Understanding visual complexity with statistical physics
Rensselaer NSF REU final report

Alex Striff
Department of Physics, Reed College, Portland OR 97202, USA

Vincent Meunier
Department of Physics, Applied Physics, and Astronomy, Rensselaer Polytechnic Institute, Troy NY 12180, USA

(Dated: August 7, 2020)

An image of natural scene may be described in a few words, yet an image of static requires describing all of the pixel values. We call this property of images visual complexity. Different perspectives for understanding why natural images have low visual complexity have wide applications in fields like computer vision and image compression. This REU project reviews conventional approaches to visual complexity related to the field of natural image statistics, and proposes two new perspectives on visual complexity which do not require such statistics: a notion of information for an statistical ensemble created by modifying an image, and a notion of how informative different coordinate systems are for describing points from a probability distribution. Taken together, these different perspectives form a framework for understanding visual complexity in a variety of images.

CONTENTS

I. INTRODUCTION

Why is it that thousands of pixels and millions of possible colors are required to represent images from the natural world on computers, yet that the scene may be reasonably described as a landscape of a lake at sunset in only these few words? In more detail, an image is worth a thousand words, yet the file size may be millions of binary machine words. We call this phenomenon *visual complexity*. We seek systematic ways to understand the information content or complexity in images. How are the handful of salient objects present in a scene related to all of the pixel values captured by a camera, or to the signals in the brain generated from the eyes? Different notions of visual complexity could aid progress in fields like computer vision and image compression. For example, if a computer could render an image from something like a human language description, file sizes could be much smaller and computer vision tasks like face recognition or quality control could be improved.

We first define what we mean by information, complexity, and entropy in Sec. II. With this backdrop in place, we then look at different perspectives on visual complexity.

We identify a conventional approach to visual complexity in the literature. This heap of ideas may be roughly separated into three interconnected perspectives. We start with the overarching *natural perspective* in Sec. III, which looks at the statistics of images from the natural world. The *psychophysical perspective* considers the human visual system in Sec. IV. Finally, we look past data on lattices like the fovea or sets of pixels, and instead hold the processes that create the data to be responsible for their complexity. The *algorithmic perspective* of Sec. V considers programs as explanations of complexity, with commonly-used computer graphics techniques serving as upper bounds on the true algorithmic complexity.

We propose a new perspective on the information content of an image modification that is inspired by statistical physics. To this end, we propose an ensemble of images that are differ-

ent from a given image in Sec. VI. This *ensemble perspective* considers the information to describe the modification to be much like the information to specify a microstate in statistical physics. The natural and ensemble perspectives usually work with grayscale images. The complications associated with considering color arise from the choice of coordinates for the color space. This is addressed by the *coordinate perspective* that we propose in Sec. VII. All of these perspectives contribute to defining the notion of visual complexity.

II. INFORMATION, ENTROPY, AND COMPLEXITY

We generally use these terms as synonyms for the same kind of idea, but with different connotations from information theory, statistical physics, and computer science, respectively.

A. Information Theory

A mathematical notion of information is provided by the work of Shannon [1]. In the view of classical information theory, *information* is a property of an event, in the sense of a random process. To this end, we consider a random variable X with support $\mathfrak{X}$ and probabilities $p(x)$ for $x \in \mathfrak{X}$. As regards communication, the information $I(x)$ required to describe the event $X = x$ should satisfy intuitive axioms:

- If $p(x) = 1$, the event is certain to happen and no information is required to describe its occurrence: $I(x) = 0$.
- If $p(x) < p(x')$, then x is less likely to happen, and ought to require more information to describe: $I(x) > I(x') \geqslant 0$. As an analogy, compare the phrases "nightly" and "once in a full moon:" The less probable event has a longer description.
- For independent events x and y, it makes no difference to describe the combined event (x, y) instead of each event individually: $I(x, y) = I(x) + I(y)$.

Given these axioms, the only solution is the *self-information* (Theorem A.1)

$$I(x) = -\log p(x), \tag{1}$$

where the base of the logarithm determines the units of information: base two (lg) gives *bits* and base e (ln) gives *nats*. The information of the entire random variable may then be defined as the average of (1) over all events, which is known as the *Shannon entropy*

$$H = -\sum_{x\in\mathfrak{X}} p(x) \log p(x). \tag{2}$$

The Shannon entropy may also be derived from intuitive axioms similar to those for the self information [1, 2]. The continuous version of (2) is known as the *differential entropy*

$$h = -\int_{\mathfrak{X}} p(x) \log p(x)\, \mathrm{d}x\,, \tag{3}$$

which is insufficient as a notion of information because it may change with different coordinates. Instead, we consider the *relative entropy* or *Kullback-Leibler divergence* from a reference distribution q to p defined by

$$D_{KL}(p \parallel q) = \int_{\mathfrak{X}} p(x) \log \frac{p(x)}{q(x)}\, \mathrm{d}x\,, \tag{4}$$

which is invariant under parameter transformations and is nonnegative [3, p. 243].

