication of regions..

Analyzing the image of Figure 3 (a) and (b) using the graph of the experimental RD curves, it can be seen that there is a variation on these curves due to the manipulation of the image. Figure 4 shows the variation, with the blue curve for Figure 3 (b) and the green curve for Figure 3 (a).

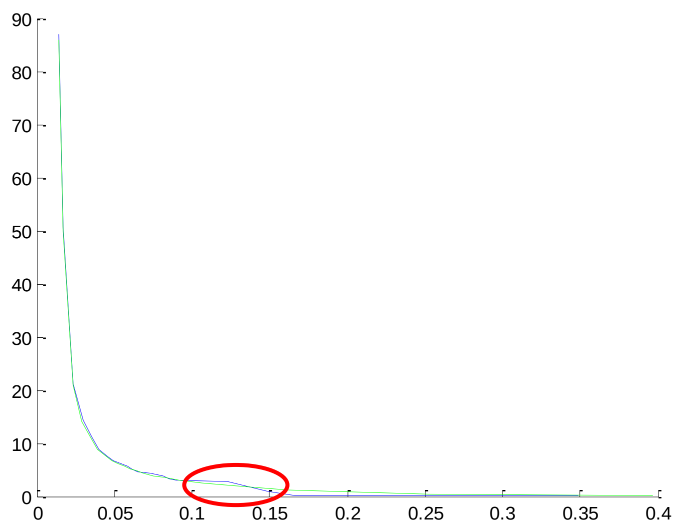


Fig. 4 Rate-Distortion Experimental curve, the horizontal axis represents the compression factor (size of the compressed image file / size of the original image file), the vertical axis represents the distortion calculated using the MSE (mean square error). Blue for Figure 3 (a) and green for Figure 3 (b), the RD experimental curve of the original image is different from the experimental curve RD for image that contains manipulations. [5].

Thus, using the variation in the experimental curve RD of images, we can detect when an image was manipulated or not.

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

V. Image Denoising

A method for filtering and denoising is presented in [6] where the authors use lossy compression techniques for computing experimental RD curve of an image while calculating the Minimum Description Length function [6] and thus obtain the minimum point and identify the image whose noise was eliminated by lossy compression.

Similarly, we may use a new calculation experimental RD curve as a measure of distortion using the Normalized Compression Distance NCD between the error E (produced by the difference between the original image X and the compressed-decompressed image Y) and the compressed-decompressed image Y. The *Rate-NCD(E,Y)* curve between the error E and the compressed-decompressed image Y should have a particular behavior as shown in Figure 5, for a moment, when there is not much compression, the compressed-decompressed image Y is almost equal to the original image X and hence the error E is little information which makes the distance between E and Y large, this distance decreases as the compression increases, because the information is going to error and this will seem a bit to the image, but there comes a point which should be more information on the error E that on the compressed-decompressed image Y thus the distance between them will increase again.

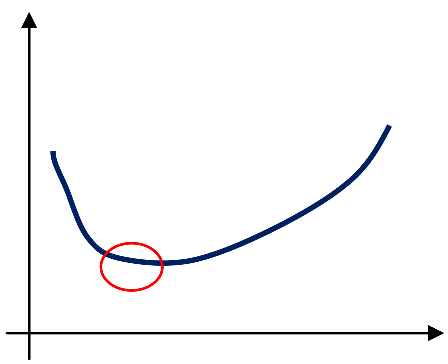


Fig. 5 *Rate-NCD* curve, hypothetical comparison between the error E and the compressed-decompressed image Y.

In this way we can identify the minimum point which would correspond to the image without noise as shown in Figure 6 (using the same image used in [6]). This process can also serve to indicate how to find the minimum amount of information needed to interpret the information, how much detail we can loss without to arrive until a not interpretable data.

