Zipf law: a tool for image characterization

Pascal Makris and Nicole Vincent
Laboratoire d'Informatique
Université de Tours (LI/E3I)
64, avenue Jean Portalis,
37200 Tours, France
makris@univ-tours.fr, vincent@univ-tours.fr

Abstract

We intend to point out the interest in the use of Zipf law in the field of image analysis. We show how it is possible to adapt it to image analysis and under which conditions it yields to best results. We tackle several possible applications of the method depending on the one hand of the images to be analyzed and on the other hand on the kind of questions to be solved. The kind of images that we have been concerned with, are more or less structured images that most of the time can be considered as the result of man activity and that we oppose to images from nature. Our goal is to underline their differences and so to open some new ways for works involving authentication problems.

Keywords

Zipf law, statistical parameters, compression, natural and geometrical images.

Introduction

The very important number of papers that are devoted to image analysis attests both that a general pattern analysis and recognition method is not yet efficient and that the activity of the researchers is still vivacious. This activity finds expression in numerous applications, specific to each kind of situation.

At the moment, image is taking a more and more important place in any document, and then, an evaluation of the image quality has often to be done, as well as a choice, relying upon some criteria rather than on others. Some processes of the images are necessary as in the case of compression for instance. Different aspects can be privileged, they may concern either the general content of the image or the presence of some details, or too the more and more frequent presence of superimposed printings, it can occur either in texts or in hypertexts as well as in video images.

We have felt that the opposition between images, on the one hand having some elements entirely independent of humans and on the other hand having some objects coming from human activity, would enable the discrimination between two large families of images. Nevertheless, a clear and easy method to achieve this classification was to be found. To obtain this goal we have made use of Zipf law.

In a first part, and after the precise meanings of the words "natural" and "artificial" have been introduced, we will present the principle of the method that relies on Zipf law.

In a second part, it will be explained how it is possible to adapt this method to image analysis, in particular how the relations that may exist between the patterns present in an image, can be analyzed and quantitatively computed.

A last part will be devoted to presenting some significant results. They have been obtained in different domains. Some future perspectives opened by this method will be pointed out.

1. The method

1.1. Natural versus Artificial

The expression is familiar and seems to need no further explanation, no complementary precision. Actually it is quite different. Of course, we consider we are working from photographs, digitized with a scanner or directly obtained with usual numerical systems such as numerical cameras, video cameras or microscopes, for instance SEM. Of course, these systems are man made, millions of identical copies are manufactured, but the images they produce, can either be man made or natural productions. By definition, a natural component is a data in the elaboration of which

Human has taken no part. It is the work of nature and it has to be opposed to the artificial aspect, which is in fact man made.

At first glance, a landscape is natural, a synthetic image is artificial, and man himself and the human face are natural, as they are part of Nature. The images of biological elements are most often natural images because man has taken no part in their creation. On the contrary, any drawing that wants to figure the human faces or the tissues or different organs is an artificial work. It is the same for all drawings, geometrical drawings or free-hand drawings; all of them belong to the set of synthetic images.

Man is working according to specific rules: the rulebook, the rules of handicraftsmen or of artists. He uses a self-given measurement system and that is relying on Euclidean geometry. This often induces the geometrical character of his productions and allows their distinction from the nature productions.

The case of the works of art, painting, sculpture, architecture is some what special. These works are man made, so they are artificial. Some persons want to give them a specific place because these works are more or less free copies of nature and because they are the expression of a talent, a gift of nature.

In another domain the texts are the origin of some specific problems. They are important both because they are so many and only a small part of them is contained in the libraries and the archives, and because they have an important place in our civilization of written texts. These texts more and more frequently appear in all kinds of documents, as images, they are to be discriminated and extracted [1]. At first view they are human works, so they are artificial works, especially as far as typed or printed texts are concerned. The copy of a printed text is strictly identical to the original. As to handwritten text, it is a unique and personal piece of work and each writer has his own way of writing. So, we have not tackled here the problems that are emerging. In the same way, we manage to omit the word "authentic" that may be ambiguous. It involves something indisputable; it can only be used in the case of a copy or a counterfeit.

There exists an other kind of images, that have not lost their proper characteristics, either natural or artificial, but that have been modified through some compression – decompression processes. The resulting images have been altered and it will be shown how Zipf law can help to detect this alteration and even to quantify it.

