The Structural Information Potential and Its Application to Document Triage

VLAD ATANASIU
Department of Informatics, University of Fribourg, Bd de Perolles 90, 1700 Fribourg, Switzerland (e-mail: atanasiu@alum.mit.edu)

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 typical of 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.

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.

In the document domain, the above observation is supported by compelling empirical evidence from the practice [1] [2] [3] [4], theory [5] [6] [7], psychology [8] [9] [10] [11] [12] [13], and history [14] [15] of document design (these references are a small representative sample from a much larger bibliography). For example, much of the effort invested in microtypography [16] [17] (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”) [18] [19]. 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 (see the six lines above) [20]. 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 accidently (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.

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 an operational 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.

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 2, “Related work”, presents advantageous and limiting factors of major approaches to information measurement, explaining the need for a new method, while introducing elements used in the proposed method. Section 3, “Method”, is the theoretical core of the article, in which the analytical formula and algorithmic components of the Structural Information Potential measurement method are described and justified; the design of structures with minimal redundancy is also discussed. Section 4, “Experiments”, integrates the proposed method into a cybernetic perspective, by demonstrating how it helps translate patterns to information in support of a decision-making process. Section 5, “Discussion”, concerns technical matters and future work. Section 6, “Conclusions”, summarizes the theoretical and practical significance of the proposed method and applications.

II. RELATED WORK

A. GENERIC QUANTIFICATION OF PATTERN IRREGULARITY

Information theory has a central place among the domains relating pattern structures to informativeness. It is rooted in the Shannon entropy, $H$ [21], 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 [22, pp. 393]. The normalized variant is given by the relative entropy, $H_r = H/H_{max}$, $H_{max} = \log_2(n)$ [22, 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 [23], [24]. 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 [25], [26]. From the field of nuclear physics comes the strength of structure [27, 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 \( n = \sum_{i=1}^{u} v_{ui} \) entities, and the entropy of the whole, containing \( n = \sum_{i=1}^{u} v_{ui} \):

\[
J = - \sum_{i=1}^{u} \sum_{j=1}^{v} p_{ij} \log_2 p_{ij} + \sum_{k=1}^{n} p_k \log_2 p_k.
\] (1)

This method has been utilized in document layout segmentation [28]. 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 [29] [30] [31]. ApEn has been refined through many parametric [32] [33] and algorithmic variants [34] [35] [36] and compares well with other methods [37] [38]. From its inception onwards, it has been successfully applied to various biosignals [37] [38], and later to a broad range of other applications such as online signature verification [39], speaker recognition [40], radar jamming [41], earthquake prediction [42], and cryptography [31] [43]. 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) [44] [45]. The method may be extended to the analysis of images using vectorized bidimensional windows [46] [47]. This approach, in addition to improved parametrization [32], was used in this article to process document images. The results are discussed in Section 4, “Experiments”, and illustrated in Fig. 6. While clustered documents are classified at one end of the spectrum, 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 [48].

Fractals [49] 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 [50]. One of the most reliable methods for determining the fractal dimension, \( D \), uses spectral analysis and defines it as the slope, \( \beta \), 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 [51, pp. 97–114]. Two requirements must be satisfied if the measured entity is to be deemed a fractal: namely, the phase must be uniformly random [51, pp. 99] and the power spectrum must follow a power law, \( 1/f^\beta \) [52] [53]. 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.

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 \( n \):

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

Variations of the definition include obtaining the power spectrum from the discrete cosine transform or by way of auto-correlation, 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 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 online Supplemental Material).

Spectral flatness, or $SF$ [68, pp. 112–115] [69], [70], 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:

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

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. This method also results in a perceptually unsatisfactory classification of images (Fig. 6).

Point set analysis and spatial statistics are concerned with the characterization of the distribution of point-like objects and events in space [75] [76] [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] [51, 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] [51, 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.

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, 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.

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) [49, pp. 78–81] [87], packing [88], tiling [89], tessellation [50], coding [90] [91], and randomness [92] [93] [94] [95].

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 INFORMATIVENESS 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 [96] [97]. Image summarization is conceptually similar but more difficult to realize; for example, representing a person by her or his face and a
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 Eq. 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

```latex
SIP = 1 − |dSIP| \quad (4)
\]
```

where $I$ is an intensity image, $B$ a binarization process, $F$ the image’s Fourier transform \cite{110}, $|·|^2$ 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, $sgn \ dSIP$.