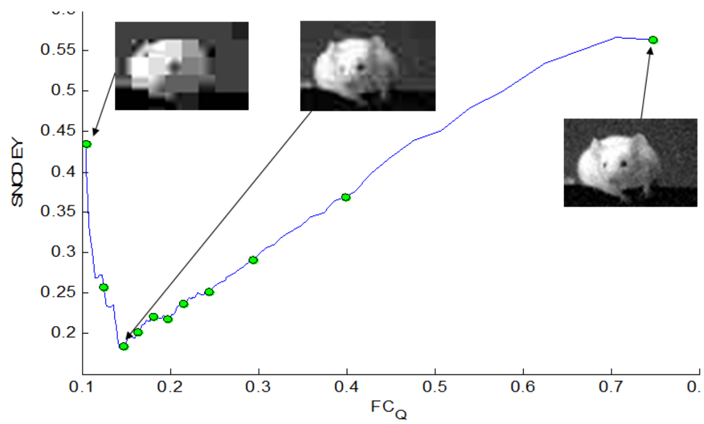


Fig. 6 Image Denoising

VI. Artifacts Detection

In the research Works [7], [8] y [9], the authors present different methods for artifacts detection in satellite images, the method that has better results is the method that uses a lossy compression to calculate the experimental RD curve presented in [9]. The idea is to examine how an artifact can have a high degree of regularity or irregularity for compression [9]. The RD analysis was done as the blocks diagram shown in Figure 7.

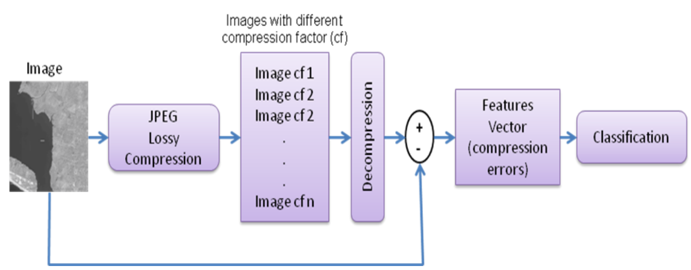


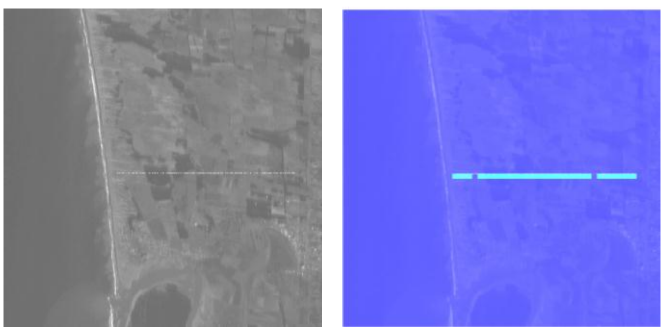
Fig. 7 Block Diagram for Rate-Distortion Analysis[9]

First the author take the image under test I, they dive the image I in different n patches X_i of 64 x 64 pixels, with $i = 1, 2, \dots, n$, thus $I = \{X_1, X_2, \dots X_n\}$. For each *i-th* patch X_i, they compress the patch with different quality q using a lossy compression, using different quality they obtain different compression factor. After that, they decompress the image and obtain a decompressed image Y_{iq}. The next step is to calculate the error for each compression factor between the original patch X_i and the compressed-decompressed patch Y_{iq}, for calculate the error; they use the Mean Square Error (MSE). Based on the errors for each compression factor q and for each patch X_i, they compose a features vector $V_i = [F_{i1}, F_{i2}, \dots F_{iq} \dots F_{iQ}]$ where $F_{iq} = MSE(X_i, Y_{iq})$. Thus, they have a matrix:

$$V = \begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1q} & \cdots & F_{1Q} \\ F_{21} & F_{22} & \cdots & F_{2q} & \cdots & F_{2Q} \\ . & . & \cdots & . & \cdots & . \\ F_{i1} & F_{i2} & \cdots & F_{iq} & \cdots & F_{iQ} \\ . & . & \cdots & . & \cdots & . \\ F_{n1} & F_{n2} & \cdots & F_{nq} & \cdots & F_{nQ} \end{bmatrix}$$

Finally, with this matrix *V*, they apply a supervised or non-supervised classification method like *SVM*, *KNN* or *KMEANS*.

An example of the application of this method based on the RD analysis is the Dropout detection shown in Figure 8, In this case it is a SPOT image containing actual artifacts, and the detection is done correctly.



(a) © CNES (b)

Fig. 8 Dropout (SPOT). (a) Some electronic losses during the image formation process create these randomly saturated pixels. The dropouts often follow a line pattern (corresponding to the structure of the SPOT sensor). (b) Artifact is detected.

