IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 6, 2021, accepted November 22, 2021, date of publication December 7, 2021, date of current version February 4, 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3133654

The Structural Information Potential and Its Application to Document Triage

VLAD ATANASIU
Department of Informatics, University of Fribourg, 1700 Fribourg, Switzerland
e-mail: atanasiu@alum.mit.edu

This work was supported by the Document, Image and Video Analysis (DIVA) Research Group, Department of Informatics, University of Fribourg, Switzerland.

ABSTRACT This article introduces structural information potential (SIP), a measure of information based on the potential of structures to be informative about their content. An example of this concept is the clustered appearance that typically characterizes the first page of scientific articles, which summarizes the article’s contents and provides additional data, yielding potentially the largest and most diverse amount of information from a single page in the shortest time with the least effort. This characteristic makes SIP particularly well-adapted to triage tasks (i.e., rapid decision-making under conditions of uncertainty and limited resources), an application illustrated by means of a case study on classifying document images. The SIP method consists in unifying the Shannon entropy, the Fourier transform, the fractal dimension, and the golden ratio into a single equation and several algorithmic components. While the application domain is document images, the concept has generic character. The method results in a mathematically and perceptually coherent pattern space, characterized by continuous transition between uniform, clustered, and regular configurations, and corresponding to a structural information potential with a well-defined maximum. The maximum SIP leads to the identification of shapes and patterns with minimal structural redundancy, termed “fluorescent objects” as a complement to regular graphs and the Platonic solids.

INDEX TERMS Information theory, structural entropy, spectral entropy, Fourier transform, fractals, graph theory, golden ratio, pattern analysis, image classification, document analysis, layout analysis, document triage, digital libraries.

> Form is the visible shape of content.
> — Ben Shahn

I. INTRODUCTION

This article introduces a new measure of information, the structural information potential (SIP). SIP defines information as the potential of structures to be informative about their nature and utility. A readily available example is the clustered appearance typical of the first page of scientific articles, such as this one; these pages summarize the article’s content and provide additional metadata, thus yielding potentially the largest and most diverse amount of information from a single page in the shortest time with the least effort. The optimality of the data transmission rate, as influenced by the structure of the communication channel represented by documents, is also a critical factor in document triage, the main application studied in this article to evaluate the SIP measurement method.

Code 1

```
function loopyloops(n); message = {‘Hello world!’,
‘I like apple pi.’}; for loop1 = 1:length( message );
for loop2 = 1:n; disp( message{loop1} ); end; end;
```

Code 2

```
function loopyloops(n);
message = {‘Hello world!’, ‘I like apple pi.’};
for loop1 = 1:length( message );
    for loop2 = 1:n;
        disp( message{loop1} );
    end;
end;
```

FIGURE 1. **Demonstration of the structural information potential principle: While the above codes are identical both content-wise and functionally, the structured layout of code 2 facilitates faster analysis by programmers, enabling them to notice at a glance that there are potentially seven instructions and two nested loops.**

In the document domain, the above observation is supported by compelling empirical evidence from the practice [1]–[4], theory [5]–[7], psychology [8]–[10], and history [11], [12] of document design (these references are

The associate editor coordinating the review of this manuscript and approving it for publication was Khmaies Ouahada.

a small representative sample from a substantially larger bibliography). For example, much of the effort invested in microtypography [13], [14] (already an aesthetic quality and commercial factor in the early days of typography) concerns the spacing of characters (kerning) and words (justification), character ligatures, hyphenation rules (ladders), trailing paragraph lines (widows and orphans), and so on, for the purpose of producing visually homogeneous pages of text (a characteristic denoted as ''page gray'') [15], [16]. The principal motivations for this painstaking effort are the desire to keep semantic units visually grouped, as well as to prevent vertical streaks (rivers) from emerging due to chance alignments of spaces or similar letters, drawing the reader's attention towards spurious shapes devoid of content-related information (Fig. 2) [17]. Conversely, an intentionally clustered page pattern is the result of graphic designers arranging distinct informational units so as to augment hierarchical and typological distinctiveness; the goal is to improve legibility, speed up access to information, and guide the reader's gaze with minimal interference. This strategy for written communication is a product of evolution, a centuries-long shift away from homogeneous layouts driven by the increasing availability of written information. Clustered patterns can also emerge as a natural part of the document life-cycle; these may be introduced in the post-production stage either intentionally (e.g., by layers of annotations) or accidentally (e.g., due to physical degradation). In the absence of specific search goals or prior knowledge about the content, the most informative documents are those with clustered patterns.

It is possible to generalize beyond documents, given that uniform and regular signals, images, objects, and events are in general less informative than structured entities. The paradigm resulting from these insights postulates a correspondence between informativeness, structure, a uniform–clustered–regular pattern continuum, scale-space filling, and structural redundancy. This article accordingly aims to devise a quantitative pattern description method for ordering patterns along said continuum, and further develops a conceptual framework to aid in identifying structures with minimal redundancy.

What this approach cannot provide is an estimation of information potential in the absence of structure, or a semantic content analysis. SIP is no substitute for text recognition and visual scene interpretation. Instead, its purpose is to characterize information at the level of structural organization.

Great strides in characterizing structural informativeness have been made in various scientific fields. However, the application of the proposed solutions to the task of classifying images, particularly text-based document images, has been found to be insufficient, as no prior approach has been able to satisfy both mathematical and perceptual desiderata. The present article substantiates this claim and presents a solution.

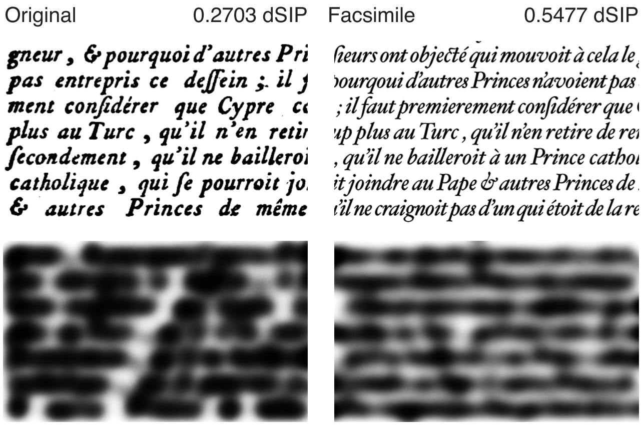

Original
0.2703 dSIP
Facsimile
0.5477 dSIP
gneur, & pourquoi d'autres Pri
pas entrepris ce deſſein ; il
ment conſidérer que Cypre c
plus au Turc, qu'il n'en reti
ſecondement, qu'il ne bailleroi
catholique, qui ſe pourroit jo
& autres Princes de même
ſieurs ont objecté qui mouvoit à cela le
pourquoi d'autres Princes n'avoient pas
; il faut premierement conſidérer que
up plus au Turc, qu'il n'en retire de re
, qu'il ne bailleroit à un Prince catho
it joindre au Pape & autres Princes de
s'il ne craignoit pas d'un qui étoit de la re

FIGURE 2. **Example of spurious information emerging from a salient visual structure, and its detrimental effect. — Left: The smooth flow of reading this text is perturbed due to the reader's attention being reflexively diverted towards three asemantical structures resulting from chance diagonal alignments of white interword spaces (a typographical "river"), commas (an "island"), and the character q (a "ridge"). — Bottom: The perceptual effect is increased when squinting (here, simulated via Gaussian blur), which acts as a low-pass filter that makes the river into a salient structure in the visual field (a structure that is also large and has a different orientation than the text lines). — Right: The artifacts in the original document are removed by creating a facsimile where the spacing has been carefully adjusted to make the distribution of ink more homogeneous overall (typeset in Adobe InDesign). — Top: The SIP measure correctly reflects in a quantitative manner the observed difference in perceptual homogeneity between the "noisy" original (0.2703 dSIP) and the artifact-free facsimile (0.5477 dSIP). However, the lower dSIP value also suggests a higher structural information potential, which from the point of view of the text content is an illusion created by the typographical artifacts. Since these artifacts are structures absent from the facsimile, the original document is indeed richer in information, but of a kind other than linguistic. This information is, for example, useful for evaluating the typographical quality of the document and the evolution of this quality across time and space (accordingly, it may be valuable to e.g. historians and antique dealers). In this respect, the high prices of "incunabula" books published by famous 15th century printers, such as Aldus Manutius of Venice, are in part due to the high typographical quality of their products, including a remarkably even page "gray" that remains a model of exquisite quality even today. The role of this historical note is to explain through concrete examples the relevance of SIP for a broad range of applications. — Credits: Baron de Zur-Lauben, *Mémoires et letters de Henri Duc de Rohan, Sur la Guerre de la Valteline*, Geneva & Paris: Vincent, 1758, vol. 1, p. lxxxvii.**

In essence, the proposed method characterizes a distribution in the scale-space domain with respect to the degree of redundancy. This characterization is achieved by unifying a number of classical concepts from the fields of information theory (Shannon entropy), pattern analysis (Mandelbrot's fractals), signal processing (the Fourier transform), combinatorics (the golden ratio), and graph theory (the chromatic concept) into a single analytic formula and several algorithmic components.

The application domain of the method is restricted in this article to images, more specifically to text-based document images. Given the generic nature of some of the core concepts, the concluding section considers its application to other media, such as video and three-dimensional data.

The relevance of the present article stems from the status of information measurement as a fundamental theoretical and practical issue across a broad range of scientific and technical fields. Its foremost contribution consists of a practical method to measure informativeness from structure, for application to images and possibly other data types. The elaboration of the pattern phase space and the identification of fundamental shapes with minimal structural redundancy are further theoretical contributions. Finally, a survey of the quantification of irregularity links various fields, and places this research in a wider scientific perspective.

This article demonstrates a practical application of SIP to document triage via a real-life case study. The task in question is a type of classification similar to triage in emergency medicine, defined as *rapid decision-making for critical matters under conditions of uncertainty and with limited resources*. While triage cannot yield optimal solutions due to the constraints under which it operates, it is a useful and sometimes necessary step before more sophisticated procedures (such as semantic document analysis) are implemented. The proposed method fulfills the task in an explainable way, with reduced algorithmic complexity and assumptions.

Section I, ''Introduction'', has acquainted the reader with the concept of SIP and some of its empirical basis. Section II, ''Related work'', presents the advantageous and limiting factors of major existing approaches to information measurement, going on to explain the need for a new method while introducing elements used in the proposed method. Section III, ''Method'', is the theoretical core of the article, in which the analytical formula and algorithmic components of the SIP measurement method are described and justified; the design of structures with minimal redundancy is also discussed. Section IV, ''Experiments'', is devoted to the quantitative and qualitative evaluation of the proposed method through a case study in document triage, as well as to providing the exemplary discussion of a few other applications. Section V, ''Discussion'', concerns technical matters and future work. Finally, Section VI, ''Conclusions'', summarizes the theoretical and practical significance of the proposed method and its applications.

II. RELATED WORK

A. PATTERN IRREGULARITY

1) INFORMATION THEORY

Information theory has a central place among the domains relating pattern structures to informativeness. It is rooted in the *Shannon entropy*, H [18], which defines the amount of information, uncertainty, and choice, and quantifies it via the well-known equation $H = -\sum_{k=1}^{n} p_k \log_2 p_k$, where p_k is the occurrence probability of data class k from among n classes [19, pp. 393]. The normalized variant is given by the relative entropy, $H_r = H/H_{max}$, $H_{max} = \log_2(n)$ [19, p. 398]. This formulation produces a data distribution in which the extrema are on one hand the uniform distribution of values over all possible classes ($H_r = 1$), and on the other hand the concentration of values into a single class ($H_r = 0$). In the context of grayscale images, these correspond respectively to a uniformly monochromatic image and an image in which the number of present gray-level values equals the number of pixels. The application domain of this equation is nominal data, i.e., independent categories, such as the values of a fair dice. Therefore, this equation (and many of its variants) is not directly applicable to ordinal and structural data; given that the measure is independent of the configuration of sampling points (such as pixels in an image), it is not suitable for the goal pursued in this article.

Nevertheless, it is worth mentioning some notable entropy definitions as a way to exemplify the extent and vitality of research in this field, to point to research directions, to underscore the interdisciplinary character, and to identify ideas related to this article's topic. The Rényi and Tsallis entropies are generalizations of the Shannon entropy and define a parametrized family of entropies [20], [21]. Aiming at quantifying biodiversity, the field of ecology has contributed to the research on entropy with rich theories of diversity, as well as with formalisms, such as allowing for weighted probability classes, which express the similarity between species [22], [23]. From the field of nuclear physics comes the *strength of structure* [24, pp. 137–144] defined in 1939 by Satosi Watanabe (then a student of Werner Heisenberg) as J, the difference between the sum of entropies of u parts of a system, each containing v_u entities, and the entropy of the whole, containing $n = \sum_{i=1}^{u} v_{u_i}$:

$$J = -\sum_{i=1}^{u}\sum_{j=1}^{v} p_{ij} \log_2 p_{ij} + \sum_{k=1}^{n} p_k \log_2 p_k. \tag{1}$$

This method has been utilized in document layout segmentation [25]. The inclusion of both parts and the whole in the characterization of structures is also important to the concepts addressed this article.

Approximate entropy (ApEn) is a measure of ''irregularity'' created to extend information entropy to structural data [26]–[28]. ApEn has been refined through many parametric [29], [30] and algorithmic variants [31]–[33] and compares well with other methods [34], [35]. From its inception onwards, it has been successfully applied to various biosignals [34], [35], and later to a broad range of other applications such as online signature verification [36], speaker recognition [37], radar jamming [38], earthquake prediction [39], and cryptography [28], [40]. In brief, ApEn is obtained by sliding a signal over itself, measuring the distance between the two within a window of given length according to the difference between the maxima, and computing a logarithmic average; this process is carried out for two different window lengths and the irregularity index is obtained as the difference between these partial results. For a finite discrete signal, there exists both a lower and upper bound; this is zero for perfectly periodic patterns, with higher values indicating greater irregularity. One very interesting aspect of this measure is that it is possible to compute number sequences with maximal irregularity (this upper bound is neither white noise nor a deterministic fractal) [41], [42]. The method may be extended to the analysis of images using vectorized bidimensional windows [43], [44]. This approach, in addition to improved parametrization [29], was used in this article to process document images. The results are discussed in Section IV, ''Experiments'', and illustrated in Fig. 15. While clustered documents are classified at one end of the spectrum,

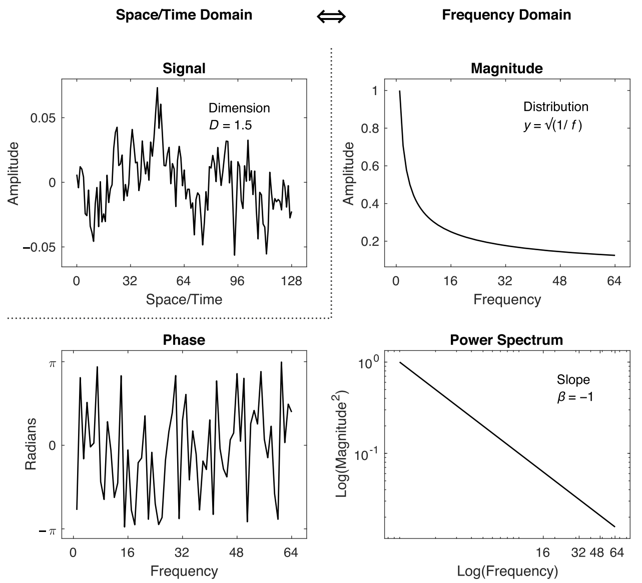


FIGURE 3. Sample fractal with dimension $D = 1.5$ represented in the space/time and frequency domains. The signal was synthesized from the magnitude and phase via inverse Fourier transform. The signal has the characteristics of $1/f$ noise, a power law distribution visible in the shape of the magnitude and confirmed by the linearity of the power spectrum on log–log scales; the phase is random. Only half of the symmetrical spectrum is displayed. The DC is set to zero, hence the zero mean amplitude of the signal.

as desired, a mix of uniform (predominately empty) and homogeneous (predominately text) patterns appear at the other end, which is undesirable, since these patterns differ both in terms of appearance and information potential. This counter-intuitive behavior has been explained in the past by the observation that ApEn is a measure of irregularity rather than complexity [45].

2) FRACTALS

Fractals [46] are deterministic or stochastic self-similar or self-affine mathematical objects with an interesting property from the point of view of information theory: their clustered structure, which results from their infinite filling of the scale-space, maximizes their information potential. This is one of the reasons why fractal-like structures abound in nature; for example, the energetic intake of plants is optimized via the organization of branches and leaves around stems according to a power law [47], [48].

The dimension of a fractal is a fractional value, D, bounded by the fractal's topological and Euclidean dimensions, $D_T < D \leq D_E$. E.g., $1 < D \leq 2$ for signals, and $2 < D \leq 3$ for images. One of the most reliable methods for determining the fractal dimension uses spectral analysis and defines it as the slope, $\beta < 0$, of the linear fit of the log-power spectrum vs log-frequency: $D = (c + \beta)/2$, where $c = 6$ for an image, and $c = 4$ for a signal [49], [50, pp. 54, 97–114]. Two requirements must be satisfied if the measured entity is to be deemed a fractal: namely, the phase must be uniformly random [49, pp. 99] and the power spectrum must follow a power law, f^{α}, assuming $\alpha = \beta$ [51]–[53] (Fig. 3). By definition, this is not the case for non-fractal structures that are nevertheless clustered, and even less so for other pattern types along the uniform–clustered–regular continuum. The use of the fractal dimensions is therefore not appropriate for use in characterizing patterns of such a broad spectrum. However, the concept of fractality does provide a useful framework for thinking about SIP, especially as concerns its maximal value.

3) SPECTRAL ANALYSIS

Spectral analysis is useful for the characterization of clusteredness. This is because, unlike spatial entropy, the frequency domain captures spatial organization in a compact manner that is amenable to mathematical manipulation, and also captures scale variation (in a similar way to the fractals). Two popular spectral analysis methods are spectral entropy and spectral flatness.

Spectral entropy, H_s, is [54] defined as the Shannon entropy of the probabilities, P, associated with the frequency components of the power spectrum, S, given by the discrete Fourier transform, $\mathcal{F}$, of the data, X, of length n:

$$S = |\mathcal{F}(X)|^2, \quad P = S/\sum_{i=1}^{n} S_i, \; H_s = -\sum_{i=1}^{n} P_i \log_2 P_i. \qquad (2)$$

Variations of the definition include obtaining the power spectrum from the discrete cosine transform or by way of autocorrelation, along with using entropies other than Shannon's. Understanding the properties of these analytical expressions so as to be able to thoroughly explain their effects on empirical data remains an active research field [55]. This is especially critical for the analysis of biomedical data, a major application domain of spectral entropy [56]–[58], where it is used in particular for the clinical interpretation of EEG signals (e.g., for monitoring depth of patient sedation during anesthesia [59]). Audio signals analysis (e.g., urban soundscape classification [60], dolphin whistle segmentation [61], abnormal milling sounds detection [62]) and speech analysis (e.g., speaker identification [63], noise quality assessment [64]) are other common application domains. Spectral entropy has been less frequently applied in image processing, but is used for image quality assessment [65], scene saliency analysis [66], and camera focus estimation [67]. The common goal is the need to distinguish between regular and irregular patterns, where the latter are usually of interest, a task for which spectral entropy has been found useful. The problem remains that "irregularity" has many possible formal definitions, and means different things for different data types, tasks, and contexts; moreover, the spectral entropy equation has its own peculiar effects on the data, with the interaction between the two not always being well understood. When applied to image classification, for example, it results in a mix of homogeneous and uniform patterns (see Section IV-A, "Comparison Of Methods").

Spectral flatness, or *SF* [68]–[70, pp. 112–115], is a widely used measure of signal structuredness [71]–[73] (e.g., as an audio descriptor in the MPEG-4 file format [74]). It is defined as the ratio of the geometric and arithmetic mean of a signal's power spectrum, S, taken over its n frequency components:

$$SF = \frac{(\prod_{i=1}^{n} S_i)^{\frac{1}{n}}}{\frac{1}{n}\sum_{i=1}^{n} S_i} = \frac{\exp\left(\frac{1}{n}\sum_{i=1}^{n}\log_2 S_i\right)}{\frac{1}{n}\sum_{i=1}^{n} S_i}. \tag{3}$$

Its principal utility comes from the bounds of the expression being zero for an impulse and one for a uniform distribution in the frequency domain, which correspond in the spatial domain to a regular signal and a impulse, respectively. *SF* relates to SIP in that both have the same bounds. However, just like spectral entropy, the spectral flatness also results in a perceptually unsatisfactory classification of images (Fig. 15).

Point set analysis and *spatial statistics* are concerned with the characterization of the distribution of point-like objects and events in space [75]–[77], while the related field of *discrepancy theory* deals specifically with characterizing the irregularities of distribution [78], [79]. Geographical information systems, ecology, astrophysics, and material science are among the typical application domains. We will herein briefly focus on the latter, as it bears a direct relation to both the topic and methods of the present article. Based on the empirical observation of the structure of physical matter, the continuous pattern space extending from uniform to clustered to homogeneous has been identified as a useful classification concept in material science to characterize such properties as surface roughness and particle dispersion. Much effort has therefore been invested in quantifying these patterns, with some of the classical methods being based on the nearest-neighbor distribution, morphological operations (dilation followed by counting), and Dirichlet tessellation [80]. To date, some of the best-performing approaches rely on spectral fractal analysis, using variations of the methods described in the preceding paragraphs [81] [82, pp. 81–98] [49, p. 108]. The utility of the fractal paradigm has however been called into question by practitioners on practical grounds, due to the difficulty of measuring fractality, as well as on theoretical grounds pertaining to its appropriateness as a model of the observed data [83] [49, p. 109]. In summary, we note that spectral analysis is a powerful method of characterizing structures, that better methods are required, and that methods and data must be compatible.

4) GRAPH THEORY

Graph theory can be applied to model discrete patterns, such as the individual pixels of digital document images, as well as the visual and logical entities of document layouts [84]. A dynamic sub-field studies irregular or color graphs, whose properties derive from the value of their edges; for example, a rainbow graph is one with distinct edge values [85], [86]. The topic relates to this article not only because of its focus on pattern irregularity (as opposed to regular structures, such as the Platonic regular bodies), but also because it deals with determining maximal irregularity, which the approaches reviewed above have yet to achieve. After an extensive literature survey (see below), however, it was not possible to find previous work on maximal irregularity applicable to patterns such as document images, even for basic shapes (such as the triangle). Furthermore, another well-known practical element makes the graph theory approach to pattern classification problematic: the degree of computational complexity for data with millions of sampling points, such as document images, creates high computational costs, especially in comparison with other methods such as spectral analysis.

Survey — We commence by stating here that our goal is the study of *the irregularity of continuous chromatic graphs*, and then proceed by commenting on the aesthetic dimension of this mathematical research, which also serves to introduce the notions vehiculated by the terminology. We next expose the historical origins of chromatic graph theory [87] and the interdisciplinary strands from which the modern research on irregular graphs emerged; finally, we conclude by discussing some particularities of our research in respect to graph theory.

There exists a little theorem that all it says is: “No graph is irregular” [88, p. 25] [89, p. 24] [85, pp. 36–37]. What it means is that there exists no graph with two or more vertices for which the vertices have distinct numbers of adjoining edges (i.e. distinct degrees). The simplicity of this statement is exquisite, moreover, and perhaps unsurprisingly, its corollary—that every graph with two or more vertices has at least two vertices with the same degree—has been voted one of the twenty most beautiful theorems of all time [88], [90], [91, p. 25]. Even the field of chromatic graph theory [92] to which these two theorems belong is not bereft of meta-mathematical charms, which exude from the colorful terminology employed to designate various types of graphs. For illustrative purposes, the author composed out of graphical terms the following pangram (a kind of linguistic color graph, in which all letters of the alphabet must appear): “Graph Euler cycles over the rainbow to join the flower snarks in a zero-vertex quasi-bramble.”

The question has been asked as to whether it is possible to color a cartographic map with no less than four colors, such that no two adjacent countries have the same color. In the form of this Four-Color Problem, graph coloring famously made its public debut on the 23th of October 1852 as an intriguing mapmaking problem among British mathematicians, albeit in the absence of any specific request from Her Majesty's Ordnance Survey [93, pp. 1–26]. Its impact would be profound, with “many of the concepts, theorems, and problems of Graph Theory” being said to “lie in the shadows of the Four-Color Problem” [93, pp. 2]. The study of graph irregularity adds new layers to the already consequential applications of coloring (e.g., timetabling, sequencing, and scheduling [94, p. xv]), such as drug design [95, p. 600] and secure intelligence and military communication networks [93, p. 76] [94, p. v–vi]. The topic remains, however, outside of the mainstream: in a hefty 1633-pages handbook on graph theory published in 2014 the word “irregular” occurs only once [96, p. 1032], and not at all in many other books of its kind.

A series of fundamental problems in *combinatorial geometry* and *geometric graph theory* were published in the 1930s and 1940s, dealing with the number of distinct colors that color the plane in such a way that no two blocks of the same color are at unit distance [94], [97]–[100]. The research grew around what became known as the Hadwiger–Nelson problem of the chromatic number of the plane, providing nourishing theoretical ground for the later development of chromatic graph theory and graph irregularity. The earliest trace of *irregular graphs* identified by this author dates back to Germany during the Second World War [101, p. 76–77]. The concept later resurfaced independently in the late 1980s in the work of the American-Iranian mathematician Yousef Alavi, and was elaborated by his colleagues at Western Michigan University, including Paul Erdős himself [102, p. 235]. Their approach to the problem is both astute and unintentionally revealing about the nature of mathematics. Rather than asking "what *is* an irregular graph?", their question is "what *could* an irregular graph be?", or in the slightly more normative verbatim: "How *should* one define an irregular graph?" (emphasis added). They followed up by stating their expectations: "In research, the goal is not only to come up with a definition that seems natural but to arrive at a class of graphs with interesting, and perhaps even some surprising, properties." [85, p. 39]. These propositions seem to advance the mutually exclusive views of mathematics as an activity of *invention* (the ludic and aesthetic motivations are explicit), and as one of *discovery*, in line with Erdős' concept of an immanent Book of mathematical proofs ("Mathematics is there." [103], [104, p. 27]). In the end, their answer is a multiplicity of definitions of graph irregularity, with various degrees of "interestingness" and "surprise". One of the researchers, Gary Chartrand, would subsequently coin the term "rainbow graph" [93], [105], [106], labor extensively in this research field [89], [105], [107], [108], popularize graph irregularity [88], and point to its practical usefulness [93].