B. The maximum entropy principle (MaxEnt)

A physicist familiar with statistical mechanics might wonder why Shannon's entropy (2) has the same mathematical form as the thermodynamic state variable for temperature

$$S = -k_B \sum_{x\in\mathfrak{X}} p(x) \ln p(x),$$

which we may call the *Gibbs entropy*. This connection between information theory and statistical physics was developed by E. T. Jaynes to produce the maximum entropy principle (MaxEnt) [2]. We would like to make predictions about systems given some macroscopic quantities that we observe. To do so, we must assign probabilities to microstates, which we ought to do in an unbiased way, subject to the constraints that average macroscopic quantities take their observed values. Jaynes argues that this unbiased assignment corresponds to maximizing the entropy, and describes how this subjective assignment can be expected to make physical predictions, while an objective assignment of probabilities is required to understand the microscopic mechanisms behind these predictions. In particular, maximizing the entropy with constrained average energy produces the canonical distribution [2]

$$p(x) = \frac{1}{Z} e^{-\beta E(x)},$$

where $\beta = 1/k_B T$ and the *partition function* is

$$Z = \sum_{x\in\mathfrak{X}} e^{-\beta E(x)},$$

with the variates x being different states of a system.

C. Algorithmic entropy

A more general notion of entropy is found in an algorithmic approach. We say a *computer* takes a finite binary string $p \in \{0, 1\}^*$ called a *program*, and either produces a finite string as output or produces an infinite string and does not halt.

Definition. Given an object x (representable as a binary string), its *Kolmogorov complexity* $K(x)$ is defined as the length of the shortest program that outputs the object when executed on a computer.

The canonical choice of computer is a universal Turing machine $\mathcal{U}$, but the complexity from different computer $\mathcal{A}$ (like a Python interpreter) differs by only a constant: $K_{\mathcal{U}}(x) \leqslant K_{\mathcal{A}}(x) + c_{\mathcal{A}}$ [3, p. 467]. That is, the universal computer $\mathcal{U}$ may compute x by simulating $\mathcal{A}$. For long strings x, the constant is insignificant. We may also define the *conditional Kolmogorov complexity* $K(x \mid l(x))$, where the computer already knows the length of x.

The generality of Kolmogorov complexity is apparent when we consider stochastic processes. Similar to in Sec. II A, consider a stochastic process $\{X_i\}$ with the X_i drawn IID from the finite set $\mathcal{X}$ with PMF $p(x)$ for $x \in \mathcal{X}$. Then [3, p. 473]

$$H(X) = \lim_{n\to\infty} \frac{1}{n} \langle K(X^n \mid n) \rangle .$$

The length of an optimal compression program approaches the entropy limit. In this sense, the Kolmogorov complexity is a generalization of entropy.

In practice, we do not use minimum-length programs or know Kolmogorov complexities, since in general they are not computable [3, p. 482]. However, they are useful in applications of Occam's razor. To avoid multiplying our explanations beyond necessity, we choose the least complex explanation that is correct.

III. NATURAL IMAGE STATISTICS

Images of the natural world are easily distinguishable from man-made images or random noise. This difference is reflected in the statistical structure of natural images, and is generally robust across variations in subject matter or lighting. For example, the power spectrum of a natural image (averaged over orientations), scales like $k^{-2+\eta}$, where k is the spatial frequency and η is small. This and the other statistics discussed in [4] establish that natural images are approximately *scale-invariant*. The scale invariance of natural images is mostly explained by occlusion [5]. Different objects appear to be different sizes and occlude one another, causing an image to depict multiple scales. However, the scale invariance of a natural image may also reflect the scale invariance of objects it depicts: nature is full of fractals [6].

IV. HUMAN VISION

The psychophysical perspective on visual complexity focuses on the light entering the eye and the various responses generated in different regions of the brain.

A. The light field

We may describe the light field from a scene impinging upon the eye with the *plenoptic function* L. This function gives the spectral radiance of light with wavelength λ along a ray specified by spherical angles θ and φ at a given point $(\boldsymbol{x}, t)$ in space and time. All of the information about how a scene looks is contained in the function $L(\boldsymbol{x}, t, \theta, \varphi, \lambda)$. The creation of *light field cameras* to capture the whole light field that would enter the eye is a topic of active research [7]. While our eyes and conventional cameras use many photosensitive elements, significant capture of the light field may be done with only one intensity-sensitive element. This is simple enough that the author previously created a proof of concept *single-pixel camera* for an undergraduate laboratory project [8, 9]. Such a *single-pixel camera* and advances in light-field cameras are made feasible by a *compressive sensing* technique which leverages the power-law statistics of natural images discussed in Sec. III [10].

B. The visual system

The early stages of processing in the visual system recognize local properties like color, edge orientation, motion, and binocular disparity. These properties arise naturally from the plenoptic function, as discussed in [11]. More precisely, one may show simple images to an animal and probe the responses of neurons in affected regions of the brain, like the lateral geniculate nucleus (LGN) and primary visual cortex (V1). By varying the stimulus, one may develop a map of how the neuron responds as a function of position in θ and φ. This map, called the *receptive field* of the neuron, is generally set up to capture an aspect of the plenoptic function. For example, the receptive fields of *simple cells* in V1 are similar to Gabor filters, and are used to detect edges with a particular orientation [11]. Van Hateren and Ruderman have shown that the independent components of natural image sequences give similarly shaped filters, supporting the hypothesis that the simple cell receptive fields are tuned to encode natural images well [12].

