Actionable Information in Vision

Stefano Soatto

Technical Report CSD090007
March 10, 2009, revised March 18, 2010

Abstract

A notion of visual information is introduced as the complexity *not* of the raw images, but of the images after the effects of nuisance factors such as viewpoint and illumination are discounted. It is rooted in ideas of J. J. Gibson, and stands in contrast to traditional information as entropy or coding length of the data regardless of its use, and regardless of the nuisance factors affecting it. The non-invertibility of nuisances such as occlusion and quantization induces an "information gap" that can only be bridged by controlling the data acquisition process. Measuring visual information entails early vision operations, tailored to the structure of the nuisances so as to be "lossless" with respect to visual decision and control tasks (as opposed to data transmission and storage tasks implicit in traditional information theory). These ideas are illustrated on visual exploration, whereby a "Shannonian Explorer" navigates unaware of the structure of the physical space surrounding it, while a "Gibsonian Explorer" is guided by the topology of the environment, despite measuring only images of it, without performing 3D reconstruction. Our operational definition of visual information suggests desirable properties that a visual representation should possess to best accomplish vision-based decision and control tasks.

1 Preamble

This paper discusses the role visual perception plays in the "signal-to-symbol barrier" problem.

The "signal-to-symbol barrier" stems from the observation that perceptual agents, from plants to humans, perform measurements of physical processes at a level of granularity that is *essentially continuous*.[1] They also perform actions in the continuum of physical space. And yet, cognitive science, primary epistemics, and in general modern philosophy, associate "intelligent behavior"

[1] The continuum is an abstraction, so here "continuous" is to be understood as existing at a level of granularity significantly finer than the resolution of the measurement device or actuator. For instance, although retinal photoreceptors are finite in number, we do not perceive discontinuities due to retinal sampling. Even when the sensory signals and the actions are discrete (e.g. due to digital encoders or transducers), the "analog-to-digital" conversion usually occurs in a manner that is independent of the signal being sampled (*e.g.* fixed-rate sampling), or dependent only on coarse phenomenological aspects of the signal (*e.g.* adaptive sampling based on frequency characteristics or sparsity constraints).

with some kind of *internal representation* consisting of discrete symbols (“concepts”, “ideas”, “objects”, “categories”) that can be manipulated with the tools of logic or probabilistic inference. But little is known about why such a “signal-to-symbol” conversion should occur, whether it would yield an evolutionary advantage, or what principles would guide such a discretization process.

Traditional Information Theory, Statistical Decision Theory, and Control Theory shed little light on this process, and indeed suggest that it may be counter-productive. If we consider biological systems as machines that perform actions or make decisions in response to stimuli in a way that maximizes some decision or control objective, then Rao and Blackwell ([64] page 88) indicate that the best possible agents would avoid “breaking down the data into pieces,” *i.e. data analysis*[2], or for that matter any kind of intermediate decision unrelated to the final task, as would instead be necessary to have a discrete internal representation.[3]

So, why would we need, or benefit from, an internal representation? Is “intelligence” not possible in an analog setting? Or is *data analysis* necessary for cognition? If so, what would be the principle that guides it?

And with respect to the academic field of Computer Vision, why have we been performing data analysis (edge detection, feature selection, segmentation, image parsing etc.) against the basic tenets of (traditional) Information and Decision Theory? The latter would suggest that, eventually, a reductionist approach where images are fed raw into a black-box decision or control machine will be most successful. Or perhaps, on the contrary, the traditional notion of Information should be revised, and this revision will point to new principles for data analysis, and validate what Computer Vision scientists have done for decades.

Yet another possibility is that data analysis is not guided by any principle, but an accident due to the constraints imposed by biological hardware, as advocated by Turing in [73], where he showed that reaction-diffusion partial differential equations (PDEs) that govern neuronal ion concentrations, although continuous in nature, exhibit discrete solutions. So, if we want to build machines that interact intelligently with their surroundings and are not bound by the constraints of biological hardware, should we draw inspiration from biology, or is it better to jettison it?

The question of representation is ill-posed outside the scope of a task. A task can be as narrow as a binary decision, such as the presence/absence of a person in a scene, or as general as “survival,” but in the context of visual perception I distinguish four broad classes of tasks, which I call the

[2] Note that I refer to *data analysis* as the process of “breaking down the data into pieces” (cfr. gr. *analyein*), *i.e.* the generally lossy conversion of data into discrete semantic entities. This is not the case for global representations such as Fourier or wavelet decomposition, or principal component analysis (PCA), that are unfortunately often referred to as “analysis” because such techniques were developed in the context of harmonic analysis, a branch of mathematics. The Fourier transform is globally invertible, which implies that there is no loss of data, and PCA consists in linear projections onto subspaces.

[3] Discretization is often advocated on complexity grounds, but complexity calls for data *compression*, not necessarily for data *analysis*. Any complexity cost could be added to the decision or control functional, and the best decision would still avoid data analysis. For instance, to simplify a segment of a radio signal one could represent it as a linear combination of a small number of (high-dimensional) bases, so few numbers (the coefficients) are sufficient to represent it in a parsimonious manner. This is different than breaking down the signal into pieces, *e.g.* partitioning its domain into subsets, as implicit in the process of encoding a visual signal through a population of neurons each with a finite receptive field. So, is there an evolutionary advantage in data analysis, beyond it being just a way to perform lossy data compression?

four “R’s” of vision: Reconstruction (building models of the geometry of the scene), Rendering (building models of the photometry of the scene), Recognition (or, more in general, vision-based decisions such as detection, localization, categorization), and Regulation (or, more in general, vision-based control such as tracking, manipulation etc.).

For Reconstruction and Rendering, I am not aware of any principle that suggests an advantage in data analysis. It is not accidental that the current best approaches to reconstruct models of the geometry and photometry of a scene from image streams recover (piecewise) continuous surfaces and radiance functions directly from the data, as opposed to the traditional multi-step pipeline[4] that was long favored on complexity grounds [52].

In this manuscript, I explore the issue of representation for decision and control tasks. I will try to avoid philosophical discourses and will not make an attempts to define “intelligent behavior” or even knowledge, other than to postulate that knowledge – whatever it is – *comes from data*, but it is *not data*. This leads to the notion of the “useful portion” of the data, which one might call “information.” So, the first step is understanding what “information” means in the context of visual perception. That is the subject of this manuscript.

What I will show is that visual perception plays a key role in understanding the signal-to-symbol barrier. Specifically, the need to be able to perform decision and control tasks in a manner that is independent of nuisance factors that affect the image formation process leads to an internal representation that is intrinsically *discrete*, and yet *lossless*, in a sense to be made clear soon. However, for this to happen the perceptual agent has to have control over certain aspects of the sensing process. This ties together inextricably sensing and control, in the sense that without the ability to control the sensing process with motion, a discrete internal representation would be a sure loss. A peculiar illustration of this phenomenon is the case of Sea Squirts, or Tunicates, shown in Fig. 1. These are organisms that possess a nervous system (ganglion cells) and the ability to move (they are predators), but eventually settle on a rock, become stationary and thence swallow their own brain.

2 Introduction

More than sixty years ago, Norbert Wiener stormed into his students’ office enunciating *“entropy is information!”* before immediately storming out. Claude Shannon later made this idea the centerpiece of his Mathematical Theory of Communication, formalizing and unifying the wide variety of methods that practitioners had been using to transmit signals through channels. The influence of Shannon’s communication theory has since spread beyond the transmission and compression of data, and is now broadly known as Information Theory. But is the entropy of the *data* really “information”? There is no doubt that the more complex the data, the more costly it is to *store* and *transmit*. But what if we want to *use* the data for *tasks* other than storage or transmission? What is the “information” that an image contains about the *scene* it portrays? What is the value of an image if we are to *recognize* objects in the scene, or *navigate* through it? (Fig. 2).