In Yugoslavia, meanwhile, members of the Zagreb Mathematical Chemistry Group developed measures of molecular complexity based on their own indices of graph irregularity, the *Zagreb indices*, which remain a mathematical staple among chemists [109]–[113]. Only in the late 1990s did the three research strands converge [114]–[119, p. 45]; after this point, the focus begins to shift towards special topics, such as determining the maximal and minimal irregularity bounds [114], [120], [121], local versus global irregularity [122], irregular graph assignments [123], and many more [124].

The measure explored by this author differs in an essential way from previous graph irregularity concepts. The Zagreb indices are variously the sum of the squared vertices degrees, the sum of products of adjacent vertices degrees, or the sum of absolute differences of adjacent vertices degrees. In other words, they represent the valence of entities, and are therefore suited to applications in chemistry, especially given the prevalence of irregular molecular graphs of interest to chemists [118, p. 222]. The rainbow measure is given by the number of distinct colors in a graph or along specific paths, which is appropriate for routing problems [107], [125], [126]. In both cases the actual distances between vertices are disregarded and the graph's topology is one of distanceless connections. There are instances, however, where edges represent not discrete label classes, but *continuous distance magnitudes*, notably in pattern and shape analysis. This is the case and particularity of the research presented in this article.

Note: Graphs embedded in the Euclidean space are *geometric graphs*, those with other topologies are *topological graphs* [99, p. 465], and both are *distance graphs* [100, pp. 429–430]. Graphs optionally embedded in a space and with values attached to the edges are *chromatic graphs* [87]. Colors may represent both *continuous* values and *categories*. We will say therefore that we study the *ir|regularity of continuous chromatic graphs*.

5) OTHER FIELDS

For the sake of completeness, it is worth mentioning other important fields that deal with irregularity (in the graph theoretical sense) and redundancy (in the information theoretical sense): *combinatorics* (particularly *combinatorial geometry*) [46], [127, pp. 78–81], *packing* [128], *tiling* [129], *tessellation* [130], *coding* [131], [132], and *randomness* [133]–[136].

Details — Irregularity can be conceptualized from many different perspectives. In its simplest form, line segmentation, it can be approached from the points of view of graph theory, point spread, and combinatorics. It is indeed an old problem of *combinatorial geometry* [137], counting among its earliest investigators the Polish mathematician Wacław Sierpiński (1882–1969; cryptanalyst during the First World War and later father of the eponymous "gasket" and "carpet" fractals [138, pp. 78–81]) and his Swiss-Russian mentee Sophie Piccard (1904–1990; author of a monograph on collinear point sets [127], [139], [140]). The combinatorial aspects of point sets were also a long-standing interest of Paul Erddős [141]–[145]. The absence of mentions of maximal dissimilarity in his publications [146], [147] leaves open the possibility that "the prince of problem solvers and the absolute monarch of problem posers" [148, p. 64] may have considered the topic a pleasant surprise, worthy of *The Book*, the imaginary object where all mathematical proofs are kept.

In the fields of *tessellation* and *tiling*—of interest to computer graphics, surface modeling, and material sciences, for example—there is a tangential interest in the irregularity of these planar graphs, albeit typically with the aim of suppressing it so as to achieve maximal regularity [129], [130].

Discrepancy theory is another field that appeared an auspicious match with the search for a model of pattern irregularity, considering its generic framing as "a measure of the deviation of a point set from a uniform distribution" [149], [150] [78, pp. 1–3]. With roots in the 19th and early 20th centuries, the mathematical theory of discrepancy has experienced an increase in research interest over the last decades (it has its own conferences and journal [151]) thanks its many links to other areas of mathematics (e.g. number theory, graph theory, and Ramsey theory), as well as the development of instruments fundamental to computer science (such as for numerical integration, Monte Carlo simulations, or randomization), which find applications in computational physics, computer graphics, mathematical finance, and cryptography, among many others [152, pp. xi–xiv] [78, pp. vii–viii, 22–35] [153, p. 1]. There exist a host of methods for computing discrepancy, including utilizing spectral analysis, be it the Fourier or the Haar transform [79, pp. 621–678] [78, pp. 213–240]. For practical reasons related to ease of computability, the most widely used method is the corner method or L_2-discrepancy, which is based on the number of points falling within rectangles anchored at each of these points and the bounding box of all points (for details see [78, pp. 2–3, 10, 12–16] [79, pp. 623–624]). Unfortunately, the pattern spectrum defined by the discrepancy measure differs from the perceptual spectrum uniform–clustered–regular targeted in this article and defined by the structural information potential. For example, maximum regularity in the sense of discrepancy is not a pattern of equally spaced points (a triangular lattice), as for SIP, but rather one with inhomogeneous density [79, p. 3].

Given the importance of the golden ratio and the most irregular triangle for the structural information potential, it was anticipated that some references would also be found in the plethoric literature on the *golden ratio* [154]–[157] and *triangle geometry* [158]–[161], both of which also bear some relevance to graph theory. One tantalizing possible avenue that might lead to the discovery of new mathematical visions is that of looking into cultures with limited global contact, such as pre-modern Japan, when the distinctive *sangaku* geometry was developed in temples [162]. This search, however, was less than fruitful.

The one domain that is almost pathologically fascinated by structural irregularity is that of the *arts* and *architecture* (Fig. 4). East Asian calligraphy, for example, is theoretically grounded in a spectrum extending from almost mechanical regularity to highly irregular brush patterns, while contemporary Western typography is built on the visual tension between homogeneous and hierarchical layouts. Even the movement of people within movie frames may be analyzed in terms of dynamic changes in graph irregularity, as a means of expressing meaning and emotion (the sleek geometrical compositions of Michelangelo Antonioni's films are good examples). Architecture makes irregularity concrete, such as in the fractal geometry of Islamic *muqarnas* (stalactite-like wall decorations) and the irregular volumes of the villas designed a century ago by Ludwig Wittgenstein and Le Corbusier (the creator of the golden ratio-based Modulor building module), and nowadays with the aid of computers by Zaha Hadid and Frank Gehry.

FIGURE 4. A photographic study of natural irregularity, in which a visual framing is found such that the pairwise length between the three rocks, along with their size differences, is maximized. The theme is reminiscent of similar experiments related to the aesthetics of perspective conducted by the American photographer Ansel Adams, as well as the spatial organization of traditional Japanese gardens. — Credits: Vlad Atanasiu, 2016, Golden Gate Bridge, San Francisco.

6) MACHINE LEARNING

This section addresses the possibility of using machine learning for classifying documents on the uniform–clustered–regular pattern spectrum and determining graphs with minimal redundancy.

Classification — If the goal is to solve the classification problem at hand, rather than to use a specific method to solve it, then machine learning is not necessary since we have already identified a solution that does not require learning. Machine learning may however be useful if it could demonstrate certain advantages over the classical approach. The author searched, but could not find a suitable solution of this kind; on the contrary, a number of issues were instead identified that have the potential to diminish the classification quality and complicate the solving process.

The first step of learning—constituting a representative dataset—is already problematic. Since the concept of structural information potential (SIP) is applicable to any kind of pattern (e.g., time series, images, movies, three-dimensional objects, and social networks), in any dimension, the typological diversity and size of the training and testing dataset is at the limit of practicality. Even when restricting ourselves to the domain of documents, there is a significant diversity of document types, including types of noise, which represent real possible impediments to knowledge generalization.

The second learning step—producing the groundtruth—is expensive and possibly suboptimal. It is expensive because the classification examples to be emulated by the machine must be generated by humans; this will result in small dataset sizes, which may reduce the learning quality. Additionally, the machine would be trained on human behavior, but physical and perceptual pattern measurement are not identical.

Moreover, the influence of perceptual variability across a wide range of human factors (e.g., age, gender, culture, motivation, human–machine interfaces) would need to be assessed and accounted for.

Minimal redundancy — The contribution of this article is not only a practical method of pattern classification, but also a theoretical work on defining minimal structural redundancy. Of course, emulation and trial-and-error are fundamental learning methods; however, there were no preexisting examples to emulate in the minimal redundancy definition process, nor a set of reward-and-penalty criteria to use. Instead, the author developed a rational argument about what ought to be a solution that is both mathematically provable and useful in practice. It was through a cognitive argumentation process, rather than learning, that it was possible to define the SIP value with maximal information potential, which is an important component of the SIP theory, the SIP value with maximal information potential, which is an important component of the SIP theory, the SIP measurement method, and the uniform–clustered–regular pattern spectrum.

7) CONCLUSIONS

As a concluding remark, it can be stated that a common goal of the reviewed perspectives is the characterization of structuredness, as distinct from uniformity, regularity, and randomness, and under the moniker "information," as a proxy for the utility that can be derived from pattern analysis. The goal is thus doubly defined in terms of methods and applications. The review has highlighted how a thorough understanding of the empirical application domain supports the development of successful theoretical methods. This insight is reflected in the work reported herein, particularly in the description of the design, psychology, history, and life cycle of documents that has shaped the theory of structural information potential.

B. LAYOUT-BASED DOCUMENT TRIAGE

The classification of documents with respect to their informativeness is a task common to a number of applications, notably document overview, retrieval, summarization, and presentation. The problem becomes more critical as the number of documents to be processed increases (e.g., within organizations, libraries, and archives), but is equally relevant for a single document, such as when browsing a digital book. Informativeness is a relational quality, in that it depends on both the stimulus and the observer. In cases where prior knowledge about the user is lacking, or where there are many users with heterogeneous interests, a user-independent estimation of document informativeness is necessary. An additional constraint is the speed with which users may acquire the information presented to them. From a statistical perspective, the most commonly applied solutions rely on some form of data summarization and sampling.

Summarization consists in beginning with a given document set or item and synthesizing a new, more compact one that retains the characteristic features of the original. For text documents, this typically involves removing text chunks so as to reduce the overall redundancy and fit the text in a smaller spatial frame, producing something like a more or less extended abstract or even a title [163], [164]. Image summarization is conceptually similar but more difficult to realize; for example, representing a person by her or his face and a wood by a single tree [165], [166], removing empty areas from document pages [167], replacing a color image with a black-and-white line art sketch, or reducing an image to its dominant colors [168]. Documents representing three-dimensional data (e.g., buildings or landscapes) and multimodal documents (e.g., videos) are far more complex to represent compactly (a high-quality movie trailer goes beyond simple cut-and-paste; it is an artistic project in itself) [169], [170].

This article's contention that spatial organization is informative has also been applied to document summarization. For example, this author has introduced the Document Towers visualization paradigm, which represents the three-dimensional structure of bounding boxes of paragraphs, images, and other entities in paginated documents as architectural wire-mesh models that resemble buildings and cities [171]. The quantified structural information potential is encoded as a color-coded "ribbon", which allows users to take stock of features such as document fragmentation, regularity, and outliers without opening the document itself. In addition to facilitating overview and navigation, this information enables document type, quality, and other insights to be inferred, often serendipitously.

Sampling differs from summarization in that it does not create new entities but instead aims to identify a limited amount of existing document parts that are representative of the whole. A further difference can be compared to the distinction between statistical expectation and the probability density function: while summarization often ends up presenting the average content or the most informative (e.g., the table of contents), sampling may provide the full range of content types (e.g., cover, text, figure, index). Semantic layout analysis is therefore a dominant method used to sample document images, as it benefits from not requiring character recognition (a substantial argument in support of its performance and quality, particularly given that noisy, historical, and/or handwritten documents still present challenges for this process) [172], [173]. Another sampling approach is pattern-driven; for example, a handwriting dataset can be compactly represented by a few "vantage point" samples [174]. This is a classical pattern classification and clustering problem, resulting in an ordering specific to an application domain and dataset. A good example of the complexity of defining interestingness is given in [9], where the interaction of factors as diverse as color, layout, content, reading ergonomy, and readership is analyzed. The contrast along multiple dimensions in terms of types and number of entities between neighboring document pages has also been used as a criterion for informativeness, in the framework of Shannon's information theoretical definition of information as the amount of "surprise" [175]. The work presented in [176] is the closest to a pure pattern-based measure of document informativeness such as that described in this article. Although it aims at being a fast, simple, and approximate method (in the spirit of the triage task), it does not address spatial organization. Instead, page informativeness is quantified as the degree of (chromatic) saturation and (achromatic) lightness computed over the connected components of binarized pages and weighted by their size.

This succinct survey reveals how the quantification of informativeness has been approached at various points on the spectrum, ranging from pattern to semantic to contextual analysis. However, the quantification of spatial organization and the derivation of insights therefrom remains a fruitful research direction for computational document analysis.

III. METHOD

The empirical insights expounded in the introduction suggest that pattern informativeness varies along a uniform–clustered–regular continuum. The goal of this section is to introduce a quantitative description method, namely the structural information potential, which facilitates the ordering of patterns along this continuum. A discussion of the correspondence between maximal SIP and minimal structural redundancy is also presented, as the latter provides a theoretical foundation for the former.

A. STRUCTURAL INFORMATION POTENTIAL

We begin by introducing the SIP equations, then explain the general rationales behind them, after which we discuss each element of the procedure in detail.

1) CORE PROCEDURE

The core mathematical machinery of the structural information potential measurement consists of the Shannon entropy of the logarithm of the power spectrum of a binary image's two-dimensional Fourier transform. Hence, it is a mixture of information-theoretical concepts and signal processing, with a fractals perspective and properties of the golden ratio playing also a role, as shall be seen. Formally, the value SIP $\in [0, 1]$ is obtained through a number of equations and algorithmic steps:

$$\mathrm{SIP} = 1 - |\mathrm{dSIP}|, \tag{4}$$

$$\mathrm{dSIP} = T\left(H_r\left(\log_2\left(1 + V\left(|\mathcal{F}\left(B\left(I\right)\right)|^2/n\right)\right)\right)\right), \tag{5}$$

where I is an intensity image, B a binarization process, $\mathcal{F}$ the image's Fourier transform [177], $|\cdot|^2$ the power spectrum (S), n the number of image pixels, V a vectorization algorithm for the input matrix (described below), H_r the relative Shannon entropy, T a transfer function used for convenience purposes (also described below), and dSIP $\in [-1, +1]$ the divergence from maximum SIP. The utility of dSIP is to concomitantly provide information about the intensity of the information potential as given by the absolute value, |dSIP|, and locate the pattern type on the uniform–clustered–regular continuum, made explicit by the signum, sgn dSIP. For the sake of simplicity, we will use the same notation ("SIP") to denote both the concept of structural information potential and its instantiation in a particular mathematical expression; when necessary, a qualifier will provide clarification.

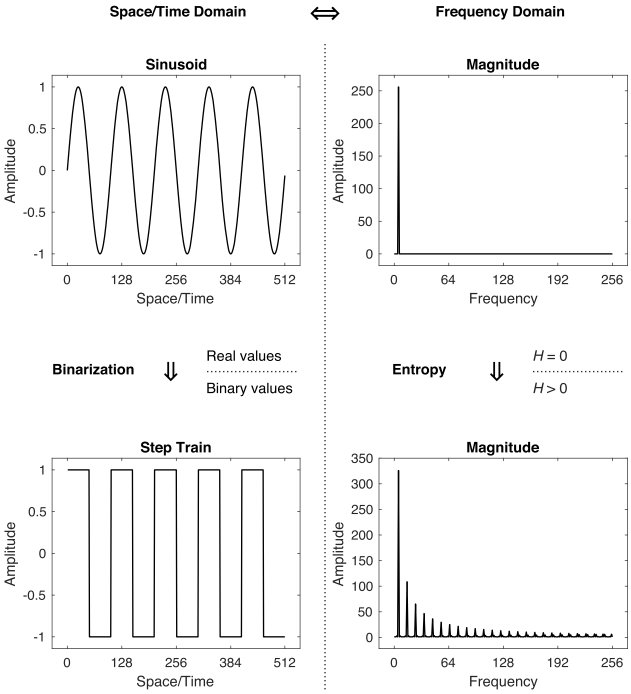


FIGURE 5. Illustration of the effect of binarization of a pure sine signal on the Shannon entropy, *H*, of its frequency domain magnitude.

2) FREQUENCY DOMAIN AND ENTROPY

The choice of the frequency domain for pattern analysis has a number of benefits [178]: it facilitates the integration of sample points across space, and thus the description of spatial structures in terms of regularity; it enables characterization of the degree of clustering via the scale-space spectrum of frequencies; finally, it allows for a pattern phase space to be obtained for which the extreme values correspond to regular and uniform patterns in the spatial domain. (Recall that the magnitude of the Fourier transform does not encode the spatial location of structures, but rather contains information about their form. For example, the magnitude represents the frequency and amplitude of the sine wave, but provides no information regarding the location of the local maxima; this information is carried by the phase component of the Fourier transform. From a SIP perspective, however, it is the form and not the location of a pattern that is of primary concern.)

The use of the logarithm of the power spectrum stems from the method (outlined above) for determining the fractal dimension and is intended to allow the characterization of clustered patterns. Normalizing the power in equation 5 makes it independent of image size; adding one to the squared magnitude avoids the logarithm of zero and negative output values. To put SIP in a fractal perspective, SIP is a measure of how far a pattern is from potentially being a fractal.

In addition to having a dimensionality reduction effect, the role of the Shannon entropy is to help devise a linear space in which patterns are ordered from uniform to clustered to regular. This can be achieved if a transform is found such that the extrema of the pattern space correspond to a vector of zeros except for one value (i.e., an impulse) and a vector with equal values (i.e., uniform), respectively; in that case, the entropy will range from zero to one.

3) BINARIZATION

A uniform signal in the time domain has as its pair in the frequency domain an impulse with frequency of zero [179, p. 33]. As this frequency is ignored in the SIP measurement method, the Shannon entropy value of the remaining power spectrum will also be zero, as desired. Considering that the Fourier transform uses sines and cosines as base functions, the sinusoid is the regular pattern with the fewest spectral artifacts (e.g., aliasing, harmonics, Gibbs effect [180, pp. 194–200, 218–222]). It becomes an impulse in the frequency domain [179, p. 29], and thus has zero entropy and represents an undesirable outcome for our goal. However, by binarizing the sinusoid, a step train is obtained, which corresponds in the frequency domain to a set of harmonics of the form $1/(f\,\pi), f \in \mathbb{N}_{\text{odd}}$ [181, pp. 102–113] [180, p. 257], the effect of which is to increase the signal's spectral entropy (Fig. 5). A slightly different, but relevant, regular signal is the pulse train, where the step length differs from the distance between consecutive steps. As the spatial length of the pulses expands and the pattern becomes more uniform within the finite signal bounds, the power distribution is increasingly compressed towards the lower frequencies [180, p. 201] and the spectral entropy decreases, as desired.

In summary: (*a*) we translate the problem of devising an analytic formulation of the pattern–informativeness space in the frequency domain in order to be able to characterize spatial structures; (*b*) we employ the entropy because its extrema are the impulse and the uniform distributions; (*c*) we apply data binarization to transform the spectral representation of regular patterns from the impulse to the uniform distribution, so that the spectral entropy yields the desired uniform–clustered–regular continuum. Taking advantage of the spectral artifacts introduced by binarization is key to obtaining measurements of the spectral entropy that order patterns in a perceptually consistent manner.

The global Otsu binarization algorithm [182] has been used to produce the SIP measurements presented in the figures of this article. This general-purpose method was appropriate for our intention to preserve document noise, a particularly important aspect of the card images presented in the case study, as noise was found to have a direct impact on the quality of the optical character recognition. In the case of the head pictures of Fig. 21, however, the background shadow was impinging on the preservation of facial details that were the focus of interest during binarization; for this reason, the locally adaptive algorithm of Raleigh was chosen [183]. Conversion from color images to grayscale is realized by converting the images to the perceptual CIELAB color space and extracting the lightness channel, $L*$ [184, pp. 30, 95, 200–212]. The binarization step of the SIP measurement is not necessary for data that is already binary.

4) DIMENSIONALITY REDUCTION

The dimensionality reduction of the two-dimensional power spectrum to a vector is performed to ensure the rotation-independent measurement of patterns. The step consists in averaging data points of identical frequency. Due to quantization, however, digital images have sparse spectral representations; for example, the lowest frequency is expressed only at two orientations, 0 and π radians, in the usual Cartesian representation of the Fourier transform. To guarantee sufficient data and avoid flattening the spectrum (which artificially increases entropy), the frequencies are therefore rounded to the nearest integer prior to averaging the corresponding power spectrum values. We also disregard the direct component (DC; it has a frequency of 0) from the computation, which has no impact on the image pattern, since it represents the mean signal power. The vectorization procedure can be formally expressed as

$$w = \text{int}(\omega + 0.5), \tag{6}$$

$$n = \text{card}(\text{unique}(w)), \tag{7}$$

$$V = \sum_{i=1}^{n} P_{w_i}/\text{card}(w_i), \tag{8}$$

where the vector ω contains the frequencies corresponding to the power spectrum values S, the matrix into which the vector is indexed, with $\omega \in \mathbb{R}_{\geq 0}$; w is defined over the integer frequencies, $w \in \mathbb{N}_{\geq 0}$; n is the index of the rounded Nyquist frequency into the unique values of w; and card(w_i) is the number (cardinality) of samples for a given integer frequency. (The Nyquist frequency for an image is half the number of pixels of the image diagonal.)

5) MAXIMAL SIP

Given that the SIP method orders patterns along the uniform–clustered–regular continuum, the following two fundamental questions arise: "What is the pattern corresponding to maximal clustering?" and "What dSIP value does this pattern have before it was calibrated to 0?" The second question will be answered here, while a tentative answer to the first question is provided in section III-C4.

Referring to the dSIP equation (5), we are interested in finding the value $\mathfrak{f}$, such that $T(\mathfrak{f}) = 0$. Whatever this value might be, it should correspond to a maximally clustered pattern for any length n of the power spectrum (S) of equation (5), including for $n = 2$. Anticipating the explanations to follow, maximal clustering corresponds to minimal redundancy; for the division of a whole into two parts (hence, $n = 2$) this is the reciprocal, Φ, of the golden ratio, ϕ, $\Phi = 1/\phi = (\sqrt{5} - 1)/2 \approx 0.6180$ (see section III-C1). We further

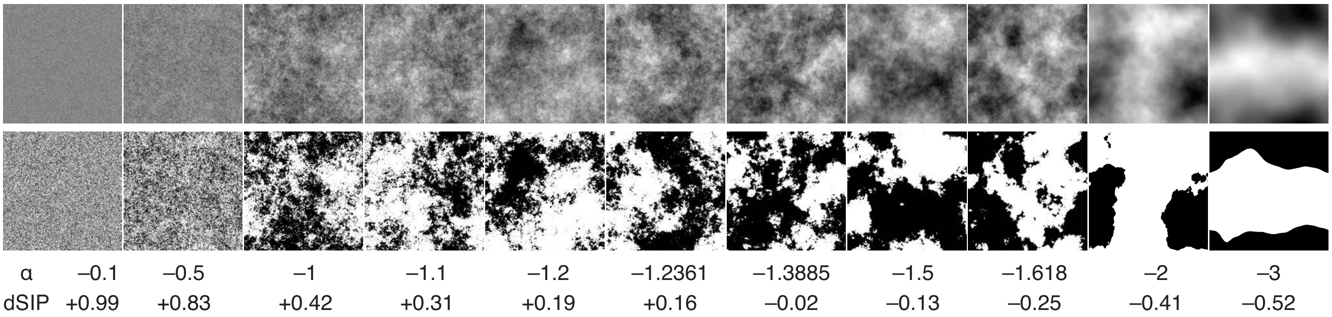

α –0.1 –0.5 –1 –1.1 –1.2 –1.2361 –1.3885 –1.5 –1.618 –2 –3
dSIP +0.99 +0.83 +0.42 +0.31 +0.19 +0.16 –0.02 –0.13 –0.25 –0.41 –0.52

FIGURE 6. Intensity (top row) and binarized images (bottom row) of random fractals, with various power spectrum exponents, α, and their corresponding dSIP values. The third value is the negative of the golden ratio ($-\phi \approx -1.618$), the fifth value is the negative of twice the reciprocal of the golden ratio ($-2\Phi = \sqrt{5} - 1 \approx -1.2361$), and the sixth value is twice the negative of the value $\mathfrak{f}$ of equation 9 ($-2\mathfrak{f} = 2\log_2 \Phi \approx -1.3885$). The fractal images are obtained by synthesis in the frequency domain, through inverse Fourier transform from $\sqrt{f^\alpha}$ magnitude, random phase, and DC = 0; a subsequent zero-level cut produces the binary images [51, pp. 49–50] [49, p. 122]. The fractal dimension, D, of the image corresponding to $\alpha = -2\mathfrak{f} \approx -1.3885$ is $D = (6 + \alpha)/2 = 2.3058$. For the "paradoxon" of images with fractional dimensions $D \notin [2, 3]$, i.e. outside the expected bounds defined by the topological and Euclidean dimensions, see [53].

