Global Method Based on Pattern Occurrences for Writer Identification

Rudolf Pareti, Nicole Vincent

Global Method Based on Pattern Occurrences for Writer Identification

Rudolf Pareti *Nicole Vincent*

Laboratory Crip5-Sip
Université René Descartes Paris 5
45, rue des Saints Pères 75270 Paris cedex 06
rudolf@pareti.org
nicole.vincent@math-info.univ-paris5.fr

Abstract

Writer identification is a difficult problem, it can be considered at different levels according to the applications, details can be observed or more globally the general aspect of a writing. We are going to present a new method to index and identify manuscript and other handwritten texts in ancient documents. Our method is based on Zipf law, originally used in mono-dimensional domains. Here same kind of relation is looked for in the image analysis domain. We use it as a model to characterize the distribution of patterns occurring in these special images that are handwritten texts. Based on this model some new features are extracted and we prove their efficiency for writer identification task.

Keywords*: writer identification, global method, power law*

1. Introduction

Manuscripts can be considered as a set of characters arranged in non random way and in this order of characters lays the sense of the text. Each person has his own writing style more or less linked to what has been impacted in the leaning phase. Upper strokes are more or less stressed, in more or less parallel directions with respect to the page position. The writing may figure a reflection of the personality, the education and the culture. Even without reading, a person is able to tell whether two documents are from the same writer or not. Writer identification is not a modern problem; nevertheless, today it is really important to authenticate a document. Digitization and technical progresses enable easy forgeries of documents and writer authentication may be a way to solve the problem. Security is not the only purpose of writer identification. Since the History times handwriting is used to the communication between humans. In ancient documents other problems occur as letters may be deteriorated by time effect and some information as the author may be lost. Some documents never contained this information for example an act play or a letter to a friend. The global style of writing gives more information about who the author is than the text content itself and it enables to show if a document is an original or a copy.

The experiments associated with our method rely on a database of manuscripts we know the identity of the scripter. Then the set of documents are divided into two parts. One that will be used as a learning base in which we know who has written the documents and a second one used as a test base. The goal is to identify the writers with a method relying only on the image analysis and on indexes computed from the first part of the database.

2. Existing methods and approaches

Many studies are tackling writer identification. Here we recall the main approaches. They can be classified in three groups, contextual approaches, non contextual ones and those using a style characterization to identify the writer.

2.1. Contextual approaches

These approaches use the text image but also the text semantic [1-2]. These approaches are too binding in our case and need to ask the writer to write the proper text. In our case document writers are no more present. Besides the documents are very heterogeneous with respect to the content so these methods do not fit our problem.

2.2. Non contextual approaches

These approaches impose fewer constraints on the writer than the contextual ones. They most often rely on the study of a histogram [3], on an analytic description of the writing [4], on neuronal-networks [5]. These approaches implement local methods. They do not take into account that humans do not need to analyze or read the text in a precise way in order to recognize a writer.

2.3. Identification by the style

The goal of these methods is to extract features from a text apart from the semantic. A set of parameter is used in [6] and the style is used to adapt a HMM model in the recognition process. Others extract invariant shapes in the writing that can be used either to adapt recognition process or to authenticate writers[7-8]. Some fractal studies have been performed to classify writers [9] or to identify them. Our own research takes place in this category of approaches.

3. The methodology

Before we come to the application itself we are going to recall the statement of Zipf law, how images can be processed and up to what degree the approach can be used for our purpose.

3.1. The Zipf law

Zipf law is an empirical law expressed fifty years ago [10]. 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 not organized in a random way. It can be observed that the apparition frequencies of the patterns M_1, M_2… M_n, we note N_1, N_2... N_n, are in relation with rank of these symbols, when sorted with respect to their decreasing occurrence frequency. The following relation can be observed:

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

$N_{\sigma(i)}$ 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 studied symbol sequence. 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. The various computations are then made easier. The value of exponent a can be 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.

