Comparison and Evaluation of Information-based Measures in Images

Jorge Miguel Silva
jorge.miguel.ferreira.silva@ua.pt
IEETA,
University of Aveiro,
Aveiro, Portugal

Diogo Pratas
pratas@ua.pt
Department of Virology,
University of Helsinki,
Helsinki, Finland

Sérgio Matos
aleixomatos@ua.pt
DETI,
University of Aveiro,
Aveiro, Portugal

Abstract

Lossless data compressors and small Turing machines can approximate the quantity of information present in a digital object. In this paper, we describe and compare these approaches of measuring unsupervised probabilistic and algorithmic information on images (2D) with different characteristics. We use the Normalized Compression (NC) employing the data compression PAQ8 and compare it with the Block Decomposition Method (BDM) and show some advantages and limitations of both measures.

1 Introduction

There are several approaches to quantify the amount of information. Kolmogorov described three, namely combinatorial, probabilistic, and algorithmic [4]. While the Kolmogorov complexity is non-computable, it can be approximated with programs for such purpose, such as data compressors, using probabilistic and algorithmic schemes. Practical applications to approximate the Kolmogorov complexity for multiple dimensional digital objects have been developed using Turing machines [6, 7] and data compressors [3]. Recently, Zenil *et al.* have shown that this methodology has a closer connection to algorithmic information than other measures based on statistical regularities [7], namely fast lossless compression methods, for sources that follow algorithmic schemes. One of the applications of information theory is to measure image information. Herein, we define an image's quantity of information as the smallest number of bits required by a model to represent an image losslessly. To perform this task, the model searches for unknown patterns of similarity between sub-regions of the image and uses this information to create this compressed representation of the image, relying exclusively on the two-dimensional pixels' patterns without using exogenous information. In this paper, we describe and compare solutions for unsupervised measures of probabilistic and algorithmic information in images (2D) of different datasets. We use the Normalized Compression (NC) employing PAQ8 data compression tool and compare it with the Block Decomposition Method (BDM) [7], and the inherent Coding Theorem Method (CTM) measures [2]. The BDM is an information-based measure that uses small Turing machines to approximate the algorithmic information, approximating to the Shannon entropy as a fallback mechanism.

2 Methods

In this section, we describe the Normalized Compression (NC) and two Block Decomposition Method (BDM) normalizations.

Normalized Compression (NC)

An efficient compressor, $C(x)$, gives a possible approximation for the Kolmogorov complexity ($K(x)$), where $K(x) < C(x) \leq |x|$ ($|x|$ is the length of string x in the appropriate scale). Usually, an efficient data compressor is a program that approximates both probabilistic and algorithmic sources. Although the algorithmic nature may be more complex to model, data compressors may have embedded sub-programs to handle this nature. For a definition of safe approximation, see [1]. The normalized version, known as the Normalized Compression (NC), is defined by $\mathrm{NC}(x) = \frac{C(x)}{|x|\log_2|\mathrm{A}|} = \frac{C(x)}{|x|}$, where x is a string, $C(x)$ is the compressed size of x in bits, $|\mathrm{A}|$ the number of different elements in x (size of the alphabet) and $|x|$ the length of x. Since we consider a binary matrix of each image, $|A| = 2, \log_2 2 = 1$. Given the normalization, the NC enables to compare the information contained in the strings independently from their sizes [5]. If the compressor is efficient, then the compressor is able to approximate the quantity of probabilistic-algorithmic information in data.

Normalized Block Decomposition Method (NBDM)

Another possible approximation to the Kolmogorov complexity is given by the use of small Turing machines, where these small computer programs approximate the components of a broader representation. The Block Decomposition Method (BDM) extends the power of a CTM, approximating local estimations of algorithmic information based on the Solomonoff-Levin's algorithmic probability theory. In practice, it approximates the algorithmic information and, when it loses accuracy, it performs like Shannon entropy. Since in this article we intend to perform a direct comparison of both measures, we first considered the normalization of the BDM (NBDM_1), given by the number of elements (length) of the digital object: $\mathrm{NBDM}_1(x) = \frac{BDM(x)}{|x|\log_2|\mathrm{A}|} = \frac{BDM(x)}{|x|}$. However, the normalization of the BDM is usually performed using a minimum complexity object (BDM_{Min}) and a maximum complexity object (BDM_{Max}). A minimum complexity object is filled with only one symbol, like a binary string of only zeros. In contrast, a maximum complexity object is an object that, when decomposed (by a given decomposition algorithm), yields slices that cover the highest CTM values and are repeated only after all possible slices of a given shape have been used once. Using these two objects, the NBDM_2 for a given string can be computed as $\mathrm{NBDM}_2(x) = \frac{BDM(x)-BDM_{Min}}{BDM_{Max}-BDM_{Min}}$, where $BDM(x)$ is the BDM value of that string, BDM_{Min} is the minimum complexity object, and BDM_{Max} is the maximum complexity object. Kolmogorov complexity is invariant only up to a constant factor, which depends on the choice of a description language $K = K' + L$, where K is the total complexity, K' is the description of the object and L is the description of the language. As such, by performing the normalization according to Equation 2, the normalization is aiming to remove the constant factor as $\frac{K-K_{Min}}{K_{Max}-K_{Min}} = \frac{K'+L-K'_{Min}-L}{K'_{Max}+L-K'_{Min}-L} = \frac{K'-K'_{Min}}{K'_{Max}-K'_{Min}}$, where K_{Max} and K_{Min} are the maximum and minimum Kolmogorov complexity objects and K'_{Max} and K'_{Min} are the maximum and minimum Kolmogorov complexity description of the objects.