V. ALGORITHMIC COMPLEXITY OF NATURAL IMAGES

The computer renderings of natural scenes that are widely used in media can be convincing to the eye without needing to simulate or model the scene at a physical level of detail. The low algorithmic complexity of a rendering reflects the low algorithmic complexity of aspects of a natural scene. In computer graphics, one usually manipulates the *geometry* of objects and the *textures* drawn on the geometry. Both of these aspects of a model can have low complexities.

A. Geometry

While most geometry in a scene is manually modeled, some natural objects like trees can be generated algorithmically. The first comprehensive approach was through the *Lindenmayer systems* studied in *The Algorithmic Beauty of Plants* [13]. This and other techniques are used in the commercial *SpeedTree* software package, which is widely used for movies and video games [14]. Another technique is to use naturalistic power-law noise to generate terrain and other

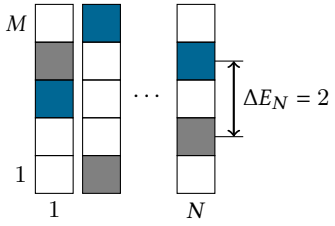

FIG. 1. The energy difference between base image pixel values (■) and modified image pixel values (■).

structures. The most common implementation is to combine band-limited noise at different scales, where the noise is either Perlin or simplex noise [15].

B. Texture

Realistic textures may also be synthesized from noise, and further improvement in realism is found from applying nonlinear transformations and turbulence to a noise texture [15, 16]. Repeating patterns like tile or weaves are also simple to generate.

C. Fractals: between geometry and texture

More aspects of the natural world may be described by fractals. A more detailed account of these ideas and how they relate to natural images may be found in *The Science of Fractal Images* [17]. An application of these ideas is the use of *iterated function systems* (IFSes) to compress natural images by taking advantage of their self-similarity [17, p. 228]. This technique can achieve compression ratios that are superior to those of common formats like JPEG, at the cost of being asymmetric. Compression by finding an IFS representation can take a long time, but the decompression by simulating the IFS is fast.

VI. AN ENSEMBLE OF IMAGES

We now neglect the spatial structure of natural image statistics and consider digital images as collections of pixels. We would like to know the information needed to specify a modification or fluctuation of the pixel values, given that the magnitude of the modification is approximately known.

A. Theory

Given a base image A with N pixels which take integer gray values $1 \leqslant a_i \leqslant M$, we define the *energy* of a different image B with gray values $1 \leqslant b_i \leqslant M$ as

$$E_A(B) = \sum_{i=1}^{N} |a_i - b_i|, \tag{5}$$

as depicted in Fig. 1.

We would like to consider all possible modified images, but focus on images with a typical value for the energy which indicates the size of fluctuations we are considering. We do this by assigning a probability distribution to the images with constrained average energy. Given the results of Sec. II B, we choose the MaxEnt distribution, which we may consider as a canonical ensemble. By thinking of our images as a physical system, we may apply tools from statistical mechanics. We would like to know the entropy of the MaxEnt distribution, which we will compute with the partition function as

$$S/k_B = \beta E + \ln Z. \tag{6}$$

In turn, we obtain the partition function

$$Z = \sum_{E \in E(\mathfrak{X})} g(E) e^{-\beta E}$$

from the number of states $g(E)$ with energy E (the *density of states*). For the case where the base image is all black ($a_i = 1$) or all white ($a_i = M$), we may explicitly count that the density of states is (Theorem A.2)

$$g(E;\, N,\, M) = \sum_k (-1)^k \binom{N}{k} \binom{N + E - Mk - 1}{E - Mk}.$$

However, the situation for general grayscale images becomes more complicated. For this reason and the ability to analyze more complex systems, we determine the density of states numerically using the Wang-Landau algorithm [18].

B. Methods

The Wang-Landau algorithm (WL) was implemented to determine the density of states for grayscale image fluctuations. Our implementation adapts the algorithm described by Wang, Landau, et al. in [18, 19] for lattice models to work on the image fluctuation model we have described (Appendix B). The offset of the log density of states was set by ensuring that the number of states from $\sum_E g(E)$ gave the total number of states M^N. We then computed the entropy from the numerical density of states with (6).

C. Results

The log density of states from WL for a black image is given in Fig. 2. Since this is indistinguishable from the exact result, we quantify the error by running 1024 simulations for a black image with the same parameters for histogram flatness and f tolerance as in [19]. The resulting relative errors are given in Fig. 3. This relative error is consistent with that in [19] for a similarly-sized 2D Ising ferromagnet, which establishes that the implementation of the algorithm is correct and has the expected error characteristics.

We now consider the desired calculation of the entropy for random grayscale images. The WL densities of states for

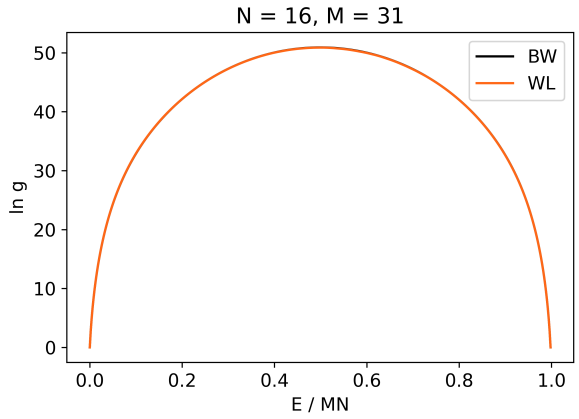

FIG. 2. The log density of states for a black image from the Wang-Landau algorithm (WL), compared to the exact result (BW). The two densities of states are indistinguishable.