[4] A sequence of “steps” including point-feature selection, wide-baseline matching, epipolar geometry estimation, motion estimation, triangulation, epipolar rectification, dense re-matching, surface triangulation, mesh polishing, texture mapping.

Figure 1: *The* **Sea Squirt**, *or Tunicate, is an organism capable of mobility, sometimes of predatorial nature, until it finds a suitable rock to cement itself in place. Once it becomes stationary, it digests its own cerebral ganglion, or "eats its own brain" and develops a thick covering, a "tunic" for self defense.*

Despite its pervasive reach today, Shannon's notion of information had early critics,[5] among who James J. Gibson, who wrote *"My theory of the available information in ambient light is radically different from [that of] Shannon. [...] My notion is that information consists of invariants underlying change"* [29].[6] Already in the fifties he was convinced that *data* is not *information*,[7] and the value of data should depend on what one can do with it, *i.e.* the task [54]. Much of the complexity in an image is due to *nuisance factors*, such as illumination, viewpoint and clutter,[8] that

[5] Even Shannon's disciples acknowledge that "information theory is a total misnomer [...] it does not deal with information at all, it deals with data" (R. Gallagher, personal communication).

[6] It is only unfortunate that, in engineering communications, the signals are heavily structured so the nuisance, often dubbed "noise," is usually assumed to be additive, zero-mean, white, and Gaussian. As a consequence, the issue of invariance was never thoroughly explored, although, interestingly, Wiener was aware of it. In ([78], p. 50) he even introduced the first (integral) moment as an invariant statistic to a group (eq. (6.01), p. 138) and called it a *gestalt*!

[7] He was nevertheless rooted in empirical epistemology and therefore assumed that *information comes from data*.

[8] I use the word "nuisance" in the standard sense of statistical inference; this does not imply that nuisance factors are dismissed or irrelevant. It just means that they affect the data, but not the task. Gibson wrote: *"Four kinds of invariants have been postulated: those that underlie change of illumination, those that underlie change of the point of observation, those that underlie overlapping samples, and those that underlie a local disturbance of structure. [...] Invariants of optical structure under changing illumination [...] are not yet known, but they almost certainly involve ratios of intensity and color among parts of the array. [...] Invariants [...] under change of the point of observation [...] some of the changes [...] are transformations of its nested forms, but [...] The major changes are gain and loss of form, that is, increments and decrement of structures, as surfaces undergo occlusion." [...] The theory of the extracting of invariants by a visual system takes the place of theories of "constancy" in perception, that is, explanations of how an observer might perceive the true color, size, shape, motion and direction-from-here of objects despite the wildly*

have little to do with the decision (perception) and control (action) task at hand. So it is intuitive that the value of data should relate to its complexity *only after the effects of nuisance factors has been discounted.*[9] Unfortunately, any constant function is an "invariant underlying change", so Gibson was missing the other facet of information that relates to its "usefulness" (sufficiency) towards the task.

The goal of this manuscript is to define an operational notion of "information" that is relevant to visual inference tasks, as opposed to the transmission and storage of image data. Following Gibson's lead, I define *Actionable Information* to be the complexity (coding length[10]) of a maximal statistic that is invariant to the nuisances associated to a given task. According to this definition, the Actionable Information in an image depends *not just* on the complexity of the data, but also on the *structure* of the scene it portrays. I illustrate this on a simple environmental exploration task, that is central to Gibson's ecological approach to perception. A robot seeking to maximize Shannon's information (a "Shannonian Explorer") is drifting along unaware of the structure of the environment, while one seeking to maximize Actionable Information (a "Gibsonian Explorer") is driven by the topology of the surrounding space. Both measure the same *data* (images), but the second is *using it* to accomplish spatial tasks.[11] This manuscript relates to work in information theory, video compression, robotics, visual recognition, as I discuss in Sect. 7. There, I also discuss the visual representations that our operative definition implies as "lossless" relative visual decision or control tasks.

fluctuating sensory impressions on which the perceptions are based." ([30], p. 310).

[9] Appealing as the idea of characterizing "invariants under change" sounds in words, a modern Computer Vision scientist would dismiss it at the outset, for it has since become known that such invariants *do not exist*. Invariants were considered "the Holy Grail of Computer Vision" in the eighties, until [14] and [20] showed that they do not exist neither for viewpoint, nor for illumination. Lacking mathematical and computational foundations that enable scientific (falsifiable) discourse and engineering applications, Gibson's ideas thus remain confined to the realm of philosophy [39] and perception psychology. However, recent developments have shown that the situation is more complex than commonly assumed: [14] refers to statistics of *point features,* not *images,* and [74] shows instead that non-trivial viewpoint invariants always exist for images of Lambertian objects of general shape. Similarly, [20] consider general illumination fields, but invariants can be constructed for simpler illumination models, such as contrast transformations [2], even though these are valid only locally. Invariance always refers to an underlying model, that is as good as the assumptions it is based on, and as useful as the ensuing algorithms are for the task of interest. Invariance to even more restricted classes of transformations is the underpinning of very simple image statistics that have recently gained significant popularity in visual recognition and categorization tasks.

[10] The relation to entropy [48] requires defining a distribution of coding lengths, as will be discussed later in this manuscript.

[11] It could be argued that Shannon would not seek to maximize the entropy of the data, but instead the mutual information between the scene and the data (which he called "equivocation"). Our exercise can be thought of a way to formalize this notion, but avoiding having to place an explicit probability distribution on the set of scenes, which is a tall order. Furthermore, while it is easy to formally define the mutual information between the scene and the image, computing it for an erasure channel (occlusions) under compositional infinite-dimensional domain warpings (viewpoint changes) and multiplicative infinite-dimensional disturbances (illumination) is not something easily done using the tools developed in classical Information Theory.

3 Preliminaries

The paper is structured in the following way.

- In the previous section, as a way of motivation, I have argued that traditional information theory, as developed with an eye towards the problem of *"reproducing"* the output of a source, is inadequate to characterize the value of an individual image for the purpose of *decision* or *control* tasks relative to the scene that the image portrays. Images are affected by *"nuisance factors"* that act on the data in a complex and highly structured fashion. Although closer to our scope, Gibson's notion of information as "invariants under change" falls short because it does not consider the counterpart of invariance, which is the *"discriminative"* (decision) or *"reachable"* (control) component of the representation.

- In Sect. 4.1, I introduce the notion of *"actionable information"* as the complexity of the maximal statistic that is invariant to a given nuisance. Similarly, I define *"complete information"* as the minimal statistic that is sufficient for a given task. When the nuisance is *"invertible"*, the two are identical, and their difference, the *"actionable information gap"* defined in Sect. 4.3, is zero.

- In the context of vision, viewpoint and illumination – away from visibility artifacts such as occlusions and cast shadows – are *invertible.* However, occlusions, cast shadows and quantization are *not.* Therefore, in general, the actionable information gap is non-zero.[12]

- The invertibility of a nuisance depends on the *control authority* of the sensor. While *occlusions and quantization* are non-invertible for a passive and static observer, they *become invertible when the observer is able to control the data acquisition process* (Sect. 4.3) for instance by changing viewpoint or accommodation (Fig. 10). Similarly, cast shadows are not invertible in grayscale images, but may become invertible when one can sample multiple spectral bands. The process of "information pickup" consists in the exploration of the environment aimed at closing the information gap (Sect. 4.4).