### Rationales

The choice of the frequency domain for pattern analysis has a number of benefits \cite{111}: it facilitates the integration of sample points across space, and thus the description of spatial structures; 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 maxima correspond to regular and uniform patterns in the spatial domain.

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 that facilitates the ordering of patterns according to their Structural Information Potential. A discussion of maximal SIP, and its correspondence with minimal structural redundancy, is also presented.

A. STRUCTURAL INFORMATION POTENTIAL

The core machinery of the Structural Information Potential measurement consists of the Shannon entropy of the logarithm of the power spectrum of a binary image’s Fourier transform. Formally, the value $SIP \in [0, 1]$ is obtained through a number of equations and algorithmic steps:

- The choice of the frequency domain for pattern analysis has a number of benefits \cite{111}: it facilitates the integration of sample points across space, and thus the description of spatial structures; 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 maxima correspond to regular and uniform patterns in the spatial domain.

- 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 Eq. 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.

**Enabling operation** — A uniform signal in the time domain has as its pair in the frequency domain an impulse with frequency of zero [112, 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 [113, pp. 194–200, 218–222]). It becomes an impulse in the frequency domain [112, p. 29], and thus has zero entropy and represents an undesirable outcome for our goal. However, by binarizing the sinusoid, a rectangular pulse train is obtained, which corresponds in the frequency domain to a set of harmonics of the form \(1/(f \pi)\), \(f \in \mathbb{N}_{\text{odd}}\) [114, pp. 102–113], the effect of which is to increase the signal’s spectral entropy. 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 [113, 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 [115] 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. 10, 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 [116]. — Conversion from color images to grayscale is realized by converting the images to the perceptual CIELAB color space and extracting the lightness channel, \(L^*\) [117, pp. 30, 95, 200–212]. The binarization step of the SIP measurement is not necessary for data that is already binary.

**Supporting elements** — The reduction of the power spectrum dimensionality 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

\[
w = \text{int}(\omega + 0.5), \quad n = \text{card}(\text{unique}(w)), \quad V = \sum_{i=1}^{n} P_{w_i}/\text{card}(w_i)
\]

where the vector \(\omega\) 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 \(\text{card}(w_i)\) is the number (cardinality) of samples for a given integer frequency. The transfer function \(T\) calibrates the Structural Information Potential value to facilitate easier mathematical and cognitive manipulation. In the next two sections, it is shown that the minimal value (= 0) of the SIP equation (eq. 5) inside the transfer function corresponds to a uniform pattern, and the maximum (= 1) to a regular pattern, while the clustered pattern with minimal redundancy has an intermediary value (\(-\log_2((\sqrt{5} - 1)/2) \approx 0.6942\)). Since the latter pattern is the most informative, the SIP values are remapped using a polynomial of order two as transfer function, such that the value of the clustered pattern becomes the center of the distribution and its range is \([-1, +1\]). Similar to the approach used for the minimal redundancy scalar, we apply the logarithm to the entire spectrum:

\[
T = g(1.83x^2 + 0.1699x - 1)
\]

\[
g(y) = \begin{cases} 
\log_2(y + 1), & \text{for } x \geq 0 \\
-\log_2(|y| + 1), & \text{for } x < 0
\end{cases}
\]

The implication is that the dSIP value is a measure of divergence from the maximum Structural Information Potential, with its sign indicating the pattern towards which it leans: positive for regular and negative for uniform.

**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. At longer
cycles, they also have entropy closer to zero due to the strong harmonic components introduced by the sharp edges. The same issue affects the second problematic pattern, a single square aligned with the image raster; due to the step function of the function having the sinc function \( \text{sinc}(a) = \frac{\sin(a\pi)}{a\pi} \) as frequency domain pair [113, pp. 212–215], its harmonics increase entropy to nearly one, contrary to what would be expected from a pattern with a large uniform surface.

**B. PATTERN PHASE SPACE**

Fig. 1 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 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. 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”). Postproduction 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. 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 [118] is appropriate for patterns with statistically constant homogeneity, such as homogeneously distributed line segments of quasi-equal length and variable orientation. In this case, the spectrum simply represents the amount of randomness or redundancy of the variable feature, which is not location and hence does not describe structural information.