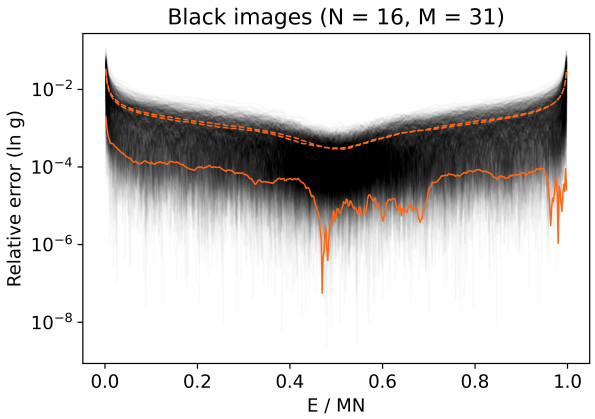

FIG. 3. The relative error in the log density of states for 1024 black image Wang-Landau simulations. The mean density of states is indicated in orange and the composite densities of states one standard deviation away are dashed.

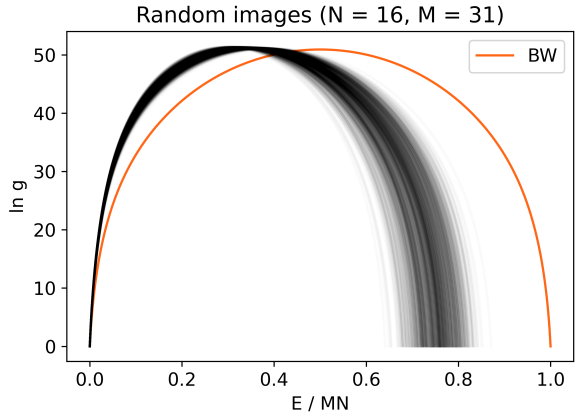

FIG. 4. The log density of states for 1024 random grayscale image Wang-Landau simulations. The black image result is provided as reference in orange (BW).

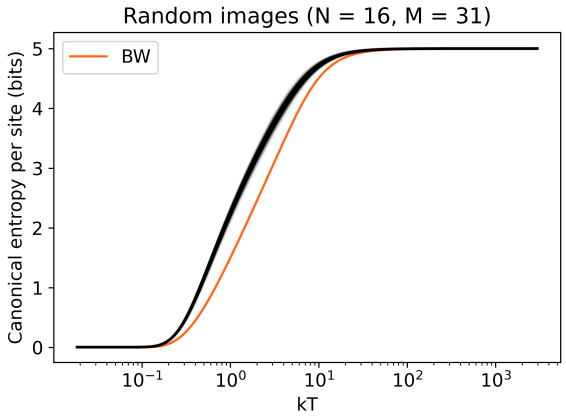

FIG. 5. The canonical entropy computed from the simulation density of states for 1024 random grayscale images. The entropy from the exact result for a black image is shown in orange (BW).

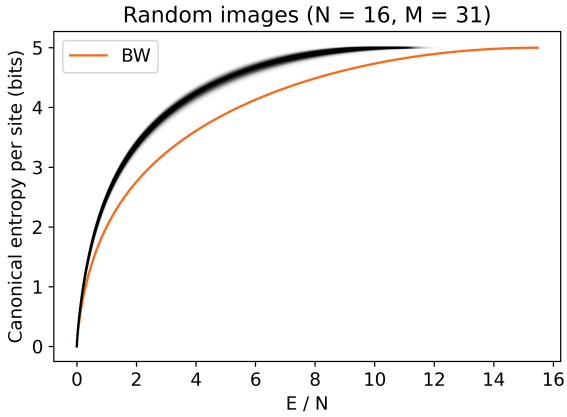

FIG. 6. The canonical entropy for grayscale images increases quickly with energy before saturating at the maximum. The entropy from the exact result for a black image is shown in orange (BW).

1024 random grayscale images is given in Fig. 4. The corresponding entropies (Fig. 5) represent the information that we seek. The entropies for different grayscale images are similar, since the local energy landscapes for gray pixels are close to the same. The entropy for a black image is lower than for a grayscale image by almost 1 bit, which reflects that the black pixel values may only fluctuate up in value, rather than in both directions for a gray pixel value. We may also see how the entropy depends upon the average energy, which is the quantity we originally considered (Fig. 6). As expected, this tends towards the log density of states as the canonical and microcanonical ensembles coincide in the thermodynamic limit of large N and M.

D. Discussion

Now that we have computed the modification information as the entropy of a fluctuating grayscale image, what can

we learn from it? If we consider the fluctuations as noise, we need a psychophysical threshold of the average energy E where the difference between images is barely noticeable. This is possible, but requires adjustment of the metric of energy to suit the Weber-Fechner law [20]. This was not done, since using the absolute deviation (5) as the energy is simpler to interpret. We may at least regard the present results as an approximate measure of the information lost in visual perception.