- While complete information, in general, cannot be measured, actionable information can be computed. I describe a representational structure that organizes two-dimensional (region statistics), one-dimensional (boundaries) and zero-dimensional (attributed points) image characteristics at all scales; its coding length measures actionable information (Sect. 5). This is a conceptual construction. Nevertheless, a "poor man's version" of this construction can be easily and efficiently computed.

- Since complete information, and therefore the actionable information gap, cannot be known in advance, perceptual exploration must proceed based on locally computable quantities. I define the "decrease in actionable information gap" as a quantity that can be computed

[12] Indeed, it is infinite, for actionable information is zero (there is no occlusion-invariant in one image) and the complete information is infinity (to compute a sufficient statistic with respect to occlusion one would have to acquire the entire light field).

instantaneously (in the presence of infinitesimal motion) and argue that its integral during exploration converges to the complete information, assuming a bounded universe (Sect. 4.3).

- Finally, in Sect. 7 I discuss some consequence of our arguments, including a suggested set of prescriptions that a visual representation should obey, and their effects on the "signal-to-symbol barrier."

Before articulating our arguments, we need to introduce certain definitions, for which we need some notation.

3.1 Notation and Conventions

An image is represented as a function $I : D \subset \mathbb{R}^2 \to \mathbb{R}^k_+;\ x \mapsto I(x)$ that is $\mathbb{L}^2$-integrable, but otherwise not necessarily continuous, taking positive values in k bands, *e.g.* $k = 3$ for color, and $k = 1$ for grayscale. A time-indexed image is indicated by $I(x,t),\ t \in \mathbb{Z}_+$, assuming a discrete temporal sampling, and a sequence is denoted by $\{I(x,t)\}_{t=1}^T$, or simply $\{I\}$. The image relates to the scene, which is represented as a collection of piecewise continuous surfaces ("shape") $S \subset \mathbb{R}^3$, possibly parameterized by x, $S : D \to \mathbb{R}^3;\ x \mapsto S(x)$, and a reflectance $\rho : S \to \mathbb{R}^k$, which is also parameterized, with an abuse of notation exploiting visibility constraints, as $\rho(x) \doteq \rho(S(x))$. I indicate points in space with capital letters $X \in \mathbb{R}^3$, and points in the image with $x \in \mathbb{R}^2$. I model illumination changes by contrast transformations, *i.e.* monotonically increasing continuous functions $h : \mathbb{R}^+ \to \mathbb{R}^+$. This is a rough approximation for Lambertian objects viewed under ambient illumination, where the radiance ρ corresponds to the diffuse albedo. Changes of viewpoint are rigid body transformations, *i.e.* elements of the Special Euclidean group $g \in SE(3)$, represented by a translation vector $T \in \mathbb{R}^3$ and a rotation matrix $R \in SO(3)$, indicated by $g = (R, T)$ [52]. As a result of a viewpoint change, points in the image domain $x \in D$ are transformed (warped) via $x \mapsto \pi(g^{-1}(\pi_S^{-1}(x))) \doteq w(x)$, where $\pi : \mathbb{R}^3 \to \mathbb{P}^2;\ X \mapsto \bar{x} = \lambda X$ is an ideal perspective projection and $\lambda^{-1} = [0\ 0\ 1]X \in \mathbb{R}_+$ is the depth along the projection ray $[x] \doteq \{X \in \mathbb{R}^3 \mid \exists\ \lambda \in \mathbb{R}_+, x = \lambda X\}$; π_S^{-1} is the inverse projection, that is the point of first intersection of the projection ray $[x]$ with the scene S. I use the notation $w(x; S, g)$ when emphasizing the dependency of w on viewpoint and shape. Without loss of generality [68], I represent changes of viewpoint with diffeomorphic domain deformations $w : D \subset \mathbb{R}^2 \to \mathbb{R}^2$. This model is viable only away from visibility artifacts (occlusions, cast shadows), which is discussed in Sect. 3.2. All un-modeled phenomena (deviation from Lambertian reflection, complex illumination effects etc.) are lumped into an additive "noise" term $n : \mathbb{R}^2 \to \mathbb{R}^k$. We finally have our image formation model:

$$\begin{cases} I(x) = h(\rho(X)) + n(x) \\ x = \pi(g(X)), \quad X \in S. \end{cases} \tag{1}$$

Summary: (Refer to Fig. 3) I call the image I, the reflectance ρ, illumination (contrast) h, warping w, which depends on the shape S and the viewpoint g. I further call the scene ξ, the collection of (three-dimensional, 3D) shape and reflectance $\xi \doteq \{\rho, S\}$, and the nuisance ν, the collection of viewpoint and illumination $\nu \doteq \{g, h\}$. In short-hand notation, substituting X in the first equation above with $g^{-1}(\pi_S^{-1}(x))$, I write (1) as

$$\boxed{I(x) = h \circ \rho \circ w(x; S, g) + n(x) \doteq f(\rho, S; g, h, n)} \tag{2}$$

or, again with an abuse of notation, as

$$\boxed{I = f(\xi; \nu).}$$

3.2 Visibility and Quantization

The model (1) is only valid away from visibility artifacts such as occlusions and cast shadows. I will not deal with cast shadows, and assume that they are either detected from the multiple spectral channels $k \geq 3$, or that illumination is constant and therefore they cannot be told apart from material transitions (*i.e.* discontinuities in the reflectance ρ). Occlusions, on the other hand, we cannot do away with. Based on empirical studies of natural intensity and range statistics [36, 56], I model occlusions as the "replacement" of f, in a portion of the domain[13] $\Omega \subset D$, by another function β having the same statistics [56].[14] Sometimes Ω is called the *background* even though, in practice, it can be in front of the object of interest, or it can be part of the object of interest itself, as in self-occlusions:

$$I(x) = \begin{cases} f(\rho, S; g, h, n) & x \in D \backslash \Omega \\ \beta(x) & x \in \Omega. \end{cases} \tag{3}$$

Digital images are spatially quantized into a discrete lattice, with each element averaging the function I over a small region $\mathcal{B}_v(x_{ij}) \subset D$ of size $v \in \mathbb{R}_+$ centered at $x_{ij} = (iv, jv),\ i, j \in \mathbb{Z}$:

$$I(i,j) = \int_{\mathcal{B}_v(x_{ij})} I(x)dx = I(x_{ij}) + n(x_{ij}) \tag{4}$$

where the quantization error is lumped into the additive noise n. In what follows, depending on the context, we may lump occlusions Ω, β, quantization and noise n among the nuisances ν.

3.3 Invariant and Sufficient Statistics

A statistic, or "feature," is a deterministic function ϕ of the data $\{I(x), x \in D\}$, taking values in some vector space, $\phi(I) \in \mathbb{R}^K$. I indicate this in short-hand notation via $\phi(I)$. A statistic is *invariant* if its value does not depend to the nuisance, *i.e.* for any $\nu, \bar{\nu}$, we have $\phi(f(\xi, \nu)) = \phi(f(\xi, \bar{\nu}))$. A trivial example of invariant feature is a constant function $\phi(I) = \text{c} \ \forall \ I$. Among all invariant statistics, we are most interested in the *largest*[15], also called *maximal invariant*

$$\boxed{\hat{\phi}(I).}$$

[13] Note that Ω is not necessarily simply connected, and this model does not impose restrictions on how many depth layers there can be.