Patterns with semantic value, such as images of faces, are obviously very important pattern types 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.

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 both “rainbow” graphs (which have edge values that are distinct, but not necessarily

**FIGURE 1.** Conceptual schema of the generic relationship between information and structure, along with the correspondence of the visual and functional organization of documents. — Legend & credits: (1) typography inspired by medieval manuscripts, by the the Arts and Crafts movement exponent William Morris; “The Golden Legend”, 1892, London, Kelmscott Press/Quaritch Books (author’s personal collection); (2) official FBI printed form, heavily annotated by handwriting and stamps; Kennedy Assassination Records Collection (National Archives and Records Administration); (3) memorandum frontpiece, including institutional header, summary, circulation list, and secrecy stamps; Claude Shannon, “A Mathematical Theory of Cryptography”, 1945 (AT&T Bell Laboratories); (4) censored newspaper front page; Austria, August 9, 1918 (Austrian National Library).
maximally distinct) and to the fact that fluorescence maximizes perceptual color discrimination. The proposed notation is $\mathcal{V}_d = \{v\}$, where $d$ is the object’s embedding dimension, $v$ the number of vertices, and the placeholder $\{\}$ may be used to specify edges. For example, $\mathcal{V}_{2d} = \{(v_{12}, e_{13}, e_{23})\} = \{x_1, y_1, ..., x_3, y_3\}$ 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.

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 extremes and is found by determining the value at the intersection of the functions representing the conditions, i.e., $f(x) = 1 - x$ and $g(x) = x/(1 - x)$, for $x \in [0, \frac{1}{2}]$ and $f(x) = x$ and $g(x) = (1 - x)/x$, 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 $x_1 = (1 - \sqrt{5})/2 + 1 \approx 0.6180$ and $x_2 = (1 + \sqrt{5})/2 - 1 \approx 0.3819$. These values are also known, respectively, as the conjugate of the reciprocal, $\Phi^2$, and the reciprocal, $\Phi$, of the golden ratio, $\phi = (1 + \sqrt{5})/2 = 1/\Phi = (1/(1 - \Phi^2)) \approx 1.6180$. A diagrammatic representation of these solutions (Fig. 2 (a)) 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: $R = \{a/(a + b), b/(a + b), \min(a, b)/\max(a, b)\}$, $SIP_{\text{max}} = \min(\max(R))$, 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. 2 (b)). The max formula can be avoided by combining ratios and entropy: $SIP_{\text{max}} = \max(H(-\log_2(R)))$ (Fig. 2 (c)). The similarity between this equation and Eq. 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.

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.

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, it 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. 3). 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.

Fig. 4 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 as the ratio of one edge taken at a time and its adjacent two.

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 to open a window into these possibilities. We conclude this section with two remarks: one on the link between fluorescence and SIP (a), and another on the human perception of fluorescence in relation to its mathematical definition (b).

a) The concept of Structural Information Potential 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 (fractal) 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 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 subdues the pattern spectrum and has a direct information theoretical meaning (Fig. 1). 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 (equal to the Nyquist frequency) entities.

The partial identity of the SIP and fluorescence equations — the outer term \( H(\log_2(\cdot)) \) is common to both — suggests an equivalence even at the algebraic level. The well-known Wiener–Khinchin theorem, which relates the power spectrum to autocorrelation [119], may provide a means of equating the inner SIP term (the power spectrum of an image) with the inner fluorescence term (the combinatorial edge ratios of a graph). From a performance point of view, however, the Fourier-based method of scale-space redundancy measurement is bound to outperform the graph-combinatorial method due to the computational complexity of the latter for images of practical size. Nevertheless, the study of graph redundancy offers a powerful and versatile instrument to understand the elemental levels from which fluorescent patterns emerge. The goal of this section is to provide precisely such a low-level explanation of the SIP method.

b) Human estimation of pattern redundancy differs from mathematically derived values, 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. 5 (a) [120]). 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. 5 (b–c)).
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.” [121, 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 modelization by the 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.” [121, 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

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

Fig. 6 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 (the online Supplemental Material shows all pages, as well as providing the quantitative evaluation). 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 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 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 are similar, but less good). Structural information potential (last block) achieves a perceptually gradual ordering of the pages from empty to multiscale to homogeneous.