A limitation of this measure is the use of a digital image instead of considering the light impinging on the retina as discussed in Sec. IV. The pixels in an image may be considered as averages of the true continuous intensity over a small solid angle. Considering a static grid of pixels in a digital image differs from our foveated imaging process, which features multiple saccades around a scene to refine points of interest. This aspect of the issue has previously been addressed by another maximum entropy method which finds optimally informative fixations based on contrast [21].

Whether the fluctuations are due to noise or varying the scene, the density of states found in Fig. 4 may be used to determine the distribution of fluctuating values at a single pixel. As indicated by Fig. 6, even a tiny 4×4 image has a canonical entropy similar to the log density of states. Any typical image with thousands of pixels and hundreds of gray values may be considered in the thermodynamic limit. In this limit, the distribution of energy for the single pixel j may be evaluated in the microcanonical ensemble as

$$P(E_j \mid E) = \frac{g(E - E_j;\, N-1,\, M)}{g(E;\, N,\, M)} \approx \frac{g(E - E_j;\, N,\, M)}{g(E;\, N,\, M)}.$$

For small fluctuations, E is small and $E_j \leqslant E$ lies on the initial linear region of the log density of states, no matter the other gray values. Thus, small fluctuations give rise to an exponential distribution of energies, which means a Laplacian distribution of gray pixel values. As the total energy E grows towards the maximum density ($T \to \infty$), the Laplacian peak of pixel values flattens out and increases in entropy.

The final issue is the arbitrary number of values M for the number of gray values, which affects the entropy value. The simplest solution is to instead consider the intensive quantity $S/\lg M$, but it is unclear how to best generalize this to the case of color, where discretizations of different color spaces may produce different results. This problem of color is avoided in the case of scotopic (night) vision.

VII. COLOR AND CHOICE OF COORDINATES

The ensemble perspective on images differs from the entropy of statistical physics in the continuous case. The conjugate nature of position and momentum in phase space make the thermodynamic differential entropy (3) invariant under a canonical coordinate transformation, but this is not so for color spaces [22].

A. Color spaces

Colors may be represented independently from how they are produced with the various *CIE color spaces* [23]. These representations are ultimately derived from the spectral sensitivities of the different human *cone cells* which detect colors. The endeavors of the CIE (International Commission on Illumination) have resulted in perceptually-uniform color spaces like the CIELAB and CIELUV color spaces. Instead of representing colors as (scaled) RGB values that control hardware pixels, the CIE color spaces use different parameters like the lightness, chromaticity, and hue coordinates of the CIELCH space (the cylindrical version of CIELUV). These spaces aim to be scaled so that the Euclidean distance between the coordinates of two colors captures the perceptual difference between them. A small enough ball around a color will contain all colors that are perceptually indistinguishable in the sense of the just-noticeable difference [20]. This is the kind of perceptual energy metric that we would like to use in the ensemble perspective, but we are confronted with the problem of enforcing an arbitrary discretization of the color space.

B. Choice of coordinates

Instead of choosing an arbitrary discretization, we propose a relative notion of information from coordinates which may guide the choice of discretization. As a motivating example, consider drawing a dot on a ruler. To guess where the dot is, one would rather know what the ruler reads at the dot location, rather than the position of the dot along the short side of the ruler. We aim to capture the sense in which the short side is less informative than the ruled side.

Consider coordinates $x_1, \ldots, x_m$ for $\mathbb{R}^m$. An example would be the spherical coordinates r, θ, φ for $\mathbb{R}^3$. In terms of probability distributions, the dot is a distribution which places most of the probability near a single location. In general, we consider relative coordinate information for locating points by an arbitrary *target* distribution p on the coordinates. The usual perspective of independent coordinates as being equally important is taken as a reference, which we describe with a uniform distribution p_u.

Definition. The *relative information of coordinate i* about a target distribution p is defined as

$$I(x_i) = \frac{D_{KL}(p(x_i) \parallel p_u(x_i))}{D_{KL}(p(x_1, \ldots, x_m) \parallel p_u(x_1, \ldots, x_m))}. \tag{7}$$

We give two simple examples to demonstrate the calculation of the relative coordinate information.

1. *Which edge of a ruler is best?*

The different sides of a ruler define Cartesian coordinates on a rectangle. Let the short side have length a and the ruled side have length b. The target distribution is a uniform distribution on square of side length $c \leqslant a$ within the ruler. With

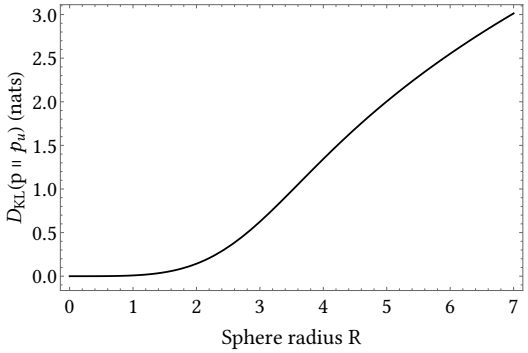

FIG. 7. The KL divergence from uniform to truncated normal as a function of ball radius.

