

# **Matrix Frequency Analysis<sup>1</sup>**

**And**

## **Its Applications to Language Classification of Textual Data for English and Hebrew**

**Joseph Uchill<sup>2</sup> and Amir Assadi<sup>3</sup>**

**Introduction.** The advent of the internet has opened a host of new and exciting questions in the science and mathematics of information organization and data mining. In particular, a highly ambitious promise of the internet is to bring the bulk of human knowledge to everyone with access to a computer network, providing a democratic medium for sharing and communicating knowledge regardless of the language of the communication. The development of sharing and communication of knowledge via transfer of digital files is the first crucial achievement in this direction. Nonetheless, available solutions to numerous ancillary problems remain far from satisfactory. Among such outstanding problems are the first few fundamental questions that have been responsible for the emergence and rapid growth of the new field of Knowledge Engineering, namely, classification of forms of data, their effective organization, and extraction of knowledge from massive distributed data sets, and the design of fast effective search engines. The precision of machine learning algorithms in classification and recognition of image data (e.g. those scanned from books and other printed documents) are still far from human performance and speed in similar tasks. Discriminating the many forms of ASCII data from each other is not as difficult in view of the emerging universal standards for file-format. Nonetheless, most of the past and relatively recent human knowledge is yet to be transformed and saved in such machine readable formats. In particular, an outstanding problem in knowledge engineering is the problem of organization and management--with precision comparable to human performance--of knowledge in the form of images of documents that broadly belong to either text, image or a blend of both. It

---

<sup>1</sup> Copy Right (University of Wisconsin-Madison). Patent Pending.

<sup>2</sup> Department of Computer Science, University of Wisconsin Madison, WI 53706 USA. E-mail jhuchill@students.wisc.edu

<sup>3</sup> Department of Mathematics, and Department of Medical Physics, University of Wisconsin Madison, WI 53706 USA. E-mail assadi@math.wisc.edu

was shown in [11] that the effectiveness of OCR was intertwined with the success of language and font recognition.

The objective of this paper and ensuing studies is to contribute to knowledge engineering of documents that are images from printed text. We have made a fundamental departure from traditional machine-learning view of data mining from texts and images. Namely, we propose to view the entire potential collections of printed text as a *space* (potentially curvilinear) with metric properties that are derived, in the first place, from human perception of images. Subsequently, they are subject to linguistic and cognitive constraints of viewing images in a local-to-global process metaphorically similar to “reading”, but not necessarily bound to the average human habits of reading. This paradigm is particularly useful for effective classification of very large heterogeneous data sets into hierarchies, thus reducing the search spaces for other algorithms to a bare minimum possible from the image properties of text documents. In the present paper, we introduce a simple, yet powerful method, to study “*the Universal Text Space*” (UTS), and describe some of its basic properties. This method, which we term Matrix Frequency Analysis, is then applied to the geometry of UTS and to knowledge engineering.

## **Prior Work in Script and Language Recognition**

Spitz [3] developed a three stage algorithm to differentiate between languages. This technique differentiated between Asian and European character sets by vertical frequencies of upward concavities (runs of pixels which spanned the distance between two unconnected runs of pixels which were immediately above them). European languages were then separated from each other using word shapes. This method, used in a prior work by Nakayama and Spitz [4], classified characters by their general peaks and lows, and used the order of these characters within separate words to identify language. Further separations were performed by analyzing the optical density of the Asian characters. Optical density was also used in a previous work by Spitz [5] to separate Japanese from English.

These methods each required preprocessing segmentation to separate characters, and normalize them to the same line. Ding et al[6] also noted that vertical concavity showed anomalies when faced with Cyrillic, however with a multiple stage method like that of their study or of Spitz’s study these anomalies were ironed out.

A variety of token techniques have been constructed aside from the word shape tokens of Spitz. Hotchberg et al. [7] used character templates characteristic to each script to identify various languages. Effective even in images containing both text and image, this characteristic shape identification method needs minimal training for effective script detection. However, token techniques offer less emphasis on analysis than other techniques, and are more focused on the goal of script identification.