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 [122]. 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. 7(c)–(h)). A total of 1.2 million cards were scanned from high-contrast black-and-white films (accessible at http://siibns.ch/
FIGURE 6. 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; $dSIP \in [-0.3109, +0.6523]$, $\max(SIP) = 0.0073$, 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.

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 [123] and the commercial Google Vision [124], revealed that the obtained text is not directly exploitable in the library’s public catalog (Fig. 7 (i)). 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 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. 8. The patterns 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. 7 (b)). 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. 7 (a) 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. 7 (b) 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. 7 (g), (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. 9 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.

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. 9 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, a classification by the amount of recognized characters be a sufficient estimate of matching quality? As illustrated by Fig. 7 (i), 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. 7 (b). 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.
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. Beyond documents, the classification into compact, clustered, and homogeneous distributions is of direct interest to materials science.

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 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 TO NON-TEXTUAL DOCUMENTS

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.

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.

Metamerism — Because of dimensionality reduction (in the case of images being reduced from two dimensions to a scalar), many differently patterned data 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).

Psychophysics — 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.

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 [113, pp. 355–358] [125], [126]; 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 [113, pp. 355–358] [127]–[130]. 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
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 Erdős 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. CONCLUSIONS
This article has introduced Structural Information Potential (SIP), a measure of information based on pattern 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.

ACKNOWLEDGMENT
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.

REFERENCES
[1] T. Samara, Making and Breaking the Grid: A Graphic Design Layout Workshop. Gloucester, MA: Rockport, 2002.
[2] K. Elam, Grid Systems: Principles of Organizing Type. New York, NY: Princeton Architectural Press, 2004.
[3] F. Franchi, Designing News. Berlin: Gestalten, 2013.

[4] B. Schewengen, Formulare gestalten. Mainz: Hermann Schmidt, 2007.

[5] K. A. Schriver, Dynamics in Document Design: Creating Text for Readers. Chichester: John Wiley & Sons, 1897.

[6] A. Dillon, Designing usable electronic text: Ergonomic Aspects of Human Information Usage, 2nd ed. Boca Raton, FL: CRC Press, 2004.

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

[8] F. Richardau, La lisibilité: Langage, typographie, signes... lecture. Paris: Gauthier, 1969.

[9] P. Cleveland, “How much visual power can a magazine take?” Design Studies, vol. 26, no. 3, p. 271–317, 2005.

[10] J. Wagemans, Ed., The Oxford Handbook of Perceptual Organization. New York, NY: Oxford University Press, 2015.

[11] J. R. Pomerantz and M. Kubovy, “Theoretical approaches to perceptual organization: Simplicity and likelihood principles,” in Handbook of Perception and Human Performance, Vol. 2: Cognitive processes and performance, K. R. Roff, L. Kaufman, and J. P. Thomas, Eds. New York, NY: John Wiley & Sons, 1986, ch. 36, p. 1–46.

[12] W. Prinz, “Quantitative versuche über die prägnanz von punktmustern,” Psychologische Forschung, vol. 29, no. 4, p. 297–359, 1966.

[13] R. Arheim, Art and Visual Perception: A Psychology of the Creative Eye. Berkeley, CA: University of California Press, 1974.

[14] A.-M. Christin, History of Writing: From hieroglyph to multimedia. Paris: Flammarion, 2004.

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

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

[17] J. Hochuli, Detail in typographie. Paris: Editions B42, 2008.

[18] URW, H2-program: Micro-typography for advanced typesetting. Hamburg: URW Software & Type GmbH, 1993.

[19] H. T. Thân, “Micro-typographic extensions to the ñEg typesetting system,” Ph.D. dissertation, Masaryk University, Brno, Oct. 2000.

[20] B. Shaw, Bernard Shaw on Modern Typograph. Cleveland, OH: The Printing Press, 1915.

[21] J. Azcel and Daroczy, On Measures of Information and Their Characterizations. New York, NY: Academic Press, 1975.

[22] C. E. Shannon, “A mathematical theory of communication,” The Bell System Technical Journal, vol. 27, no. 3, p. 379–423, 1948.

[23] A. Rényi, “On measures of entropy and information,” in Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, 41, Berkeley, CA, USA, 1961, p. 547–561.

[24] T. Leinster and M. M. W., “Maximizing diversity in biology and beyond,” Proc. IEEE 12th Intl. Conf. on Electronic Measurement & Instruments (ICEMI ’2015), Shenzhen, China, 2012, p. 917–922.