the coordinates $0 \leqslant x \leqslant a$ and $0 \leqslant y \leqslant b$, the uniform reference distribution is $p_u(x, y) = (ab)^{-1}$ and the target distribution is $p(x, y) = c^{-2}[0 \leqslant x \leqslant c,\ 0 \leqslant y \leqslant c]$, where $[P]$ is 1 if P is true and 0 otherwise. From (4) and (7) we compute that

$$I(x) = \frac{\log(a/c)}{\log(a/c) + \log(b/c)} \quad \text{and}$$
$$I(y) = \frac{\log(b/c)}{\log(a/c) + \log(b/c)}.$$

Since the coordinates are independent, $I(x) + I(y) = 1$. We check that the extreme cases are intuitive. For $c = a$, $I(x) = 0$ and $I(y) = 1$, which is that all the information is held by the coordinate y. This is expected since the x-coordinate is irrelevant for locating a point when $c = a$. As $c \to 0$, $I(x)$ and $I(y)$ tend to $1/2$. In this limit, we see that both coordinates become equally informative for locating a single point.

2. The sphere

We may also consider relative coordinate information for coordinate systems with nontrivial Jacobians, like spherical coordinates. That is, the uniform distribution on a sphere of radius R in Cartesian coordinates is

$$p_u(x, y, z) = \frac{3}{4\pi R^3},$$

but on spherical coordinates pulls back to

$$p_u(r, \theta, \varphi) = \frac{3r^2 \sin\theta}{4\pi R^3}.$$

The spherical coordinates form independent random variables with marginal distributions

$$p_u(r) = \frac{3r^2}{R^3}, \quad p_u(\theta) = \frac{\sin\theta}{2}, \quad \text{and} \quad p_u(\varphi) = \frac{1}{2\pi},$$

whereas the Cartesian coordinates form dependent random variables with conditional distributions

$$p_u(x_i \mid x_{j\neq i}) = \frac{1}{2}\left(R^2 - \sum_{j\neq i} x_j^2\right)^{-1/2} \qquad (i, j = 1, 2, 3).$$

Similarly, the truncated normal distribution with unit variance on the ball pulls back in spherical coordinates to

$$p(r, \theta, \varphi) = \left[\operatorname{erf}\left(\frac{R}{\sqrt{2}}\right) - R\sqrt{\frac{2}{\pi}}e^{-R^2/2}\right]^{-1} \frac{e^{-r^2/2}}{\sqrt{8\pi^3}}\, r^2 \sin\theta.$$

We take the truncated normal as our target distribution. While $p(\theta) = p_u(\theta)$ and $p(\varphi) = p_u(\varphi)$, the marginal distribution of the radius is

$$p(r) = \left[\operatorname{erf}\left(\frac{R}{\sqrt{2}}\right) - R\sqrt{\frac{2}{\pi}}e^{-R^2/2}\right]^{-1} \frac{e^{-r^2/2}}{\sqrt{2\pi}}\, 2r^2.$$

The KL divergence from the uniform to the truncated normal distribution over the ball is

$$D_{KL}(p \parallel p_u) = \int_{r=0}^{R}\int_{\theta=0}^{\pi}\int_{\varphi=0}^{2\pi} p \ln\frac{p}{p_u}\, \mathrm{d}r\, \mathrm{d}\theta\, \mathrm{d}\varphi\,,$$

independent of using spherical coordinates (Fig. 7). But since the spherical coordinates form independent random variables for both distributions, this separates as

$$\begin{aligned} D_{KL}(p(r, \theta, \varphi) \parallel p_u(r, \theta, \varphi)) &= D_{KL}(p(r) \parallel p_u(r)) \\ &+ D_{KL}(p(\theta) \parallel p_u(\theta)) \\ &+ D_{KL}(p(\varphi) \parallel p_u(\varphi)). \end{aligned}$$

As only the radial distribution is different,

$$D_{KL}(p(r, \theta, \varphi) \parallel p_u(r, \theta, \varphi)) = D_{KL}(p(r) \parallel p_u(r)),$$

as expected from the symmetries. Thus $I(r) = 1$ and $I(\theta) = I(\varphi) = 0$. This is consistent with the intuition that the radius alone controls the probability of a point. Put negatively, knowing angles θ and φ for a point gives no information about if it is probable.

3. Relevance to color spaces

We have now set up the machinery to compare different coordinates on color spaces. Future work could assess the probability distribution on a color space from natural images, as in [24], or consider the visible gamut of colors as a uniform distribution over a subset of a CIE color space. A computation of relative coordinate information for a color space with one of these two target distributions would indicate which coordinates of the color space are perceptually important. For example, we might expect that luminance would be important, given the importance of identifying edge contrast in vision. The target distributions could also be used to find an optimal entropy coding which would reduce the information needed to express natural colors.

VIII. CONCLUSION

The idea of visual complexity has been explored through conventional perspectives rooted in natural image statistics,

as well as through the new perspectives from an image ensemble and relative coordinate information. These notions of visual complexity have varied applications to disciplines like computer vision, compressive sensing, image compression, and computer graphics. Taken together, the different perspectives that we have discussed model much of the visual complexity in the natural world in much the same way that the simple harmonic oscillator acts as a model for natural phenomena like bobbing in water or the swaying of trees.

