Ancient Initial Letters Indexing

Rudolf Pareti, Nicole Vincent
Laboratoire CRIP5, Université Paris 5
45, rue des Saints Pères 75270 Paris Cedex 06 France
rudolf@pareti.org, nicole.vincent@math-info.univ-paris5.fr

Abstract

Discrimination of images is necessary in many tasks, either understanding or indexing for example. Here we are concerned by indexing. More precisely we are working about initial letters extracted from early Renaissance printed documents as an application. Different observation levels can be considered according to the applications, either details can be observed or more globally what could be called the style. Here we are concerned with a global view of image. Then we are going to present a new method to index ornamental letters in ancient books. We show how the Zipf law, originally used in mono-dimensional domains can be adapted to the image domain. We use it as a model to characterize the distribution of patterns occurring in these special images that are initial letters. Based on this model some new features are extracted and we show their efficiency for image indexing and retrieval.

1. Introduction

Most of the books in the Middle Age period are reproductions of ancient and religious texts, realized in monasteries in which copyists work by dictation. The invention of printing led to the multiplication of the books that tried to attain the quality and the beauty of the ornamented presentation of previous period manuscripts. Books are rich and well ornamented.

Some artistic schools were created and some copies were done. Some illuminators copied one another and it is not rare an illuminator realized almost the same illuminations as some realized in other places. Nevertheless they differ by some details from the original one and a close look at the image is needed to conclude whether some initial letters come from the same printer or not. Besides, in spite of the similarities, the styles can be different and discriminated in a global way by the experts. Since a book contains several media and comprises both a support and content, the document interests many experts in different domains. The text in itself does not contain the same information as can provide the study of the illuminations and particularly the ornamental initial letters. Indexing these pictures will permit us to know whether a book has been made by the same artistic school, or if someone has cribbed from another artist. Even better, books are often damaged and questions arise: who is the author, when it has been written, is it an original version or some copy, what edition is being read. To answer these questions and many others, the fonts used in the book can be observed, the figures can be used too [1]. Here we are concerned with the initial letters. In the Renaissance period, the period we are interested in, the initial letters were black and white. Lots of methods exist to classify black and white pictures, the most simple relies on the use of the histogram and others rely on segmentations and pattern matching techniques [2]. We propose a method more robust than other existing methods. A statistical approach seems necessary to solve the problem. Indeed the structural approaches are still not mature enough to handle small differences we cannot model well. Then we have chosen an approach that would use a mathematical model to characterize the distribution of patterns in an image. The Zipf law model we are using is defined by different parameters. This model is performing well in many observed 1D phenomena, and we apply it to this type of pictures.

In a first part we are going to recall the statement of Zipf law, as developed in the study of 1D signal, then we will show how to test it for the purpose of image analysis and more especially for the specific images we are studying. Finally results will be presented.

Our process can be divided into two steps, the computing phase and the comparison phase. In the first step, we are going to characterize each image of a data set and store the indexes in a database. Then in the second step, we are going to take any ornamental initial letter and recognize its style according to the classifier developed. More precisely the reference images most similar to the unknown image are extracted. Therefore the output of the system comprises two results one is the most believable style

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

of the initial letter and the second is a list of 5 similar images.

2. Zipf Law and images

2.1. Statement

Zipf law is an empirical law expressed fifty years ago [3]. It relies on a power law. The law states that in phenomena figured by a set of topologically organized symbols, the distribution of the occurrence numbers of n-tuples named patterns is organized in such a way that the occurrence frequency of the patterns M_1, M_2… M_n, noted N_1, N_2... N_n, are in relation with rank of these symbols when sorted with respect to their occurrence frequency. The following relation holds:

$$N_{\sigma(i)} = k \times i^a \quad (1)$$

$N_{\sigma(i)}$ represents the occurrence number of the pattern with rank i. k and a are constants. This power law is characterized by the value of the exponent a. k is more linked to the length of the symbol sequence studied. The relation is not linear but a simple transform leads to a linear relation between the logarithm of N and the logarithm of the rank. Then, the value of exponent a can be easily estimated by the leading coefficient of the regression line approximating the experimental points of the 2D graph ($\log_{10}(i)$, $\log_{10}(N_{\sigma(i)})$) with *i=1 to n*. Further on, the graph is called Zipf graph. One way to achieve the approximation is to use the least square method. As points are not regularly spaced, the points of the graph are re-scaled along the horizontal axis.

The validity of this law has been observed in many domains but rather for mono dimensional signals [4].

In order to study images, we are going to adapt the concepts introduced in the statement of Zipf law to two dimensional data.

2.2. Images

In this section, we are to point out some problems that may occur when images are concerned.