[25] M. Spindler, “On concepts of randomness in finite sequences,” Ph.D. dissertation, Swiss Federal Institute of Technology (ETHZ), Zürich, 2009.

[26] A. Carpi, “Uniform words,” Advances in Applied Mathematics, vol. 32, no. 3, p. 485–522, 2004.

[27] S. B. H. and S. M. Pincus, “Irregular arrays and randomization,” Proceedings of the National Academy of Sciences (USA), vol. 95, no. 4, p. 1363–1368, 1998.

[28] C. J. Moore, “A threshold structure metric for medical image interrogations: The 2d extension of approximate entropy,” in Proc. 20th Intl. Conf. Information Visualization (IV’2016), Lisbon, Portugal, 2016.

[29] A. L. Goldberger, C.-K. Peng, and L. A. Lipsitz, “Neurobiology of aging,” What is physiologic complexity and how does it change with aging and disease?, vol. 23, no. 1, p. 23–26, 2002.

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

[31] T. Okabe, “Biophysical optimality of the golden angle in phyllotaxis,” Scientific Reports, vol. 5, no. 15358, 2015.

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

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

[34] A. Clauset, C. R. Shalizi, and M. E. J. Newman, “Power-Law Distributions in Empirical Data,” SIAM Review, vol. 51, no. 4, p. 661–703, 2009.

[35] J. C. Russ, and I. C. Percival, “A spectral entropy method for distinguishing regular and irregular motion of hamiltonian systems,” Journal of Physics A: Mathematical and General, vol. 12, no. 11, p. 2053–2071, 1979.

[36] V. Jantti and S. Alahulahti, “Spectral entropy—what has it to do with anesthesia, and the egg?” British Journal of Anaesthesia, vol. 93, no. 1, p. 150–152, 2004.

[37] C. Nicolina, G. Forcellina, L. Minati, and A. Bifone, “Scale-resolved analysis of brain functional connectivity networks with spectral entropy,” NeuroImage, vol. 211, p. 116603, 2020.

[38] V. Zakeri, “Automatic identification of hard and soft bone tissues by spectral entropy,” in Proc. ICT 2013, Thuckalay, Tamil Nadu, India, 2013, p. 917–922.

[39] A. Anier, T. Lipping, R. Ferenets, P. Puumala, E. Sonkajärvi, I. Rätsep, V. Zackeri, “Automatic identification of hard and soft bone tissues by spectral entropy,” in Proc. 20th Intl. Conf. Information Visualization (IV’2016), Lisbon, Portugal, 2016.

[40] A. Clauset and M. E. J. Newman, “Power-Law Distributions in Empirical Data,” SIAM Review, vol. 51, no. 4, p. 661–703, 2009.

[41] J. C. Russ, “Fractal Surfaces,” New York, NY: Springer, 1992.

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

[43] T. Okabe, “Biophysical optimality of the golden angle in phyllotaxis,” Scientific Reports, vol. 5, no. 15358, 2015.

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

[45] H.-O. Peitgen and D. Saupe, The Science of Fractal Images. New York, NY: Springer, 1992.

[46] A. Clauset, C. R. Shalizi, and M. E. J. Newman, “Power-Law Distributions in Empirical Data,” SIAM Review, vol. 51, no. 4, p. 661–703, 2009.

[47] J. C. Russ, “Fractal Surfaces,” New York, NY: Springer, 1992.

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

[49] T. Okabe, “Biophysical optimality of the golden angle in phyllotaxis,” Scientific Reports, vol. 5, no. 15358, 2015.

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

[51] H.-O. Peitgen and D. Saupe, The Science of Fractal Images. New York, NY: Springer, 1992.

[52] A. Clauset, C. R. Shalizi, and M. E. J. Newman, “Power-Law Distributions in Empirical Data,” SIAM Review, vol. 51, no. 4, p. 661–703, 2009.

[53] J. C. Russ, “Fractal Surfaces,” New York, NY: Springer, 1992.

[54] A. Clauset, C. R. Shalizi, and M. E. J. Newman, “Power-Law Distributions in Empirical Data,” SIAM Review, vol. 51, no. 4, p. 661–703, 2009.
[115] N. Otsu, “A Threshold Selection Method from Gray-Level Histograms,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, p. 62–66, 1979.