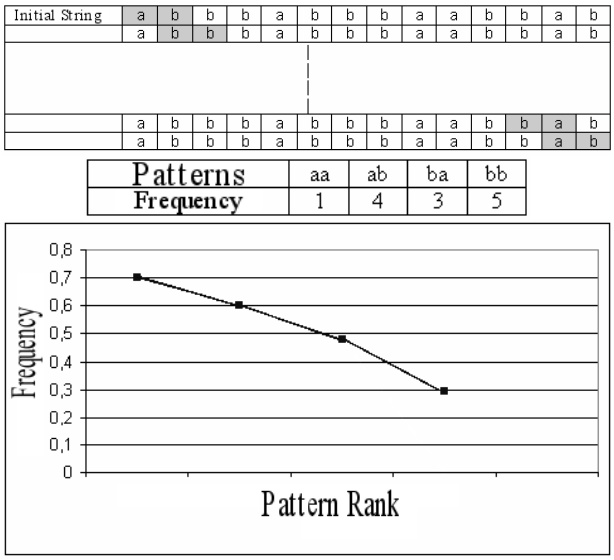

Fig. 1. Zipf curve building

The validity of this law has been observed in many domains but rather for mono dimensional signals [11].
In order to study images, we are going to adapt the concepts introduced in the statement of Zipf law to two dimensional data.

In order to study manuscripts, we are going to look whether Zipf law holds for such images. Therefore to adapt the concepts introduced in the statement of Zipf law have to be adapted to two dimensional data.

3.2. Application to images

The hand written texts we are studying have been scanned as grey level images where each pixel is encoded with 8 bits (256 different levels). The intensity is the information encoded. In spite of the black and white nature of the text, we have preferred to keep grey level information to achieve the study. The noisy paper and the deformation due to ancient paper need this precision. Indeed, the relative grey levels are more significant than the absolute values.

In this section, we are to point out some of the differences occurring when processing 2D-images. In the case of the mono dimensional data, the observed patterns were contained in masks limited to successive symbols. 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 masks. They figure some neighborhood of a pixel in a 2D space.

Then the principle remains the same. A relation concerning the number of each pattern is looked for. Nevertheless as 256 symbols are used to code pixels, the grey levels, there are theoretically 256^9 different patterns. This number is much larger than the number of pixels in an image. All patterns happen to be rare and the frequencies would not be reliable, the statistics would lose there significance. For example a 640x480 image contains only 304 964 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 in order to label the patterns with a reasonable number of tags. 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.

3.2.1. Coding of the patterns

Some studies have shown Zipf law was holding in the case of images with different encoding processes [12]. The images used were landscapes or large sceneries. Here we are studing hilghly non natural images as they figure writing, a typicaly human made concept. We are looking for coding process that gives models apt at discriminating the images we are studying. In our case, this would qualify as effective a coding process. Two writings that look alike from a style point of view should verify similar distribution of the patterns. We are going to present different coding methods we have tested.

According to the remarks previously done, the number of different possible patterns must be decreased.

This can be done decreasing the number of pixels involved in the mask or decreasing the number of tags associated with a pixel and its neiborhood.

The general ranks

Here our motivation is to respect the vision of a scene that relies more on grey levels differences than on the absolute values. So, within the mask, the grey level values are replaced by their ranks when sorted according to the grey level values. The method affects the same rank when the grey levels have the same value. Then the maximum number of values involved in the mask is 9 and this leads to a very large decrease in the number of different possible patterns.

Image Pattern 1:

2	8	6
21	31	31
32	32	32

Image Pattern 2:

130	136	134
149	159	159
160	160	160

Coded pattern:

0	2	1
3	4	4
5	5	5

Fig. 2. patterns coded using general rank method

The number of symbols used to represent the grey levels is limited to 9. It can be noticed that the patterns in figure 1 (a) and (b) are different, yet, the coded pattern (c) that is generated is the same in both cases. It is one of the limitations of this coding. Of course information is lost. Further the method relying on this coding process will be called general rank method.

3.2.2. Grey level quantization

The previous method has led to 9 possible values associated with a pixel in the pattern considered. A simpler way would be to consider only k grey levels to characterise the intensity level of the pixels. When k is well chosen, it is sufficient to observe an image. More over the images we are dealing with are essentialy black and white. A quantisation in k equal classes would lead to unstable results, so we have chosen to use a classsification method of the grey levels into k classes by way of a k-mean algorithm. Further the method relying on this coding process will be called **k-mean** [13]. We have experimented different values of k. In figure 2 we are presenting an example with 9 clusters.

Grey level	**0-20**	**21-76**	**77-96**	**97-120**	**121-146**	**147-174**	**175-203**	**204-229**	**230-255**
center	15	56	86	107	133	159	190	217	242
Class	0	1	2	3	4	5	6	7	8

Fig. 3. K-mean example

The example depends on the images and considers 9 classes. We can see that some classes may be smaller than others. We experimented that 3 classes were giving the best results.