3 Results and Discussion

In order to compare NC with BDM, we performed three types of tests. Namely, we compared the robustness of both measures according to increasing rates of random pixel changes in paintings, tested their application on different types of images, and made an assessment of the minimal information bounds. In the first test, we assessed the impact of an increasing rate of pixel editions using a pseudo-random uniform distribution and compared both information-based measures. This approach is not identical to image noise, but rather a pure edition of pixels. For the purpose, we selected a painting from three authors (Theodore Gericault, Marc Chagall, and Rene Magritte), making 50 adulterated copies of each painting with increasing edition rate (from 1 to 50%). Finally, we measured the NC (Eq. 2), the NBDM_1 (Eq. 2), and NBDM_2 (Eq. 2) in all the paintings. Figure 1 (A) depicts the values obtained for the NC and BDM. The results show that, when using the same type of normalization, NC is more robust to the increment of pixel edition than NBDM (NBDM_1). On the other hand, whereas NBDM_1 considers the normalization by the length of the input object, NBDM_2 performs a normalization that aims to mimic the removal of the constant factor related to Kolmogorov complexity (see Eq. 2). Since the NBDM_2 normalization does not take into account the constant of the description language, it shows a more robust behavior than

$NBDM_1$, which increases rapidly with the increase of pixel edition. Since NC and $NBDM_1$ have the same type of normalization, we will focus on comparing these normalizations from now on.

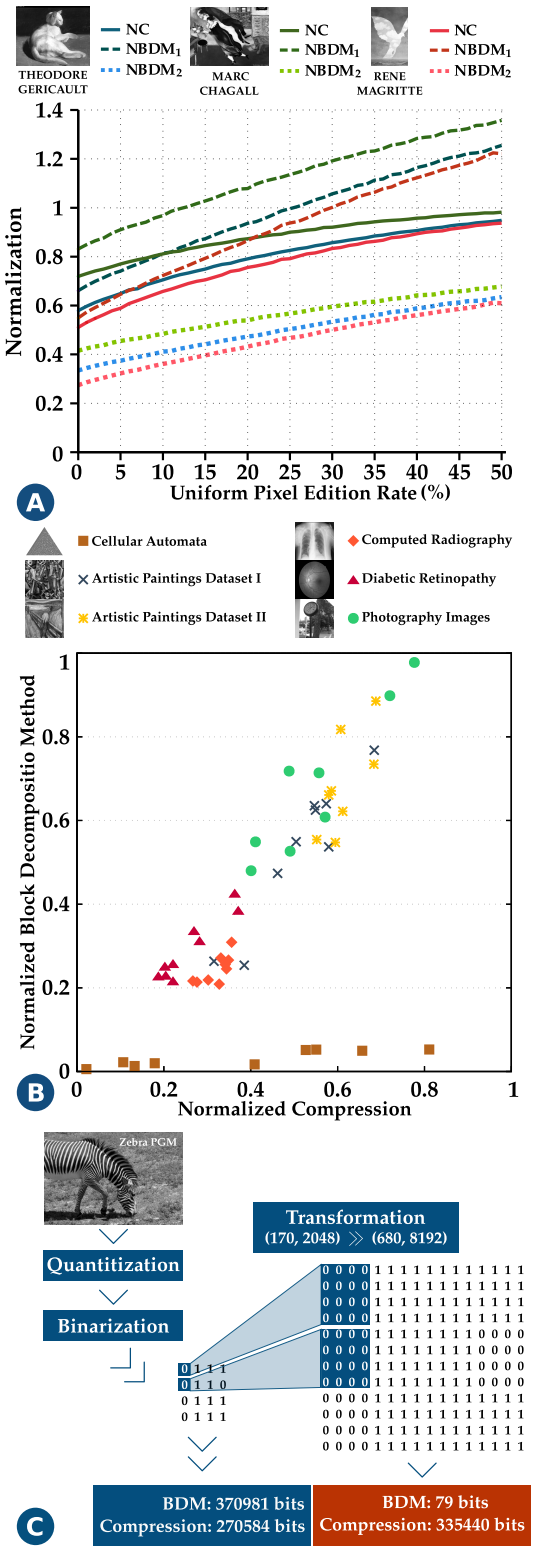

Figure 1: Evaluation of Information-based measures. (**A**) Impact of increasing pseudo-random substitution on information-based measures: NC (approximated using the PAQ8 algorithm) and two BDM normalizations ($NBDM_1$ and $NBDM_2$). (**B**) NC and $NBDM_1$ for different types of images. (**C**) Image transformation pipeline leading to BDM underestimation of the amount of information contained in the transformed object.