[116] D. Bradley and G. Roth, “Adapting Thresholding Using the Integral Image,” *J. of Graphics Tools*, vol. 12, no. 2, p. 13–21, 2007.

[117] M. D. Fairchild, *Color Appearance Models*, 3rd ed. Chichester: John Wiley & Sons, 2013.

[118] F. Attneave, “Stochastic composition processes,” *The Journal of Aesthetics and Art Criticism*, vol. 17, no. 4, p. 503–510, 1959.

[119] Wikipedia contributors. (2021) Wiener–Khinchin theorem. (2021.03.18). [Online]. Available: https://en.wikipedia.org/wiki/Wiener%28E2%80%93Khinchin_theorem

[120] G. Jewell and M. M. E., “Pseudoneglect: a review and meta-analysis of performance factors in line bisection tasks,” *Neuropsychologia*, vol. 38, no. 1, pp. 93–110, 2000.

[121] H. Cartier-Bresson, *L’imaginaire d’après nature*. Paris: Fata Morgana, 1996.

[122] B. Wolf, M. Jung, A. Fischer, G. Waerber, and V. Atanasiu, “Analyse “Automatisierte Integration des alten Sachkatalogs in Alma”,” HES-SO/Fribourg, iCoSys, Fribourg, Switzerland, Tech. Rep., Dec. 2020.

[123] Wikipedia contributors. Tesseract (software). (2021.01.31). [Online]. Available: https://en.wikipedia.org/wiki/Tesseract_(software)

[124] Google. Vision AI. (2021.01.31). [Online]. Available: https://cloud.google.com/vision/

[125] D. L. Wang and J. S. Lim, “The unimportance of phase in speech enhancement,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-30, no. 4, p. 619–681, 1982.

[126] A. V. Oppenheim and J. S. Lim, “The importance of phase in speech enhancement,” *Speech Communication*, vol. 53, no. 4, p. 465–494, 2011.

[127] M. H. Hayes, “The reconstruction of a multidimensional sequence from the phase or magnitude of its fourier transform,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-30, no. 2, p. 140–154, 1982.

[128] A. V. Oppenheim and J. S. Lim, “The importance of phase in signals,” *Proc. of the IEEE*, vol. 69, no. 5, p. 529–541, 1981.

[129] B. C. Hansen and H. R. F., “Structural sparseness and spatial phase alignment in natural scenes,” *J. of the Optical Society of America*, vol. 24, no. 7, p. 1873–1885, 2007.

[130] P. Kovess, “Image features from phase congruency,” *Videre: Journal of Computer Vision Research*, vol. 1, no. 3, p. 2–26, 1999.

[131] A. W. Lohmann, D. Mendlovic, and G. Shabtai, “Significance of phase and amplitude in the fourier domain,” *J. of the Optical Society of America*, vol. 14, no. 11, p. 2901–2904, 1997.

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

[133] K. Perlin, “An image synthesizer,” *ACM SIGGRAPH Computer Graphics*, vol. 19, no. 3, p. 287–296, 1985.

---

**VLAD ATANASIU** is a doctoral candidate in the Informatics Department of the University of Fribourg, Switzerland. His research topic is document analysis and visualization. He passed the student admission examinations at the Faculty of Telecommunications of the Polytechnic University, Timisoara, Romania in 1989. He received a B.A. degree in Middle Eastern Studies and a M.A. degree in Arabic Linguistics from the University of Provence, Aix-en-Provence, France, in 1994 and 1995, respectively, and a Ph.D. degree in Arabic Paleography from the École pratique des Hautes Études, Paris, France, in 2003.

Between 2003 and 2005, he was Postdoctoral Fellow in the Architecture Department and the Brain and Cognitive Science Department at Massachusetts Institute of Technology, Cambridge, MA, USA. From 2006 to 2009, he was at the Visualization Commission of the 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 Télécom-ParisTech, Paris, France, and the back- and front-end of an online cartography atlas of Iran for the French National Centre for Scientific Research, Paris, France. He taught information visualization at the University of Graz, Austria, and the University of Fribourg, Switzerland, from 2011 to 2015. Between 2019 and 2021, he was with the University of Basel, Switzerland, where he created software for the legibility enhancement of ancient papyri. 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).