Wood et al. [8] used projection profiles to identify scripts. These studies totaled the number of pixels for each y value in graphed line of text. Graphs showed unique traits for separate scripts. Large-scale normalization was required, including making sure lines of text were of sufficient length to obtain a reasonable distribution, and standardization of characters' placement on a line.

Chaudhury and Sheth [9], as well as Peak and Tan [10], used Gabor filters to identify script and language. Normalization in these studies involved text-padding to standardize lines of text to a reasonable length, and uniformity of a character's size.

The goal of *Matrix Frequency Analysis* (MFA) is to use easily available, yet fundamental, statistical information to evaluate an image. Viewed as a matrix, every image is hierarchically composed of numerous smaller submatrices. This idea has been successfully exploited in vector quantization algorithms, but is yet to be used as a source of additional information about an image. The general form of MFA attempts to provide a hierarchical organization of image information encoded within the data in n-by-n matrices. It turns out that for the study in this paper, the portion of information in MFA depending on four by four matrices of pixel values is sufficient. Since text has only two colors, there are 65536 possible monochrome four by four matrices. A random image composed of a random selection of pixels will exhibit a distribution of these matrices that is most likely uniform, provided we have shown no biases in selecting the values of the pixels. In contrast, a conventional image file will exhibit a distribution that draws matrices from a small subset of these 65536. Informally speaking, conventional images have a vastly uneven distribution of submatrix constituents. This distribution maintains certain constancy features over images of similar type; for instance, it is invariant under change of image scale.

**Parameterization of Textual Data by Geometric Structures.** In this section, we review some results from Assadi-Uchill [12] in the appropriate context. The technical details are relegated to the reference [12]. In many areas of mathematics and recently, in several

branches of theoretical physics, topological methods and global geometric structures have proved to be indispensable. We shall adapt this point of view to image processing and knowledge engineering, and show how to take advantage of this geometric outlook. Let us briefly review the relevant aspects of an informal language from which the somewhat unfamiliar geometric constructions and concepts are derived. To study a particular object of interest, for example the trajectory of movement of a particle between two points A and B subject to an external force field and minimizing the work, it is useful to consider a much more general space  $X(A,B)$  whose points correspond to various paths between A and B. The geometric and topological properties of the space  $X(A,B)$  are crucial in finding the minimum of the function defined on  $X(A,B)$  that encodes the physical condition of the problem. Thus, first we represent each object of interest (namely a path between A and B) by a point in a space  $X(A,B)$ . Second, we encode the geometric relationship between different paths, say in terms of their closeness to each other, by means of a topology or a metric structure on  $X(A,B)$ . Third, we construct the function  $f$  on  $X(A,B)$  that assigns to a point  $P \in X(A,B)$ , the work for the particular path that is represented by P, thereby defining a function compatible with the metric or topological structure of  $X(A,B)$ . At this point, the problem of finding a path from A to B along which work is minimized is translated to the problem of finding the point of P belonging to  $X(A,B)$  for which the minimum of the function  $f$  is achieved. The latter problem is well-known in calculus of variation or even just calculus of function with several variables if we are willing to accept an approximate solution. This method, in a more sophisticated form, goes back to Bernhard Riemann who studied the entire collection of complex structures on a surface of given topological type. Riemann called the parameters that described uniquely the points on the space of complex structures as the *moduli*, and the space of the complex structures was called the *moduli space*.

Since our problem is closer to the situation that Riemann studied, we provide a brief account of Riemann's method of moduli. The complex structure on a torus is responsible for the geometry of the torus as a metric surface; e.g. the way angles between two curves would change if one transforms the surface with the same rule (i.e. topologically the same formulas). In this case, the complex structure is also known as the conformal structure, reminding us of the significance of the theory of conformal functions in complex analysis. Different complex (conformal) structures on the same torus can be realized as two dimensional real surfaces that can be brought in the three-dimensional space in a piece-by-piece manner, so that their partial realizations in the three-dimensional space looks quite different, depending on differences in their complex (or conformal) structure. The process of visualization of such surfaces in the three-dimensional space is not very