VII. Conclusions

As has been seen, there are many studies which use data compression techniques for image processing purposes.

In this paper we have tried to present the different developments made by numerous research groups around the world about the applications of data compression, specifically in image processing. Each group uses different techniques, different types of compression methods and different approaches for different applications but all of them related to two important points: the compression and its application to image analysis in order to make this task easier, understandable and with good results.

The use of compression techniques in image processing, makes this task easier and more accessible, since everybody know to use a compressor, everybody use a compressors in your daily lives to reduce the size of the files and to transmit or transport them more efficiently. No specific knowledge is needed to use a compressor.

Through this article, we provide the background and fundaments to better understand why we can use data compression on digital image analysis and thus continue to develop more applications and improving existing ones.

It was clearly observed that the application of data compression for image processing is entirely feasible. It is possible because data compression is fully related to the information theory in the two approaches. In the image processing and image analyzing, the important thing is the image information content, based on this information we can reach a correct analysis.

The results obtained in the different applications are very encouraging, so it is necessary to continue in this line of research and develop more applications.

VIII. References

[1] A. Roman-Gonzalez, “Digital Inages Analysis”, Revista ECIPeru, vol. 9, N° 1, 2012, pp. 61-68.

[2] B.J.L. Campana y E.J. Keogh, “A Compression Based Distance Measure for Texture”, University of California, Riverside, EEUU 2010.

[3] M. R. Quispe-Ayala, K. Asalde-Alvarez, A. Roman-Gonzalez, “Image Classification Using Data Compression Techniques”; 2010 IEEE 26th Convention of Electrical and Electronics Engineers in Israel – IEEEI 2010; Eilat – Israel; November 2010, pp. 349-353.

[4] A. Roman-Gonzalez, “Clasificación de Datos Basado en Compresión”, Revista ECIPeru, vol. 9, N° 1, 2012, pp. 69-74.

[5] A. Roman-Gonzalez, C.J. Reynaga-Cardenas, “Implementacion de un Método General para la Detección de Imágenes Alteradas Utilizando Técnicas de Compresion”, Engineering Thesis, Universidad Andina del Cusco, 2012.

[6] S. Rooij, P. Vitanyi, “Approximating Rate-Distortion Graphs of individual Data: Experiments in Lossy Compression and Denoising”, IEEE Transaction on Computers, vol.61, N° 3, March 2012, pp. 395-407.

[7] A. Roman-Gonzalez, M. Datcu, “Satellite Image Artifacts Detection Based on Complexity distortion Theory”, IEEE International Geosciences and Remote Sensing Symposium – IGARSS 2011, Vancouver – Canada, July 2011, pp. 1437-1440.

[8] A. Roman-Gonzalez, M. Datcu, “Data Cleaning: Approaches for Earth Observation Image Information Mining”, ESA-JRC-EUSC Image Information Mining: Geospatial Intelligence from Earth Observation Conference, Ispra – Italy, March 30 – April 1, 2010, pp. 117-120.

[9] A. Roman-Gonzalez, M. Datcu, “Parameter Free Image Artifacts Detection: A Compression Based Approach”, 2010 SPIE Remote Sensing, vol. 7830, 783008, Toulouse – France, September 2010.

[10] M. Li and P. Vitányi, “The Similarity Metric”, IEEE Transaction on Information Theory, vol. 50, N° 12, 2004, pp. 3250-3264.

[11] R. Cilibrasi, P. M. B. Vitanyi; “Clustering by Compression”, *IEEE Transaction on Information Theory*, vol. 51, N° 4, April 2005, pp 1523 - 1545.

[12] A. Roman-Gonzalez, M. A. Veganzones, M. Graña, M. Datcu, “A Novel Data Compression for Remote Sensing data Mining”, ESA-JRC-EUSC Image Information Mining: Geospatial Intelligence from Earth Observation Conference, Ispra – Italy, March 30 – April 1, 2010, pp. 101-104.

[13] F. Tussell, “Complejidad Estocástica”, San Sebastián 1996.

[14] D. McKay, “Information Theory, Inference, and Learning Algorithms”, Cambridge University Press, 2003.

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)