1.2. Method principle

Our study is relying on application of an empirical law that has been stated 50 years ago: Zipf law [2]. Many verifications have been achieved in different domains [3, 4]. It may be stated as follows. In a set of topologically organized symbols, the n-tuples of symbols are not randomly organized. It may be observed that the appearance frequencies $N_1,...N_n$ of the present n-tuples $M_1,...M_n$ are depending on these patterns. More precisely, if patterns are ordered according to frequency decreasing order, the sequence $(N_{(1)},...N_{(n)})$ where i=1 to n, verifies fundamental formula:

$$N_{(i)} = k.i^a \qquad (1)$$

If such a power law can be proved, it is mainly characterized by the "a" value of the power, and of course, this value is negative.

The most convenient way to estimate this "a" value is to study the link between the respective logarithms of $N_{(i)}$ and i. When the law holds, the two quantities are linked through a linear relation and the "a" value can be deduced from the director parameter of the regression straight line that approximates, in the mean square sense, the couples $[\ln i, \ln(N_{(i)})]$. Practically, on a graph where abscissa indicates ln i and y-coordinate, $\ln(N_{(i)})$, it is easily observed these points are aligned. Then, the slope of the straight line gives the "a" value.

To achieve a classification of the studied phenomenon, either the "a" exponent value or the quality of the alignment, that is to say the adequation of the Zipf law to the studied phenomenon, can be used.

To compare two representations of an only phenomenon, graphs associated to the images have to be compared. They are to be matched locally rather than using a global comparison of the slopes.

From this law point of view, the field of mono-dimensional signals has been most studied. Now, our purpose is to apply it to image analysis. Of course, our goal being different, we are obliged to adapt the concept, to introduce some parameters corresponding to the choices we are now to present.

2. Adaptation to image analysis

It is obvious this adaptation leads to delicate problems, more or less difficult to be solved according to the more or less important structure of the image to be studied. The interpretation is easier when some invariant links exist between the different elements of the image; otherwise many difficulties may arise. Hence the problem is to find a method to analyze the structure of the image, using on the one hand some masks and on the other hand some coding method.

2.1. Mask choice

An image is a set of pixels organized on a plane, in a

matrix. Then, implicitly a neighborhood relation exists between pixels. Of course, the choice of the masks must respect this topological notion.

The natural symbols used to code an image are the gray levels of each pixel. (2 for a binary image and 256 for most of gray level images). Then, the n-tuples mentioned in the statement of Zipf law are chosen as the sequence of pixel gray levels, present in masks of various shapes. In our applications, 3x3 square masks are evident. They are considered as a neighborhood of the central pixel. Some larger masks can be chosen too.

The mask can as well be considered as a model of the local zone the eye integrates during a global observation of the image. Nevertheless, the shape of the zone is influenced by the global characteristics of the image. For instance, the block aspect due to some image processes will have some influence in the vertical and horizontal directions.

Then the image is viewed as a set of occurrences of such masks. Their number is equal to the number of pixels. Their accumulation contributes to the building of the global perception of the image.

2.2. Symbol and mode of coding choice

The rough coding

The mask shape being set up, the alphabet to be used to describe the pixels within the mask has to be fixed, that is to say the different possible patterns to achieve the description.

The number of these patterns may vary a lot according to the masks and to the description symbols used.

The simplest coding uses the 256 gray levels of the image. In this case, the number of the different possible patterns in a 3x3 mask is equal to 256^9, that is to say about 10^{21}. This is far too much to try and consider all of them. Besides, the number of patterns, either distinct or identical, that a 512x512 image can contain is only about 10^6.

In this last case, the solution is to store only the patterns that are encountered during the scanning of the image. This allows the software to satisfy two fundamental constraints, that is to say optimizing the memory space and reducing the processing time.

It can be noticed that in a random image, a pattern has but a very low probability to occur in an image (one chance out of 10^{15}). Then, when a pattern is met with significant frequency, the given information indicates the image presents a clear structure.

Whereas it is convenient to increase the occurring probability of each pattern, nevertheless, the decrease of the number of possible patterns it implies, must be taken into account. In this study, it has been chosen to reduce the number of symbols used in each pattern. This number has been fixed less or equal to mask size. For a 3x3 mask, the symbol number may at most be equal to 9.

Rank methods

All the coding methods using rank and specially the distinct rank method, the general rank method and the method with filtering, present interesting aspects but also some inconvenient. The general rank method has been chosen in our study.