useful beyond a first experience with the metric view of parts of the surfaces. It is more useful to find a space whose points correspond to different complex structures on the surface. For example, the donut-shaped surface known as *torus* could be parameterized as the set of points in a square  $S$  spanned by the standard basis  $\{e_1, e_2\}$  whose opposite sides are “glued together” (that is, identified as the same points). Since the 2-dimensional plane is the same as the set of complex numbers, the torus inherits the structure of a complex manifold from the interior points of the square  $S$ , provided we pay attention to how the points in the boundary of the square are identified. This description of the complex structure for the surface is cumbersome, and at best subject to confusion. A more systematic method is to tile the entire plane with the copies of  $S$  that are translations of its set of its points by vectors  $pe_1$  and  $qe_2$  to different parts of the plane, in a way that the translated squares overlap only on their boundary (perimeter) points. Now, different tiling of the plane provides different complex structures for the torus in terms of their geometric properties. Roughly, tiling the plane by a square with sides  $\{e_1, e_2\}$  is equivalent to the tiling by  $\{u_1, u_2\}$  if there is a rigid motion that shows the parallelograms spanned by  $\{e_1, e_2\}$  and  $\{u_1, u_2\}$  are congruent. Informally, rigid equivalence of tiling is quite close to equivalence of complex structures, though some technical assumptions are needed to make this rigorous. As an analogy: the topological structure corresponding to the torus corresponds to the geometric structure  $\mathbf{X} = \mathbf{X}_{\text{English}}$  underlying images of documents written in the English language with its standard alphabet. The different complex structures on the torus correspond to the actual image spaces that occur in some large feature space that we construct from Matrix Frequency Analysis applied to given data files, and we call a “*textual structure*”. To make the analogy simpler, assume that the feature space is a Euclidean space of dimension  $d$ , and with coordinate axes  $\{t_1, t_2, \dots, t_d\}$ . Then a description of the textual structure on  $\mathbf{X}_{\text{English}}$  is provided by the probability density functions that are estimated from the  $d$  projections  $\{t_j : \mathbf{X}_{\text{English}} \rightarrow \mathbf{R} \mid j = 1, 2, \dots, d\}$  regarded as  $d$  random variables determined by the arbitrary choice of the document images given by the data file. These are called probability density functions ( $f_1, f_2, \dots, f_d$ ) (abbreviated to pdfs). This collection provides a point on a “*statistical manifold*”  $\mathbf{M}$  whose points are defined as occurrences of all such  $d$ -tuple of pdfs in each data file in English language, and printed in any font, after “completion”, a technical mathematical operation that takes the closure of the set of actual points  $(f_1, f_2, \dots, f_d)$  with respect to the appropriate information metric; the infinitesimal form of this appropriate information metric is the Fisher metric, such as the metric constructed from the Kullback-Leibler divergence). Thus, the operation of closure adds the limits of all convergent sequences of pdfs with respect to the information-theoretic metric to the sample of points  $(f_1, f_2, \dots, f_d)$  that estimate the pdfs from a given collection of data. The statistical manifold  $\mathbf{M}$  is an

interpolation of all actual textual structures on  $\mathbf{X}_{\text{English}}$  created by adding points that information-theoretically refine the actual estimates, and in terms of practical computation, it is essentially determined by any sample of data files that is sufficiently rich to reflect the image properties of the variety of texts that occur in the English language documents. In practice, classification of textual structures and recognition of such structures from among other textual structures and non-text images is subject to resolution limitation, that is, subject to bounds on computation errors that determine satisfactory versus unsatisfactory performance. Therefore, information-theoretically, the process of completion of the actual set of pdfs to its nearest manifold does not pose any obstacle for design of algorithms that are intended for real-world problems. The cost of higher precision and smaller error is simply accommodation of a much larger collection of samples, something that has to be decided empirically based on the concrete circumstances at hand. In summary, the statistical manifold  $\mathbf{M}$  is the moduli of textual structures on the image space  $\mathbf{X}_{\text{English}}$ , and mathematical study of geometric properties of the pairs  $\{ (\mathbf{X}_{\text{Language}}, \mathbf{M}_{\text{Language}} : \text{Language} = \text{English, Hebrew, Sanskrit, Cyrillic, ...} ) \}$  is the fundamental theoretical tool for design of algorithms in knowledge engineering and data analysis related to text documents that are presented as images.