Cross mean method

An other way to decrease the number of different patterns is to limit the number of pixels in the mask. 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 4.

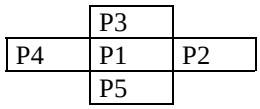

Fig. 4. Five pixel mask associated with P1

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. A k-mean with k equal to 3 has been used 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-mean classification makes the method independent on the illumination of the scanned image and the printing conditions. Further, the method relying on this coding process will be called **crossmean method**.

The method we are proposing is invariant under the geometrical transforms that leave invariant the shape of the mask. The dependency to change of scale that can occur when images are scanned in different conditions is intrinsically linked to the method itself but the statistical aspect of the method makes the approach rather robust.

3.3. Zipf curve construction

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.

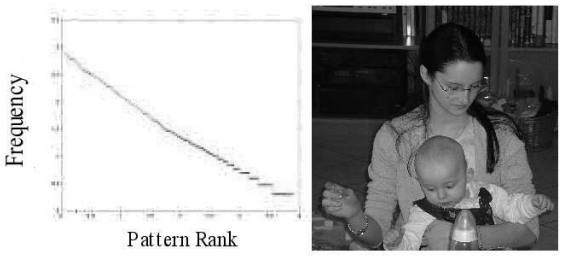

Fig. 5. Zipf curve associated with an image

Whatever the encoding process used, even though Zipf law does not hold, Zipf curves can be built. An example is shown in figure 5. Now in order to study a family of images, these plots have to be compared. A close look at the curves obtained when processing images of handwritten texts shows they are not always globaly linear. That is to say Zipf law does not hold for the whole image. Several structures are involved in the images. Of course, they depend on the chosen coding process.

Nevertheless Zipf curve can be approximated by some straight line segments. According to the coding process used these zones can be interpreted. Some parts may refer to regions in the image whereas other parts give information on the contours present in the image. Indeed, some structures take a larger part in the image than others. It depends on the characteristics of the writing, Then the corresponding patterns are more frequent. Thus, we can extract some structure indication at different levels 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 6 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.

The Zipf curve has been drawn with respect to a logarithmic scale, therefore, it is necessary to begin with a resampling of the curve in order to have a fare regression approach with the least square method.

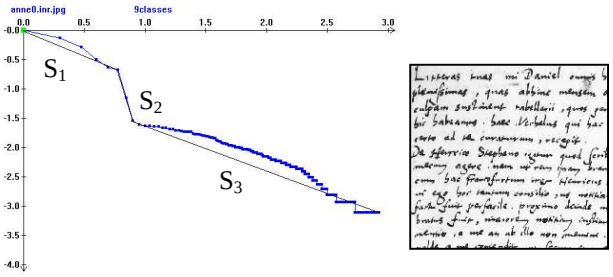

Fig. 6. Manuscript example and its Zipf plot where are indicated the different straight zones extracted

Then the output of the process is made of 3 meaningful values associated with the picture (the manuscript figure 6). They correspond to the 3 slopes, leading coefficients.

4. Similarity measure between texts

Here we are going to define a measurement between handwritten texts and to present the results of the experiments we performed.

4.1. Comparison

According to the previous study, we have decided to index the handwriting images with the three exponent values extracted from the three power laws highlighted in the model. The exponents are not sufficient to characterize the writer as the different structure may take more or less place in the writing of each one. The same structure can be linked to a more or less important number of patterns. Then we are taking into account the length of each segment extracted in the Zipf curve. Then an image and may be a writer is assumed to be characterized either in a 3 or in a 6 dimensional space. The three slopes and the abscissa of their extremities are considered. These abscissa are linked to the rank of ghe patterns whereas the vertical axis gives information on their frequency that is widely depending on the the image size. Then abscissa is more reliable.

In any case we have chosen the Hamming distance in the parameter space to compare two images.

$$\text{distance}(I, I') = \sum_{i=1}^{d} \left| S_i - S'_i \right| \qquad (2)$$

4.2. Evaluation

Our system can be used in two different ways:

- From a request manuscript the user can ask for n most similar manuscripts contained in the data base to be extracted.
- With a new manuscript of an unknown style, the system can indicate the corresponding writer, then the decision is relying on the text writers occurring as nearest neighbors in our database.

4.2.1. image retrieval

In figure 7 we present an illustration of the first application. Here five images are presented.