We have been aware that, even in the rough coding method, the different possible patterns are still too many and may figure very near perceptions, that is to say, distinct patterns may correspond to very near visual perceptions. Are to be privileged totally different patterns. That is why, to code a pattern, we use the ranks of the pixel gray levels within the mask. In a 3x3 mask, the number of patterns using distinct ranks is about 4.10^5. It is much lower than the 10^{21} previous possibilities. In spite of this reduction in the number of patterns, it is possible to express the diversity of the perceptions occurring in a local zone. The following example, TAB.1 shows a 3x3 mask with a particular pattern that is given using rough coding. Its code under the so-called distinct rank method is presented in TAB. 2.

TAB.1: pattern based on the gray levels.

255	255	240
255	0	18
20	0	20

TAB.2: pattern based on the distinct ranks.

6	7	5
8	0	2
3	1	4

Even though this rank method presents the advantage of a less important number of possible patterns, besides it is not entirely satisfying as when, within the mask, several pixels have the same gray level the code is affected in a quite random mode. So a somewhat different method can be used, the general rank method.

The general rank method

To reduce the loss of information and its deformation, the equality between gray levels has to be taken into account when it occurs within the mask. In this case, the

previous coding of the mask is modified. The new code of this pattern is indicated in TAB. 3.

TAB. 3: pattern based on general ranks.

4	4	3
4	0	1
2	0	2

This method, called general rank method [5], enables to keep the original profile of the mask. Of course the number of patterns is greater than the number obtained in the distinct rank method, but in the same time, this number is less than the number of rough patterns. For example, in the 3x3 mask there are about 7.10^6 distinct patterns.

This coding method has been chosen both for binary images and for gray level images. It leads to the elaboration of the pattern occurrence frequency graph.

2.3. Pattern occurrence frequency graph

To ensure the verification of Zipf law, the immediate application of the method previously presented implies the computation of the couples sorted according to the i value ($\ln i$, $\ln N_{(i)}$). Of course, are taken into account only the patterns that are present in the image and that appear more than once. The graph of the pattern frequency has to be computed.

Practically, to obtain the elements of the graph, for each possible pattern, the number of its occurrences in the image has to be counted. The scanning of the image is achieved with covering masks, this ensure the independence of the results toward a low magnitude translation.

These data, figured in the plane, give a set of points. By experience, we can say the linear aspect of the set is not as obvious as one could expect from the law. It can be seen in the example of

FIG. 2 on which it seems possible to extract several straight-line segments.

Though the position of the occurrences of the patterns is missed, the numerical organization of the pattern distribution, nevertheless gives interesting information that is more precise than when considering only gray level histogram, because the mask around each point introduces a local knowledge, but not a punctual one.

Let us now come to the results the method allowed to perform on various images. We will present here only two examples of problems that can be solved using Zipf law. In all the cases presented here, 3x3 masks have been used, because we wanted to make no special hypothesis on the studied images.

3. Results

Let us recall it, the goal of this study is to allow some comparison between the results obtained with different kinds of images. The analysis of these results brings much knowledge on the nature of the contents of the studied images.

The underlying information is expressed by the curve figuring the occurring frequency of the most important patterns that constitute the image. The three images that we consider belong to very general families. They allow to highlight two important problems:

- On the one hand the effect of "compression - decompression" on a natural image,
- On the other hand the opposition between a natural image and an artificial image with geometrical characteristics.

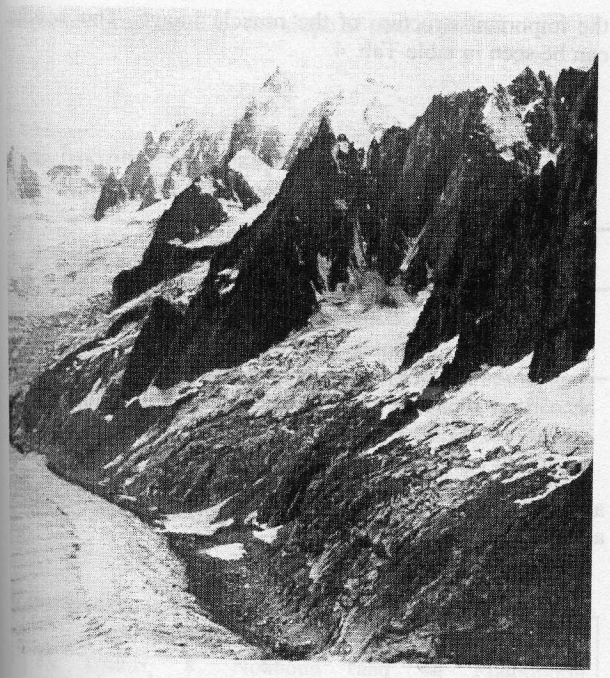
3.1. The choice of the images

The first image, that can be qualified as natural, is a photo of an alpine landscape. The second one is the image of the same landscape after compression – decompression in a ratio of about 20. Then comes a well-known image of the structure of a muscle, image that of course is eminently natural. The last image, entirely artificial, is the photo of a canvas from Vasarely painter, it has been chosen because of its geometrical character.