**Matrix Frequency Analysis.** Assume that  $\Omega$  is a given class of monochrome images with common physical features. For instance,  $\Omega$  could be the class of English texts, or Sanskrit, or Hebrew texts. The *matrix frequency analysis* of files in  $\Omega$  is performed as follows. Each monochrome image is encoded as a matrix of size  $m \times n$  with values in a range of numbers, say  $F$ . In dealing with text, we assume the number of columns of the image matrices are fixed to be no greater than a maximum integer  $N_{\max}$  and no less than  $N_{\min}$ . The number of rows could be variable, corresponding to variable length of text. We define a “generalized window”  $W$  to be a collection of contiguous entries in the square matrix  $N_{\min} \times N_{\min}$ . *A priori*,  $W$  need not have any particular shape. What we are interested in is to select the best collection of windows that describe an estimate to the “*topological structure*” of the text files in the given collection  $\Omega$ , and we denote this set by  $W$ , suppressing all references to various intermediate notations. The reference to a *topological structure* for images is informal, and indeed, an abuse of terminology borrowed from geometry. The point is not so much as to find a basis for the topology of the structures in  $\Omega$ ; rather, to single out a collection of subsets of the digitized square regions that capture statistically the neighborhood and affinity structures of pixels in the images, and encode them in terms of a hierarchy of statistical manifolds that govern the laws of probability for

occurrences of the windows. For a class of discrete subsets of the corresponding distribution will belong to statistical manifolds that exhibit the probabilistic laws, their range of variability, and stochastic characteristics as they might occur from distribution of submatrices in collections of text document images, respectively from English, Sanskrit, or Hebrew. This observation leads us to conjecture that the Universal Text Space can be approximated by a geometric structure composed of pieces that are Riemannian manifolds, and that the Riemannian metric on each piece can be estimated from the statistical manifold that describes the range and variation of the probability density functions that account for occurrences of submatrix images in the textual data images. Accordingly, our task is to elucidate such geometric pieces and describe their geometry. The problem of design of algorithms for classification of textual data is, henceforth, translated into more familiar questions in computational geometry and information geometry, and amounts to instruction sets for calculation of basic Riemannian invariants of the constituent pieces.

**Advantages of Matrix Frequency Analysis.** Most methods required sizable normalization to succeed Algorithms like horizontal projection profiling and upward concavity require letters to be positioned on the same line, and lines of text to be of substantial length to provide accurate results. Pixel density requires letters to be of the same size in each instance. MFA works independent of scale, and independent of line positioning.

MFA operates using a random sample of submatrices. While larger samples provide exemplary results, smaller samples still provide good results. At 2000 matrices, MFA's success rate was over 98%. With a quarter of that sample size, MFA was only 6% less effective. Even a sample size of 2000 is miniscule. There are 88209 submatrices per square inch. Using the sampling methods described below there are 2742 submatrices in the letters 'MFA' written in 12 point Times New Roman at 300 by 300 dots per inch.

### **Relationship of MFA to Vector Quantization.**

Matrix frequency analysis falls under the broader category of Vector Quantization (VQ). In VQ, objects are compressed via a reduced codebook of vectors of clustered pixels. VQ has been used for token-based recognition, segmentation and low-level classification [1] in the past, but never for analysis of particulate information within an image.

While VQ analysis methods have, in the past, dealt with hunting for an expected vector, MFA is concerned with the frequencies of existing vectors. This allows for greater generalization of the method.

### Matrix Frequency Analysis Algorithm

The Matrix Frequency Analysis algorithm requires an image length and width. Combine each value along the X axis of the set  $\{X_1, X_2, X_3 \dots X_{(\text{width}-3)}\}$  with a value on the Y axis  $\{Y_1, Y_2, Y_3, \dots Y_{(\text{length}-3)}\}$  to form a set of every point on the image that can be the upper left corner of a four by four matrix. This yields a set of points of size  $(\text{length} - 3) * (\text{width} - 3)$ . For each point within this set view the four by four matrix for which this point is the upper left hand corner, and add one to the tally of the number of occurrences of this matrix.