While we have largely considered images as entities in their own right, it should be noted that their visual complexity is rooted in the complex phenomena of the natural world. Many of the ideas covered here are relevant to understanding the complexity of physical laws as well as vision. Thought in physics along these lines may be found in proceedings of the Santa Fe Institute [25] or in P. W. Anderson's maxim that "more is different" [26]. We may only hope that the enrichment to applications in vision embodied by the success of image compression algorithms and computer rendering due to understanding visual complexity has an analog in physics and other natural sciences.

IX. ACKNOWLEDGEMENTS

Most of all, I thank Prof. Vincent Meunier for proposing and mentoring the project. Our near-daily meetings became highlights of the long days spent inside during the summer of COVID-19. I also thank Profs. Peter Persans and K. V. Lakshmi for all their effort in coordinating an unusual physics REU. Funding for this project was provided as a part of the physics research experience for undergraduates program of the National Science Foundation at Rensselaer Polytechnic Institute.

Appendix A: Theorems

Theorem A.1. *The only twice continuously differentiable function $I(x)$ that satisfies the axioms in Sec. II A is the self-information $I(x) = -\log p(x)$.*

Proof. Consider independent events x and y with probabilities p and p'. The axioms only concern the probabilities of the events, so we may express the information as $I(x) = \tilde{I}(p(x))$. Then as proposed,

$$I(x, y) = \tilde{I}(pp') = \tilde{I}(p) + \tilde{I}(p')$$

by independence. Taking the partial derivative with respect to p gives

$$p'\tilde{I}'(pp') = \tilde{I}'(p),$$

and then taking the partial derivative with respect to p' gives

$$\tilde{I}'(pp') + pp'\tilde{I}''(pp') = 0.$$

We may then define $q = pp'$ to obtain the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}q}\left(q\tilde{I}'(q)\right) = 0,$$

which has solution

$$\tilde{I}(q) = k \log q$$

for real k. The condition that $\tilde{I}(q) \geq 0$ requires $k > 0$, which is equivalent to a choice of base for the logarithm. □

Theorem A.2. *The number of tuples $(a_1, \ldots, a_N)$ with $0 \leqslant a_i \leqslant M-1$ and $\sum_i a_i = E$ is*

$$g(E) = \sum_k (-1)^k \binom{N}{k} \binom{N + E - Mk - 1}{E - Mk}.$$

Proof. Ordinary generating functions provide the solution [27]. We represent the sum E as the exponent of a integer polynomial in z in the following way. For the tuple (x_1, x_2), we represent x_1 as z^{x_1} and x_2 as z^{x_2}. Together, we have $z^{x_1} z^{x_2}$, which gives the monomial $z^{x_1+x_2} = z^E$ for this tuple. We may then find $g(E)$ as the coefficient of z^E in

$$\left(1 + \cdots + z^{M-1}\right)^N.$$

Expanding using the binomial theorem gives

$$\begin{aligned}\left(\frac{1 - z^M}{1 - z}\right)^N &= \sum_{k=0}^{N} (-1)^k \binom{N}{k} z^{Mk} \sum_{j=0}^{\infty} (-1)^j \binom{-N}{j} z^j \\ &= \sum_{k=0}^{N} \sum_{j=0}^{\infty} (-1)^k \binom{N}{k} \binom{N + j - 1}{j} z^{Mk+j}.\end{aligned}$$

The value of j for z to have exponent E is $j = E - Mk$, so the coefficient of z^E in the polynomial is

$$g(E) = \sum_k (-1)^k \binom{N}{k} \binom{N + E - Mk - 1}{E - Mk},$$

where the limits of summation are set by the binomial coefficients. □

Appendix B: Wang-Landau algorithm implementation

The relevant core of the Wang-Landau algorithm implementation is reproduced below. For the full code, see the REU project repository [28], which includes both the code and a notebook of all progress, including some other approaches to visual complexity than those described in this report.

```
def simulation(system, Es,
               max_sweeps = 1_000_000,
               flat_sweeps = 10_000,
               eps = 1e-8,
               logf0 = 1,
               flatness = 0.2
              ):
    """
```

```
    Run a Wang-Landau simulation on system with energy bins Es to determine
    the system density of states g(E).

    Args:
        system: The system to perform the simulation on (see systems module).
        Es: The energy bins of the system to access. May be a subset of all bins.
        max_sweeps: The scale for the maximum number of MC sweeps per f-iteration.
            The actual maximum iterations may be fewer, but approaches max_sweeps
            exponentially as the algorithm executes.
        flat_sweeps: The number of sweeps between checks for histogram flatness.
            In AJP [10.1119/1.1707017], Landau et. al. use 10_000 sweeps.
        eps: The desired tolerance in f. Wang and Landau [WL] use 1e-8 in the original
            paper [10.1103/PhysRevLett.86.2050].
        logf0: The initial value of ln(f). WL set to 1.
        flatness: The desired flatness of the histogram. WL set to 0.2 (80% flatness).

    Returns:
        A tuple of results with entries:
        Es: The energy bins the algorithm was passed.
        S: The logarithm of the density of states (microcanonical entropy).
        H: The histogram from the last f-iteration.
        converged: True if each f-iteration took fewer than the maximum sweeps.

    Raises:
        ValueError: One of the parameters was invalid.
    """
    if (max_sweeps ≤ 0
        or flat_sweeps ≤ 0
        or eps ≤ 1e-16
        or not (0 < logf0 ≤ 1)
        or not (0 ≤ flatness < 1)):
        raise ValueError('Invalid Wang-Landau parameter.')

    # Initial values
    M = max_sweeps * system.sweep_steps
    flat_iters = flat_sweeps * system.sweep_steps
    logf = 2 * logf0 # Compensate for first loop iteration
    logftol = np.log(1 + eps)
    converged = True
    steps = 0

    E0 = Es[0]
    Ef = Es[-1]
    N = len(Es) - 1
    S = np.zeros(N) # Set all initial g's to 1
    H = np.zeros(N, dtype=np.int32)
    i = binindex(Es, system.E)

    while logftol < logf:
        H[:] = 0
        logf /= 2
        iters = 0
        niters = int((M + 1) * np.exp(-logf / 2))
        while (iters % flat_iters ≠ 0 or not flat(H, flatness)) and iters < niters:
            system.propose()
            Ev = system.Ev
            j = binindex(Es, Ev)
            if E0 ≤ Ev < Ef and (
                S[j] < S[i] or np.random.rand() ≤ np.exp(S[i] - S[j])):
                system.accept()
                i = j
            H[i] += 1
            S[i] += logf
            iters += 1
        steps += iters
        if niters ≤ iters:
            converged = False

    return Es, S, H, steps, converged
```