3.2. The problem of compression - decompression

The effect of the "compression – decompression" is made evident in FIG. 1 and in

FIG. 2 where is figured the pattern frequency graph.



(a)

(b)

FIG. 1: Natural images (initial image (a) and decompressed image (b))

The difference between the quality of these two landscapes can be noted. To the initial image on the left is opposed the image on the right, the quality of which is inferior.

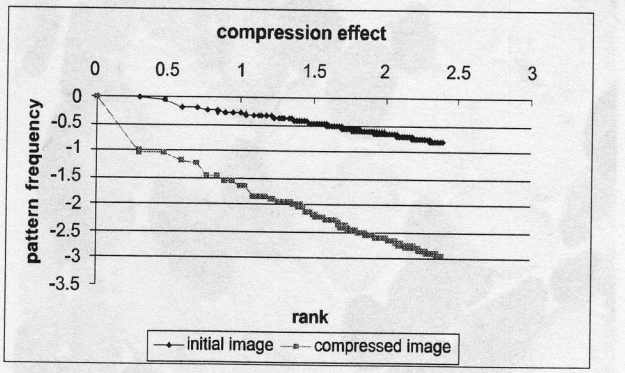


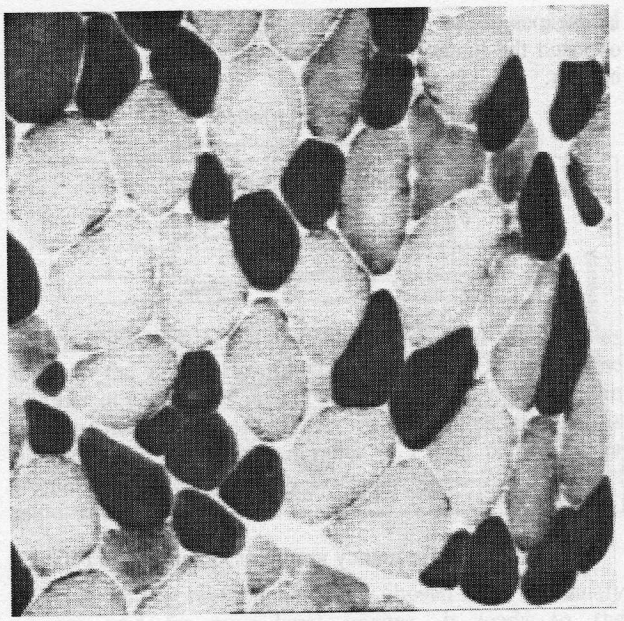
FIG. 2: Pattern frequency graphs associated with FIG. 1 images

The graphs have been normalized to make easier a comparison between curves associated with images of different sizes. More precisely, the graph of [ln i, ln ($N_{(i)}$ / $N_{(1)}$)] is figured here for each studied image.

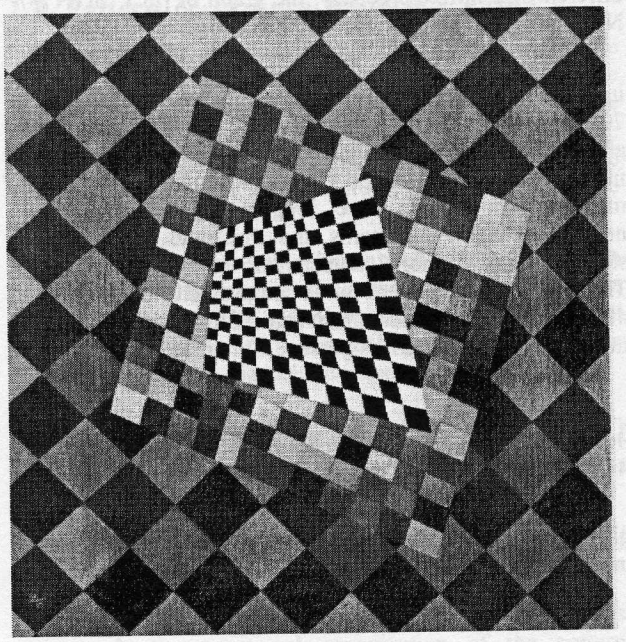
Besides, the number of patterns present in the initial image is much larger than in the decompressed image (174 701 versus 60 891). The curves emphasize the good aligment of the dots associated with the patterns occurring in the image. The second curve, the bottom one, much more sloppy, much more irregular figures the lower number of different patterns and illustrates some rough slope transitions more characteristic of artificial images. The most frequent pattern in the compressed image appears 40 366 times, it is associated with uniform zones. In the initial image, no pattern is present more than 61 times.