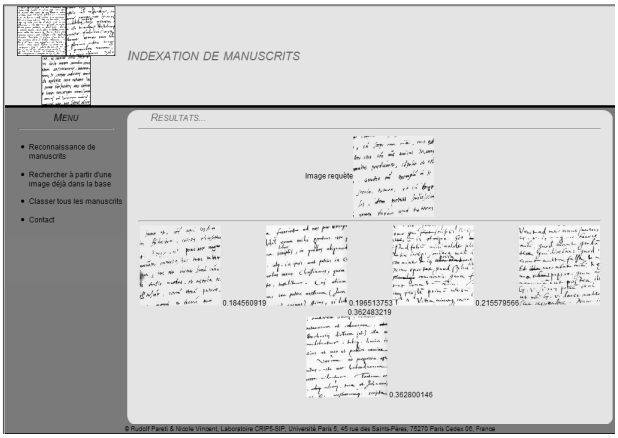

Fig. 7. Images most similar to the request image

4.2.2. Writer identification

The evaluation of the identification method is largely depending on the documents present in the database. Here the documents all date from the sixteenth century and the writer styles differ much less than now a day. We obtain a writer identification rate of **62%** as we show in table 1 with the k-mean method (k=3) when considering only 3 parameters.

Table 1. Writers identification rate in regard to a k-ppv classification

K=	2	3	4	5
Rate	55%	62%	58%	50%

But analyzing the manuscript we observe some errors due to manuscripts from different writers can be very close in style as we show in Figure 7.

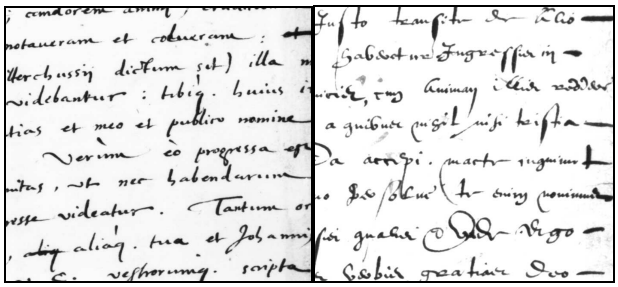

Fig. 7. Part of two letters from two different writers

When we look at the slopes in the Zipf graph we see that they are also very close. The difference between these two graphs comes from the different lengthes of the segments. Then, we have different results when we take into account not only the leading coefficient but also the break point values of each segment. It adds to our features characterising the handwritten text, three values. On figure 8 the breaking points extracted form Zipf graphs of two different documents are indicated. We can see the first segment is longer in the first case than in the second one.

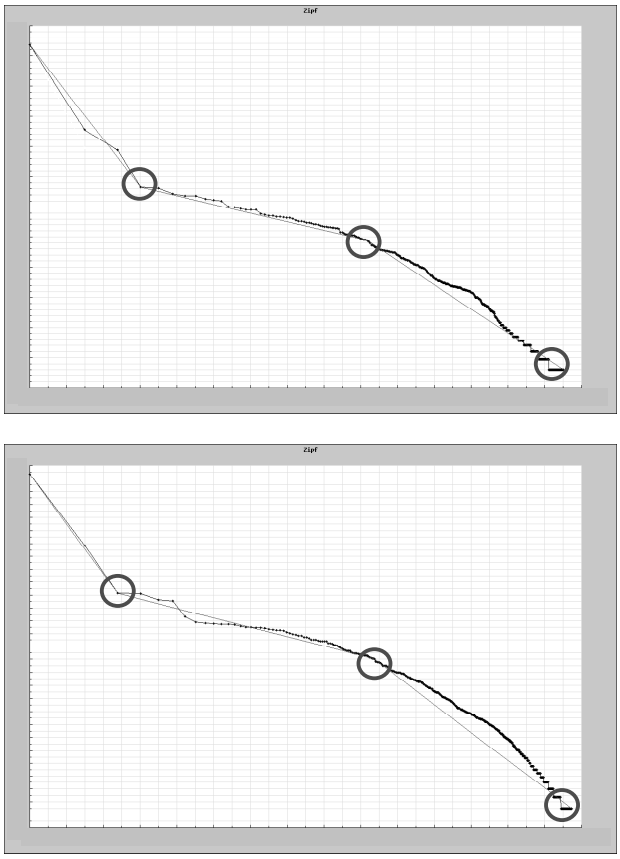

Fig. 8. Zipf curves from two different documents

Break point values can seem to be very close but the two images are very close and the differences between the values are now sufficient to discriminate between the two writers. In table 2 the numerical results with respect to these images are indicated.