In the second test, we applied both measures to six datasets with distinct nature (9 images each) to understand how $NBDM_1$ and NC behave with different types of images. The six datasets were: artistic images from 2 different datasets; cellular automata images; diabetic retinopathy images; chest computed radiography (CR) images and photographic images. The results are depicted in Figure 1 (B). Overall, the majority of the datasets show similar behavior regarding the NC and $NBDM_1$. The exceptions to this are the CR and cellular automata datasets, which exhibit a more algorithmic behavior. The latter dataset is constituted by images created with small programs with simple rules. Whereas the compressor has difficulty compressing this type of images, the BDM can determine their algorithmic nature and thus attribute them with minimal value. This outcome shows the importance of the BDM in the detection of simple output programs embedded into data. In the last test, we selected one of the most complex images identified by the NBDM in the last subsection to test if the BDM could accommodate specific data alterations. This test is depicted in Figure 1 (C). After the binarization process, we performed a super-sample image transformation where each char was amplified to a 4x4 representation. This value was selected since the BDM has the default block size value of 4x4 in 2D structures. After this operation, the BDM was computed for the original and the super-sampled image. While the original image was measured with 370981 bits, the super-sampled image had only 79 bits. This abrupt decrease in the complexity value indicates that the BDM underestimates the amount of information contained in the object. The BDM analyses object information in blocks instead of looking at the whole object. Specifically, blocks analysed by the BDM (default block size value of 4x4 in 2D structures) have the same size as the super-sample image transformation (each char was amplified to a 4x4 representation); therefore, the complexity attributed to each block is approximately zero (since each block is composed of all zeros or ones), and hence the overall value attributed to the complexity of the object will drop dramatically. This analysis shows that BDM is not prepared to deal with the information associated with the choice of the model, unlike the NC. The NC relies on the use of a lossless data compressor, bounded by a maximum information channel capacity.

4 Conclusion

The results show that, when using the same type of normalization, NC is more robust to the increment of pixel edition than NBDM ($NBDM_1$). On the other hand, BDM can determine the algorithmic nature of images created with small programs with simple rules. Whereas the compressor has difficulty compressing this type of image, the BDM can determine their algorithmic nature and attribute them with minimal value. Finally, BDM is not prepared to deal with the information associated with the model's choice, unlike NC. The NC relies on using a lossless data compressor, bounded by a maximum information channel capacity. From these three tests, we can notice some advantages and limitations of both measures. Ranking these measures is not a fair task because they have different characteristics and nature.

References

[1] Peter Bloem, Francisco Mota, Steven de Rooij, Luis Antunes, and Pieter Adriaans. A safe approximation for Kolmogorov complexity. In *International Conference on Algorithmic Learning Theory*, pages 336–350. Springer, 2014.

[2] Jean-Paul Delahaye and Hector Zenil. Numerical evaluation of algorithmic complexity for short strings: A glance into the innermost structure of randomness. *Applied Mathematics and Computation*, 219 (1):63 – 77, 2012. ISSN 0096-3003. doi: https://doi.org/10.1016/j.amc.2011.10.006. Towards a Computational Interpretation of Physical Theories.

[3] Ming Li, Jonathan H. Badger, Xin Chen, Sam Kwong, Paul Kearney, and Haoyong Zhang. An information-based sequence distance and its application to whole mitochondrial genome phylogeny . *Bioinformatics*, 17(2):149–154, 02 2001. ISSN 1367-4803.

[4] Ming Li, Paul Vitányi, et al. *An introduction to Kolmogorov complexity and its applications*, volume 3. Springer, 2008.

[5] Diogo Pratas and Armando J Pinho. On the approximation of the Kolmogorov complexity for DNA sequences. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 259–266. Springer, 2017.

[6] Fernando Soler-Toscano and Hector Zenil. A computable measure of algorithmic probability by finite approximations with an application to integer sequences. *Complexity*, 2017, 2017.

[7] Hector Zenil, Santiago Hernández-Orozco, Narsis A Kiani, Fernando Soler-Toscano, Antonio Rueda-Toicen, and Jesper Tegnér. A decomposition method for global evaluation of Shannon entropy and local estimations of algorithmic complexity. *Entropy*, 20(8):605, 2018.

*/