[14] One cannot distinguish an occluding boundary from a material transition in a single pin-hole image, unless the sensing process enables changes of viewpoint or accommodation (Fig. 10).

[15] In the sense of inclusion of sigma-algebras generated by the statistics.

A statistic is *sufficient* for a particular *task*, specified by a risk functional R associated to a control or decision policy u and loss function L, $R(u|I) \doteq \int L(u, \bar{u}) dP(\bar{u}|I)$, $R(u) \doteq \int R(u|I) dP(I)$, if the risk based on a policy computed using such a statistic is the same as the risk based on the raw data, *i.e.* $R(u|I) = R(u|\phi(I))$. A trivial example of sufficient statistic is the identity $\phi(I) = I$. Among all sufficient statistics, of particular interest is the smallest, or *minimal*, one

$$\boxed{\phi^\vee(I).}$$

Note that, in general, $R(u|I) \leq R(u|\phi(I))$ for any measurable function ϕ (Rao & Blackwell, [64] page 88, a.k.a. *data processing inequality*), with equality defining ϕ as a sufficient statistic. When a nuisance acts as a *group* on the data, it is always possible to construct invariant sufficient statistics (the orbits, or the quotient of the data under the group). In that case, the policy u is called an *equivariant* classifier (for decisions) or controller (for actions) ([64], Theorem 7.4). It would be possible, as an alternative, to define sufficient statistics in terms of Fisher's Information, although I prefer to tie the sufficiency to a particular task.

4 Placing the Ecological Approach to Visual Perception onto Computational Grounds

4.1 Actionable Information

I define *Actionable Information* to be the complexity H (coding length) of a maximal invariant,

$$\boxed{\mathcal{H}(I) \doteq H(\hat{\phi}(I)).} \tag{5}$$

When the maximal invariant is also a sufficient statistic, we have *complete information*

$$\mathcal{I} \doteq H(\phi^\vee(I)) = \mathcal{H}(I). \tag{6}$$

In this case, the Actionable Information measures all and only the portion of the data that is relevant to the task, and discounts the complexity in the data due to the nuisances. As is discussed in Sect. 4.3, invariant and sufficient statistics are, in general, different sets, so we have an *"information gap."* In Sect. 5 I show how to compute Actionable Information, which for unknown environments requires spatial integration of the information gap.

In the next section I show that for some nuisances (invertible), the gap can be reduced to zero, whereas for other nuisances (non-invertible), the gap can be infinite.

4.2 Invertible and Non-invertible Nuisances

Viewpoint g and contrast h act on the image as *groups*, away from occlusions and cast shadows, and therefore can be *inverted* [68]. In other words, the effects of a viewpoint and contrast change, away from visibility artifacts, can be "neutralized" in a single image, and an invariant sufficient statistic can, at least in principle, be computed [68]. Note that the notion of sufficient statistic in this case is with respect to any distribution, since it is possible to reconstruct individual realization of the scene regardless of the nuisance. Fig. 5 illustrates this, and [2] and [74] prove it for contrast and viewpoint respectively. It may be appear puzzling that the statistics that are invariant to contrast (the geometry of the level lines of the image [17]) are not invariant to viewpoint, and those that are invariant to viewpoint (the intensity of the image warped onto a canonical domain [74]) are not invariant to contrast. The conundrum was recently solved in [68] where it was shown that the Attributed Reeb Tree (ART) of a (portion of an) image is the viewpoint-contrast invariant sufficient statistic. The ART stores the label (maximum, minimum, saddle), relative ordering and connectivity of extrema of the function f in (3) and is supported on a zero-measure subset of the image domain, so it is somewhat surprising that this "thin" object is actually equivalent to the entire image but for the effects of viewpoint and contrast changes. If these were the only nuisances, we would be in business. Unfortunately, this is of little help, as visibility and quantization are *not* groups, and once composed with changes of viewpoint and contrast, the composition cannot be inverted. An occlusion cannot be "undone" from one image, and "occlusion-invariant statistics" make patently no sense. Or do they? I will address this issue in Sect. 4.3, but not before I have described how to compute viewpoint and illumination invariants that are non-committal with respect to visibility and quantization.

When a nuisance transformation is not a group, its effects cannot be eliminated via pre-processing, and instead must be dealt with as part of the decision or control process: The risk functional R depends on the nuisance, $R(u|f(\xi;\nu))$, which can be eliminated either by extremization, $\max_\nu R(u|f(\xi;\nu))$ following a maximum-likelihood (ML) aproach, or by marginalization $\int R(u|f(\xi,\nu))dP(\nu)$ following a maximum a-posterior (MAP[16]) approach, if a probability measure on the nuisance $dP(\nu)$ is available.[17] In either case, the decision should not be based on direct comparison of two invariant statistics, $\phi(I_1) = \phi(I_2)$ computed separately on the training/template data I_1 and on the testing/target data I_2 in a pre-processing stage. Instead, a costly optimization (ML) or marginalization (MAP) is necessarily performed at run-time. The most one can hope from pre-processing is to pre-compute as much of the optimization or marginalization functional as possible. For the case of occlusion and quantization, this leads to the notion of *texture segmentation* as follows.

4.2.1 Segmentation as Redundant Lossless Coding

An occlusion $\Omega \subset D$, $\beta : \Omega \to \mathbb{R}^k$ is a region that exhibits the same (piece-wise spatially stationary [56]) statistics of the unoccluded scene (3). It can be multiply-connected, generally has piecewise smooth boundaries, and is possibly attached to the ground. Even if we could detect discontinuities in the image, which is a tall order, we would still not know which are the occlud-

[16] Invariant classification is problematic in a Bayesian setting, as one has to use improper uninformative priors; the issue is discussed at length in [62].

[17] Consider for example the binary decision of whether two images I_1 (training, or template image) and I_2 (testing, or target image) portray the same scene ξ. If the nuisance ν involves occlusions, so that $I_1 = f(\xi;\nu_1)$ and $I_2 = f(\xi;\nu_2)$, a decision can be performed by "searching" for all possible scenes ξ and occlusions ν_1, ν_2 that generate both images to within a specified accuracy (threshold). This is equivalent to implicitly "reconstructing" the scene ξ and "registering" the nuisances ν_1, ν_2.

ing boundaries, as opposed to material transitions or cast shadows. Furthermore, we do not know the statistics of the occluder region, as different quantization scales can cause image structures (extrema and discontinuities) to appear and disappear. Fig. 5 illustrates this phenomenon. In the absence of quantization and noise, one would simply detect all possible discontinuities, store the entire set $\{f(\xi,\nu),\ \forall\,\nu\}$, leaving the last decision bit (occlusion vs. material or illumination boundary) to the last stage of the decision or control process, performed either by extremization (ML) or marginalization (MAP). Occluders connecting to the ground (such as the tree in the "Flower Garden" sequence [22]) where no occlusion boundary is present would have to be "completed" as advocated by Gestalt psychologists [76], leading to a *segmentation*, or partitioning, of the image domain into regions with smooth statistics. Unfortunately, quantized signals are everywhere discontinuous, making the otherwise trivial detection of discontinuities all but impossible. One could salvage this approach by setting up a cost functional (a statistic) $\psi_\Omega(I)$, that implicitly defines a notion of "discrete continuity" within Ω but not across its boundary, making the problem of segmentation self-referential (*i.e.* defined by its solution) and therefore unfalsifiable. But while no single segmentation is "right" or "wrong," the set of *all possible segmentations*, defined for *all possible quantization scales*, may be *useful.* It does not reduce the complexity of the image (in fact, it is highly redundant), but it may reduce the run-time cost of the decision or control task, by rendering it a choice of regions and scales that match across images. In Sect. 5 I show how to compute actionable information based on a scale-space segmentation tree.