In the case of the mono dimensional data, the mask was limited to successive characters. When images are concerned, the mask has to respect the topology of the 2D space the data is imbedded in. The natural choice is to use 3x3 mask as neighborhood of a pixel in a 2D space.

Then the principle remains the same, the number of occurrences of each pattern is computed. Nevertheless as 256 symbols are used to code pixels, there are theoretically 256^9 different patterns. This number is much larger than the number of pixels in an image. Indeed, if all patterns are rare, the model deduced from Zipf curve would not be reliable, the statistics would loose there significance. For example a 640x480 image contains only 304964 patterns. Then it is necessary to restrict the number of perceived patterns to give sense to the model. The coding is decisive in the matter.

Then several problems have to be considered. What are the properties we want to make more evident? How many classes of patterns are to be considered? This will be solved through a new encoding of the image.

According to the coding process and to the image content, Zipf curve general shape can vary a lot. When it differs from a straight line, the model of a single power law is not suited for the global image modeling. Nevertheless, if several straight segments fit the curve we can conclude several phenomena are mixed. Different codings allow to make more evident different properties of an image.

3. Characterization of the image

In this section the aim is to find the most suitable encoding way in order to define indexes apt at discriminating initial letters style. This would qualify as effective a coding process. As already mentioned, we are not interested in recognizing the nominal letter itself.

Some studies have shown Zipf law was holding in the case of images with different encoding processes [5]. Two images that look alike from a style point of view should be modeled by similar power laws.

The ornamental initial letters we are studying have been scanned as grey level images. Each pixel intensity is encoded with 8 bits (256 different levels).

According to the previous remarks, the number of different symbols must be decreased. Two ways to achieve the decrease are possible, either the number of symbols used to characterize the pixel is decreased or the number of the pixels involved in the pattern is diminished.

A simple way would be to consider only k grey levels to characterize the intensity level of the pixels. Most often it is sufficient to observe an image. A quantization in k equal classes would led to unstable results, so we have chosen to use a clustering of the grey levels in k classes by way of a k-means algorithm. Further the method relying on this coding process will be called k-mean. We have experimented different values of k.

An other way to decrease the number of different patterns is to limit the number of pixels in the pattern. To remain coherent with the 2D topology we have chosen to consider a smaller neighborhood of the pixel, it defines 4-connectivity. It is precised in figure 1. In this case we have too achieved a drastic decrease in the

number of grey levels as we have considered only 3 levels. This number is in fact issued from the nature of the images we are working on. They are rather black and white images.

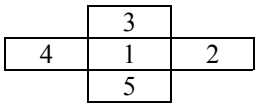

figure 1. 5 pixel mask associated with pixel number 1

A k-means clustering with k equal to 3 has been applied on each image. The number of possible patterns is therefore equal to 3^5=243, that is about the same as the initial number of grey levels but the information contained in the values is more local than ponctual. The k-means classification makes the method independent of the illumination of the scanned image and the printing conditions.

4. Zipf curve construction

Whatever the encoding process used, Zipf curves can be built. An example is shown in figure 2.

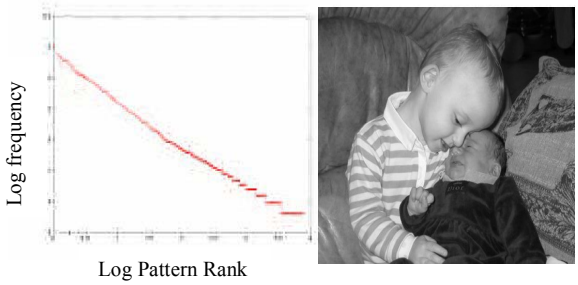

Figure 2. Zipf curve associated with an image

Now in order to study a family of images, these plots have to be compared. A close look at the curves shows they are not always globaly linear, that is to say Zipf law does not hold for the whole patterns. It depends on the coding process. Nevertheless some straight line segments can be observed. According to the coding process used these zones can be interpreted. In the case of a k-means applied to grey levels, the left part of the graph is concerned with the regions in the image whereas the right part gives information on the contours present in the image. Indeed, the area of the regions is larger than the area of the contour pixels, then the corresponding patterns are more frequent. Thus, we can extract on the one hand some structure indication of the regions and on the other hand the structure of the contours within the images.

Then we have chosen to consider in each curve up to three different linear zones. They are automatically extracted as shown in figure 3 using a recursive process. The splitting point in a curve segment is defined as the furthest point from the straight line linking the two extreme points of the curve to be split. We can say the image carries a mixture of several phenomena that are highlighted by the process. Several power laws are involved and then several exponent values can be computed.