[1] C. E. Shannon, A mathematical theory of communication, The Bell system technical journal **27**, 379 (1948).

[2] E. T. Jaynes, Information theory and statistical mechanics, Phys. Rev. **106**, 620 (1957).

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Wiley Series in Telecommunications and Signal Processing (Wiley-Interscience, 2006).

[4] D. L. Ruderman, The statistics of natural images, Network: Computation in Neural Systems **5**, 517 (1994).

[5] A. B. Lee, D. Mumford, and J. Huang, Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model, International Journal of Computer Vision **41**, 35 (2001).

[6] M. F. Barnsley, *Fractals everywhere* (Academic press, 2014).

[7] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, Compressive light field photography using overcomplete dictionaries and optimized projections, Acm Trans. Graph. **32**, 10.1145/2461912.2461914 (2013).

[8] A. Striff, jfjhh/csjlab.jl: Submitted project (2020).

[9] T. A. Kuusela, Single-pixel camera, American Journal of Physics **87**, 846 (2019).

[10] D. L. Donoho, Compressed sensing, Ieee Transactions on Information Theory **52**, 1289 (2006).

[11] E. H. Adelson, J. R. Bergen, *et al.*, *The plenoptic function and the elements of early vision*, Vol. 2 (Vision and Modeling Group, Media Laboratory, Massachusetts Institute of Technology, 1991).

[12] J. H. van Hateren and D. L. Ruderman, Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex, Proceedings of the Royal Society of London. Series B: Biological Sciences **265**, 2315 (1998).

[13] A. Lindenmayer, P. Prusinkiewicz, *et al.*, *The algorithmic beauty of plants*, Vol. 1 (New York: Springer-Verlag, 1990).

[14] C. King and M. Sechrest, Speedtree.

[15] K. Perlin, An image synthesizer, in *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*, Siggraph '85 (Association for Computing Machinery, New York, NY, USA, 1985) p. 287–296.

[16] K. Perlin and E. M. Hoffert, Hypertexture, in *Proceedings of the 16th Annual Conference on Computer Graphics and Interactive Techniques*, Siggraph '89 (Association for Computing Machinery, New York, NY, USA, 1989) p. 253–262.

[17] M. F. Barnsley, R. L. Devaney, B. B. Mandelbrot, H.-O. Peitgen, D. Saupe, R. F. Voss, Y. Fisher, and M. McGuire, *The science of fractal images* (Springer, 1988).

[18] F. Wang and D. P. Landau, Efficient, multiple-range random walk algorithm to calculate the density of states, Phys. Rev. Lett. **86**, 2050 (2001).

[19] D. P. Landau, S.-H. Tsai, and M. Exler, A new approach to monte carlo simulations in statistical physics: Wang-landau sampling, American Journal of Physics **72**, 1294 (2004).

[20] G. T. Fechner, D. H. Howes, and E. G. Boring, *Elements of psychophysics*, Vol. 1 (Holt, Rinehart and Winston New York, 1966).

[21] R. Raj, W. S. Geisler, R. A. Frazor, and A. C. Bovik, Contrast statistics for foveated visual systems: fixation selection by minimizing contrast entropy, J. Opt. Soc. Am. A **22**, 2039 (2005).

[22] V. Hnizdo and M. K. Gilson, Thermodynamic and differential entropy under a change of variables, Entropy **12**, 578 (2010).

[23] J. Schanda, *Colorimetry: understanding the CIE system* (John Wiley & Sons, 2007).

[24] M. A. Webster and J. Mollon, Adaptation and the color statistics of natural images, Vision Research **37**, 3283 (1997).

[25] W. H. Zurek, *Complexity, entropy and the physics of information* (CRC Press, 2018).

[26] P. W. Anderson, More is different, Science **177**, 393 (1972).

[27] H. S. Wilf, *Generatingfunctionology*, 3rd ed. (A. K. Peters, 2006).

[28] A. Striff, jfjhh/rpi-reu-notebook: Midterm report (2020).