4.2.2 Quantization and Texture

For any scale $s \in \mathbb{R}_+$, minimizing $\psi_\Omega(I|s)$ yields a different segmentation $\Omega(s) \doteq \arg\min_\Omega \psi_\Omega(I|s)$. Because image "structures" (extrema and discontinuities) can appear and disappear at the same location at different scales,[18] one would have to store the entire continuum $\{\Omega(s)\}_{s\in\mathbb{R}_+}$. In practice, $\psi_\Omega(\cdot|s)$ will have multiple extrema *(critical scales)* that can be stored in lieu of the entire scale-space. This is different than (single) scale selection, as advocated in the scale-space literature. In between such critical scales, structures become part of aggregate statistics that is called *textures*. See Fig. 5. To be more precise, a texture is a region $\Omega \subset D$ within which some image statistic ψ, aggregated on a subset $\omega \subset \Omega$ is spatially stationary. Thus a texture is defined by two (unknown) regions, *small* ω and *big* Ω, an (unknown) statistic $\psi_\omega(I) \doteq \psi(\{I(y), y \in \omega\})$, under the following conditions of *stationarity* and *non-triviality*:

$$\psi_\omega(I(x+v)) = \psi_\omega(I(x)),\ \ \forall\ v \mid x \in \omega \Rightarrow x+v \in \Omega; \quad \bar\Omega\backslash\Omega \neq \emptyset \Rightarrow \psi_{\bar\Omega}(I) \neq \psi_\Omega(I). \quad (7)$$

The small region ω, that defines the intrinsic scale $s = |\omega|$, is minimal in the sense of inclusion: If $\bar\omega$ satisfies the stationarity condition, then $\exists\ v \mid x \in \omega \Rightarrow x+v \in \bar\omega$.[19] Note that, by definition, $\psi_\omega(I) = \psi_\Omega(I)$. A texture segmentation is thus defined, for every quantization scale s, as the so-

[18] Two-dimensional signals do not obey the "causality principle" of one-dimensional scale-space, whereby structure cannot be created with increasing scale [50].

[19] $\{I(x), x \in \omega\}$ is sometimes called a *texton* [42], or *texture generator*. This definition applies to both "periodic" or "stochastic" textures. Regions with homogeneous color or graylevel are a particular case whereby ω is a pixel, and do not need separate treatment.

lution of the following optimization with respect to the unknowns $\{\Omega_i\}_{i=1}^N, \{\omega_i\}_{i=1}^N, \{\psi_i\}_{i=1}^N, N(s)$

$$\min \sum_{i=1}^{N(s)} \int_{\Omega_i} \|\psi_{\omega_i}(I(x)) - \psi_i\|^2 dx + \Gamma(\Omega_i, \omega_i) \tag{8}$$

where Γ denotes a regularization functional.[20]

In Sect. 5 I show how to use a (multi-scale) texture segmentation algorithm to compute actionable information.

As described in Sect. 4.1, *in general one cannot compute statistics that are at the same time invariant and sufficient, because occlusion and quantization nuisances are not invertible.* Or are they?

4.3 The Actionable Information Gap

As I have hinted at in Sect. 3.2, whether a nuisance is invertible depends on the image formation process: Cast shadows are detectable (hence invertible) if one has access to different spectral bands. Similarly, occluding boundaries cannot be detected from a single image captured with a pin-hole camera, but they can be detected if one can control accommodation or vantage point. So, if the sensing process involves *control* of the sensing platform (for instance accommodation and viewpoint), then both occlusion and quantization become *invertible* nuisances.[21] This simple observation is the key to Gibson's approach to ecological perception, whereby *"the occluded becomes unoccluded"* in the process of "Information Pickup" [28].

To make this concrete, recall from Sect. 4.1 the definition of *complete information* and note that – because of the non-invertible action of the nuisances – it must now depend[22] on the *scene* ξ. I indicate this with $\mathcal{I} \doteq H(\phi_\xi^\vee(I))$. When a sequence of images $\{I\}$ capturing the entire light-field of the scene is available, it can be used in lieu of the scene to compute the complete information as follows:

$$H(\phi_\xi^\vee(I)) \geq H(\phi_\xi^\vee(I)|I(x,0)) \geq H(\phi_\xi^\vee(I)|I(x,0), I(x,1)) \geq \ldots \geq H(\phi_\xi^\vee(I)|\{I(x,t)\}_{t=0}^\infty) = 0 \tag{9}$$

so, if the complete light field $\{I\}$ (also known as *Plenoptic Function*) were available, we would have

$$\boxed{\mathcal{I} \doteq H(\phi^\vee(\{I\})).}$$

Note that, although it may seem impossible or irrelevant to attempt to capture the complexity of the light field, there are instead computational approaches to measure it [23]. I define the Actionable

[20] This optimization is a tall order. A bare-bone version pre-computes the statistics ψ_i on a fixed domain ω, and aggregates statistics using a mode-seeking algorithm that enables model selection with respect to scale s. The downside is that boundaries between regions Ω_i are only resolved to within the radius of ω, generating spurious "thin regions" around texture boundaries. For the purpose of this study, this is a consequence we can live with, so long as we know that a sound model exists, albeit computationally challenging.

[21] Want to remove the effect of an occlusion? Move around it. Want to resolve the fine-structure of a texture, removing the effects of quantization? Move closer.

[22] Actionable information also depends on the scene ξ, but only through the image $I = f(\xi, \nu)$. Complete information, on the other hand, depends on the scene in ways that are independent of the measured image I.

Information Gap (AIG) as the difference between the Complete Information and the Actionable Information

$$\boxed{\mathcal{G}(I) \doteq \mathcal{I} - \mathcal{H}(I).} \tag{10}$$

The process of Information Pickup, therefore, is one of reducing the AIG to zero. Note that, in the presence of occlusion and quantization, *the gap can be only be reduced by moving within the environment.* In order to move, however, the agent must be able to compute the effects of its motion on the AIG, ideally without having to know the complete information $\mathcal{I}$, even if the data $\{I\}$ or the statistics $\phi^\vee(\{I\})$ were available from memory of previous explorations. To this end, I define an incremental occlusion $\Omega(t) \subset D$ between two images $I(x,t), I(x,t+dt)$ as a region which is visible in one image but not the other. In a causal setting, only the part of $I(x,t+1)$ that is not visible at time t matters (uncovered region) whereas the part of $I(x,t)$ that is not visible at $t+1$ (occluded region) is already explained. Given the assumptions implicit in the model (1), we have[23]

$$\begin{aligned}\Omega(t) = \arg\min_{\Omega, w(\cdot,t)} \int_{D\backslash\Omega} (I(x,t+dt) - I(x+w(x,t)dt,t))^2 dx + \\ + \mu_1 \int_D \|\nabla w\|_{\ell^1} dx + \mu_2 \int_\Omega \left(1 + \|\nabla I\|^2\right) dx\end{aligned} \tag{11}$$

where μ_1, μ_2 are multipliers that weight the regularizers (priors). In [5], it is shown that this functional can be written as the sum of three terms defined on the same domain as follows:

$$\Omega(t) = \arg\min_{e(\cdot)=\chi_\Omega(\cdot), w(\cdot,t)} \int_D \|\nabla I w(x,t) + I_t(x,t) - e(x)\|^2 dx + \mu_1 \int_D \|\nabla w(x,t)\| dx + \mu_2 \int_D |e(x)| dx, \tag{12}$$

where e is the mollified characteristic function of Ω: $e(x) = \chi_\Omega(x)$. This functional is *convex*, so a unique, globally optimal solution exists, for both $\Omega(t)$ and $w(x,t)$, as well as computationally efficient algorithms to compute it [5]. The functionals above trades off the cost of explaining portions of a new image $I(x,t+dt)$ with an encoding of a previously seen image $I(x,t)$ deformed by w, and the cost of encoding it anew, similarly to what is done in video coding.[24] Once an incremental occlusion has been found, for instance using the algorithm proposed in [5], the Decrease in Actionable Information Gap (DAIG) can be measured by the Actionable Information it unveils:

$$\boxed{\delta\mathcal{G}(I,t) = \mathcal{H}(I(x,t)_{|_{x\in\Omega(t)}})} \tag{13}$$