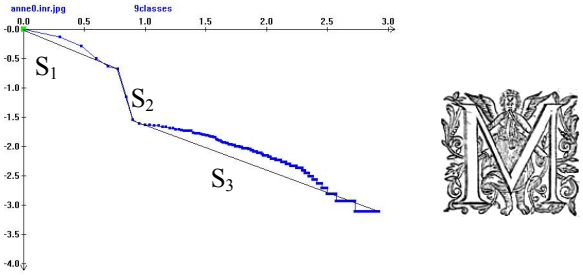

Figure 3. Initial letter example and its Zipf plot where are indicated the different straight zones extracted

Then the output of the process is made of 3 meaningful values associated with the picture (the ornamental initial letter figure 3). They correspond to the 3 slopes, leading coefficients.

5. Implementation

5.1. Indexing

We experiment on a data base made of more than 200 images coming from the **C**entre d'**E**tude **S**upérieur de la **R**enaissance of Tours. We have indexed all of them using a k-means quantization with k=3, our pattern is a 3x3, one the most usually used neighborhood in a 2D space. In order to verify the efficiency of the exponents associated with the Zipf models presented, we calculate all the ornamental initial letters Zipf curves. To have more efficient results we decide to apply an histogram normalization filter on the initial images. It leads to a better use of the image spectrum. Each initial letter image is represented by 3 numbers corresponding to the 3 slopes S1, S2 and S3 defined in the previous section.

Besides, each initial letter has been labeled by the experts who indicates, let us say the "style" of the illumination. In our database we count three different styles and some that cannot be classified because their number is not significant enough.

Then our system can be used in many different ways:

• From a request initial letter the user can ask for n most similar ornamental letters contained in the data base.

• With a new initial letter of an unknown style, the system can indicate the corresponding style; the decision is relying on the styles occurring in the nearest neighbors and their numbers.

• Finally the system can be used on a set of initial letters which have not been labeled in order to define the clusters and may be compared with some already expertised initial letters. Achieving these different goals relies on the choice of a distance between initial letters. We have chosen the Hamming distance in the parameter space. In figure 4 we present an illustration of the first application. Here six images are presented.

Figure 4. Images most similar to the request image

5.2. Evaluation

Indeed, evaluation of an indexation system is very subjective. Then we prefer to evaluate our proposal refering to the expert labels that are the "styles". In table 1 we have indicated results obtained through a cross validation approach.

Table 1. Result of "style" determination with 3-mean quantization

N	Style 1	Style 2	Style 3
1	92%	70%	84%
3	100%	94%	91%
5	100%	100%	97%

The percentages are given within each style. We made vary the number N of nearest images.

More samples of style 1 are present in the base that explains the better rate nexvertheless the nature of our indexes makes the method efficient in the other cases.

In figure 5 we show an example of each of the 3 styles concerned in the experiment.

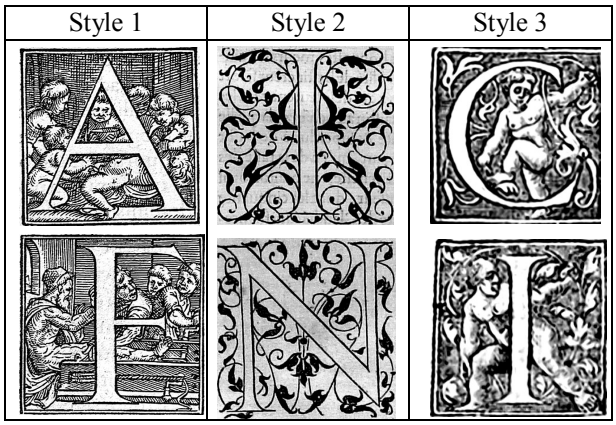

Figure 5. Initial letter styles

6. Conclusion

Here we show the modeling of pattern distribution is more efficient than histogram use in the field of image indexing. Zipf law allows to define global parameters based on details. According to the type of encoding used, the nature of information differs. Other encoding processes can be experimented and the method can be applied to other problems.

8. References

[1] J.P. Eakins, Content base image retrieval – can we make it deliver ? 2nd UK Conference on image retrieval, Newcastle upon tyne, February 99

[2] K.Melessanaki, V.Papadakis, C.Balas, D.Anglos, Laser induced breakdown spectroscopy and hyper-spectral imaging analysis of pigments on an illuminated manuscript, Spectrochimica Acta Part B 56 (2001) 2337-2346

[3] G.K. Zipf, Human Behavior and the Principle of Least Effort, Addison-Wesley, 1949

[4] E Dellandrea, P Makris, N Vincent (2004), "Zipf analysis of audio signals" Fractals, 12(1):73-85.

[5] Y. Caron, H. Charpentier, P. Makris, N. Vincent, Power Law Dependencies to Detect Regions Of Interest, DGCI 2003, Naples, Italy, November 2003