Fig. 1

Matrix Frequency Analysis Algorithm

	A	B	C	D	E
a	X	X	X	X	X
b	X	X	X	X	X
c	X	X	X	X	X
d	X	X	X	X	X
e	X	X	X	X	X

(Assume we start from point (A, a).

- Use the current point as (1,1) in a new, 4 by 4 matrix.  
 The first matrix will look like this
 

A,a	B,a	C,a	D,a
A,b	B,b	C,b	D,b
A,c	B,c	C,c	D,c
A,d	B,d	C,d	D,d
- Add one to the tally of how frequently this matrix occurred.
- Increment to the next functional starting point (in the first case, B,a). There will be  $(\text{length} - 3) * (\text{width} - 3)$  matrices to tally.

The frequencies have a bias towards matrices of all white or all black. Since the matrices are small enough not to carry information of their own, every single tone component of an image provides large numbers of these all white or all black matrices. Since the size of the background of the informational portion of the image is transient, analysis programs can ignore the all white, or all black matrices. The remaining ratio of distribution provides a ‘fingerprint’ for an image. These fingerprint distributions provide a clue of the classification of the image. While fingerprints are not unique to images that provide them, they are characteristic to classifications of images.

## **Language Classification**

Fingerprints for English were found for large juxtaposed English texts taken from a combination of country names, the first chapter of *Alice in Wonderland*, The Declaration of Independence, The Constitution, The Gettysburg Address, the first chapter of Genesis and a circulating text file titled “The Hackers Handbook”, and fingerprints for Hebrew were taken from chapters in the JPS Bible in Hebrew. English is composed of numerous representations of the same printed characters. Serifs, the horizontal embellishments on the bottoms of some fonts, cause a structural difference in these fonts. Similarly, Hebrew has deferring forms of embellishment. To increase accuracy, the language classifier accounted for two models of English characters and two models of Hebrew characters. The UTS space that a language occupies may be split into non-contiguous regions.

Test text images were generated in six new fonts per language. These images used a separate text, *Around the World In Eight Days*. Fingerprints were generated using a small random sample of matrices where pixels were not all the same tone. Using 100 tests per font per sample size, a least-squares algorithm compared the test sample fingerprint to the two English and two Hebrew model fingerprints. If the test fingerprint was closer to one of the correct models, the test was scored as a success. As we show bellow, small samples of matrices can produce outstanding accuracy. With as little as 200 samples, some fonts had their language correctly identified 100% of the time. The combined effectiveness at this sample size was 85%, increasing the sample size to 500 boosted the effectiveness to 92.5%, and at a sample size of 2000 the effectiveness was over 98.4%. Accuracy varied on a logarithmic scale from sample size to sample size.

Table 2  
Percent Accuracy in Trials of MFA Language Classification Using Different Sample Sizes

	Matrices	200	500	2000
<b>English Font</b>				
1		74%	83%	94%
2		71%	84%	95%
3		79%	88%	99%
4		95%	99%	100%
5		74%	98%	100%
6		94%	99%	100%
<b>Hebrew Font</b>				
1		100%	100%	100%
2		75%	87%	94%
3		75%	83%	99%
4		100%	100%	100%
5		100%	100%	100%
6		83%	89%	100%
<b>Total</b>		85%	92.5%	98.41667%

### Effects of Scale

MFA needs little in the way of character normalization. Aside from skew detection, the image does not need any changes. While mechanisms for techniques such as runlength histograms require that letters be of the same, standardized size and reside on the same line, MFA does not require either. The effects of scale on using MFA were tested by analyzing the text of Lewis Carroll's *Jabberwocky* in various sizes, and finding the Cartesian distance between the fingerprints. There was only minimal distance between the most extreme difference in scale, size 9 font and size 72 font. These were closer together than all but the two closest fonts were when the same test was run for separate fonts in the same passage.