The aim of environmental exploration is to maximize the DAIG, until a stopping time T is reached when, ideally,

$$\int_0^T \delta\mathcal{G}(I,t)dt = -\mathcal{H}\left(\phi^\vee(\{I\}_{t=0}^T)\right) = -\mathcal{H}\left(\{\hat{\phi}(I)\}_{t=0}^T\right) \tag{14}$$

and therefore $\mathcal{G}(I) = 0$. Note that if there are no occlusions (*i.e.* the environment is concave, *e.g.* the inside of a room, and the sensor is omnidirectional), then $\delta G(I,t) = 0 \; \forall \; t$, and so $\mathcal{G}(I) = 0$

[23] The choice of name for the region Ω, the same used for texture-based segmentation, is not accidental.

[24] Note that an uncovered region where $\langle \nabla I, w \rangle = 0 \; \forall \; x \in \Omega$ is a *latent occlusion*, as is easily explained with a vector field w and is therefore not "visible." This phenomenon is known as the *aperture problem.*

with just one measurement, with no need for exploration. In the most general sense, however, when occlusions are present, we have $\mathcal{I} = \infty$, and perceptual exploration does not end until we visit the entire universe. In practice we must restrict our attention to a bounded universe, so as to have $\mathcal{I} < \infty$.

A variational technique has developed in [5] for detecting occlusions based on the solution of a partial differential equation (PDE) that has the minimum of (11) as its fixed point. It is based on the observation that the cost functional is convex, and therefore a unique global optimum can be found efficiently using Nesterov's algorithm [57], and is described in detail in [5]. A hurried man's solution to (11), and the ensuing computation of (13), can be found by block matching followed by run-length encoding of the residual, as customary in MPEG. Efficient algorithms, including hardware implementations, are readily available for this task.[25] The shortcoming of this approach is that, in general, it yields a loss of actionable information, so that $\mathcal{G}(I) > 0$, whereas the optimal solution to (11) guarantees, at least in theory, that no actionable information is lost.

4.4 Information Pickup

To study the process of "Information Pickup" by means of closing the Actionable Information Gap, I specify a simple model of an "agent," that is a Euclidean reference frame in physical space, *i.e.* a *viewpoint* $g(t) \in SE(3)$, moving under the action of a control, which I assume can specify the instantaneous velocity $u(t) \in \mathbb{R}^6$. This kinematic model neglects masses, inertias and other dynamic fancy. The agent simply moves by integrating its velocity, *i.e.* the ordinary differential equation $\dot{g}(t) = \widehat{u}(t)g(t)$ starting from some initial position, which for convenience I assume to be the origin $g(0) = e$.[26] The agent measures an image at each instant of time, $I(x,t)$:

$$\boxed{\begin{cases} \dot{g}(t) = \widehat{u}(t)g(t) \quad g(0) = e \\ I(x,t) = f\left(\rho(x), S(x); g(t), h(t), n(x,t)\right). \end{cases}} \tag{15}$$

A myopic control would simply maximize the DAIG:

$$\hat{u}(t) = \arg\max_{u} \delta\mathcal{G}(I,t) \quad \text{subject to (15)} \tag{16}$$

and quickly converge to local minima of the Actionable Information Gap: The agent would stop at "interesting places" forever. To release it, one can devise a variety of search strategies, including jump-diffusion processes [34], or introduce a "boredom function" that increases with the time spent at any given location. Still, the agent can get trapped by its own trajectories, as soon as it is surrounded by spots it has already visited. A simple "forgetting factor" can restore the reward exponentially over time.

A more sophisticated controller, or *explorer*, would attempt to close the information gap by planning an entire trajectory:

$$\{\hat{u}(t)\}_{t=1}^{T} = \arg\sup_{u(\cdot)} \int_0^T \delta\mathcal{G}(I,t)dt \quad \text{subject to (15)} \tag{17}$$

[25] Alternate solutions, for instance using graph-cut techniques, are too slow to be realistic for applications in on-line real-time visual exploration for the foreseeable future.

[26] Here $\widehat{u}$ indicates the operator that transforms a linear and angular velocity vector u into a *twist* $\widehat{u} \in se(3)$ [52].

until (14) is satisfied, in addition to energy/efficiency requirements. Our goal is to study instantaneous control strategies (16) that converge to $\mathcal{I}$ in an efficient manner (a dumb observer with only contact sensors can explore the space, eventually, as shown in Sect. 6). The process is depicted in Fig. 4.

If our models are sensible, the explorer would attempt to *go around occlusions*, and *resolve the structure* in textured regions. Thus, when placed in an unknown environment, its motion would be guided by the structure of the *scene*, not by the structure of the *image*, despite only measuring the latter. This would not be the case for an explorer who is unaware of the role and nature of the nuisances, and instead treats as "information" the complexity of the raw data. I test this hypothesis in the experimental section 6.

Remark 1 (The Actionable Information Paradox) *Consider the task of recognizing an object that can exhibit significant reflectance variability, such as chameleon, or a passenger vehicle on the road. What determines the identity of the object is its three-dimensional shape, not its appearance. Naturally one wants a representation of shape that is viewpoint-invariant. But viewpoint cannot be "undone" from an image alone, so one would have to store the entire image and defer dealing with viewpoint as part of the matching process. If, however, one moves relative to the object of interest, then three-dimensional shape is observable, and one can infer, and store, a 3D model of the geometry of the object, and discard photometry (reflectance and illumination), thus effectively reducing the storage requirement below the size of a single image (assuming piecewise smooth surfaces). This yields the apparent paradox whereby more data yield a smaller storage requirement. It also means that, in order to "extract information" we have to "throw away some of the data," which has epistemological implications discussed in Sect. 7.*

Remark 2 *It is interesting to notice that in the traditional communication scenario the only nuisance under consideration is an additive disturbance. Since there is no invariant to the additive disturbance, the invariant is trivial (it is the set of constant functions), and so are the sufficient statistics (the statistics of the source before being corrupted by the channel). Therefore, the AIG is the information content of the data. More in general, when the noise is assumed to be zero-mean, the canonization procedure suggests a way to build a descriptor, which consists in simply computing the average of the data (detector) and subtracting it from the data (descriptor) to arrive at a canonized representation. This process of removing the mean (or assuming it is zero) is often present in classical communication, but seen as an afterthought, whereas the theory of Actionable Information affirms that this is precisely the process of building a maximal invariant.*

5 Representational Structures

I now describe the computation of Actionable Information and the representation it dictates.