Table 2. Images Features

	Slope 1	Slope 2	Slope 3	Break 1	Break 2	Break 3
Img 1	-3,9	-0,71	-2	3,61	2,12	0,78
Img 2	-3,85	-0,73	-2,1	3,05	2,72	0,52

The use of these new features in our first application gives more efficient results as we can see in the figure 9 in the case of image retrieval.

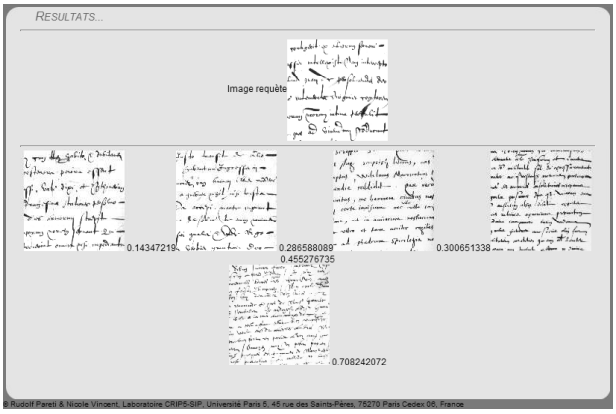

Fig. 9. Images most similar from the request image

This method seems to be more discriminating. Some troubles we have seen before disappear, for example if two writers have the same gap between lines the application detects more easily that it isn't the same writing. Analyzing the pattern and their location we notice that the pattern linked to the first slope segment correspond to background pixels and the two others represent the outline pixels.

In our second application the results are more efficient too, as we can see in table 3.

Table 3. Writers identification rate in regard to the k-ppv classification

K=	2	3	4	5
Rate	60%	**80%**	61%	57%

5. Conclusion

The results are not as good as could be though but they are encouraging. The documents we have used are not fare writings but they are degraded by time, and the process includes scan of photos.

Here we show the use of a model developed in the field of 1D phenomena can give good results in case of images. This 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 involving manuscripts. The method is invariant under any rotation. So in our case it's really important because letters and writing documents can come from different sources with different resolutions, more or less digitalized in good conditions.

Of course these global parameters can be mixed with others coming from different approaches and fusion can be made either at parameter or decision levels. This would lead to even better results.

References

[1] F. Mihelic, N. Pavesic, L. Gyergyek, "Recognition of writers of handwritten texts", International Conference On Crime Countermeasures, p 237-240 1977

[2] R.-D. Naske, "Writer recognition by prototype related deformation of handprinted characters", ICPR New New –York 1982 p 819-822

[3] B. Arazi, "Handwriting identification by means of run-length measurements", IEEE Transactions on Systems, Man and Cybernetics, SMC-7, n°12, p878-881 Dec. 1977

[4] W. Kuckuck, B. Rieger, K. Steinke, "Automatic writer recognition", Proceedings of the 1979 Carnahan Conference on Crime Countermeasures, Lexington, Kentucky, USA p 57-64 1979

[5] U.-V. Marti, R. Messerli, H. Bunke, « Writer identification using text line based features » ICDAR 2001 USA p 101-105

[6] M. Gilloux, "Writer adaptation for handwritten word recognition using hidden Markov models", ICPR 1994 USA p 135-139 vol. 2

[7] A. Nozary, L. Heutte, T. Paquet, Y. Lecourtier, "Difining writer's invariants to adapt the recognition task", ICDAR '99, Bangalore, India, pp. 765-768. 1999.

[8] A. Bensefia, A. Nozary, L. Heutte, T. Paquet, "Writer identification by writer's invariants" IWFHR'02, Niagara on the Lake, Canada, pp. 274-279, 2002.

[9] N. Vincent, V. Bouletreau, R. Sabourin,H. Emptoz, "How to use fractal dimensions to qualify writings and writers", Revue Fractals, World Scientific, Vol 8, n°1, p 85-97, 2000.

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

[11] E Dellandréa, P. Makris, N. Vincent, "Wavelets and Zipf Law for Audio Signal Analysis", 7th International Symposium on Signal Processing and its Applications (ISSPA 2003), Paris (France), Vol. 2, p. 483-486, Juil. 2003.

[12] Y. Caron, H. Charpentier, P. Makris, N. Vincent, Power Law Dependencies to Detect Regions Of Interest, 11th International Conference DGCI 2003, Naples, Italy, November 2003.

[13] J. A. Hartigan, M. A. Wang, "K-mean clustering Algo", JSTOR revue p 100-108.