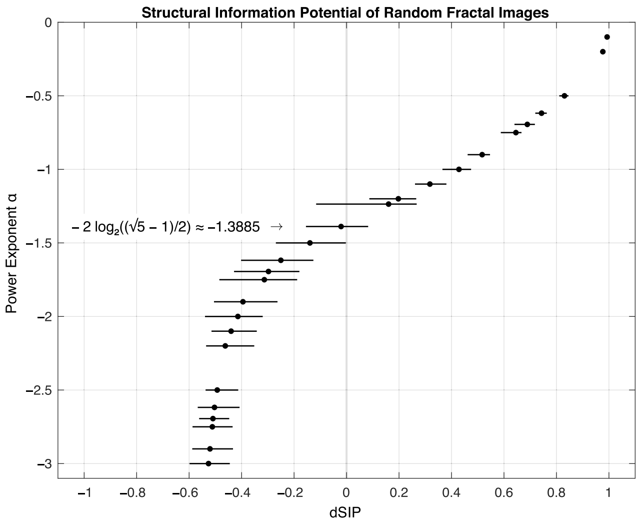


FIGURE 7. Median (dot markers) and range (bar markers) of dSIP values of one hundred binary random fractal images, a sample of which are shown in Fig. 6, vs the exponent α of the fractals' power spectrum power law distribution, f^α. The median dSIP of the images with $\alpha \approx -1.3885$ is -0.0208.

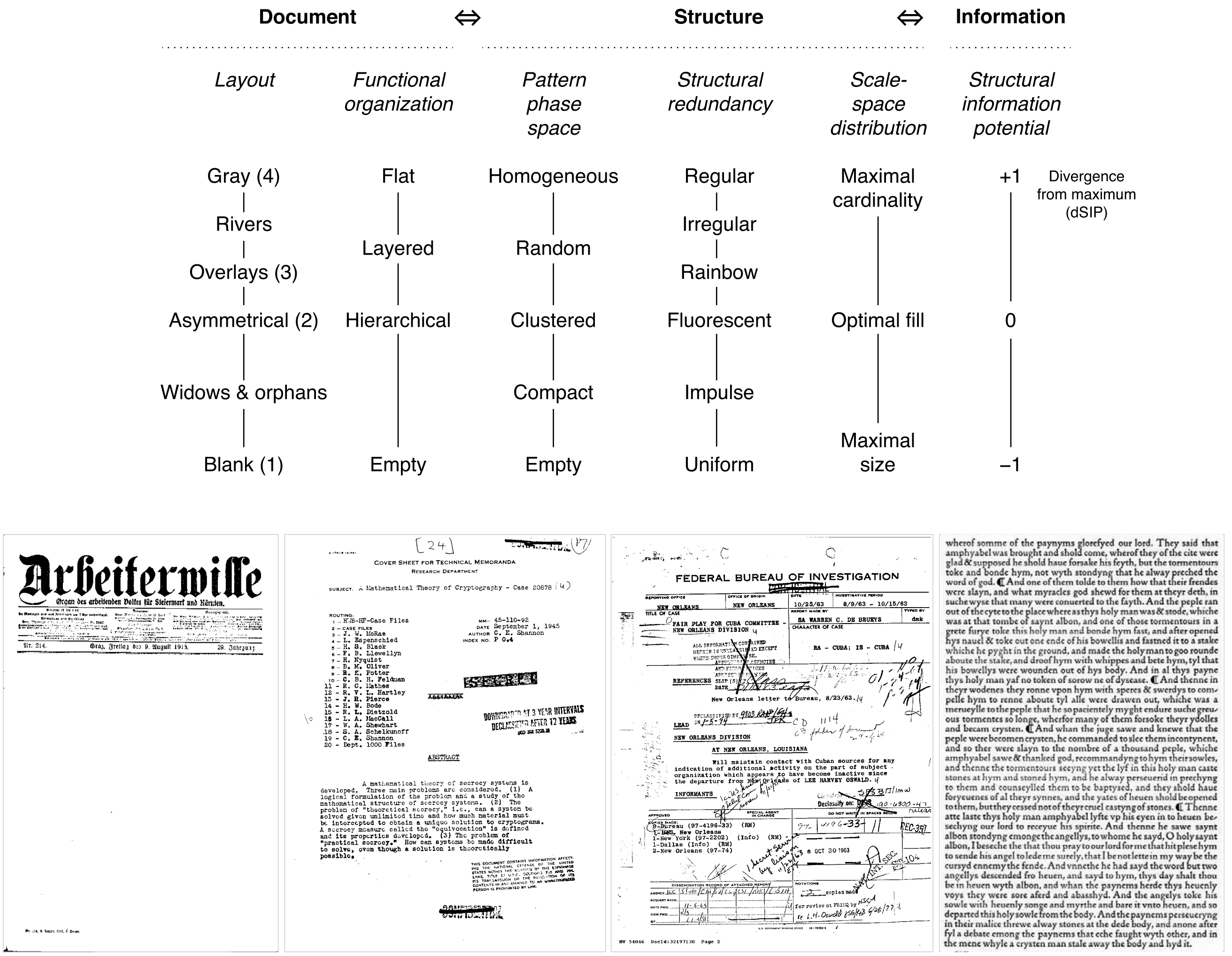
note that the input to T is modified logarithmically by the relative entropy H_r. Thus, we define $\mathfrak{f}$ (let it be called the "*fluorescent number*", as related to the most conspicuously clustered patterns, and following the graph theory terminology applied to the rainbow family of graphs) as the logarithm of the reciprocal of the golden ratio, or, equivalently, as the logarithm of the golden ratio:

$$\begin{aligned}\mathfrak{f} &= -\log_2 \Phi = -\log_2((\sqrt{5} - 1)/2) \\ &= \log_2 \phi = \log_2((\sqrt{5} + 1)/2) \\ &\quad 0.6942. \end{aligned} \tag{9}$$

We have empirically validated this logic multiple times with the numerous datasets we have processed. The reader may ascertain for her- or himself that the patterns with a dSIP close to 0 do indeed present the highest degree of clustering. A good example reference is the cover drawing of the magazine shown in Fig. 15 (framed in red), which is highly clustered due to the variety of shape sizes and the perspective view. It has a dSIP value of +0.0339, which is close to minimal, as expected. Another example drawn from the case study is the card with minimal dSIP and high clustering in Fig. 18e and 19. Furthermore, the argument is also supported from an analytical point of view by the following observation. If we consider the epitome of clustered patterns — the fractals — then we can observe that by using the value $-2\mathfrak{f}$ as the exponent of their defining power law distribution of the power spectrum, $f^{-2\mathfrak{f}}$, we can obtain fractal images with a median dSIP close to zero, dSIP $= -0.0208$ (the doubling of $\mathfrak{f}$ is due to images having an exponent twice the value it has for signals [51, pp. 49–50] [49, p. 99, 105, 122]; Fig. 6, Fig. 7).

6) TRANSFER FUNCTION

The transfer function T calibrates the structural information potential value to facilitate easier mathematical manipulation and cognitive interpretation. However, the following issue arises: due to the use of the relative entropy H_r in equation 5, the values within the transfer function T are bound to the [0, 1] range, the extrema of which correspond to uniform and regular patterns (see section III-B). As explained above, the pattern with maximal clustering (and, therefore, maximal structural information potential) has the value $\mathfrak{f} \approx 0.6942$. However, $\mathfrak{f}$ is not the center of the [0, 1] range, and the numerical pattern informativeness levels computed with the SIP method before calibration (say 0.4936 and 0.8107) are neither immediately comprehensible nor directly comparable. The solution we propose is to use a transfer function, T, to remap the values such that maximal clustering is attained for 0, maximal uniformity for -1, and maximal regularity for $+1$. We will then say that dSIP is a measure of *divergence* from the maximum structural information potential, with its sign indicating the pattern towards which it tends: positive for a more regular pattern and negative for a more uniform pattern.



(a) Uniform (b) Clustered (c) Random (d) Regular

FIGURE 8. **Conceptual schema of the generic relationship between information and structure, along with the correspondence of the visual and functional organization of documents. — (a) Uniform pattern, −0.106 dSIP; censored newspaper front page; Austria, August 9, 1918 (Austrian National Library). — (b) Clustered pattern, +0.173 dSIP; memorandum frontispiece, including institutional header, summary, circulation list, and secrecy stamps (Claude Shannon, "A Mathematical Theory of Cryptography", 1945, AT&T Bell Laboratories) [185], [186]. — (c) Random pattern, +0.276 dSIP; official FBI printed form, heavily annotated by handwriting and stamps (Kennedy Assassination Records Collection, National Archives and Records Administration). — (d) Regular pattern, +0.424 dSIP; typography inspired by medieval manuscripts, by the Arts and Crafts movement exponent William Morris. ("The Golden Legend", 1892, London, Kelmscott Press/Quaritch Books; author's personal collection).**

Furthermore, to honor the linguistic expression "maximal (or minimal) structural information potential"—which is awkwardly 0 (respectively both [*sic*] −1 and +1) when computed via dSIP—we will define SIP as given by equation 4, which is maximal at 1 and minimal at 0, representing a more intuitive situation.

To map the values from the range [0, ƒ, 1] to [−1, 0, +1], we fit a second-order polynomial to the three value pairs and obtain the following coefficients for an input value x (where x represents the input into T in equation 5) and output y:

$$y = 1.83\,x^2 + 0.1699\,x - 1. \tag{10}$$

As a further refinement, we counterbalance the non-linearity resulting from equation 10 by taking the logarithm of the output y for the final definition for the transfer function T:

$$T = \begin{cases} \log_2(y+1), & \text{for } y \geq 0. \\ -\log_2(|y|+1), & \text{for } y < 0. \end{cases} \tag{11}$$

B. PATTERN PHASE SPACE

Consider a raster image in which the pixels take the values of zero and one. The most *homogeneous* pixel distribution is obtained for a checkerboard pattern where every pair of adjacent pixels have distinct values; in other words, when the pattern is *regular*. Any disturbance of this regularity will decrease the homogeneity, including *random* disturbances. When the disturbance is not deterministically or stochastically uniformly distributed across the image, then pixel *clusters* will emerge, with maximal clustering attained when the clusters have sizes at all scale-space levels (under these circumstances, the entropy of the scale-space will also be maximal). Aside from the regular pattern that exhibits pixel objects of only a single size, and is thus located at a single scale-space level (and has zero entropy), another pattern with this characteristic is the *uniform* pattern. It appears that when pixels of identical value "coagulate" into a *compact* mass, the surrounding of this mass becomes uniform. These distributions define a unidimensional space of pattern "phases" that vary from regular to random to clustered to compact to uniform.

Fig. 8 provides a practical example of the pattern phase space described above in theoretical terms. It shows the SIP classification of sample document images along a unidimensional pattern space. The progression from regular pattern to clustered to uniform is readily observable and conforms to the problem requirements. Note in the upper part of the image the parallel between the terminologies from various fields (preeminently typography, signal processing, and material sciences). The conceptual system introduced here, and further discussed below, may be used as a model to describe the correspondence between the visual and functional organization of documents. At the paragraph level, the penmanship and typographical ideals are the production of a homogeneous pattern (appearing as gray when observed from a distance), while the functional hierarchy is reflected in a clustered layout (the technical term is "asymmetrical"). Post-production annotations and degradations disturb the intended (ir)regularity, introducing randomness. Blank pages (etymologically French for "white", another chromatic term in the field of documents) mark the end of a functional unit, as well as containing a reduced amount of semantic information.

The uniformly *random* pattern is located between the clustered and the regular, since it does not comprise large uniform areas. As such, their dSIP value may be very high (Fig. 6). Random patterns can occur due to signal noise or annotations (in the case of text documents). Note that other pattern spaces are possible: the uniform–random spectrum [187] is appropriate for patterns with statistically constant homogeneity, such as homogeneously distributed line segments of quasi-equal length and variable orientation.

Patterns with *semantic* value, such as images of faces, are obviously very important for human beings, and can be located anywhere in the proposed pattern space. Because these patterns require semantic analysis and contextual information, they cannot be considered from a low-level computer vision perspective such as that described here, just as the Shannon entropy casts information in purely mathematical terms.

C. GRAPH STRUCTURES WITH MINIMAL REDUNDANCY

While previous sections dealt with the measurement of SIP, this section concerns the definition and design of mathematical objects with minimal structural redundancy; that is, those that maximize the structural information potential.

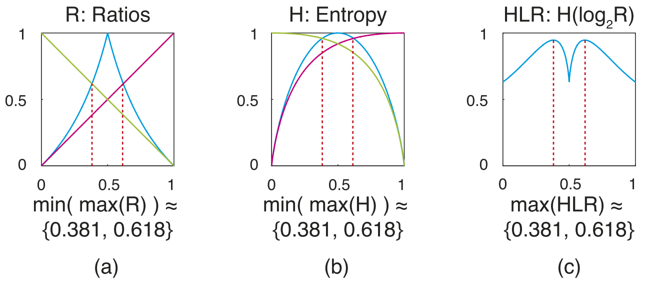


FIGURE 9. (a) Given the unit segment *c* and its division in two segments *a* and *b*, the figure shows the values of the ratios *a*/*c* (red), *b*/*c* (green), and *a*/*b* for *a* ∈ [0, 0.5] and *b*/*a* for *a* ∈ [0.5, 1]. The minima of the maximum of the ratios correspond to the reciprocal and the reciprocal conjugate of the golden ratio (dotted lines). — (b) Relative Shannon entropy of the segments taken pairwise. — (c) Relative Shannon entropy of the logarithm of the pairwise ratios.

While measurement may be employed independently of design, a discussion of the latter supports a firmer understanding of the former. In particular, such discussion provides a formal rationale for the calibration of SIP values, as well as visual evidence and quantitative characterization of objects with maximal SIP value. Furthermore, it reveals links between SIP and some interesting mathematical concepts, thereby creating an opening for generalizing SIP to structures other than images.

We will refer to objects with minimal structural redundancy as "*fluorescent*" objects, in reference to the "rainbow" graphs (which have edge values that are distinct, but not necessarily maximally distinct), to the name of the symbol, *phi*, denoting the golden ratio (which minimal redundancy properties as discussed below), and to the fact that fluorescence maximizes perceptual color discrimination. The proposed notation is $\mathfrak{F}^{v\mathrm{V}}_{d\mathrm{D}}\{\cdot\}$, where d is the object's embedding dimension, v the number of vertices, and the placeholder $\{\cdot\}$ may be used to specify edges. For example, $\mathfrak{F}^{3\mathrm{V}}_{2\mathrm{D}}\{\mathrm{e}_{12}, \mathrm{e}_{13}, \mathrm{e}_{23}\}$: $\{x_1, y_1, \ldots, x_3, y_3\}$ designates a fluorescent triangle and the vertices' coordinates.

In the following, the argument will proceed from one-, to two-, to three-dimensional objects, from shapes to patterns, and will conclude with some remarks of a more general order.

1) ONE-DIMENSIONAL FLUORESCENCE FOR THREE POINTS

How might a whole be divided into two parts so as to maximize the difference between the parts, while concomitantly maximizing their respective sizes? The first condition of this problem is satisfied when either of the parts vanishes in the limit, while the second condition corresponds to the two parts being equal. The overall solution lies in between these extrema and is found by determining the value at the intersection of the functions representing the conditions, i.e., $f(x) = 1 - x$ and $f(x) = x/(1-x)$, for $x \in [0, \frac{1}{2}]$ and $f(x) = x$ and $f(x) = (1-x)/x$, for $x \in [\frac{1}{2}, 1]$, or solving the equation $x^2 - 3x + 1 = 0$ and $x^2 + x - 1 = 0$. For the definition domains, these have the solutions $x_1 = (1-\sqrt{5})/2+1 \approx 0.3819$ and $x_2 = (1+\sqrt{5})/2-1 \approx 0.6180$.

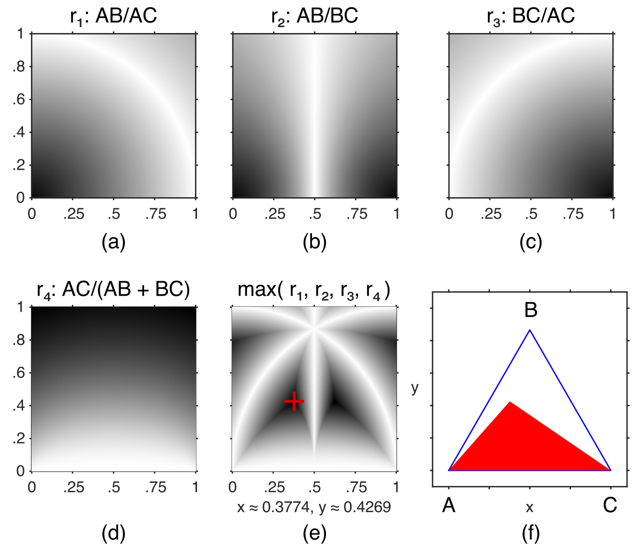


FIGURE 10. Design process of the fluorescent triangle. — (a)–(c) Pairwise edge ratios of a triangle ABC, with unit base AC, for all locations of vertex B within a unit square. Grayscale tones encode values between 0 (black) and 1 (white). — (d) Ratio of base and sum of opposing edges. — (e) Maximum of the surfaces defined by the edge ratios. The minimum of this surface for $x \in [0, \frac{1}{2}$ is marked by a + symbol, and its coordinates are specified next to the ordinate. — (f) The triangles with the most and least dissimilar edge lengths, i.e., the isosceles triangle (blue outlines) and the fluorescent triangle (solid red). Basic mensurations: apex: $x \approx 0.3774$, $y \approx 0.4269$; edge lengths: $AB \approx 0.5698$, $BC \approx 0.7548$, $AC = 1$; angles: $\alpha \approx 45.5199°$, $\beta \approx 97.0399°$, $\gamma \approx 34.4400°$; area: $T \approx 0.2134$.

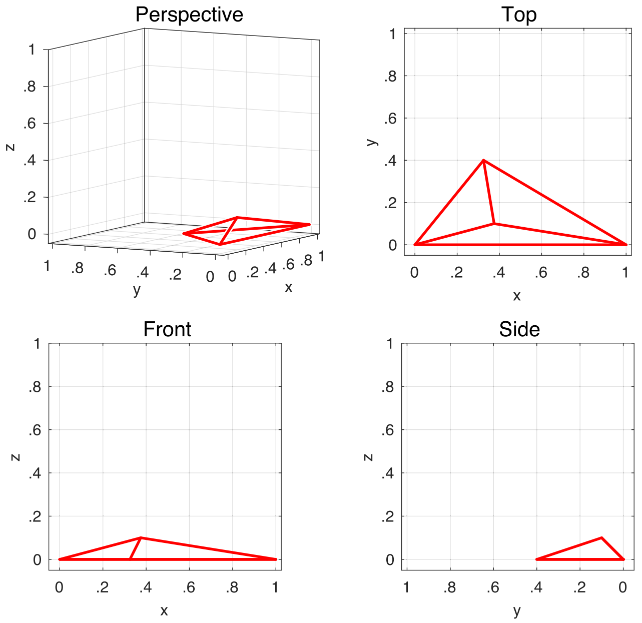


FIGURE 11. Fluorescent tetrahedron based on minimization of pairwise edge ratios and ratios of one edge and two adjacent edges. Coordinates of base triangle apex on a grid with resolution of 0.025 units: $x = 0.325$, $y = 0.4$, $z = 0$; tetrahedron apex: $x = 0.375$, $y = 0.1$, $z = 0.1$.

These values are also known, respectively, as the conjugate of the reciprocal, Φ', and the reciprocal, Φ, of the golden ratio, $\phi = (1 + \sqrt{5})/2 = 1/\Phi = 1/(1 - \Phi') \approx 1.6180$. A diagrammatic representation of these solutions (Fig. 9a) allows us to state the criterion for determining the minimal structural redundancy (or maximal SIP) as the minimum of the maximum of the ratios of the part and the whole: $R = \{a/(a+b), b/(a+b), \min(a, b)/\max(a, b)\}$, $\text{SIP}_{max} = \min(\max(R))$, where $a+b = 1$, and $(a, b) \in [0, 1]$. In other words, maximal SIP corresponds to the minimum of the range of the ratios of a system's components. The use of these ratios is equivalent to using the relative Shannon entropy, H, for each edge pair (Fig. 9b). The max formulation can be avoided by combining ratios and entropy: $\text{SIP}_{max} = \max(H(-\log_2(R)))$ (Fig. 9c). The similarity between this equation and equation 5 indicates that maximal SIP is expected for $-\log_2(\Phi)$, which is the reason why this value was used for calibration in the SIP measurement algorithm.

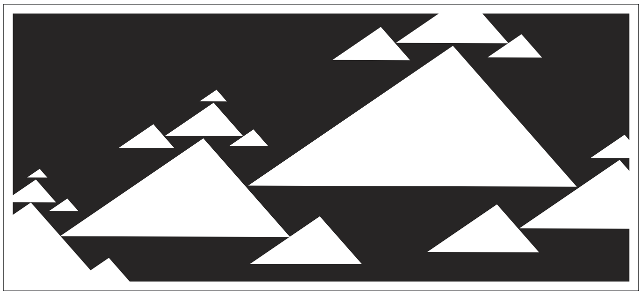


FIGURE 12. Fluorescent pattern generated from a fluorescent triangle. Note the similarity with the naturally occurring patterns of seashells [188].

Remark — Note the inclusion of the whole as a third element in the measure of the sizing of the two segments. This is explained by the whole representing the highest scale of the scale-space domain defined by the segments, thus making it part of the system. This is the case in many application domains, particularly those that are subject to human factors, such as document design and perception. For instance, the sizing according to the golden ratio of the width of a text column and a figure placed next to each other ensures maximal legibility for each while giving prominence to one of them.

2) TWO-DIMENSIONAL FLUORESCENCE FOR THREE POINTS

Let us now extend the problem to two dimensions and ask the following question: *What is the triangle with the most dissimilar edges?* Above, we identified the solution for three collinear points that define a degenerate triangle. While this triangle has minimal redundancy between its parts when taken pairwise, it also has zero area, unlike common triangles. More importantly, the degenerate triangle has maximal redundancy between a part and a subset: the triangle base AC equals the sum of the other edges, AB and BC. Therefore, we introduce the ratio of the base and the sum of the other edges as part of the problem formalism, with the effect of increasing the size of the triangle to a certain equilibrium point below that of an equilateral triangle, for which redundancy is maximal. We now solve the equations $AB/BC = BC/AC = AC/(AB + BC)$, given the coordinates x and y of the triangle apex, with $AB = \sqrt{x^2 + y^2}$, $BC = \sqrt{(1 - x)^2 + y^2}$, and $AC = 1$. The solutions in the abscissa interval $[0, \frac{1}{2}]$ yield $x \approx 0.3774$ and $y \approx 0.4269$. This is the elemental shape with maximal SIP, as per our definition (Fig. 10). A remarkable trait of this triangle is that it extends an essential property of the golden ratio—that of minimal redundancy between parts and whole—to two dimensions.

3) THREE-DIMENSIONAL FLUORESCENCE FOR FOUR POINTS

Fig. 11 illustrates the extension of the fluorescence principle to three dimensions, here for a tetrahedron with minimal edge redundancy. This tetrahedron represents a pendant of the equilateral tetrahedron, one of the five regular Platonic solids. The locations of the vertices of this graph have been computationally determined by placing vertices v_1 and v_2 at two adjacent nodes of a unit cube, then placing vertices v_3 and v_4 in turn at all locations of an orthogonal grid with resolution of 0.025 units. To avoid triangles with identical shape but different sizes and orientations (metamers), locations for which the length of edges ending in vertices v_1 or v_2 is more than unity were excluded. The measure of redundancy was based on the ratio of pairwise edges, as well as the ratio of one edge taken at a time and its adjacent two edges. The latter criterion is adopted from the triangle case discussed above, and its effect is to avoid facets with zero area. This benefits the fluorescence measurement, as it neatly facilitates the quantification of redundancy across dimensions and structure types, such as mesh surfaces and solids. (This observation highlights the possibility of using criteria other than the ratio of edge subsets, such as the perimeter and the volume, to avoid shape degeneration; these would lead to different solutions, but would be more loosely consistent with the problem definition in terms of edge ratios alone. Instead, they may serve special applications, such as maximization of the volume generally resulting in convex solids.)

The inquiry into minimal structural redundancy can be pursued for graphs with an arbitrary number of edges, such as polygons and mesh surfaces, or of arbitrary structure, such as circular, tree-like, or network-shaped graphs. In this article, the goal is limited to open a window into these possibilities.

4) TWO-DIMENSIONAL PATTERN FLUORESCENCE

We investigated minimal structural redundancy for only the most basic graphs (line bisection, triangle, and tetrahedron), stopping short at identifying the fluorescent pattern(s) with arbitrary size. A family of fractals is the likely answer—but is there a fractal dimension more apt to produce fluorescence than another? Furthermore, considering deterministic fractals (such as the Koch curve [46, pp. 87–91]), it could be argued that there may exist a single fractal that minimizes structural redundancy. I propose that the design of such a pattern may consist in taking the simplest structure of a given dimension (e.g., the fluorescent triangle for two dimensions) and replicating it infinitely with a certain size change rate (Fig. 12). The result consists, in fact, in an affine transform of the module of well-known fractals (Cantor dust, Sierpiński gasket and tetrahedron, and Julia set [46, pp. 65–79, 120–123]), so as to morph them into a fluorescent shape.