5.1 Computing Actionable Information

For each image, we first compute a viewpoint-contrast invariant as follows: First, we perform (over-) segmentation at all possible scales: Starting from a 5-dimensional vector of color channels and positions, I use Quick Shift [75] to construct in one shot the tree of all possible segmentations (Fig. 6 top). I then consider the finest partition (a.k.a. “superpixels”) to be the elementary unit, and construct the adjacency graph, then aggregate nodes based on the histogram of vector-quantized intensity levels and gradient directions in a region ω of 8×8 pixels and arrive at the *texture adjacency graph* (TAG) (Fig. 7 top-right). Two-dimensional regions with homogeneous texture (or color) are represented as nodes in the TAG. I then represent one-dimensional boundaries between texture regions as edges in the TAG, or equivalently pairs of nodes (Fig. 6 top-right and Fig. 7 bottom-left). Ridges sometimes appear as boundaries between textured regions, or as elongated superpixels. Finally, I represent zero-dimensional structures, such as junctions or blobs (Fig. 6 bottom), as faces of the TAG, or equivalently pairs of edges (Fig. 7 bottom). This structure is the *Representational Graph*, $\mathcal{R}$, whose run-length encoding measures Actionable Information. In particular, $\mathcal{H}(I)$ is computed by summing, over the number of nodes $N(s)$ of the representational graph over all stored scales s, the coding length of the texture histograms associated to each node, but not the shape or size of the regions ω_i; the strength associated to each edge, corresponding to the probability of detection of an edge and ridge detector and the proximity to a superpixel boundary, but not the shape or length of the boundary, using our implementation of [50] (Fig. 6 top-right); the presence of an attributed point region associated to a face and its descriptor, but not the position of the point. I use a SIFT detector for Difference-of-Gaussian blobs from VLfeat (http://www.vlfeat.org), and Harris-Affine from [55] (Fig. 6 bottom-right and bottom-left respectively). Although in theory we should also store, for each of these regions, the ART [68], in practice, in the experiments reported in Sect. 6, I forgo this step. Computing Actionable Information is time-consuming in our current rendition, requiring approximately 30 seconds per each 640×480 image. The DAIG, however, is fast.

5.2 Computing the Actionable Information Gap

I compute the DAIG (13) by solving, at each time instant, (11) starting from a generic initialization as in [19], and using the best estimate at time t as initialization for the optimization at time $t + dt$. On the occluded region $\Omega(t)$, I compute the actionable information as described above. Since $\Omega(t)$ is usually very small (unless the robot moves very fast), this can be done in a fraction of a second after the initial convergence. For a real-time implementation, one can consider a coarse approximation of the DAIG, as simply the run-length encoding of the residual after block-wise motion compensation, as discussed in Sect. 4.3.

In order to compute the AIG, the complete information is necessary. This is, in general, not available unless one has had the opportunity to inspect the environment beforehand, and has, for instance, the entire light-field stored in memory. The process of Information Pickup hinges on the hypothesis that, by integrating the DAIG, one would eventually converge to complete information, hence rendering the AIG equal to zero. In the next section I validate this assumption empirically.

6 Empirical Consequences of the Definitions

In this section I test our hypothesis that an agent guided by Gibson would seek to “go around occlusions" and "resolve textures," whereas one guided by Shannon would be unaware of the topological structure of the environment, despite using the same *data*.

In the first indoor experiment (Fig. 8), a simulated robot is given limited control authority $u = [u_X,\ u_Y,\ 0,\ 0,\ 0,\ 0]^T$ to translate on a plane inside a (real) room, while capturing (real) color images with fixed heading and a field-of-view of 90^o. The robot is capable of computing both Entropy and the DAIG at the current position as well as at immediately neighboring ones. Under these conditions, the agent reduces to a point (the vantage point) $g = (Id, T)$ where $T = [T_X, T_y, 0]^T$. I indicate the vantage point with $X = [T_X, T_Y]^T$, consistent with the nomenclature introduced in Sect. 3, and the control with $V = [u_X,\ u_Y]^T$.

In the second outdoor experiment (Fig. 11), the robot is Google's StretView car,[27] over which we have no independent control authority. Instead, I assume that it has an intelligent (Gibsonian) driver aboard, who has selected a path close to the optimal one. The robot measures omnidirectional panoramas at each instant of time, so the data is symmetric with respect to forward or backward traversal. In this case, we cannot test independent control strategies. Nevertheless, we can still test the hypothesis that traditional information, computed throughout the sequence, bears no relation to the structure of the environment, unlike actionable information, and in particular the DAIG. For the purpose of validation, I have used standard tools from multiple-view geometry [52] to reconstruct the trajectory of the vehicle and its relation with the 3D structure of the environment[28] (Fig. 12).

6.1 Exploration via Information Pickup

The "ground truth" Entropy Map (Fig. 8 top) and Complete Information Map (Fig. 8 middle) are computed from (real) images collected with a fixed-heading camera with 90^o field-of-view regularly sampled on a 20cm grid and up-sampled/interpolated to a 40×110 mesh (sample views are shown in Fig. 8 overlaid on the map of the room). Complete Information is computed as a sufficient statistic of the light field, that is as the actionable information of each image computed at each position in space. The traversable space here is restricted to the inside of the room, so the explorer is not allowed to go outside; however, openings due to doors and window extend the universe to the adjacent rooms and the vegetation outside the window.

The first agent considered is a **Brownian Explorer**, that follows a random walk governed by the stochastic differential equation (SDE)

$$\begin{cases} dX(t) = V(t)dt; \quad X(0) \sim \mathcal{U}(S \subset \mathbb{R}^2); V(0) = 0 \\ dV(t) = dW(t) \quad \text{a Wiener Process w/ cov. } \sigma^2 \end{cases} \tag{18}$$

to be integrated in the $\hat{\text{I}}$to sense.[29] In practice, we can make do with the discrete-time stochastic

[27] Data courtesy of Google, INC.

[28] A poor man's version of this experiment would use Google's pseudo-ground truth for the trajectory, and trust Google Earth to portray images suggestive of the three-dimensional structure of the environment.

[29] See [45] (p. 6) for a definition and characterization of a Wiener process, and [43] (Chapt. 2 and 5, in particular eq. (2.1) and the rest of Ch. 5.2) for the meaning of the SDE.

process generated by

$$\begin{cases} X(t+dt) = X(t) + V(t)dt; & X(0) \sim \mathcal{U}(S \subset \mathbb{R}^2) \\ V(t+dt) = V(t) + W(t)dt; & W(t) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I) \end{cases} \tag{19}$$

with $V(0) = 0$. The trajectory charted by the Brownian Explorer (e.g. the Roomba vacuum cleaner) is shown in Fig. 9 (top).[30] Clearly, one can do better with vision. For the **Shannonian Explorer** I consider directly the discrete-time model, with a temporal evolution of the entropy $H(I(x,t)|g(t) = X)$ of the image I captured at time t in position X, which I indicate in short form with $H(X,t) \doteq [H(X,0) - \mathcal{B}(X,t)]_+$ where $[\cdot]_+$ is a mollified rectifier to enable the computation of the gradient:

$$\begin{cases} X(t+dt) = X(t) + \nabla H(X(t),t)dt \\ \mathcal{B}(X,t+dt) = \alpha\mathcal{B}(X,t) + \beta\mathcal{N}(X|X(t),\sigma^2)dt \end{cases} \tag{20}$$

where $\mathcal{N}(x|m,\sigma^2)$ is a Gaussian kernel with mean m and isotropic variance $\sigma^2 I$; the coefficients $\beta > 0$ and $0 < \alpha \leq 1$ trade off boredom and forgetfulness respectively.