3.3. Opposition between natural image and artificial image

The problem is illustrated in FIG. 3 and in FIG. 4 is given the pattern frequency graph associated with each image.



(a)



(b)

FIG. 3: natural image of a muscle and a geometrical artificial image

To FIG. 4 could have been joined the pattern frequency graph associated with the initial alpine landscape in order to make a comparison. Both curves show the same regularity but the general slopes are different, they figure the important structure of the muscle image. The results can be seen in table Tab. 4.

Tab. 4: power law exponent value in Zipf law

	landscape	muscle	Vasarely
Zipf exponent	-0.36	-0.69	-0.16

In FIG. 3 are then opposed a natural image and an artificial image characterized by an especially geometrical aspect.

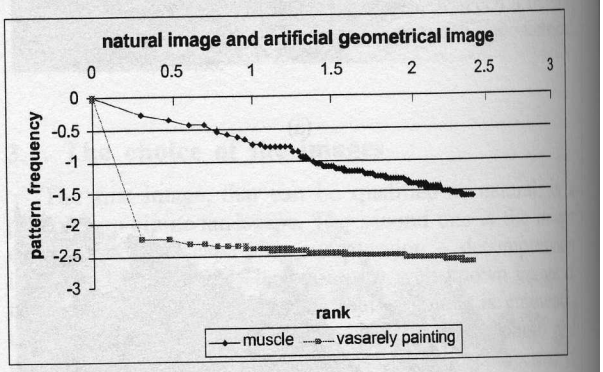


FIG. 4: pattern frequency graph of the FIG. 3 images

The pattern frequency graphs in FIG. 4 are well showing this opposition. The curve associated with the geometrical image is quite different from the other. The left zone figures the rough transition from some very frequent patterns to an other type of behavior that is indicated by the right part.

Of course the low number of present patterns can be noted.

Such properties in the curve characterize artificial images and are not present with natural images. More globally, the trial has been performed with a hundred images. This has enabled the definition of some criteria leading to an automatic classification of artificial and natural images. These criteria have been successfully tested on 20 images. The parameters that have been used are the slopes of the regression straight lines that approximate the frequency curves. Only the two first left segments have been considered, so only two parameters are computed, C_1 the slope of the first segment, in the left zone of the

frequency graph, and C_2 the slope of next segment.

Conclusion and perspectives

The adaptation of Zipf law to the image analysis problems has been shown and its use can be extended to many domains. It can contribute to the extraction of some parameters that can characterize the structure of an image, specially are important the presence or the absence of horizontal zones in the frequency graph, the slope of the regression lines and the number of patterns. In some way, all this data gives a signature to each image. It is then an interesting tool in order to distinguish between the origins of different images or to determine specific zones within an image. It is to be thought that Zipf law can contribute a lot in a large field of activity: image analysis.

References

[1] S. Djeziri, F. Nouboud and R. Plamondon, "Discrimination d'images de documents", 1er Colloque International Francophone sur l'Ecrit et le Document CIFED'98, Québec, 11-13 mai 1998, pp. 60-65, 1998.

[2] G.K. Zipf, *Human Behavior and the principle of "Least Effort"*, Addison-Wesley, New York, 1949.

[3] S. Havlin, S.V. Buldyrev, A. L. Goldberger and AL, "Statistical Properties of D.N.A Sequences", *Fractal Reviews in the Natural and Applied Sciences,* Chapman & Hall, pp.1-11, 1995.

[4] A. Cohen, R.N. Mantegna and S. Halvin, "Numerical analysis of word frequencies in artificial and natural language texts", *Fractals*, World Scientific Publishing Company, vol. 5 n°1, pp. 95-104, 1997.

[5] D. Bi, J.-P. Asselin de Beauville, and M. Mraghni, "Spatial grey level distribution based unsupervised texture segmentation". 3rd International Conference of Signal Processing (ICSP'96), Pekin (Chine), 14-19 October 1996.