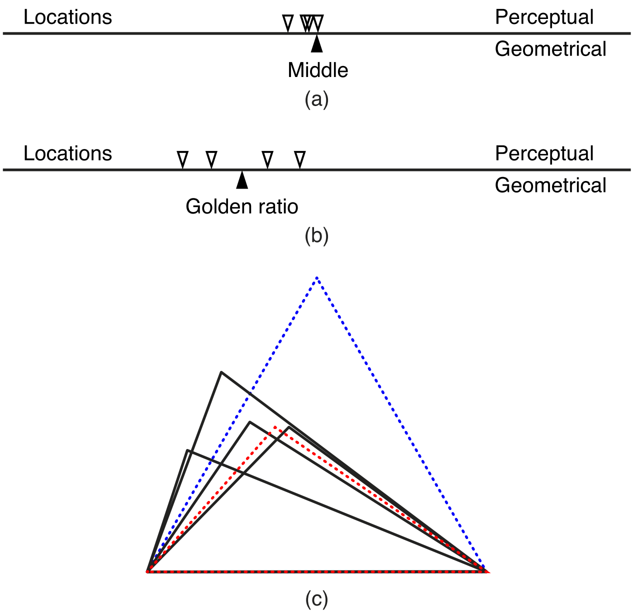


FIGURE 13. Comparison of mathematical and perceptual fluorescence. — (a) Human experimental participants tendentially place the middle of a line left of its geometrical location. Here, an example from four male computer scientists, aged late 20s to late 40s, in which the outlined markers above the line represent their choices. — (b) The same group was tasked to segment a line into the two most dissimilar segments with respect to each one, as well as to the whole line, i.e. according to the golden ratio. — (c) In this task the participants are instructed to draw a triangle with the most dissimilar edges, i.e. a fluorescent triangle, shown in dotted red, as opposed to the equilateral dotted blue triangle.

5) FLUORESCENCE AND SIP

Let us examine the link between SIP and minimal graph redundancy, two major themes of this section.

The concept of structural information potential, on one hand, posits a correlation between structure and information, based on the scale-space distribution of the structure, which defines a pattern spectrum ranging from compact to homogeneous via clustered and random. This perspective on patterns is quantified by the SIP method in the frequency domain using the Fourier transform and the Shannon entropy. The concept of fluorescence, on the other hand, attempts to characterize structures in terms of component redundancy, and to this end adopts a graph theoretical formalism of the ratio of combinations of edge subsets.

The integration of these two perspectives is realized through the concept of scale-space redundancy, which subtends the pattern spectrum and has a direct information theoretical meaning (Fig. 8). Specifically, a pattern that maximizes both the *number* and the *size* of constituent entities will fill the scale-space in an optimal manner while having maximal structural information potential and minimal structural redundancy. This is the fluorescent clustered pattern, which stands in opposition to the uniform pattern of a maximally sized single entity, as well as the homogeneous pattern of minimally sized and maximally numerous entities.

In conclusion, the study of graph redundancy offers a powerful and versatile instrument to understand the elemental levels from which clustered and structurally informative patterns emerge. One goal of this section is to provide precisely such a low-level explanation of the SIP method.

6) FLUORESCENCE AND HUMAN PERCEPTION

Human estimation of pattern redundancy differs from mathematically derived values [189], [190], which is an important factor to consider when measuring man-made patterns (such as documents) or evaluating psychophysically formal methods of redundancy measurement. A well-known example is the systematic bias in line bisection (Fig. 13a) [191]. The location of the fluorescent line partition and the drawing of the fluorescent triangle exhibit similar variability and divergence from the mathematically defined shapes (Fig. 13b, c). Psychological aspects, furthermore, interact with cultural and social ones, forming a much more complex and dynamic ecosystem of constraints upon informativeness than the basic, low-level uniform–clustered–regular pattern-informativeness spectrum. The subjective and contextual dimensions of human pattern perception may nevertheless be directly relevant to the evaluation of the SIP measurement method; for instance, to explain why observers might disagree about the redundancy of a given pattern. A converse issue is that of human-designed fluorescent configurations, such as the text/figure sizing in document layout mentioned above, whose divergence from the mathematical model may even be intentional. In particular, this may often be the case with regard to the uncritical application of rules often shunned by artists, keen observers of form, as illustrated by the words of the illustrious French photographer Henri Cartier-Bresson: “I hope that we will never see the day when the merchants will sell [golden ratio] diagrams engraved on camera displays.” [192, pp. 26–27]. The quote exemplifies the need to develop models of layout irregularity that integrate human factors. When applied to documents, the desideratum has been carried out in the design of the SIP formalism, which emerged from empirical observation and reflects low-level perception. At such levels of complexity, cognitive automatisms are more amenable to modelization by the pattern–informativeness spectrum. The task of document triage, with its emphasis on fast temporal information processing, is a good case study for testing the model and shall be discussed next.

7) REPRESENTATION

The opposing states of minimal and maximal redundancy may be represented in a single view. Fig. 14 presents the panchromatic prism, whose top is an equilateral triangle (i.e., “gray” in chromatic graph theory terminology), while the base is a triangle with most dissimilar sides (i.e., “fluorescent”). This mathematical object is visualized in the manner of Leonardo da Vinci’s drawings of the five Platonic regular bodies [193].

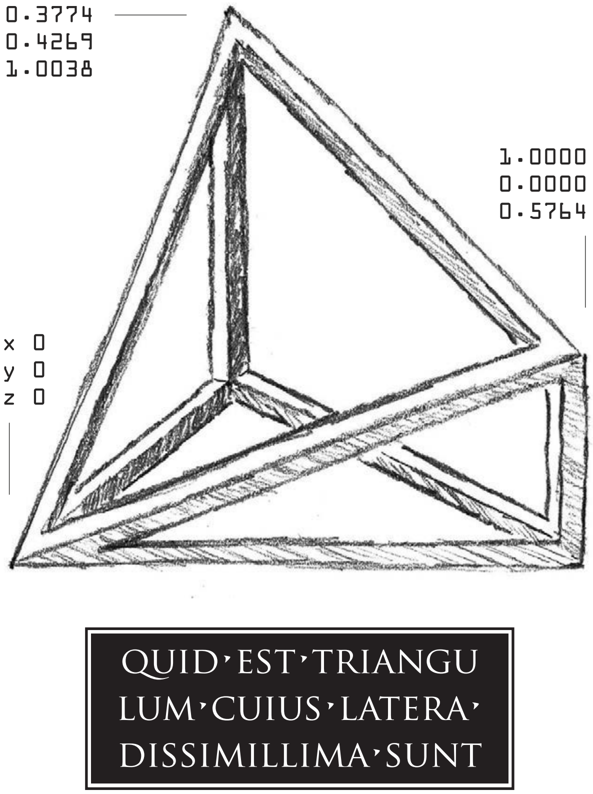


FIGURE 14. The panchromatic prism.

IV. EXPERIMENTS

In this section, we apply the method of document ordering developed above to a small set of representative real data, then compare the results to those of other methods. Next, we use a case study to test the robustness of our approach on a large dataset with respect to the triage task. Finally, we demonstrate that the proposed concept may be generalized to objects beyond text-based document images and applications other than triage.

A. COMPARISON OF METHODS

1) VISUAL EVIDENCE

Fig. 15 juxtaposes document pages ordered using the major methods discussed in this article: structural information potential, approximate entropy, spectral flatness, and the ratio of ink pixels to page area. The data are sampled from the 104 binarized pages of an issue of the *New Yorker* magazine (please refer to Fig. 26 to 31 to view all pages). This particular dataset was chosen for display among the hundred analyzed owing to its diversity of text, drawing, and image patterns, which provides exemplary illustration of the entire uniform–clustered–regular spectrum against which we want to test our methods.

We first observe that the ink/page ratio method leads to a mix of structurally unrelated patterns, and to a split of

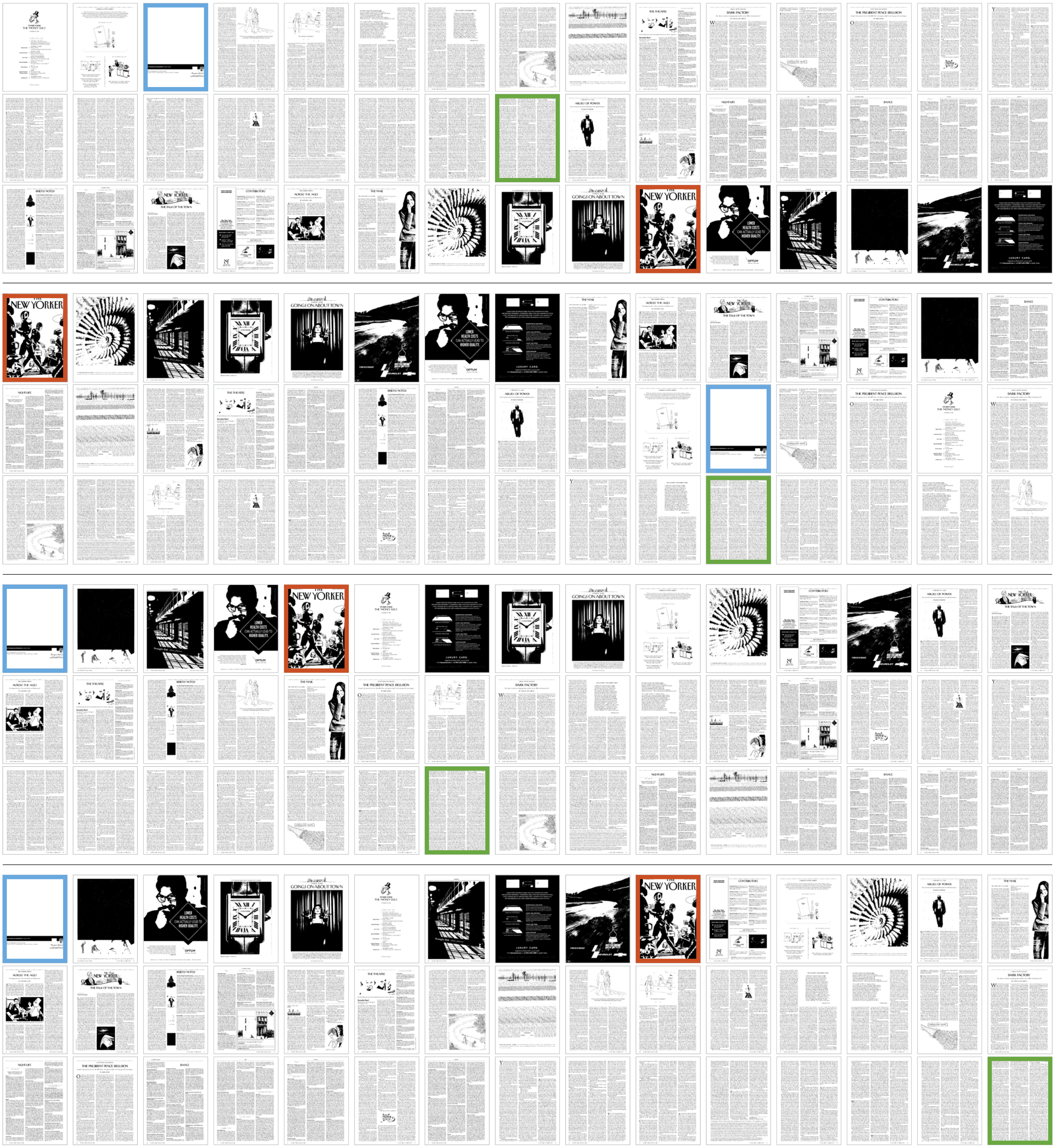


FIGURE 15. Ordering (left to right and top to bottom) using the ink-to-page-area ratio (first top block), approximate entropy (second block), spectral flatness (third block), and structural information potential (last block; *dSIP* ∈ [−0.3109, +0.6523], min(*dSIP*) = −0.0073, corresponding to the eighth image in the first row). Colors mark the most uniform (blue) and homogeneous (green) pages, and the highly clustered cover page (red). — Credit: *The New Yorker*, vol. 93, no. 33, October 23, 2017; Condé Nast.

homogeneous clusters, such as the mostly empty pages (the third image and the third from last image of the top-most block in the figure). Approximate entropy is very good at clustering multiscale patterns, such as the first three pages of the sequence in the second block, which depict either entities with different absolute sizes (e.g., people and robots in the street, the rays of the trilobite), or a diversity of sizes resulting from a perspective view (rows of bars in a prison hallway). However, the method fails to group together homogeneous patterns (text-only pages are interspersed with text and illustration pages) and empty pages. Spectral flatness (third block) groups empty pages and follows them with multiscale patterns, but fails to consistently group homogeneous patterns (the result of spectral entropy is similar, but less good). Structural information potential (last block) achieves a perceptually gradual ordering of the pages from empty to multiscale to homogeneous.

2) QUANTITATIVE EVALUATION

The previous section has presented visual evidence that the SIP method provides a better ordering of document pages on the uniform–clustered–regular pattern spectrum than the other methods considered. We now quantify the difference in suitability between the methods.

Method — Let the monotonously increasing or decreasing ranking of n items $P = \{p_1 \prec p_2 \prec \ldots \prec p_n\}$ and $P' = \{p_n \succ p_{n-1} \succ \ldots \succ p_1\}$, respectively, be the baseline orders of comparison. Any permutation Q of the items will be a less desirable ordering to a degree in accordance with a certain criterion of desirability. In the framework of the pattern spectrum, this is the case when two patterns from the opposite sides of the spectrum appear next to each other following permutation; that is, when the difference between their original rankings is maximal. For example, given an ordering of achromatic chips from white to gray to black, placing the black chip next to the white will disturb the monotonicity (Fig. 16). In terms of document pages, this corresponds to placing a homogeneous text page next to a mostly empty page. Our goodness criterion is, therefore, the difference between two adjacent ranks of the permutation Q of the order P (or its reflection P').

We repeat the process of flanking items from opposing extremities of the monotonous ranking and obtain the following permutation sequences, which represent the least desirable orderings: $R = \{p_{k+2}, p_k, \ldots, p_5, p_{n-4}, p_3, p_{n-1}, p_1, p_n, p_2, p_{n-3}, p_4, \ldots, p_{k-1}, p_{k+1}\}$, and its reflection R', where $k = n/2$ (Fig. 16d). Having established the best and worst orderings, we next set about developing a quantitative measure of ordering goodness.

Given that the objective is to maximize the difference between adjacent items of a permutation Q, we generalize and compute the sum of the differences between the ranks r_i of consecutive items, $D^1 = \sum_{i=1}^{n-1} r_{i+1} - r_i$; this allows us to distinguish between monotonous and other orderings, since the former have a computed value of n, $D^1(P) = n$, and zero for a second-order differencing, $D^2(P) = D_{\min} = 0$. To distinguish between the remaining patterns we apply the differencing process $n - 1$ times, yielding the scalar D^{n-1}. In other words, we compute at increasingly higher scales the length of a shape with coordinates given by the order indices in Q and the ranking values in P. The permutations with the highest absolute value will correspond to the least desirable permutation R, and its isomorphisms, defined above: $D_{\max(n)} = D^{n-1}(R)$.

It is clear that this is the case considering that the n-minus-one-th order differencing of even n and $k = n/2$ has the form $D^{n-1} = -c_1p_1 + c_2p_2 + \ldots - c_{n-1}p_{n-1} + c_np_n = (c_{k+1}p_{k+1} + c_{k-1}p_{k-1} + c_{k+3}p_{k+3} + \ldots + c_2p_2 + c_np_n) - (c_1p_1 + c_{n-1}p_{n-1} + c_3p_3 + c_{n-3}p_{n-3} + \ldots + c_{k+2}p_{k+2} + c_kp_k)$; here c denotes the binomial coefficients for order n, with $c_{k+1} = c_k > c_{k-1} = c_{k+2} > \ldots > c_n = c_1 = 1$; odd indices are located in the first half of the series, and even indices in the second half. The series is maximized for the ordering R and its reflection $R' = \{p_{k+1}, p_{k-1}, p_{k+3}, \ldots, p_2, p_n, p_1, p_{n-1}, p_3, p_{n-3}, \ldots, p_{k+2}, p_k\}$. For example, for $n = 10$, we have $D^9 = (126p_6 + 84\,p_4 + 36\,p_8 + 9\,p_2 + 1\,p_10) - (1p_1 + 9\,p_9 + 36\,p_3 + 84\,p_7 + 126\,p_5)$. Since $p_6 > p_4 > p_8 > p_2 > p_10 > p_1 > p_9 > p_3 > p_7 > p_5$ and $p_np_i \in \mathbb{N}$, $i \in [1, 6]$, then $p_6 = 10, p_4 = 9, p_8 = 8, p_2 = 7, p_10 = 6, p_1 = 5, p_9 = 4, p_3 = 3, p_7 = 2, p_5 = 1$, yielding the sequence $R = \{5, 7, 3, 9, 1, 10, 2, 8, 4, 6\}$, whose $D = 1930$ is maximal for all permutations. To account for isomorph permutations with the same value but opposing signs, we take the absolute value, $|D|$; while to compensate for the exponential growth induced by the binomial coefficients that affects the distribution homogeneity, we take the logarithm, $\log(|D| + 1)$.

We are now able to characterize in the interval $[0, 1]$ the goodness D of any permutation Q of length n, following normalization by $D_{\max(n)}$ computed from the R of length n:

$$D(Q_n) = \log(|D^{n-1}(Q_n)/D^{n-1}(R_n)| + 1). \tag{12}$$

One limitation of this method is that it is only valid for an even number of items. For odd numbers of items, the permutations with maximal D will include a permutation for which there is a difference of one between adjacent items, which is not the maximal departure from monotonicity. For large n, this is a benign limitation, since one item can be safely removed if selected so that it has the least change in rank across the permutations to be compared.

The author created this ordering distance measure after being unable to find a suitable solution in the literature on non-parametric statistics. However, in the field of combinatorics, topological entropy exists as a measure of the complexity of dynamical systems in the one-dimensional interval, for which the extreme values correspond to the same permutations for which our measure D is minimal and maximal, respectively [47; 3; 72; 1]. The need to develop a new method for apparently the same result is based on the different justifications made by the two methods as to why permutation R is the ''worst'' case: there is no a priori rationale for considering that the permutation maximizing adjacent differences is the same as that which maximizes complexity, as defined by topological entropy. A further reason is that our method is more straightforward than computing the topological entropy, which involves deriving a peculiar ''induced'' matrix from the permutation and finding its maximum eigenvalue. For large datasets, such as all permutations for a given n, the differencing method is also substantially faster, since it is easily vectorizable. Nevertheless, the identical results in respect to the extreme permutations suggest an equivalence between the

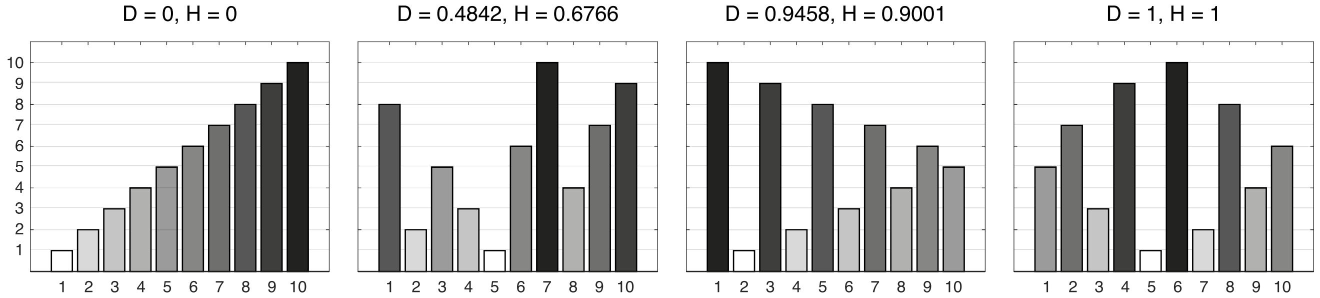


FIGURE 16. Various permutations and their departure from monotonicity, computed with the differencing (*D*) and topological entropy (*H*) methods.

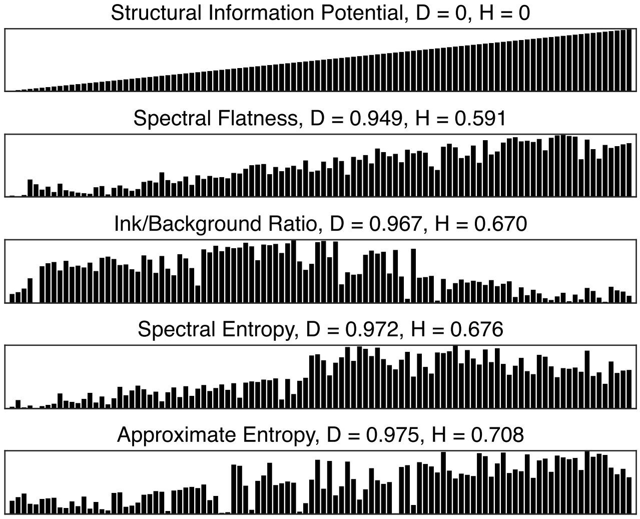


FIGURE 17. Departure from the ordering induced by the SIP pattern classification method of four comparison methods based on the document pages of Fig. 27–31.

n-minus-one-th-order differencing and eigen-analysis, perhaps in the scale-space characterization of patterns that both undertake. It may also be noted that topological entropy is not equivalent to SIP or minimal redundancy as discussed in this article. This can be determined by considering that topological entropy (and adjacency difference) increases when two items a and b are maximally different (that is, when $a \overset{\lim}{\longrightarrow} 0$ and $b \overset{\lim}{\longrightarrow} \infty$), while SIP is maximal when both items and their sum are maximally different ($a/b = b/(a+b)$).

I have made both the differencing and entropy methods available as open-source Matlab functions [6].

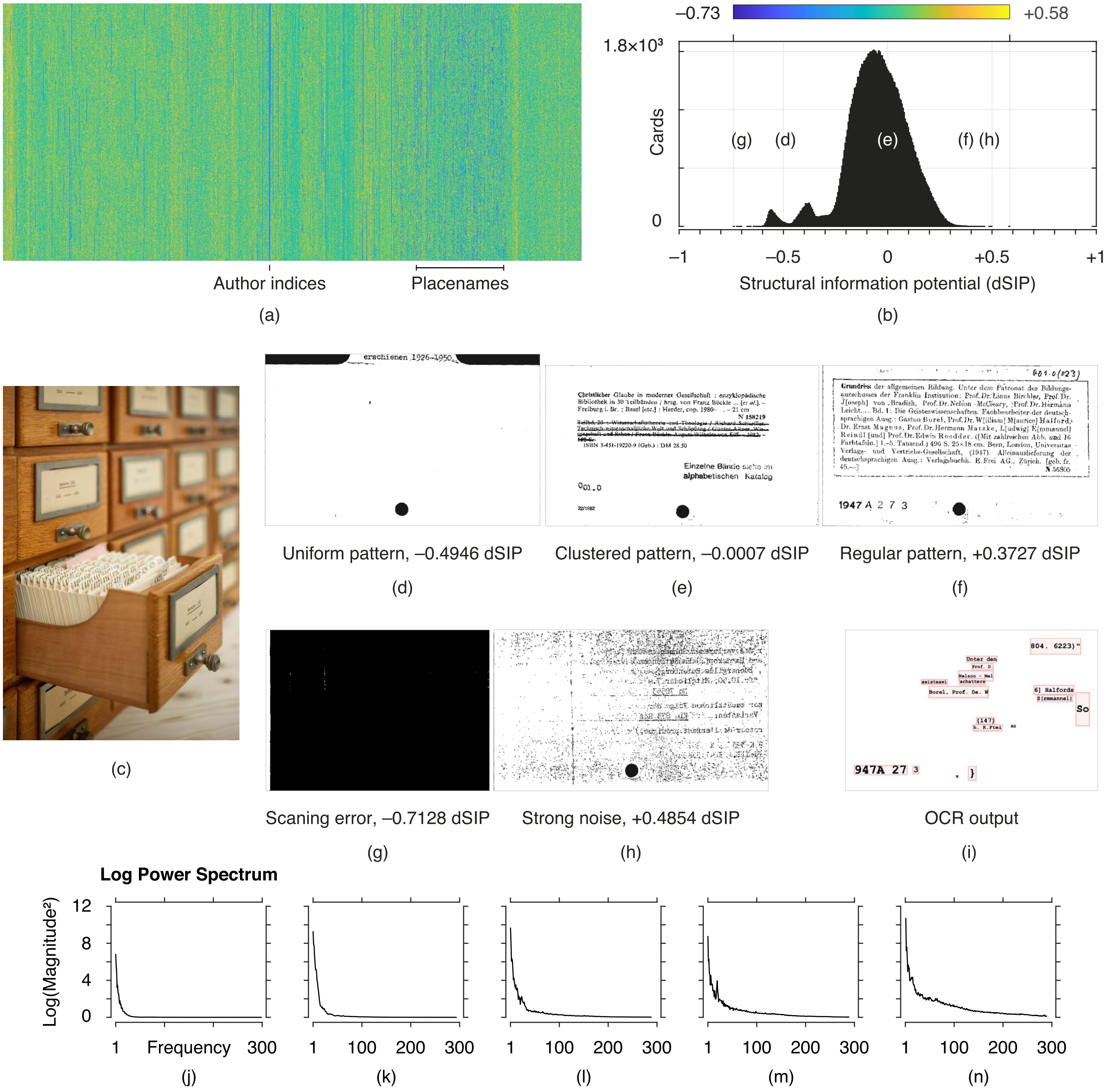
Results — Fig. 17 provides a graphical representation of the shuffling of the document pages ordered with the SIP method, as realized by the four comparison methods. The numerical values indicate the departure from monotonicity of the respective orderings. Approximate Entropy induces the ordering most different from SIP, while that of Spectral Flatness is the most similar. We can observe the clear juxtaposition of pages from the opposite sides of the uniform–clustered–regular spectrum in the Ink/Background and Approximate Entropy orderings, along with the shifting of page blocks in the case of Ink/Background and Spectral Entropy, which disturb monotonicity at small and large scales, respectively. Both distance measuring methods indicate that, overall, there is a substantial ordering difference between SIP and the other methods.