Table 3  
Relative Euclidian Distance between Fingerprints of the Same Passage in Different Fonts

Font	Ariel	Impact	Courier	Times
Ariel	0	3.3341	5.9242	0.7655
Impact	3.3341	0	12.0140	3.5448
Courier	5.9242	12.0140	0	9.2797
Times	0.7655	3.5448	9.2797	0

Table 4  
Relative Euclidian Distance between Fingerprints of the Same Passage at Different Font Sizes

Font Size	9	12	16	72
9	0	0.3207	0.6029	1.5457
12	0.3207	0	0.1536	0.6541
16	0.6029	0.1536	0	0.2975
72	1.545	0.6541	0.2975	0

Rationally, this can be explained by the effect of increasing the scale on a standard shape, like a square. As scaling increases, squares will always have four corners, but as the perimeter increases linearly, the area increases exponentially. The area, however, does not figure into the final analysis because matrices with only one pixel color are removed from the final fingerprint. Matrices whose occurrences do not increase at a linear rate and are included in the distribution (such as matrices corresponding to corners), are already dwarfed by the relative high frequency of the linearly increasing matrices (like those encoding the sides of a square).

## Conclusion

Consideration of the geometry of spaces of textual images leads to description of ensembles of text classes via representation of such image data using a multi-scale multi-resolution method. Rather than using the prevailing multi-resolution methods, we introduce a direct approach to statistics of textual image data using 4 by 4 matrices of pixel values in the image of the document. The mathematical approach to analysis of occurrences of such pieces of a text image is Matrix Frequency Analysis (MFA). We have shown that MFA provides effective information on the classification of files using small sample sizes. Since the information it provides is scale independent, there is only minimal need to prime an image before being used. We provided the results of a number of numerical experiments on different texts.

## Future Work

Matrix frequency analysis provides useful information for ‘on-the-fly’ VQ compression methods, and noise reduction. Matrix frequency is also being tested in classification of other types of data, such as environmental images.

## References

1. Cosman, P.C., Oehler, K.L., Risken, E.A., and Gray, R. M, “Using Vector Quantization for Image Processing,” Proceedings of the IEEE, Vol. 81, Iss. 9, Pages 1326-1341, 1993
2. Gray, R.M. and Neuoff, D.L., “Quantization,” IEEE Transactions on Information Theory, Vol. 44, Pages 2325 – 2383, Oct 1998
3. Spitz, A.L., “Determination of the Script and Language Content of Document Images,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, Iss. 2, Pages 235 -245, 1997
4. Nakayama, T. and Spitz, A.L, “European Language Determination from Image,” Proceedings of the International Conference on Document Analysis and Recognition, Pages 159 – 162, Oct 1993
5. Spitz, A.L., “Multilingual Document Recognition, “ Electronic Publishing, Document Manipulation, and Typography, R. Futura, ed. Cambridge University Press, pages 193-206, 1990.
6. Jie Ding, Lam, L., and Suen, C.Y, “ Determination of Oriental and European Images Using Characteristic Features,” Proceedings of the Fourth International Conference on Document Analysis and Recognition, Vol. 2, Pages 1023-1027, 1997.
7. Hotchburg, J., Kelly, P., Thomas, T., Kerns, L., “Automatic Script Identification From Document Images Using Cluster-Based Templates,” Proceedings of the Third International Conference on Document Analysis and Recognition, Pages 378-381, 1995.
8. Wood, S., Yao, X., Krishnamurthi, K, “Language Identification for Printed Text Independent of Segmentation,” International Conference on Image Processing, Pages 428-431, Oct 1995
9. Chaudhury, S. and Sheth, R., “Trainable Script Identification Strategies for Indian Languages,” Proceedings of the Fifth International Conference on Document Analysis and Recognition, Pages 657-660, 1999
10. Peake, G.S. and Tan, T.N, “Script and Language Identification from Document Images” Proceedings. Workshop on Document Image Analysis 1997, Pages 10-17, 1997.
11. Nobile, N.; Bergler, S.; Suen, C.Y.; Khoury, S., “Language identification of on-line documents using word shapes,” Proceedings of the Fourth International Conference on Document Analysis and Recognition 1997, Volume: 1 , Page(s): 258 -262 vol.1, 1997..
12. 1. Assadi, Amir, and Uchill, Joseph. Application of Multi-scale and Multi-resolution Statistical Analysis in Classification and Compression of Digital Documents (In preparation).