B. CASE STUDY

This section presents a real-world application of the structural information potential measurement method. The beneficiary is the Swiss National Library, and the goal is to merge the information of corresponding bibliographical records in electronic and analog formats [194]. The case study covers the pilot phase of an ongoing project, in which computational methods are assessed in view of deciding on the next steps.

The analog records are palm-sized paper cards, typewritten or printed, with handwritten annotations, crossings-out, stamps, bar codes, rulings, and other graphical elements (Fig. 18c–h). A total of 1.2 million cards were scanned from high-contrast black-and-white films (accessible at http://siibns.ch/french/cat1_frame.htm), at a resolution of circa 500 by 300 pixels; this introduces various artifacts, such as background noise and an irregular border around the cards. Some cards contain no records, but only captions for a sequence of cards in the wooden trays that were accessed by library patrons searching the catalog. The texts often mix together two or more of the four national languages of Switzerland (French, German, Italian, and Romanche), as well as English, Latin, and other languages. The majority describe monographs and serial publications, but also maps; the content is inconsistently semantically structured, is generally not made up of full sentences, abounds in entity names, alphanumeric shelf marks, ISBNs, price tags in various currencies, and other codified data, and is rich in typographical formatting of logical entities.

Such a wide typological variety of information, brevity, and visual complexity presents a challenge for automatic recognition. Attempts to perform optical character recognition (OCR) on the whole dataset, using the open-source Tesseract software [195] and the commercial Google Vision [196], revealed that the obtained text is not directly exploitable in the library's public catalog (Fig. 18i). It was therefore envisaged to provide users with digital images of cards along the electronic record in the same graphical interface window. However, which analog



erschienen 1926–1950.

Christlicher Glaube in moderner Gesellschaft ; enzyklopädische Bibliothek in 30 Teilbänden / hrsg. von Franz Böckle ... [*et al.*]. – Freiburg i. Br. ; Basel [etc.] : Herder, cop. 1980– . – 21 cm
N 158219
~~Teilbd. 20- : Wissenschaftstheorie und Theologie / Richard Schaeffler. Technisch-wissenschaftliche Welt und Schöpfung / Günter Altner. Wissenschaft und Ethos / Franz Böckle, August Wilhelm von Eiff. – 1982. – 168 S.~~
ISBN 3-451-19220-9 (Geb.) : DM 28.50

Einzelne Bände siehe im alphabetischen Katalog

001.0

22/1982

001.0(023)

Grundriss der allgemeinen Bildung. Unter dem Patronat des Bildungsausschusses der Franklin Institution: Prof. Dr. Linus Birchler, Prof. Dr. J[oseph] von Bradish, Prof. Dr. Nelson McCleary, Prof. Dr. Hermann Leicht.... Bd. 1: Die Geisteswissenschaften. Fachbearbeiter der deutschsprachigen Ausg.: Gaston Borel, Prof. Dr. W[illiam] M[aurice] Halford, Dr. Ernst Magnus, Prof. Dr. Hermann Matzke, L[udwig] E[mmanuel] Reindl [und] Prof. Dr. Edwin Roedder. ([Mit zahlreichen Abb. und 16 Farbtafeln.] 1.–5. Tausend.) 496 S. 25×18 cm. Bern, London, Universitas Verlags- und Vertriebs-Gesellschaft, (1947). Alleinauslieferung der deutschsprachigen Ausg.: Verlagsbuchh. E. Frei AG., Zürich. [geb. fr. 45.—]
N 56805

1947 A 2 7 3

FIGURE 18. (a) 1.2 million library index card images, at one image per pixel, color-coded by their structural information potential, in physical sequential order, rearranged for space reasons top-down and left to right, and shown at reduced size. The markers indicate thematic ranges. — (b) Histogram of dSIP values and colormap for the data range. The letters indicate the approximate locations of the five cards shown above. — (c) Paper cards in the library catalog trays. (Credit: Swiss National Library) — (d) – (f) Sample cards of the main visual pattern types, with their dSIP values. — (g), (h) Two outlier types, found at the extrema of the dSIP distribution. (i) Tesseract OCR output for the card (f). — (j) – (n) Log power spectra of the card images (g), (d), (e), (f), and (h) (to obtain one-dimensional spectra, the two-dimensional spectra are averaged and rounded to the nearest integer frequency).

record corresponds to which electronic record is unknown, and the two sets do not overlap. Consequently, the technical project objective of matching document images and electronic texts is preceded by a feasibility analysis. This requires a fast and appropriate classification of the cards. Here, "fast" concerns readying the technical resources, the human interaction with the data, and the lax requirements for classification quality and sophistication—hence, a *triage* task.

The dSIP values of the dataset images were measured and the cards located in the pattern–informativeness space. The ordering of the 1.2 million items was checked visually for perceptual consistency and found to be satisfactory; a sample of the results are presented in Fig. 19. The patterns

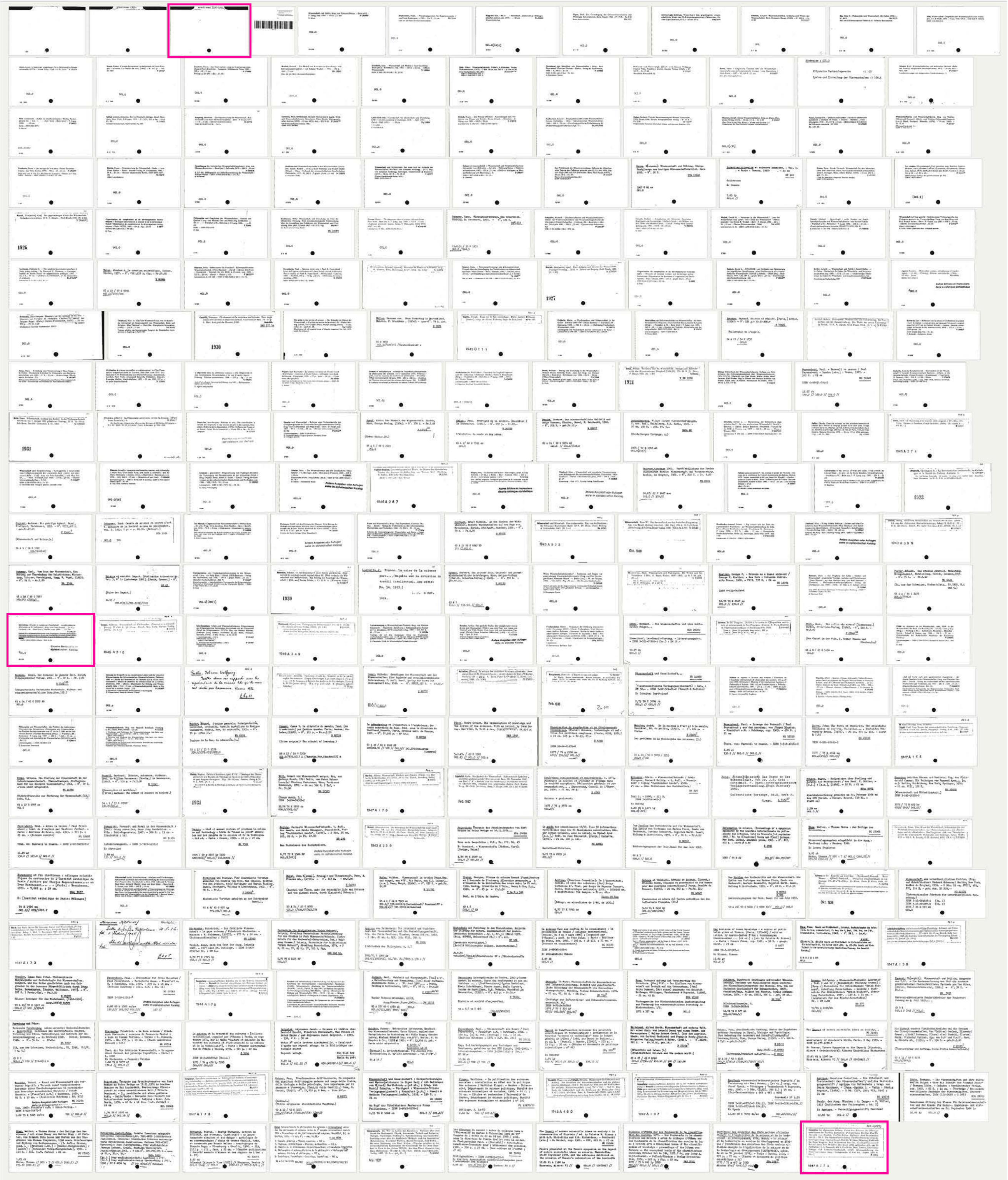


FIGURE 19. Cards of one of the library catalog thematic sections, classified according to their structural information potential. The three highlighted cards are shown in higher resolution in Fig. 18d–f. The second card is the closest to 0 dSIP, i.e. considered as the most clustered.

in this sample are representative (*a*) in semantic terms, as the sample represents an entire bibliographical index section, containing both various header cards and reference cards (section CDU 00000001), (*b*) in document typological terms, as it contains printed and handwritten text, noisy images, and other significant features, and (*c*) in statistical terms, as the sample is roughly uniformly distributed over the pattern spectrum (see the location of the three highlighted cards in the overall distribution of Fig. 18b). The ordering is meaningful as a typological categorization of bibliographical cards and suggests several potential matching strategies between physical and electronic records. The cards with clustered visual patterns are so because, usually, they exhibit greater typological information variety; in such cases, the matching algorithm could be tuned to privilege the logical structure as the matching criterion over linguistic matching. Cards with homogeneous patterns are more likely to contain longer coherent linguistic sequences, such as sentences, for the matching of which syntactic analysis would be preferable. Cards with low SIP value are either empty, are non-record section headers, or contain very little text; as a result, they may yield very poor matching quality, and therefore may best be visually inspected and excluded from matching.

The classification of the cards using the measure of structural information potential was appreciated by both software engineers and library managers for multiple reasons. It is independent of the OCR output; it produces useful results under conditions of uncertainty about the cards' content; moreover, it allows for rapid overview of a large digital image collection that would otherwise remain largely invisible, and further involves the human in the matching process. Thus, card triage by SIP became a complement to OCR in decision-making on tasks such as go/no-go for automatic record matching, selection of a card subset with an expected matching quality, estimation of expected matching quality, and evaluation of project resources (e.g., costs and duration of groundtruthing and quality control).

Fig. 18a shows the color-coded dSIP values for the 1.2 million bibliographical cards in their physical sequence in the original library card trays. Runs of similar dSIP values are apparent, such as the conspicuous blue (i.e., low dSIP) central column corresponding to author indices and mostly empty cards. The heterogeneously colored area on the right-hand side corresponds to cards indexing placenames, which have varying degrees of visual density and clustering due to the variety of information sources and the continuous updating of the card information, involving many writing technologies and annotation layers. The visualization's pixels are interactively linked to the card images, so that a visual investigation may reveal that the blue run represents author index cards and can thus be removed from the matching process. This visualization is useful for a context-oriented analysis of the card dataset. If, for example, an equivalence has been established between dSIP values and expected matching quality, then it may be used to predict the matching quality of the various thematic classes of the cards.

Fig. 18b represents the dSIP values of the card dataset in histogram form. This aggregated data yields several insights. By extracting sample cards along the distribution to determine the patterns to which the values correspond, a visual and quantitative estimation can be made regarding the amount of various card types and the expected matching quality. It can be observed that the bulk of the cards have a clustered appearance, while cards with homogeneous information distribution are not predominant. A further operationalizable insight enabled by the histogram pertains to outlier detection (Fig. 18g, h). In this case, a first observation concerns the extreme dSIP values: scanned images with the objects of interest partially out of frame are found in the lower values, while the higher values contain images with strong noise. Second, the left-most cluster of the distribution predominantly comprises section headers. While these do not contain bibliographical information, and should thus be excluded from the matching process to avoid a detrimental impact, these cards may be useful for matching, since they identify the topic to which the subsequent cards belong. This topic may then be related to those extracted from the electronic records, thereby increasing the probability of correct matches. The bivariate plot in Fig. 20 supports a finer analysis of the histogram, specifically that it is the result of a mixture of pattern classes characterized by different ink densities, as well as non-linear and loose covariation between the amount and distribution of ink on the cards.

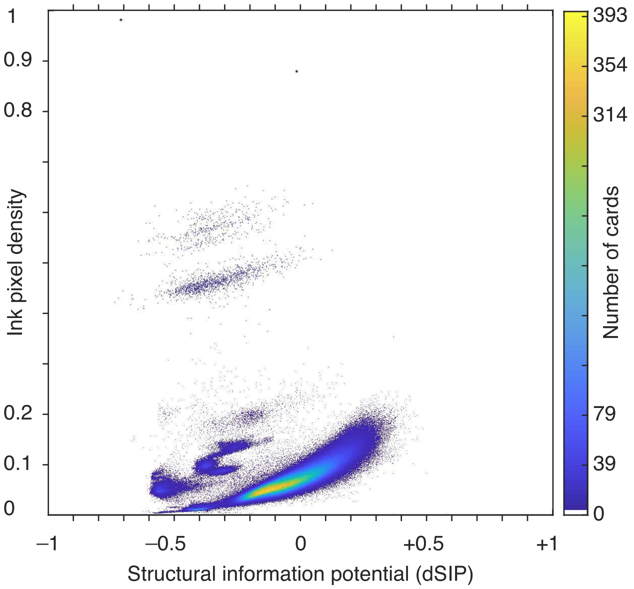
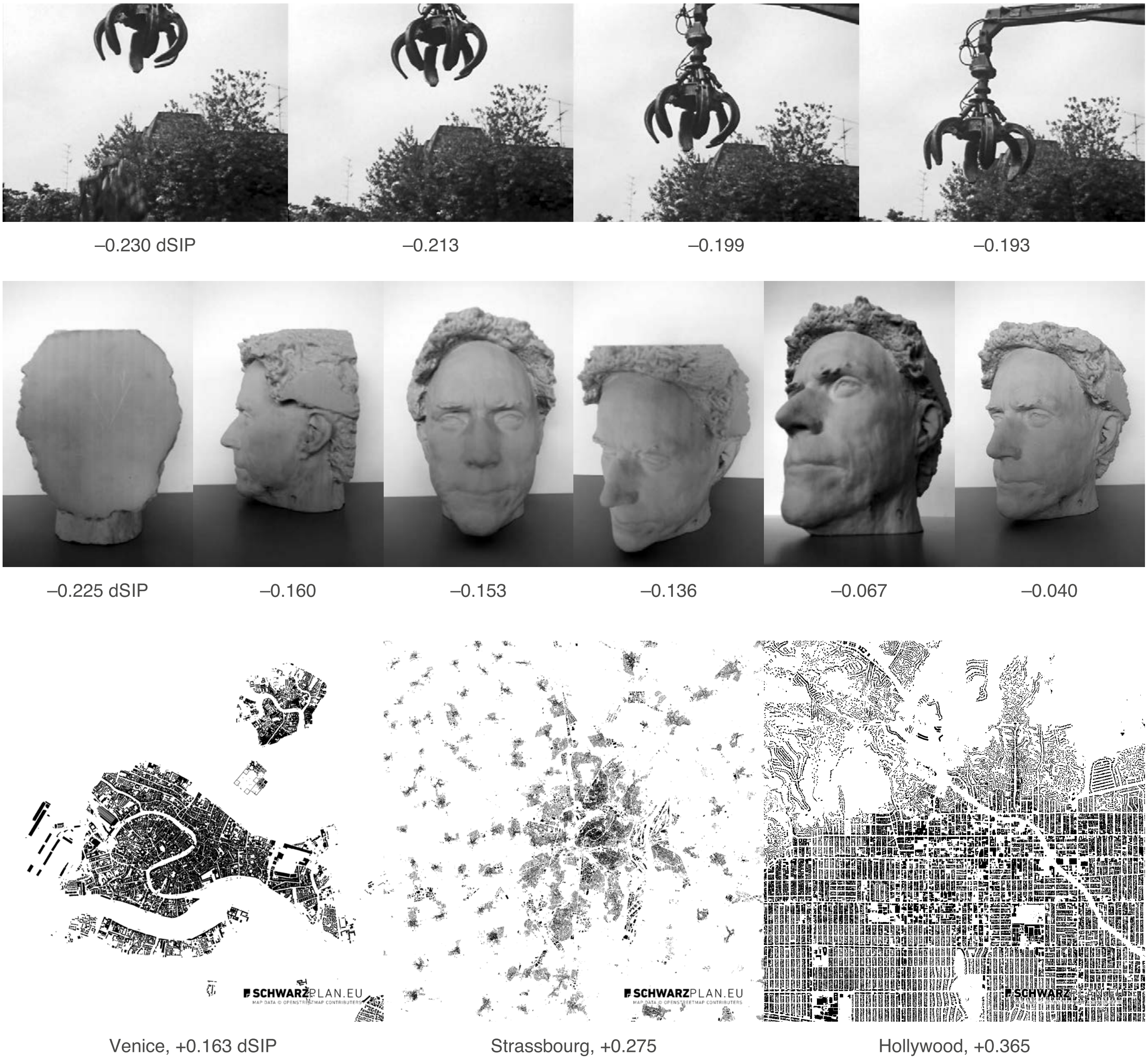


FIGURE 20. This graphic shows the density distribution of the structural information potential plotted against the ink pixel density in the digital images of the library card dataset. An "ink pixel" represents an inked area of the card surface, as opposed to the non-inked writing substrate; in the cards illustrating this article, ink pixels appear in black; noise modifies the groundtruth value of pixels. For legibility purposes, pixels above the density level of 0.4 ink pixels have been slightly dilated. Note the two pixels above level 0.8. The 1.2 million data points were aggregated for visualization in a 1000-by-1000-pixel raster.

The question might be asked as to whether simply classifying the binarized card images by the amount of black pixels (i.e. ink) would not provide insights similar to the structural information potential. Fig. 20 demonstrates that this is not the case: documents with identical ink density can have very different ink distributions, and vice-versa, as a result of distinct clusters with irregular shape and spread. May, then,



–0.230 dSIP –0.213 –0.199 –0.193

–0.225 dSIP –0.160 –0.153 –0.136 –0.067 –0.040

Venice, +0.163 dSIP Strassbourg, +0.275 Hollywood, +0.365

FIGURE 21. Application of the structural information potential to video frames (top), three-dimensional objects (middle), and urban layouts (bottom). — Credits: video: Trigon Film; sculpture: unknown artist; plans: SCHWARZPLAN.eu.

a classification by the number of recognized characters be a sufficient estimate of matching quality? As illustrated by Fig. 18i, this would not be effective either; even cards with rich linguistic contextual information can result in very few characters being identified.

From a machine learning perspective, the classification of document images according to SIP may be utilized to optimize the sampling of the training and test images. For example, a uniform random sampling of the card dataset (i.e. the typical sampling procedure) may result in an underrepresentation of less numerous but semantically important classes, such as scanning errors and cards with strong noise or that are largely empty, which are visible at the distribution extrema of the dataset SIP visualization in Fig. 18b. However, by using the SIP distribution in conjunction with visual exploratory data analysis, one can apply different sampling rates to different parts of the dataset, thereby acquiring sufficient samples for training the algorithms with the best possible accuracy given the data.

In conclusion, the measurement of structural information potential, coupled with interactive data analysis, provided computer scientists with an inexpensive tool to support record matching, particularly for sample selection and expected quality estimation. It also aided library managers in making faster and more informed decisions regarding an information technology project on issues such as go/no go, expected

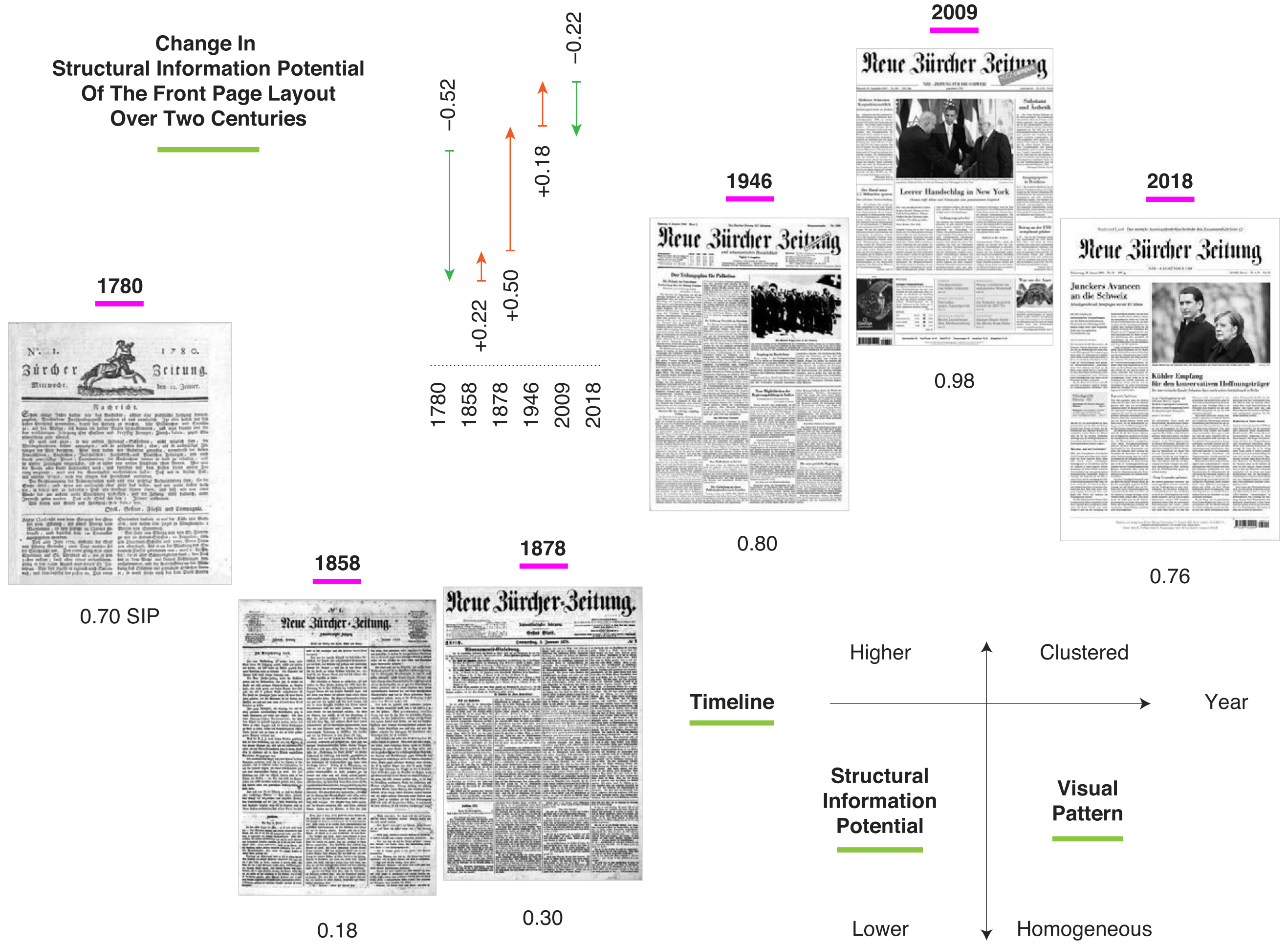


FIGURE 22. Timeline of the layout evolution of the Swiss *Neue Zürcher Zeitung* newspaper, with the SIP value of the front page. — Credits: Neue Zürcher Zeitung.

quality, and resource estimation. This case study is an example of the usefulness of SIP for both quickly extracting information from patterns and acting upon this information.

C. GENERALIZATION

The generic nature of the structural information potential concept as defining a pattern distribution space allows it to be applied to a variety of tasks and data types. Its main utility advocated in this article is as a quantitative measure of informativeness, which makes it appropriate for triage tasks, as discussed above. Notably, however, the above case study found that SIP was also useful as part of a decision-making process, involving the estimation of OCR quality and needed resources, and as a measure of image quality, given its ability to distinguish images with uniform noise and erroneously framed scans. Furthermore, SIP offers a simple computational solution for classifying document pages according to the predominance of text, the presence of illustrations, the number of post-production visual artifacts (e.g., annotations, stamps, signatures) and the amount of noise (i.e., traces of document degradation); these abilities may be translated into document navigation functionalities and implemented in document readers. The characterization of document layouts is also valuable to historians, as it supports a quantitative analysis of the evolution of written communication.

Fig. 21 illustrates the application of SIP to data types other than documents and tasks other than triage. According to the SIP concept, the patterns with dSIP values closest to 0 are the most informative. The top row illustrates the case of image retrieval from a large set of very similar samples. The SIP-based automatic extraction of keyframes from videos facilitates the identification of frames with potentially high information content. The frame with minimal dSIP value indeed shows more visual details than other frames, such as text, cables, and pistons. In the middle row, SIP is used to determine the most informative point of view of a three-dimensional object, perhaps by an examining robotic camera. For a head sculpture this is the three-quarter view, which integrates anatomical elements of both face and profile in a single shot. The bottom row shows how SIP may

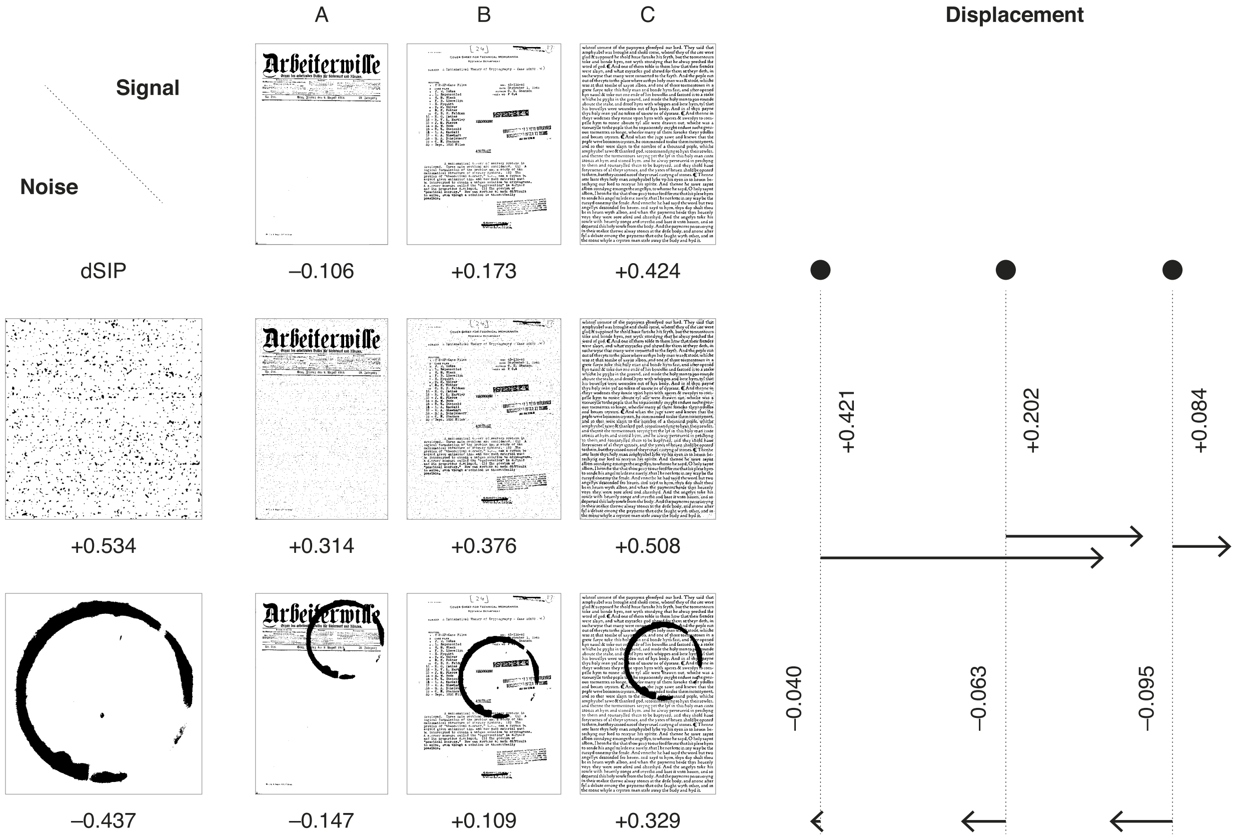


FIGURE 23. The figure illustrates the impact of two common document noise types on the dSIP values of basic document patterns. — Top row: Three noise-free documents, with uniform (A), clustered (B), and regular (C) patterns. — Left column: (1) Digital camera noise (obtained by taking a picture with the camera lens covered, due to which the pattern registered in the digital file is not that of incident light, but rather the results of electronic, thermal, software, and other similar artifacts; postprocessing consisted in median filtering and binarization). (2) Coffee stain. Both left-most images are enlarged for clarity. — Center: Pair-wise mixing of signals and noise, with dSIP values indicated. — Right: The arrows show the direction and magnitude of the change in dSIP due to the addition of noise to the signals.

be used for classifying aerial images of urban settlements on the uniform–clustered–regular continuum. The impact of the natural topography and cultural–historical factors on the hierarchical urban clustering come readily to mind when contemplating this classification; for example, the meandering water channels of Venice versus the commercial hub of medieval Strasbourg reflected in a spiderweb road network versus the plain of Los Angeles making a flat urban grid affordable.

Fig. 22 showcases the usefulness of the SIP measure to research in humanities on the history of newspaper layout. The figure shows selected front pages of the leading Swiss newspaper *Neue Zürcher Zeitung*, first published in 1780. The principal transformation of the layout concerns the increase in visual and semantic hierarchical structuring through textual elements, the addition of pictures, and greater use of empty space. Some of the evolutionary factors are clearly technological, such as the invention of photography, while others are societal. For example, the accelerating rhythm of modern living encourages the development of visual communication methods that facilitate faster navigation of information, prompting the shift from homogeneous to hierarchically clustered layouts. SIP provides a means to quantify and compare changes both within a single newspaper and between different publications.

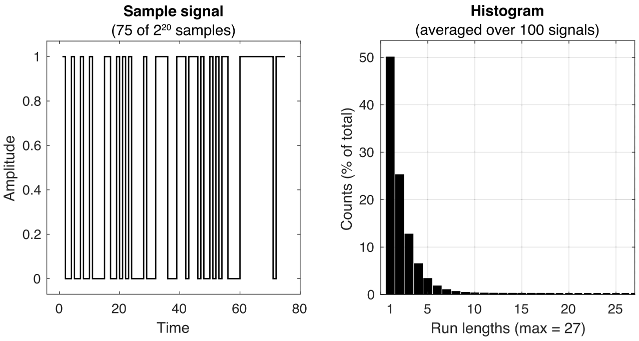
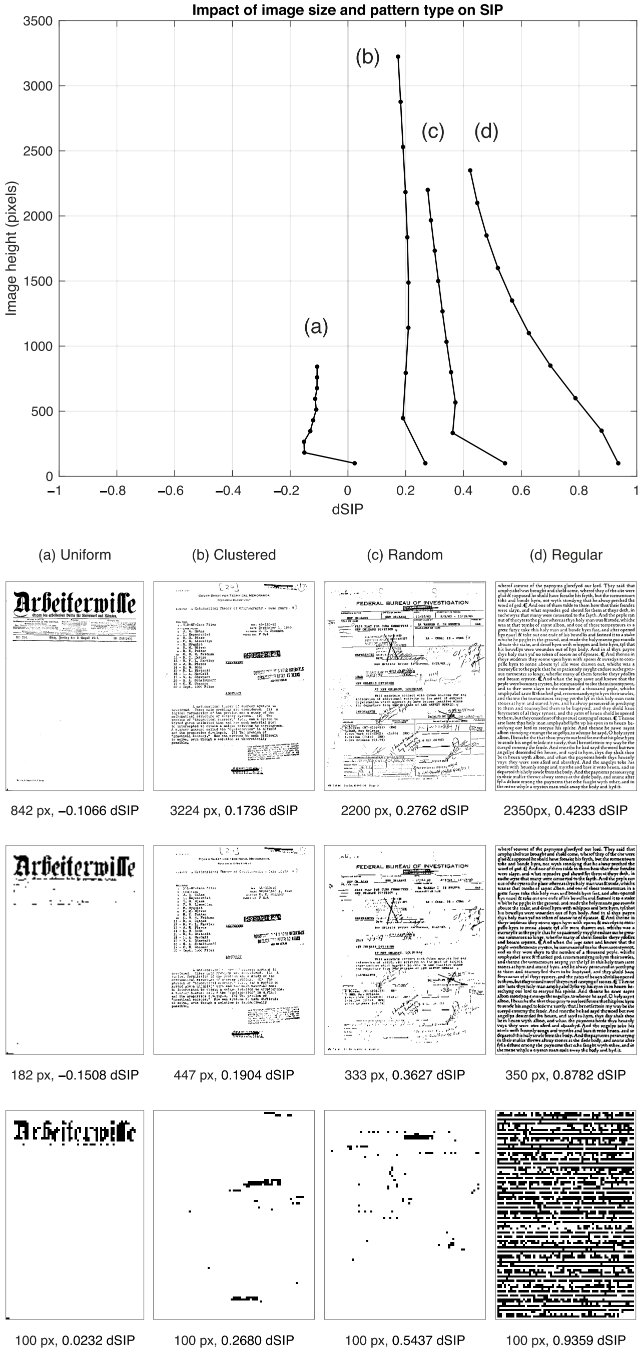


FIGURE 24. Detail of binary white noise signal (left), and the averaged histogram of the lengths of consecutive samples with identical values (runs) obtained from 100 such signals (right).

Moving beyond documents, the classification into compact, clustered, and homogeneous distributions is of direct interest to materials science. In telecommunications, the spatial or temporal distance between three senders/receivers—a basic configuration of networks that recalls our investigation



(a) Uniform (b) Clustered (c) Random (d) Regular

842 px, −0.1066 dSIP 3224 px, 0.1736 dSIP 2200 px, 0.2762 dSIP 2350px, 0.4233 dSIP

182 px, −0.1508 dSIP 447 px, 0.1904 dSIP 333 px, 0.3627 dSIP 350 px, 0.8782 dSIP

100 px, 0.0232 dSIP 100 px, 0.2680 dSIP 100 px, 0.5437 dSIP 100 px, 0.9359 dSIP

FIGURE 25. The above diagram illustrates how downscaling raster images affects their dSIP according to their pattern type. The images displayed are those with the highest and two lowest heights.

of triangles in the preceding section—affects data transmission quality and the propagation patterns of messages. This ‘‘layout’’ phenomenon is fundamentally similar in other types of networks, such as transportation, social, or epidemic. Symmetry plays a great role in physics (for example, for the growth of crystals); however, irregular configurations (minimally redundant) may be equally interesting to study. As for a table with minimally redundant edge lengths, to take an example from psychology, this is a recipe for emotional stress for a party of three sitting at its respective corners.

D. DECISION-MAKING

After having encountered a great variety of applications of structural information potential throughout the article, let us here briefly systematize how this concept might be the basis of ‘‘rapid decision-making for critical matters under conditions of uncertainty and with limited resources’’, as stated above in the introduction.

The structural information potential enables decisions to be made through the classification of patterns in the uniform–clustered–regular space, which relates to their degree of informativeness. What the users may do with this information is, however, largely task-specific, as suggested by the diversity of applications and contexts encountered in this article.

SIP-based decision-making may be rapid, first of all, from the semantic point of view, since the patterns are ordered by level of informativeness, which is desirable for the triage task. The SIP approach is also rapid from a technical perspective; this is because its implementation and use require limited resources, and further avoids expensive data preprocessing and other frequently unreliable types of processing, such as document recognition. Finally, SIP is rapid because it may be used to automatically select data with specific levels of informativeness or pattern types.

Structural information potential is a quantitative and visual tool for exploratory data analysis, i.e. situations in which the data content and task specification are uncertain. For such cases, there is no single preexisting quantitative decision-making formula that can be applied; serendipity needs to be acknowledged and actively sought, and insights emerge from the subjective interaction between humans and data. Within its limits, SIP enables users to act efficiently in an uncertain environment.

V. DISCUSSION

This section provides an in-depth discussion of some important rationales and implications of the SIP measurement. It thus helps better understand the behavior of this instrument, and opens directions for future research.

A. NOISE

How the addition of noise to a pattern affects its structural information potential depends on a number of factors, primarily the type, extent, and location of the noise, as well as the pattern type of the signal. Fig. 23 illustrates the interaction between two types of noise common in documents and the three basic patterns: uniform, clustered, and regular. Imaging artifacts, such as camera noise (depicted here), may introduce quasi-uniform, quasi-random noise that extends over the entire document and has high spatial frequency. This noise type invariably increases the regularity of the pattern and its dSIP value. Stains and blotches have limited extent and lower spatial frequency, such that their impact is largely determined by their location: they may make a clustered

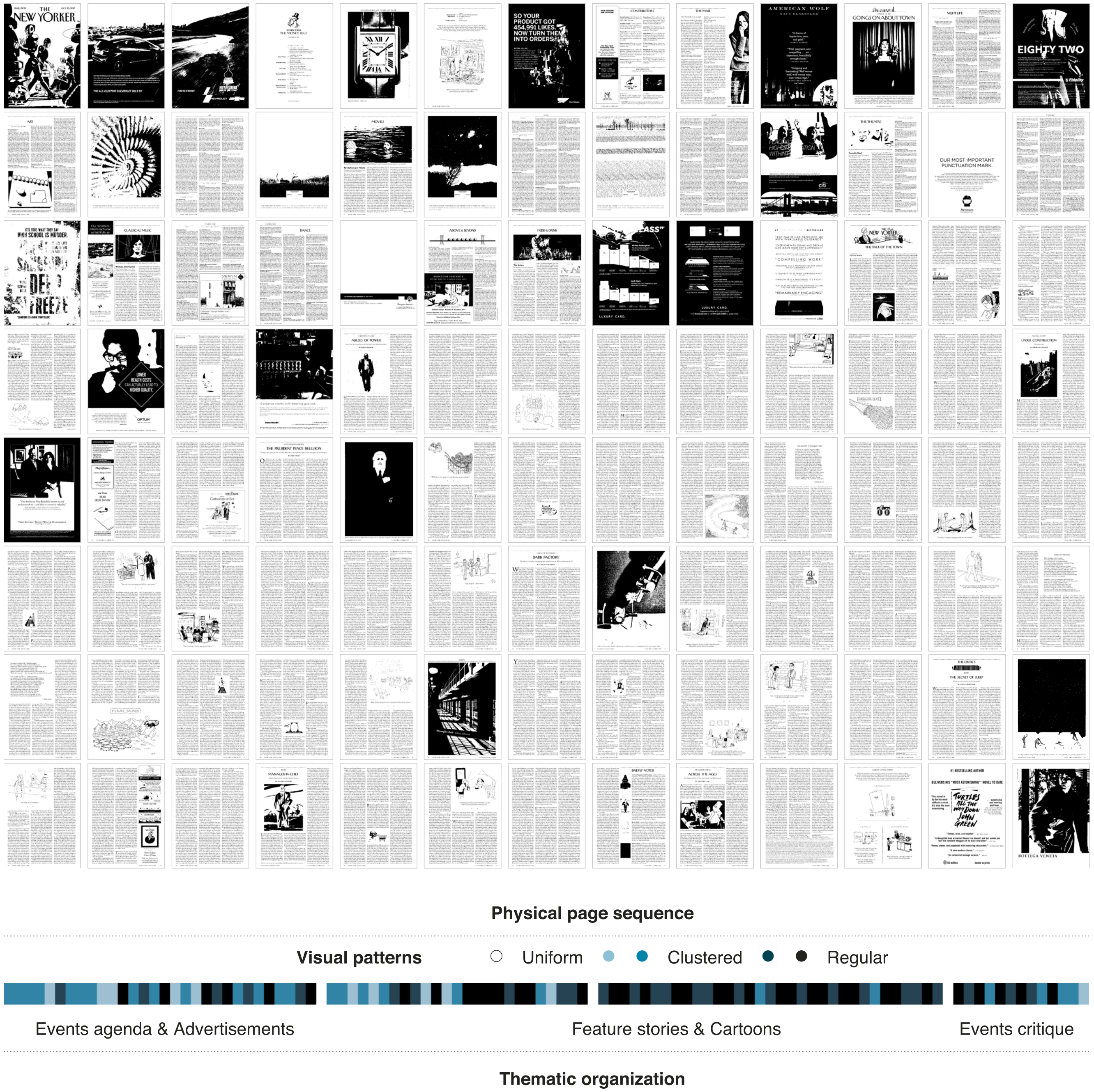


FIGURE 26. Original page order and document structure visualization — Reading order: left to right, top to bottom. The color-coded ribbon indicates the pattern type of each page according to its structural information potential. Observable is a stretch extending to a third of the document made from mixed uniform, clustered, and homogeneous page patterns, followed by a sequence where homogeneous pages predominate, and ending with several clustered pages. These visual patterns correspond to semantic patterns with different granularity levels: many short agenda items and full-page advertisements, text-based stories interspread with cartoons, and a critique section on multiple topics. The cover is a typical example of a clustered pattern, due to objects of different sizes, and a perspectival view.

pattern appear even more clustered and a uniform pattern more compact. When appearing in conjunction with a regular pattern, however, they invariably increase the clusteredness of the pattern irrespective of location. As a practical take-away message, understanding the interaction between noise and signals aids in interpreting SIP values and selecting data.

B. RANDOMNESS

The attentive reader may have noticed that random patterns have remarkably high dSIP values. Fig. 6, for instance, shows an example where dSIP is practically maximal (0.99). Given the expectation that a regular pattern has maximal value (1), how is this phenomenon possible, and what does it mean for the theory and practice of SIP?

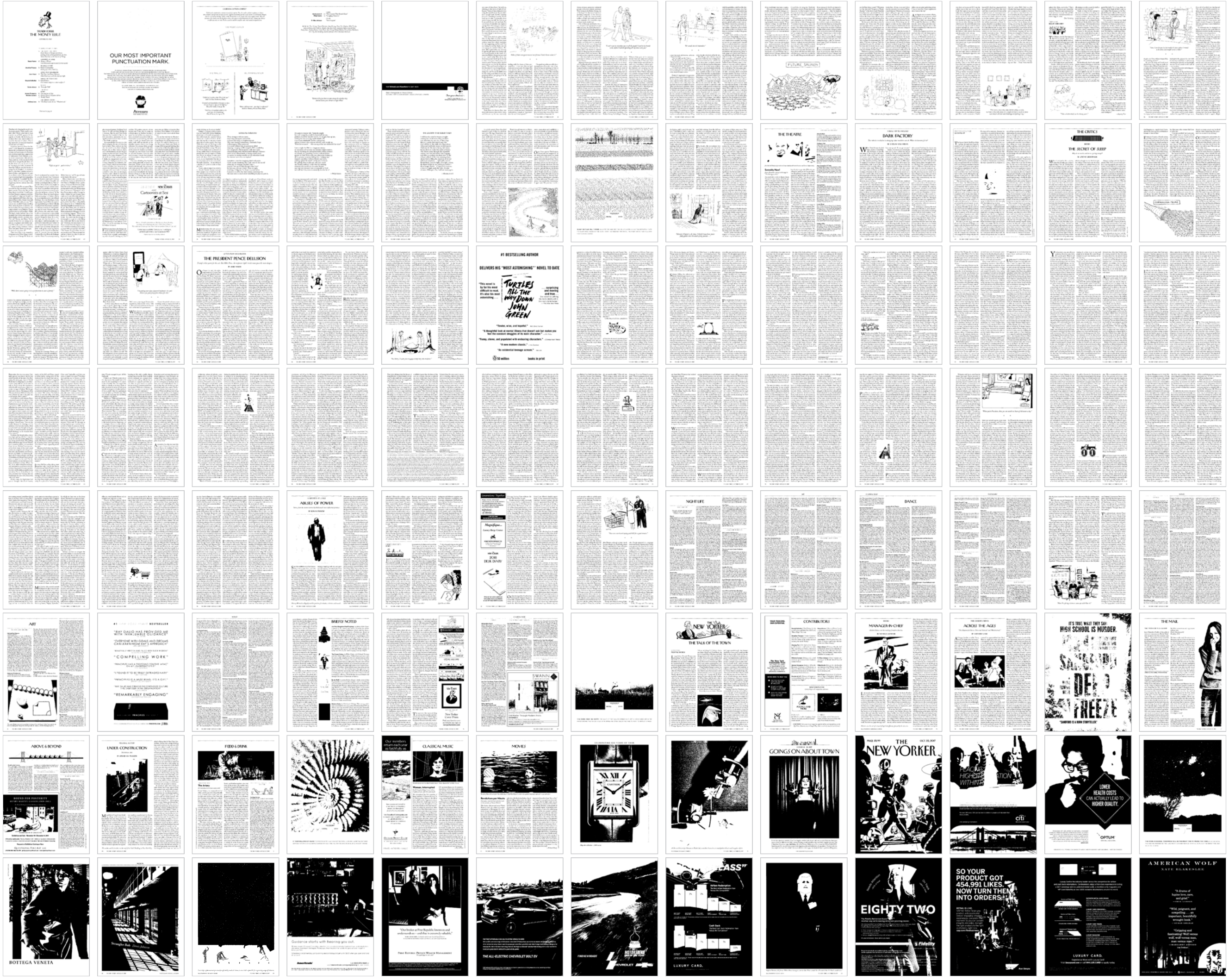


FIGURE 27. Page order by ratio of ink pixels and page area.

An infinite signal with uniform random binary amplitude values and uncorrelated samples (white noise) has an equal amount of each possible ''run length'', i.e. consecutive samples with identical value. However, a finite signal exhibits a different number of runs of equal length. Specifically, the distribution is defined as:

$$R_i \approx N/2^i\,, \quad \sum_{i=1}^{k} R_i = N\,, \tag{13}$$

where R is the number of runs of length i, N is the signal length, and k is the maximum run length (Fig. 24). This means that, theoretically, half of all two consecutive samples — a substantial share — have alternating values, just as in the case of a regular pattern: $0, 1, 0, 1, 0, 1, \ldots$. Moreover, a finite signal has a maximal run length, which is quite small; e.g., 18 for 1000 samples, and 27 for 2^{20} samples. Consequently, the extent of uniform sequences is very limited in respect to the signal length, further increasing the homogeneity and redundancy of the pattern. This phenomenon is the underlying reason for random patterns having dSIP values very close to those of regular patterns.

C. SIZE

It is an inherent aspect of quantization that modifying the resolution of raster images affects the shape of the pixel structures in the images; in other words, unlike vector graphics, shapes in bitmap graphics are not scale-invariant. Therefore, their SIP will also change with scale. The nature of this change depends on the pattern type: downscaling increases the homogeneity of homogeneous patterns (dSIP $>$ 0), has a relatively low impact on clustered patterns (dSIP $\approx$ 0), and increases the uniformity of uniform patterns (dSIP $<$ 0; Fig. 25). These changes are the most pronounced at small image resolutions, i.e. those below about 200 pixels. When image downsampling is performed for performance reasons, it is preferable to use images with height or width beyond the said threshold.

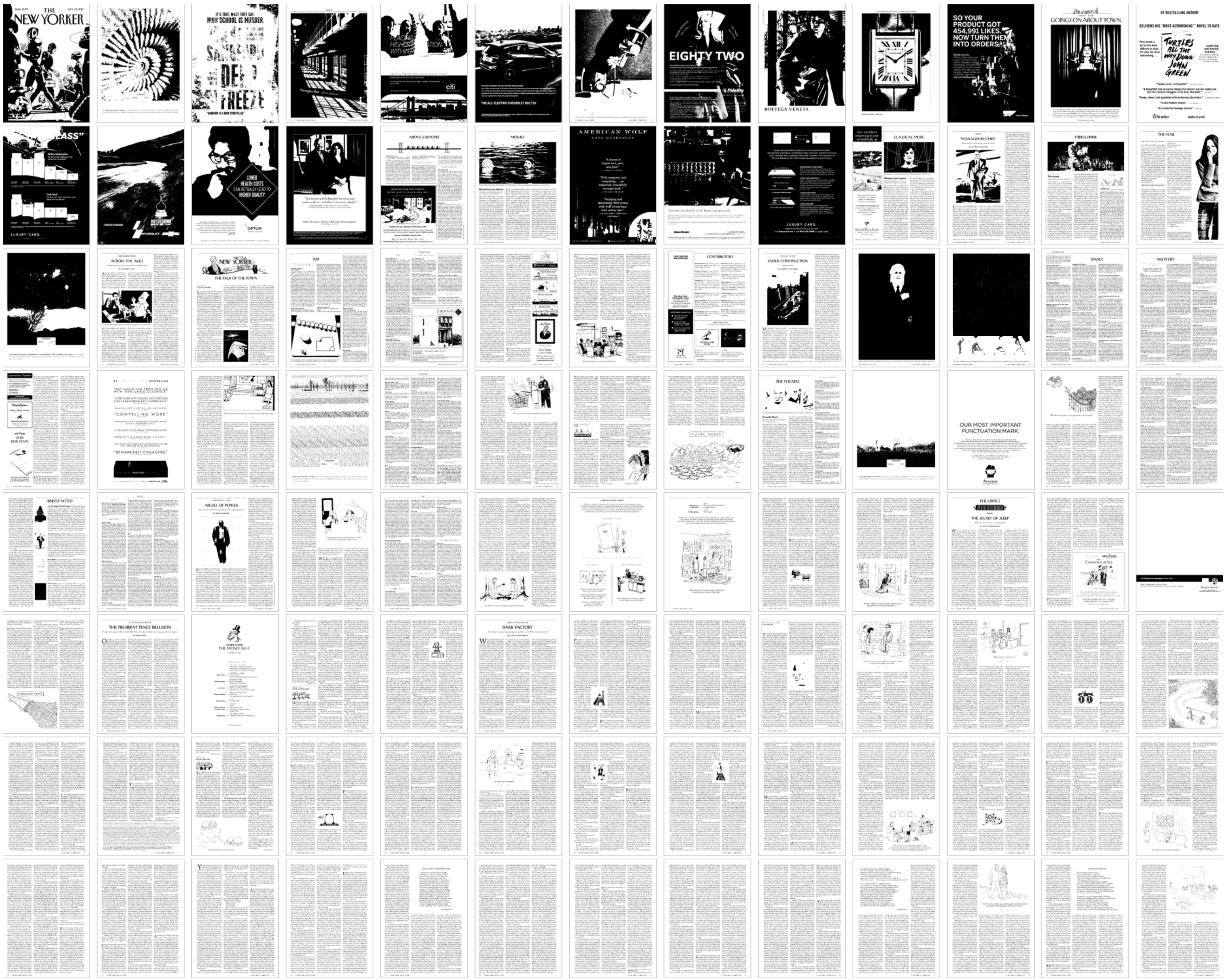


FIGURE 28. Page order by approximate entropy.

D. LIMITATIONS

Two patterns generate aberrant SIP values. These are, however, peculiar enough to not be of concern in most practical cases. The first is the perfectly regular structure of the checkerboard and stripes when aligned with the orthogonal image raster. At half-cycles of one and two pixels, their frequency spectra are impulses, hence the entropy has value zero rather than the expected value one. The second problematic pattern is a single square aligned with the image raster; due to the step function of the function having the sinc function $\mathrm{sinc}(a) = sin(a\pi)/(a\pi)$ as frequency domain pair [177, pp. 212–215], its harmonics increase entropy to nearly one, contrary to what would be expected from a pattern with a large uniform surface.

E. BINARIZATION

It may be useful to develop a SIP measurement without binarization. One reason is that it would remove a certain degree of uncontrollable data distortion inherent in the strong reduction of its dynamic range. Another reason is that it would facilitate the application of the SIP method to signals, for which binarization constitutes a drastic distortion. However, developing a method for data with large dynamic ranges creates the need for a second method for binary data (a common data type in many application domains, including document processing). Here we can observe an advantage of the binarization step, which enables both intensity and binary data to be handled with a single method.

F. METAMERISM

Because of dimensionality reduction (from two dimensions to a scalar in the case of images), many different patters will have an identical measured value (the metamerism effect). This is one reason why triage is a good application for the SIP method presented here, namely because it is tolerant to imprecision. One can imagine incorporating parameters into the SIP method to allow the description of various pattern features (for example, pixel density).

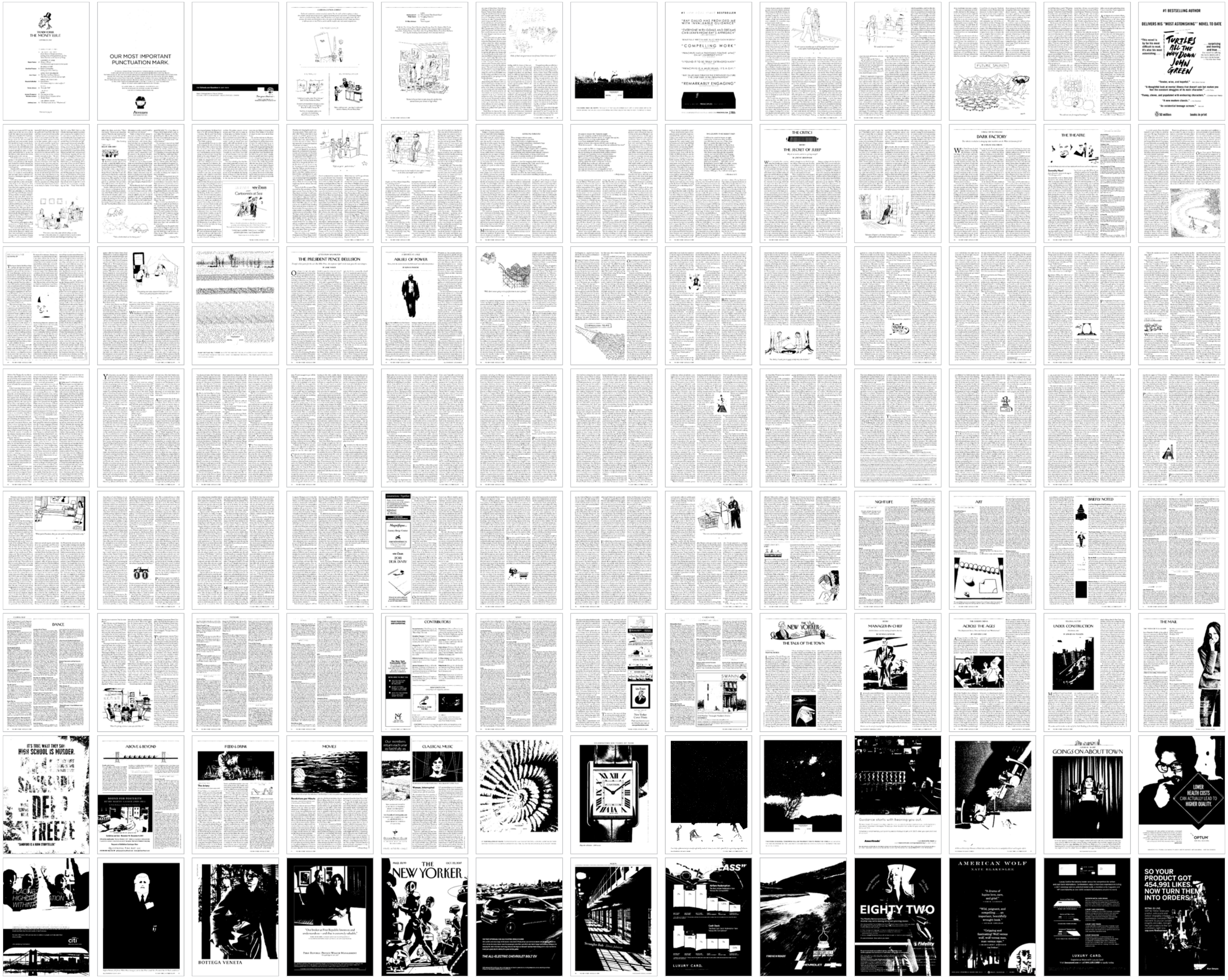


FIGURE 29. Page order by spectral entropy.

G. PSYCHOPYSICS

As the pattern ordering resulting from the SIP is primarily intended to have humans as end users, it is desirable to investigate the relevance of modifying the method to account for human pattern perception. Configurations are, however, difficult psychophysical research stimuli; this is more so the case for real data, such as documents, where semantics, aesthetics, and user experience (among many other factors) play an important role in their evaluation.

H. PHASE

The reader may have noticed that the SIP measurement method disregards the phase information arising from the image transform from the spatial to the frequency domain. This may be inconsequential for speech, where information is largely carried by the frequency spectrum [180], [197], [198, pp. 355–358]; however, phase is critical for image processing, in that radically different patterns result from identical frequency spectra with different phases, while edge location and strength are moreover strongly phase-dependent [180], [199]–[202, pp. 355–358]. This difference may explain to some extent why the spectral methods discussed in the related work section are successful in audio processing but less commonly used for image classification. From a practical point of view, however, the examples presented in this article demonstrate the robustness of the method when phase is discarded. This surprising fact demands a theoretical explanation.

As a first remark, it has been experimentally demonstrated that some binary image types may be reconstructed from frequency magnitude only and zero or random phase, and that this may further depend on the presence or absence of a single pixel [203]. While this phenomenon has yet to be elucidated, it constitutes evidence in favor of magnitude-only image analysis. Second, the extrema of the uniform–clustered–regular pattern space are phase-independent

FIGURE 30. Page order by spectral flatness.

[177, pp. 74–81, 106–107], which reduces the problem to one of explaining why clustered patterns occur around the center of the pattern space. We may recall that fractals (the patterns that maximize clustering) depend in their overall shape on the exponent of the power distribution of the frequency spectrum and a random phase. In fact, any clustered image pattern will exhibit some degree of power magnitude distribution. Moreover, the multitude of harmonics introduced by binarization create a phase that is increasingly better modelled by a random distribution. Taken together, these aspects suggest that within the limits of finite and discrete images, patterns approach fractality at and in the vicinity of maximal SIP or minimal spatial redundancy. This may be an important reason why it is practically possible to robustly order images along the uniform–clustered–regular pattern space on the basis of magnitude alone.

I. EPISTEMOLOGY

One fascinating mathematical aspect of minimal structural redundancy is how one might go about thinking about it; in other words, its epistemology. Specifically, it is richer to conceive the problem as the *design* of fluorescent structures than their *discovery*. Asking ‘‘What might the definition of minimal structural redundancy be?’’, rather than ‘‘Which is the structure with the least redundancy?’’, explicitly embeds the possibility of multiple answers into the inquiry process, as well as contextualizing the problem in respect to the questioner, the data, and the application. For example, a fluorescent graph becomes regular in terms of edge length when the number of vertices tends to infinity! If this behavior is not desired, a new definition of minimal redundancy may be created (for example, stipulating only local fluorescence). This same experimental conceptualization of mathematical definitions was followed by Gary Chartrand, Paul Erddős and Ortrud Oellermann in their article ‘‘How to Define an Irregular Graph’’: ‘‘In research, the goal is not only to come up with a definition that seems natural but to arrive at a class of graphs with interesting, and perhaps even some surprising, properties.’’ [85, pp. 39].

VI. CONCLUSION

This article has introduced structural information potential (SIP), a measure of information based on pattern

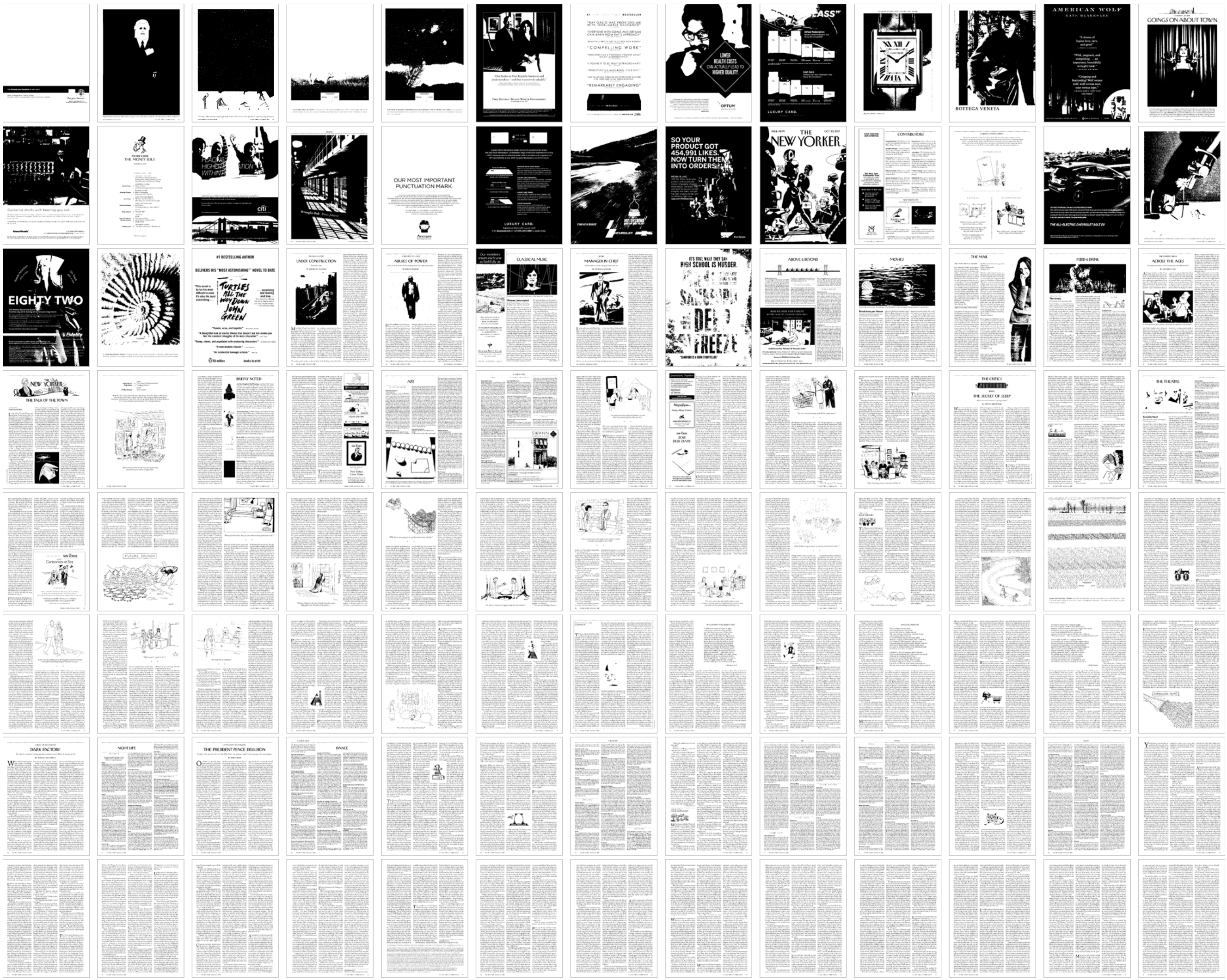


FIGURE 31. Page order by structural information potential.

configuration. Its utility was illustrated through a real-life case study for the task of document image triage. On this task, SIP performs better than other methods in both mathematical and perceptual terms.

The main theoretical significance of the work consists in (*a*) the development of a formalism that defines the uniform–clustered–regular pattern-informativeness space, which organizes fundamental pattern types in a mathematically and perceptually coherent fashion and relates them to an information potential, and (*b*) the development of a conceptual basis and analytical methods for the identification of shapes and patterns with minimal structural redundancy.

In practical terms, structural information potential is a useful classification method for triage-like conditions, characterized by decision-making under conditions of uncertainty and time pressure, when it becomes efficient to generate information about content through the analysis of structures. The generic nature of SIP makes it appropriate for many other applications, such as image quality assessment (to detect noisy and erroneously imaged data), predicting OCR output quality before applying OCR, identifying informative keyframes in video streams, and as a document navigation functionality. Given the generic nature of the structural information potential, the author believes that applications to fields as diverse as mathematics, physics, telecommunication, and psychology may be discovered in the future.

ACKNOWLEDGMENT

The author is grateful to Rolf Ingold, of the Informatics Department, University of Fribourg, Switzerland, for having facilitated the funding of the Open Access publication of this article. The author would like to thank Andreas Fischer of the University of Applied Sciences and Arts Western Switzerland (HES-SO), Fribourg, for sharing many ‘‘reflective reflections’’, and for having offered to work on the library card dataset. Irvin Schick, formerly with Bolt Beranek and Newman, Cambridge, MA, USA, now at the École des Hautes Études en Sciences Sociales, Paris, France, has been, over the years and continents, a reliable devil’s advocate for testing the author’s hypotheses. The anonymous reviewers are equally

thanked for having contributed through their comments to improve the manuscript.

REFERENCES

[1] T. Samara, *Making and Breaking the Grid: A Graphic Design Layout Workshop*. Gloucester, MA, USA: Rockport, 2002.

[2] K. Elam, *Grid Systems: Principles of Organizing Type*. New York, NY, USA: Princeton Architectural Press, 2004.

[3] F. Franchi, *Designing News*. Berlin, Germany: Gestalten, 2013.

[4] B. Schwesinger, *Formulare Gestalten*. Mainz, Germany: Hermann Schmidt, 2007.

[5] K. A. Schriver, *Dynamics in Document Design: Creating Text for Readers*. Chichester, U.K.: Wiley, 1997.

[6] A. Dillon, *Designing Usable Electronic Text: Ergonomic Aspects of Human Information Usage*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2004.

[7] R. Pettersson, *Information Design Theories*. Tullinge: Institute for infology, 2016.

[8] F. Richaudeau, *La lisibilité: Langage, Typographie, Signes Lecture*. Paris, France: Gauthier, 1969.

[9] P. Cleveland, "How much visual power can a magazine take?" *Des. Stud.*, vol. 26, no. 3, pp. 271–317, May 2005.

[10] J. R. Pomerantz and M. Kubovy, "Theoretical approaches to perceptual organization: Simplicity and likelihood principles," in *Handbook of Perception and Human Performance: Cognitive Processes and Performance*, vol. 2, K. R. Boff, L. Kaufman, and J. P. Thomas, Eds. New York, NY, USA: Wiley, 1986, ch. 36, pp. 1–46.

[11] A.-M. Christin, *A History of Writing: From Hieroglyph to Multimedia*. Paris, France: Flammarion, 2002.

[12] R. Kinross, *Modern Typography: An Essay in Critical History*, 2nd ed. London, U.K.: Hyphen Press, 2004.

[13] G. Dowding, *Finer Points in the Spacing & Arrangement of Type*, 3rd ed. Vancouver, BC, Canada: Hartley & Marks, 2008.

[14] J. Hochuli, *Detail in Typography*. Paris, France: Éditions B42, 2008.

[15] *Hz-Program: Micro-Typography for Advanced Typesetting*. Hamburg, Germany: URW Software & Type GmbH, 1993.

[16] H. T. Thành, "Micro-typographic extensions to the TEX typesetting system," Ph.D. dissertation, Masaryk Univ., Brno, Czech Republic, Oct. 2000.

[17] B. Shaw, *Bernard Shaw on Modern Typography*. Cleveland, OH, USA: The Printing Press, 1915.

[18] J. Aczél and Daróczy, *On Measures of Information and Their Characterizations*. New York, NY, USA: Academic, 1975.

[19] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[20] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. Probab.*, Berkeley, CA, USA, vol. 4.1, Jun./Jul. 1960, pp. 547–561.

[21] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," *J. Stat. Phys.*, vol. 52, nos. 1–2, pp. 479–487, 1988.

[22] A. E. Magurran, *Measuring Biological Diversity*. Oxford, U.K.: Blackwell, 2003.

[23] T. Leinster and M. Meckes, "Maximizing diversity in biology and beyond," *Entropy*, vol. 18, no. 3, p. 88, Mar. 2016.

[24] S. Watanabe, *Pattern Recognition: Human and Mechanical*. New York, NY, USA: Wiley, 1985.

[25] Y. Y. Tang, *Document Analysis and Recognition With Wavelet and Fractal Theories*. Singapore: World Scientific, 2012.

[26] S. M. Pincus, "Approximate entropy as a measure of system complexity," *Proc. Nat. Acad. Sci. USA*, vol. 88, pp. 2297–2301, Mar. 1991.

[27] S. Pincus, "Approximate entropy (ApEn) as a complexity measure," *Chaos*, vol. 5, no. 1, pp. 110–117, Mar. 1995.

[28] S. Pincus and R. E. Kalman, "Not all (possibly) 'random' sequences are created equal," *Proc. Nat. Acad. Sci. USA*, vol. 94, no. 8, pp. 3513–3518, Apr. 1997.

[29] K. H. Chon, C. G. Scully, and S. Lu, "Approximate entropy for all signals," *IEEE Eng. Med. Biol. Mag.*, vol. 28, no. 6, pp. 18–23, Nov./Dec. 2009.

[30] F. Kaffashi, R. Foglyano, C. G. Wilson, and K. A. Loparo, "The effect of time delay on approximate & sample entropy calculations," *Phys. D, Nonlinear Phenomena*, vol. 237, no. 23, pp. 3069–3074, Dec. 2008.

[31] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *Amer. J. Physiol.-Heart Circulatory Physiol.*, vol. 278, no. 6, pp. H2039–H2049, Jun. 2000.

[32] W. Chen, Z. Wang, H. Xie, and W. Yu, "Characterization of surface EMG signal based on fuzzy entropy," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, no. 2, pp. 266–272, Jun. 2007.

[33] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Phys. Rev. Lett.*, vol. 89, no. 6, Jul. 2002, Art. no. 068102.

[34] Z. Liang, Y. Wang, X. Sun, D. Li, L. J. Voss, J. W. Sleigh, S. Hagihira, and X. Li, "EEG entropy measures in anesthesia," *Frontiers Comput. Neurosci.*, vol. 9, no. 16, pp. 1–17, 2015.

[35] M. Borowska, "Entropy-based algorithms in the analysis of biomedical signals," *Stud. Log., Grammar Rhetoric*, vol. 43, no. 1, pp. 21–32, 2015.

[36] O. A. Rosso, R. Ospina, and A. C. Frery, "Classification and verification of handwritten signatures with time causal information theory quantifiers," *PLoS ONE*, vol. 11, no. 12, Dec. 2016, Art. no. e0166868.

[37] H. Aronowitz and D. Burshtein, "Efficient speaker recognition using approximated cross entropy (ACE)," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 7, pp. 2033–2043, Sep. 2007.

[38] Y. Xingwei, Z. Jun, L. Dawei, and W. Jianwei, "Radar jamming detection based on approximate entropy and moving-cut approximate entropy," in *Proc. IET Int. Conf. Inf. Sci. Control Eng. (ICISCE)*, Shenzhen, China, Dec. 2012, pp. 1–6.

[39] R. Hernández-Pérez, L. Guzmán-Vargas, A. Ramírez-Rojas, and F. Angulo-Brown, "Pattern synchrony in electrical signals related to earthquake activity," *Phys. A, Stat. Mech. Appl.*, vol. 389, no. 6, pp. 1239–1252, Mar. 2010.

[40] G. S. Raman and R. T. Subhalakshmi, "Active steganalysis based on adapted lempel-ziv complexity and approximate entropy estimation," in *Proc. IEEE Conf. Inf. Commun. Technol. (ICT)*, Thuckalay, India, Apr. 2013, pp. 917–922.

[41] M. Spindler, "On concepts of randomness in finite sequences," Ph.D. dissertation, Swiss Federal Inst. Technol. (ETHZ), Zürich, Switzerland, 2009.

[42] A. Carpi and A. de Luca, "Uniform words," *Adv. Appl. Math.*, vol. 32, no. 3, pp. 485–522, Apr. 2004.

[43] B. H. Singer and S. Pincus, "Irregular arrays and randomization," *Proc. Nat. Acad. Sci. USA*, vol. 95, no. 4, pp. 1363–1368, Feb. 1998.

[44] C. J. Moore, "A threshold structure metric for medical image interrogation: The 2D extension of approximate entropy," in *Proc. 20th Int. Conf. Inf. Vis. (IV)*, Lisbon, Portugal, Jul. 2016, pp. 336–341.

[45] A. L. Goldberger, C.-K. Penga, and L. A. Lipsitz, "What is physiologic complexity and how does it change with aging and disease?" *Neurobiol. Aging*, vol. 23, no. 1, pp. 23–26, 2002.

[46] H.-O. Peitgen, H. Jürgens, and D. Saupe, *Chaos and Fractals: New Frontiers of Science*, 2nd ed. New York, NY, USA: Springer, 1992.

[47] T. Okabe, "Biophysical optimality of the golden angle in phyllotaxis," *Sci. Rep.*, vol. 5, no. 1, p. 15358, Dec. 2015.

[48] T. Okabe, "Vascular phyllotaxis transition and an evolutionary mechanism of phyllotaxis," 2012, *arXiv:1207.2838*.

[49] J. C. Russ, *Fractal Surfaces*. Boca Raton, FL, USA: Plenum Press, 1994.

[50] H. E. Schepers, J. H. G. M. van Beek, and J. B. Bassingthwaighte, "Four methods to estimate the fractal dimension from self-affine signals (medical application)," *IEEE Eng. Med. Biol. Mag.*, vol. 11, no. 2, pp. 57–64, Jun. 1992.

[51] H.-O. Peitgen and D. Saupe, *The Science of Fractal Images*. New York, NY, USA: Springer, 1988.

[52] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, 2009.

[53] S. E. Hough, "On the use of spectral methods for the determination of fractal dimension," *Geophys. Res. Lett.*, vol. 16, no. 7, pp. 673–676, Jul. 1989.

[54] G. E. Powell and I. C. Percival, "A spectral entropy method for distinguishing regular and irregular motion of Hamiltonian systems," *J. Phys. A, Math. Gen.*, vol. 12, no. 11, pp. 2053–2071, Nov. 1979.

[55] V. Jäntti and S. Alahuhta, "Spectral entropy—What has it to do with anaesthesia, and the EEG?" *Brit. J. Anaesthesia*, vol. 93, no. 1, pp. 150–152, Jul. 2004.

[56] C. Nicolini, G. Forcellini, L. Minati, and A. Bifone, "Scale-resolved analysis of brain functional connectivity networks with spectral entropy," *NeuroImage*, vol. 211, May 2020, Art. no. 116603.

[57] V. Zakeri and A. J. Hodgson, "Automatic identification of hard and soft bone tissues by analyzing drilling sounds," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 2, pp. 404–414, Feb. 2019.

[58] Z. Longfu, S. Yi, H. Sun, L. Zheng, H. Dapeng, and H. Yonghe, "Identification of bowel sound signal with spectral entropy method," in *Proc. 12th IEEE Int. Conf. Electron. Meas. Instrum. (ICEMI)*, Qingdao, China, Jul. 2015, pp. 798–802.

[59] A. Anier, T. Lipping, R. Ferenets, P. Puumala, E. Sonkajärvi, I. Rätsep, and V. Jäntti, "Relationship between approximate entropy and visual inspection of irregularity in the EEG signal, a comparison with spectral entropy," *Brit. J. Anaesthesia*, vol. 109, no. 6, pp. 928–934, Dec. 2012.

[60] T. Giannakopoulos, G. Siantikos, S. Perantonis, N.-E. Votsi, and J. Pantis, "Automatic soundscape quality estimation using audio analysis," in *Proc. 8th ACM Int. Conf. Pervasive Technol. Rel. Assistive Environ. (PETRA)*, Corfu, Greece, Jul. 2015, pp. 1–9.

[61] G. Qiao, T. Ma, S. Liu, N. Zheng, Z. Babar, and Y. Yin, "Spectral entropy based dolphin whistle detection algorithm and its possible application for biologically inspired communication," in *Proc. OCEANS*, Marseille, France, Jun. 2019, pp. 1–6.

[62] Y. Ji, X. Wang, Z. Liu, Z. Yan, L. Jiao, D. Wang, and J. Wang, "EEMD-based online milling chatter detection by fractal dimension and power spectral entropy," *Int. J. Adv. Manuf. Technol.*, vol. 92, nos. 1–4, pp. 1185–1200, Sep. 2017.

[63] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Expert Syst. Appl.*, vol. 171, Jun. 2021, Art. no. 114591.

[64] G. Mittag and S. Möller, "Quality estimation of noisy speech using spectral entropy distance," in *Proc. 26th Int. Conf. Telecommun. (ICT)*, Hanoi, Vietnam, Apr. 2019, pp. 197–201.

[65] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Process., Image Commun.*, vol. 29, no. 8, pp. 856–863, 2014.

[66] K. T. Park, M. S. Park, J. H. Lee, and Y. S. Moon, "Detection of visual saliency in discrete cosine transform domain," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Las Vegas, NV, USA, Jan. 2012, pp. 128–129.

[67] M. Kristan, J. Perš, M. Perše, and S. Kovačič, "A Bayes-spectral-entropy-based measure of camera focus using a discrete cosine transform," *Pattern Recognit. Lett.*, vol. 27, no. 13, pp. 1431–1439, Oct. 2006.

[68] J. I. Makhoul and J. J. Wolf, "Linear prediction and the spectral analysis of speech," Bolt, Beranek Newman, Cambridge, MA, USA, Tech. Rep. BBN-TR-2304, Aug. 1972.

[69] T. Ziyaee, "Unsupervised denoising via Wiener entropy masking in the STFT domain," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Baltimore, MD, USA, Oct. 2014, pp. 467–472.

[70] N. Madhu, "Note on measures for spectral flatness," *Electron. Lett.*, vol. 45, no. 23, pp. 1195–1196, Nov. 2009.

[71] N. Obin and M. Liuni, "On the generalization of Shannon entropy for speech recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Miami, FL, USA, Dec. 2012, pp. 97–102.

[72] J. J. Burred and A. Lerch, "Hierarchical automatic audio signal classification," *J. Audio Eng. Soc.*, vol. 52, nos. 7–8, pp. 724–739, 2004.

[73] O. Tchernichovski, F. Nottebohm, C. E. Ho, B. Pesaran, and P. P. Mitra, "A procedure for an automated measurement of song similarity," *Animal Behav.*, vol. 59, no. 6, pp. 1167–1176, Jun. 2000.

[74] H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. Chichester, U.K.: Wiley, 2005.

[75] A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp, Eds., *Handbook of Spatial Statistics*. Boca Raton, FL, USA: CRC Press, 2014.

[76] P. J. Diggle, *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*, 3rd ed. Boca Raton, FL, USA: CRC Press, 2014.

[77] T. Wiegand and K. A. Moloney, *Handbook of Spatial Point-Pattern Analysis in Ecology*. Boca Raton, FL, USA: CRC Press, 2014.

[78] J. Matoušek, *Geometric Discrepancy: An Illustrated Guide*. Berlin, Germany: Springer, 2010.

[79] W. Chen, A. Srivastav, and G. Travaglini, Eds., *A Panorama of Discrepancy Theory*. Cham, Switzerland: Springer, 2014.

[80] M. T. Shehata, "Characterization of particle dispersion," in *Practical Guide to Image Anal.*, J. J. Friel, Ed. Materials Park, OH, USA: ASM International, 2000, ch. 6, pp. 129–144.

[81] F. Nekka and J. Li, "Various mathematical approaches to extract information from textures of increasing complexities," in *Fractals in Engineering: New Trends in Theory and Applications*, L.-V. Jacques and L.-V. Evelyne, Eds. London, U.K.: Springer, 2005, pp. 255–270.

[82] D. J. Whitehouse, *Handbook of Surface and Nanometrology*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2011.

[83] D. J. Whitehouse, "Fractal or fiction," *Wear*, vol. 249, nos. 5–6, pp. 345–353, Jun. 2001.

[84] A. Fischer, K. Riesen, and H. Bunke, "Graph similarity features for HMM-based handwriting recognition in historical documents," in *Proc. 12th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Kolkata, India, Nov. 2010, pp. 253–258.

[85] G. Chartrand, P. Erdös, and O. R. Oellermann, "How to define an irregular graph," *College Math. J.*, vol. 19, no. 1, pp. 36–42, Jan. 1988.

[86] A. Ali, G. Chartrand, and P. Zhang, *Irregularity in Graphs*. Cham, Switzerland: Springer, 2021.

[87] G. Chartrand and P. Zhang, *Chromatic Graph Theory*. Boca Raton, FL, USA: CRC Press, 2009.

[88] A. Benjamin, G. Chartrand, and P. Zhang, *The Fascinating World of Graph Theory*. Princeton, NJ, USA: Princeton Univ. Press, 2015.

[89] G. Chartrand, L. Lesniak, and P. Zhang, *Graphs & Digraphs*, 6th ed. Boca Raton, FL, USA: CRC Press, 2015.

[90] D. Wells, "Which is the most beautiful?" *Math. Intelligencer*, vol. 10, no. 4, pp. 30–31, Sep. 1988.

[91] D. Wells, "Are these the most beautiful?" *Math. Intell.*, vol. 12, no. 3, pp. 37–41, 1990.

[92] Wikipedia. (Jul. 24, 2017). *Glossary of Graph Theory Terms*. [Online]. Available: https://en.wikipedia.org/wiki/Glossary_of_graph_theory_term

[93] G. Chartrand, G. L. Johns, K. A. McKeon, and P. Zhang, "The rainbow connectivity of a graph," *Networks*, vol. 54, no. 2, pp. 75–81, Sep. 2009.

[94] T. R. Jensen and B. Toft, *Graph Coloring Problems*. New York, NY, USA: Wiley, 1995.

[95] E. Estrada, "Randić index, irregularity and complex biomolecular networks," *Acta Chim. Slovenica*, vol. 57, pp. 597–603, Jan. 2010.

[96] J. L. Gross, J. Yellen, and P. Zhang, *Handbook of Graph Theory*. Boca Raton, FL, USA: CRC Press, 2014.

[97] A. Soifer, *The Mathematical Coloring Book: Mathematics of Coloring and the Colorful Life of Its Creators*. New York, NY, USA: Springer, 2008.

[98] J. Pach and P. K. Agarwal, *Combinatorial Geometry*. New York, NY, USA: Wiley, 1995.

[99] J. Pach, "The beginnings of geometric graph theory," in *Erdős Centennial*, L. Lovász, I. Z. Ruzsa, and V. T. Sós, Eds. New York, NY, USA: Springer, 2013, pp. 465–484.

[100] A. M. Raigorodskii, "Coloring distance graphs and graphs of diameters," in *Thirty Essays on Geometric Graph Theory*, J. Pach, Ed. New York, NY, USA: Springer, 2013, pp. 429–460.

[101] L. Collatz and U. Sinogowitz, "Spektren endlicher Grafen," *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg*, vol. 21, no. 1, pp. 63–77, 1957.

[102] Y. Alavi, G. Chartrand, F. R. K. Chung, P. Erdös, R. L. Graham, and O. R. Oellermann, "Highly irregular graphs," *J. Graph Theory*, vol. 11, no. 2, pp. 235–249, 1987.

[103] P. Hoffman, *The Man Who Loved Only Numbers: The Story of Paul Erdős and the Search for Mathematical Truth*. London, U.K.: Fourth Estate, 1999.

[104] M. Aigner and G. M. Ziegler, *Proofs From the Book*, 5th ed. Berlin, Germany: Springer, 2014.

[105] G. Chartrand, G. L. Johns, K. A. McKeon, and P. Zhang, "Rainbow connectivity of cages," in *Proc. 38th Southeastern Int. Conf. Combinatorics, Graph Theory Comput.*, Boca Raton, FL, USA, no. 184, Mar. 2007, pp. 209–222.

[106] G. Chartrand, G. L. Johns, K. A. McKeon, and P. Zhang, "Rainbow connection in graphs," *Math. Bohemica*, vol. 133, no. 1, pp. 85–98, 2008.

[107] G. Chartrand, M. S. Jacobson, J. Lehel, O. R. Oellermann, S. Ruiz, and F. Saba, "Irregular networks," *Congressus Numerantium*, vol. 64, pp. 355–374, 1988. [Online]. Available: https://www.researchgate.net/publication/265701559_Irregular_networks/references

[108] G. Chartrand, L. Lesniak, D. W. VanderJagt, and P. Zhang, "Recognizable colorings of graphs," *Discussiones Mathematicae Graph Theory*, vol. 28, no. 1, pp. 35–57, 2008.

[109] I. Gutman, B. Ruščić, N. Trinajstić, and J. C. F. Wilkox, "Graph theory and molecular orbitals. XII. Acyclic polyenes," *J. Chem. Phys.*, vol. 62, no. 9, pp. 3399–3405, 1975.

[110] I. Gutman and N. Trinajstić, "Graph theory and molecular orbitals. Total ϕ-electron energy of alternant hydrocarbons," *Chem. Phys. Lett.*, vol. 17, no. 4, pp. 535–538, 1972.

[111] S. Nikolić, G. Kovačević, A. Miličević, and N. Trinajstić, "The zagreb indices 30 years after," *Croatica Chem. Acta*, vol. 76, no. 2, pp. 113–124, 2010.

[112] G. H. Fath-Tabar, "Old and new Zagreb indices of graphs," *MATCH Commun. Math. Comput. Chem.*, vol. 65, no. 1, pp. 79–84, 2011.

[113] D. Janežič, A. Miličević, S. Nikolić, and N. Trinajstić, *Graph-Theoretical Matrices in Chemistry*. Boca Raton, FL, USA: CRC Press, 2015.

[114] F. K. Bell, "A note on the irregularity of graphs," *Linear Algebra Appl.*, vol. 161, pp. 45–54, Jan. 1992.

[115] M. O. Albertson, "The irregularity of a graph," *Ars Combinatoria*, vol. 46, pp. 219–225, 1997. [Online]. Available: http://www.combinatoire.ca/ArsCombinatoria/TOC.html and http://www.combinatoire.ca/ArsCombinatoria/Vol46.html

[116] V. Nikiforov, "Eigenvalues and degree deviation in graphs," *Linear Algebra Appl.*, vol. 414, no. 1, pp. 347–360, Apr. 2006.

[117] M. Krivelevich and R. Yuster, "The rainbow connection of a graph is (at most) reciprocal to its minimum degree," *J. Graph Theory*, vol. 63, no. 3, pp. 185–191, 2010.

[118] I. Gutman, P. Hansen, and H. Mélot, "Variable neighborhood search for extremal graphs. 10. Comparison of irregularity indices for chemical trees," *J. Chem. Inf. Model.*, vol. 45, no. 2, pp. 222–230, Mar. 2005.

[119] J. A. D. Oliveira, C. S. Oliveira, C. Justel, and N. M. M. D. Abreu, "Measures of irregularity of graphs," *Pesquisa Operacional*, vol. 33, no. 3, pp. 383–398, Nov. 2013.

[120] A. Frieze, R. J. Gould, M. Karoński, and F. Pfender, "On graph irregularity strength," *J. Graph Theory*, vol. 41, no. 2, pp. 120–137, Oct. 2002.

[121] S. Mukwembi, "On maximally irregular graphs," *Bull. Malaysian Math. Sci. Soc.*, vol. 36, no. 3, pp. 717–721, 2011.

[122] D. Rautenbach and L. Volkmann, "How local irregularity gets global in a graph," *J. Graph Theory*, vol. 41, no. 1, pp. 18–23, Sep. 2002.

[123] G. Ebert, J. Hemmeter, F. Lazebnik, and A. Woldar, "On the number of irregular assignments on a graph," *Discrete Math.*, vol. 93, nos. 2–3, pp. 131–142, Nov. 1991.

[124] S. Mukwembi, "A note on diameter and the degree sequence of a graph," *Appl. Math. Lett.*, vol. 25, no. 2, pp. 175–178, Feb. 2012.

[125] M. Bača, S. Jendroľ, M. Miller, and J. Ryan, "On irregular total labellings," *Discrete Math.*, vol. 307, nos. 11–12, pp. 1378–1388, May 2007.

[126] G. Chartrand, O. R. Oellermann, and M. Schultz, "Distance: A graphical tour," in *Proc. 2nd Quadrennial Int. Conf. Theory Appl. Graphs*. San Francisco, CA, USA: San Francisco State Univ., Jul. 1989, pp. 441–458.

[127] S. Piccard, *Sur les Ensembles de Distances des Ensembles de Points d'un Espace Euclidien*. Neuchâtel, Switzerland: Univ. Neuchâtel, 1939.

[128] T. Aste, *The Pursuit of Perfect Packing*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2008.

[129] C. Zong, "Packing, covering and tiling in two-dimensional spaces," *Expositiones Mathematicae*, vol. 32, no. 4, pp. 297–364, 2014.

[130] A. Okabe, B. Boots, W. Laurier, K. Sugihara, and S. Chiu, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. Chichester, U.K.: Wiley, 2000.

[131] S. M. Moser and P.-N. Chen, *A Student's Guide to Coding and Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[132] P. D. Grünwald, *The Minimum Description Length Principle*. Cambridge, MA, USA: MIT Press, 2007.

[133] R. Weatherford, *Philosophical Foundations of Probability Theory*. London, U.K.: Routledge & Kegan Paul, 1982.

[134] R. S. Nickerson, "The production and perception of randomness," *Psychol. Rev.*, vol. 109, no. 2, pp. 330–357, 2002.

[135] M. Challinor, "Change, chance and structure: Randomness and formalism in art," *Leonardo*, vol. 4, no. 1, pp. 1–11, 1971.

[136] Y. B. Sanderson, "Color charts, esthetics, and subjective randomness," *Cognit. Sci.*, vol. 36, no. 1, pp. 142–149, Jan. 2012.

[137] L. Lovász and K. Vesztergombi, "Geometric representations of graphs," in *Paul Erdős and His Mathematics*, vol. 2, G. Halász, L. Lovász, M. Simonovits, and V. T. Sós, Eds. Berlin, Germany: Springer, 2002, pp. 471–498.

[138] H.-O. Peitgen, H. Jürgens, and D. Saupe, *Chaos and Fractals: New Frontiers of Science*. New York, NY, USA: Springer, 1992.

[139] Wikipedia. (Jul. 1, 2017). *Sophie Piccard*. [Online]. Available: https://fr.wikipedia.org/wiki/Sophie_Piccard

[140] Y. Voegeli. (Aug. 2, 2017). *Suffrage Féminin. Dictionnaire Historique de la Suisse*. [Online]. Available: http://www.hls-dhs-dss.ch/textes/f/F10380.ph

[141] P. Erdös, "On sets of distances of n points," *Amer. Math. Monthly*, vol. 53, no. 5, pp. 248–250, May 1946.

[142] P. Erdös, "On sets of distances of *n* points," *Amer. Math. Monthly*, vol. 77, no. 7, pp. 739–740, 1970.

[143] P. Erdös, "Set theoretic, measure theoretic, combinatorial, and number theoretic problems concerning point sets in Euclidean space," *Real Anal. Exchange*, vol. 4, no. 2, pp. 113–138, 1978.

[144] E. Paul and R. K. Guy, "Distinct distances between lattice points," *Elemente Math.*, vol. 25, no. 6, pp. 121–123, 1970.

[145] E. J. Makai, J. Pach, and J. Spencer, "New results on the distribution of distances determined by separated point sets," in *Paul Erdős and His Mathematics*, vol. 2, G. Halász, L. Lovász, M. Simonovits, and V. T. Sós, Eds. Berlin, Germany: Springer, 2002, pp. 499–511.

[146] zbMATH. (Jul. 1, 2017). *Erdős, Paul. Zentralblatt MATH*. [Online]. Available: https://zbmath.org/authors/?s=0&q=au%3Apaul+erdo

[147] P. Erdős. (Aug. 2, 2017). *Paul Erdős' Papers*. [Online]. Available: http://login.math-inst.hu/~p_erdos/Erdos.htm

[148] L. Babai and J. Spencer, "Paul Erdős (1913–1996)," *Notices Amer. Math. Soc.*, vol. 45, no. 1, pp. 64–73, 1998.

[149] E. W. Weisstein. (Nov. 28, 2017). Discrepancy. MathWorld. [Online]. Available: http://mathworld.wolfram.com/Discrepancy.html

[150] E. W. Weisstein. (Nov. 28, 2017). Discrepancy theorem. MathWorld. [Online]. Available: http://mathworld.wolfram.com/DiscrepancyTheorem.html

[151] (Dec. 4, 2021). *Uniform Distribution Theory*. [Online]. Available: https://sciendo.com/journal/UDT

[152] B. Chazelle, *The Discrepancy Method: Randomness and Complexity*. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[153] P. Hellekalek and G. Larcher, *Random and Quasi-Random Point Sets*. New York, NY, USA: Springer, 1998.

[154] M. Ghyka, *The Geometry of Art and Life*. New York, NY, USA: Dover, 1977.

[155] R. A. Dunlap, *The Golden Ratio and Fibonacci Numbers*. Singapore: World Scientific, 1998.

[156] N. J. A. Sloane. (Jul. 24, 2017). *A001622: Decimal Expansion of Golden Ratio. The On-Line Encyclopedia of Integer Sequences*. [Online]. Available: https://oeis.org/A00162

[157] M. Neveux and H. E. Huntley, *Le Nombre d'or. Radiographie d'un Mythe Suivi de La Divine Proportion*. Paris, France: Points, 2014.

[158] G. W. Berkhan and W. F. Meyer, "Neuere Dreiecksgeometie," in *Encyklopädie der Mathematischen Wissenschaften mit Einschluß Ihrer Anwendungen*, vol. III.1.2 AB 10, W. F. Meyer and H. Mohrmann, Eds. Leipzig, Germany: B. G. Teubner Verlag, 1914, pp. 1173–1276.

[159] P. Baptist, *Die Entwicklung der Neueren Dreiecksgeometrie*. Mannheim, Germany: BI-Wissenschaftsverlag, 1992.

[160] P. J. Davis, "The rise, the fall, and possible transfiguration of triangle geometry: A mini-history," *Amer. Math. Monthly*, vol. 32, no. 3, pp. 204–214, 1995.

[161] P. Romera-Lebret, "Die neue Dreiecksgeometrie: Der Übergang von der Mathematik der Amateure zur unterrichteten Mathematik," *Mathematische Semesterberichte*, vol. 59, no. 1, pp. 75–102, 2012.

[162] H. Fukagawa and T. Rothman, *Sacred Mathematics: Japanese Temple Geometry*. Princeton, NJ, USA: Princeton Univ. Press, 2008.

[163] P. Mehta and P. Majumder, *From Extractive to Abstractive Summarization: A Journey*. Singapore: Springer, 2019.

[164] J.-M. Torres-Moreno, *Automatic Text Summarization*. London, U.K.: Wiley, 2014.

[165] D. Simakov, Y. Caspi, E. Shechtman, and M. Irani, "Summarizing visual data using bidirectional similarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, Jun. 2008, pp. 1–8.

[166] S. Le Moan, I. Farup, and J. Blahová, "Towards exploiting change blindness for image processing," *J. Vis. Commun. Image Represent.*, vol. 54, pp. 31–38, Jul. 2018.

[167] K. Berkner, E. L. Schwartz, and C. Marle, "SmartNails—Image and display dependent thumbnails," *Proc. SPIE*, vol. 5296, pp. 54–65, Jan. 2004.

[168] F. Brodbeck. (2011). *Cinemetrics*. Accessed: Jun. 9, 2021. [Online]. Available: http://cinemetrics.fredericbrodbeck.de

[169] M. A. Smith and T. Kanade, *Multimodal Video Characterization and Summarization*. New York, NY, USA: Kluwer, 2005.

[170] Z. Xiong, R. Radhakrishnan, A. Divakaran, Y. Rui, and T. S. Huang, *A Unified Framework for Video Summarization, Browsing & Retrieval: With Applications to Consumer and Surveillance Video*. San Diego, CA, USA: Elsevier, 2006.

[171] V. Atanasiu and R. Ingold, "Document towers: A MATLAB software implementing a three-dimensional architectural paradigm for the visual exploration of digital documents and libraries," *SoftwareX*, vol. 14, Jun. 2021, Art. no. 100684.

[172] D. S. Bloomberg and F. R. Chen, "Document image summarization without OCR," in *Proc. 3rd IEEE Int. Conf. Image Process.*, Lausanne, Switzerland, Sep. 1996, pp. 229–232.

[173] R. P. Futrelle, "Summarization of diagrams in documents," in *Advances in Automated Text Summarization*, I. Mani and M. T. Maybury, Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 403–421.

[174] A. Brink, L. Schomaker, and M. Bulacu, "Towards explainable writer verification and identification using vantage writers," in *Proc. 9th Int. Conf. Document Anal. Recognit. (ICDAR)*, Curitiba, Brazil, Sep. 2007, pp. 824–828.

[175] K. McCall and K. Piersol, "Enhancing accuracy of jumping by incorporating interestingness estimates," U.S. Patent 20 070 180 355, Aug. 2, 2007.

[176] F. Matulic, "Automatic selection of visually attractive pages for thumbnail display in document list view," in *Proc. 3rd Int. Conf. Digit. Inf. Manage.*, London, U.K., Nov. 2008, pp. 221–226.

[177] R. N. Bracewell, *The Fourier Transform and Its Applications*, 3rd ed. Boston, MA, USA: McGraw-Hill, 2000.

[178] R. L. J. Easton, *Fourier Methods in Imaging*. Chichester, U.K.: Wiley, 2010.

[179] D. C. Champeney, *Fourier Transforms and Their Physical Applications*. London, U.K.: Academic, 1973.

[180] S. W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing*, 2nd ed. San Diego, CA, USA: California Technical Publishing, 1999.

[181] E. W. Kamen and B. S. Heck, *Fundamentals of Signals and Systems Using the Web and MATLAB*, 3rd ed. Harlow, U.K.: Pearson, 2014.

[182] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[183] D. Bradley and G. Roth, "Adapting thresholding using the integral image," *J. Graph. Tools*, vol. 12, no. 2, pp. 13–21, 2007.

[184] M. D. Fairchild, *Color Appearance Models*, 3rd ed. Chichester, U.K.: Wiley, 2013.

[185] C. E. Shannon, "A mathematical theory of cryptography," Bell Laboratories, Murray Hill, NJ, USA, Tech. Memorandum 45-110-92, Sep. 1945.

[186] N. J. A. Sloane and A. D. Wyner, Eds., *Claude Elwood Shannon Miscellaneous Writings*. Murray Hill, NJ, USA: Mathematical Science Research Center, AT&T Bell Laboratories, 2013.

[187] F. Attneave, "Stochastic composition processes," *J. Aesthetics Art Criticism*, vol. 17, no. 4, pp. 503–510, 1959.

[188] H. Meinhardt, *The Algorithmic Beauty of Sea Shells*, 4th ed. Berlin, Germany: Springer, 2009.

[189] J. Wagemans, Ed., *The Oxford Handbook of Perceptual Organization*. New York, NY, USA: Oxford Univ. Press, 2015.

[190] W. Prinz, "Quantitative Versuche über die Prägnanz von Punktmustern," Psychologische Forschung, vol. 29, no. 4, pp. 297–359, 1966.

[191] G. Jewell and M. E. McCourt, "Pseudoneglect: A review and meta-analysis of performance factors in line bisection tasks," *Neuropsychologia*, vol. 38, no. 1, pp. 93–110, Jan. 2000.

[192] H. Cartier-Bresson, *L'imaginaire d'après nature*. Paris, France: Fata Morgana, 1996.

[193] L. Pacioli, *De Divina Proportione*. Venice, Italy: A. Paganius Paganinus, 1509. [Online]. Available: https://archive.org/details/diuinaproportion00paci/

[194] B. Wolf, M. Jung, A. Fischer, G. Waeber, and V. Atanasiu, "Analyse 'automatisierte integration des alten sachkatalogs in alma,'" HES·SO//Fribourg, iCoSys, Fribourg, Switzerland, Tech. Rep., Dec. 2020.

[195] Wikipedia Contributors. (Jan. 31, 2021). *Tesseract (Software)*. [Online]. Available: https://en.wikipedia.org/wiki/Tesseract_(software)

[196] Google. (Jan. 31, 2021). *Vision AI*. [Online]. Available: https://cloud.google.com/vision/

[197] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, no. 4, pp. 619–681, Aug. 1982.

[198] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, Apr. 2011.

[199] M. Hayes, "The reconstruction of a multidimensional sequence from the phase or magnitude of its Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, no. 2, pp. 140–154, Apr. 1982.

[200] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proc. IEEE*, vol. 69, no. 5, pp. 529–541, May 1981.

[201] B. C. Hansen and R. F. Hess, "Structural sparseness and spatial phase alignment in natural scenes," *J. Opt. Soc. Amer.*, vol. 24, no. 7, pp. 1873–1885, 2007.

[202] P. Kovesi, "Image features from phase congruency," *Videre, J. Comput. Vis. Res.*, vol. 1, no. 3, pp. 2–26, 1999.

[203] A. W. Lohmann, D. Mendlovic, and G. Shabtay, "Significance of phase and amplitude in the Fourier domain," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 14, no. 11, pp. 2901–2904, 1997.

VLAD ATANASIU passed the student admission examinations at the Faculty of Telecommunications, Polytechnic University, Timişoara, Romania, in 1989. He received the B.A. degree in middle eastern studies and the M.A. degree in Arabic linguistics from the University of Provence, Aix-en-Provence, France, in 1994 and 1995, respectively, and the Ph.D. degree in Arabic paleography from the École Pratique des Hautes Études, Paris, France, in 2003. He is currently pursuing the Ph.D. degree with the Informatics Department, University of Fribourg, Switzerland, with a focus on document analysis and visualization.

From 2003 to 2005, he was a Postdoctoral Fellow with the Architecture Department and the Brain and Cognitive Science Department, Massachusetts Institute of Technology, Cambridge, MA, USA. From 2006 to 2009, he was at the Visualization Commission, Austrian Academy of Sciences, Vienna, Austria, where he developed image processing methods for paper structure analysis and geographical information system tools to support paper history, and further co-coordinated the European computational paper history project "Bernstein." From 2010 to 2011, he developed handwriting analysis and visualization software for Télecom-ParisTech, Paris, and the back- and front-end of an online cartography atlas of Iran for the French National Centre for Scientific Research, Paris. He taught information visualization at the University of Graz, Austria, and the University of Fribourg, from 2011 to 2015. From 2019 to 2021, he was with the University of Basel, Switzerland, where he created software for the legibility enhancement of ancient papyri. Currently, he is doing data analysis of natural language datasets for the University of Bern, Switzerland. He does freelance typographical and graphic design work. He is the author of the books *On Letter Frequency and Its Influence on Arabic Calligraphy* (Paris: L'Harmattan, 1999) and *Expert Bytes: Computer Expertise in Forensic Documents* (Boca Raton: CRC Press, FL, 2014).