The Gibsonian Explorer seeks to maximize Actionable Information $\mathcal{H}(I(x,t)|g(t) = X) \doteq G(X,t)$, or reducing the Actionable Information Gap, by trading off boredom and forgetfulness in $G(X,t) \doteq [G(X,0) - \mathcal{B}(X,t)]_+$

$$\begin{cases} X(t+dt) = X(t) + \nabla G(X,t)dt \\ \mathcal{B}(X,t+dt) = \text{as in (20)}. \end{cases} \tag{21}$$

Representative sample trajectories of the Shannonian and Gibsonian explorations are shown in Fig. 9 (left and right column respectively). The Shannonian Explorer loves wallpaper, complex texture and generally operates regardless of the 3D structure of the scene. The Gibsonian Explorer is claustrophobic: It prefers apertures and attempts to go through windows and doors; the simulation does not allow that, hence it bounces off like a fly on glass. Both explorers briefly dwell in regions that yield complex images before moving on. The exploration stops when enough area has been covered that the boredom factor renders the reward function flat. Note that the Gibsonian explorer is not given the complete information, and can therefore only plan its action based on the DAIG. The goal of this experiment is to show that local control strategies, based on image computations, yield space exploration that is compatible with the complete information, that is considered ground truth. In addition to the forgetting factor and the driving noise process, randomness can be inserted in the planned path by performing saccades, that change visibility when the camera has a finite field-of-view. In the experiment in Fig. 9 the camera had a fixed heading, so this phenomenon was not explored. In Fig. 10 I experimented with the dilemma between complexity due to near-field texture versus far-away structure, two situations indistinguishable from an image alone. In the ecological approach to perception, however, accommodation is actively controlled, so one can discriminate between complexity due to nearby texture (near-field focus), or to far-away structure (far-field focus).

[30] The behavior of our naïf Brownian explorer around the boundaries (Fig. 9) is dictated by a simplified reflection processes: We just invert the component of velocity that causes the crossing of the boundary. A proper simulation would instead use shadow paths following the Reflection Principle as described in [43] (Sect. 2.6.A, p. 79).

6.2 Exploration via Minimization of the DAIG

Our working hypothesis is that the Actionable Information Gap computed along a spatial trajectory (whether actively controlled or not) is related to the structure of the scene, and in particular to its topology (openings and occlusions). Since there is no analogous notion in classical Information Theory, I will compare the Actionable Information Gap to the same Entropy gradient considered in the previous experiment. It is unsurprising, and patent in Fig. 11, that neither Entropy nor its gradient bear any relation to the structure of the scene. In Fig. 12, I show the top view of a 250 frame-long detail with the trajectory and point-wise structure computed from point correspondences using standard tools from multiple-view geometry. For comparison, I also show the pseudo-ground truth provided with the dataset (yellow push-pins). The color-coded trajectory on the bottom shows the entropy gradient, with enhanced color-coding (red is high, blue is low). On the top I show the same for the Actionable Information Gap. It shows peaks at turns and intersections, when large swaths of the scene suddenly become visible. Note that the peaks are both before and after the intersection, as the omni-directional viewing geometry makes the sequence symmetric with respect to forward and backward directions. For the same reason, there is a constant "creation/distruction of data" in the direction of motion due to quantization. The "ground truth" coordinates are rather imprecise, as they would have the vehicle crossing lanes into opposing traffic and into buildings. Trees, and vegetation in general, attract both the Shannonian and the Gibsonian explorers, as they are photometrically complex, but also geometrically complex because of the fine-scale occlusion structure, visible in the last part of the sequence (right-hand side of the plot; images are shown in Fig. 11). Similar considerations hold for highly specular objects such as cars and glass windows. Although this experiment does not entail active exploration, but only passive motion, it shows that the DIAG, computed from pairs of adjacent images, strongly relates to the structure of the scene, and in particular to its topology.

7 Summary and Discussion

I have presented a characterization of visual information for the purpose of decision and control tasks. It stands in opposition to the traditional notion of information as entropy or coding length of the data, which is tailored to the tasks of storage and transmission. Actionable Information is defined as the complexity of the maximal statistic that is invariant to the nuisances. Specifically, I have considered viewpoint and illumination variations, which have been recently shown to admit invariant sufficient statistics. In addition, I have considered occlusion and quantization artifacts, that cannot be inverted, and therefore induce an "information gap" that can be filled by controlling the data acquisition process.

While in traditional Information Theory "all data matters," in the context of Actionable Information at least a portion of the data is irrelevant, and *the process of "extracting information from data" requires a control action*. This tie between sensing and control is very prominent in Actionable Information.

I have illustrated these ideas on a simulated exploration task, guided by visual measurements. Whereas a Shannonian Explorer is guided by the complexity of the *data*, a Gibsonian Explorer is guided by the topology of the physical space surrounding it. In both cases, the data consists of images, and no 3D reconstruction, stereo or structure-from-motion is necessary.

This work relates to visual navigation and robotic localization and planning [70, 8, 37]. In particular, [77, 11, 67] propose "information-based" strategies, although by "information" they mean localization and mapping uncertainty based on range data. Range data are not subject to illumination and viewpoint nuisances, which are suppressed by the active sensing, *i.e.* by flooding the space with a known probing signal (*e.g.* laser light or radio waves) and measuring the return. There is a significant literature on vision-based navigation [12, 80, 58, 60, 69, 65, 27, 24, 63, 41], and our experimental section could be characterized simply as occlusion-driven navigation [46, 47, 7]. In most of the literature, stereo or motion are exploited to provide a three-dimensional map of the environment, which is then handed off to a path planner, separating the *photometric* from the *geometric and topological* aspect of the problem. Not only is this separation unnecessary, it is also ill-advised, as the regions that are most informative are precisely those where stereo provides no disparity. Our navigation experiments also relate to Saliency and Visual Attention [38], although there the focus is on navigating the *image*, whereas I are interested in navigating the *scene*, based on image data. In a nutshell, robotic navigation literature is "all scene and no image," the visual attention literature is "all image, and no scene." I bridge the gap by proposing an approach that allows to go "from image to scene, and vice-versa" in the process of Information Pickup. The relationship between visual incentives and spatial exploration has been a subject of interest in psychology for a while [15].

This is not a paper on visual recognition, although it does propose a representation (the Representational Graph) that integrates structures of various dimensions into a unified representation that can, in principle, be exploited for recognition. In this sense, it presents an alternative to [35, 72], that could also be used to compute Actionable Information. However, the rendition of the "primal sketch" [53] in [35] does not guarantee that the construction is "lossless" with respect to any particular task, because there is no underlying task guiding the construction. Our work also relates to the vast literature on segmentation, particularly texture-structure transitions [79]. Alternative approaches to this task could be specified in terms of sparse coding [59] and non-local filtering [13]. I stress the fact that, while no single segmentation is "right" or "wrong," the collection of all possible segmentations, with respect to all possible statistics pooled at all possible scales, is "useful" in the sense of providing pre-computation of the optimization or marginalization functional implicit in any recognition task. This paper also relates to the literature of ocular motion, and in particular saccadic motion. The human eye has non-uniform resolution, which affects motion strategies in ways that are not tailored to engineering systems with uniform resolution. One could design systems with non-uniform resolution, but mimicking the human visual system is not our goal.

Our work also relates to other attempts to formalize "information" including the so-called Epitome [40], that could be used as an alternative to our Representational Structure if one could compute it fast enough. Furthermore, the Epitome does not capture compactness and locality, that are importantly related to the structure of the scene (occlusions) and its affordances (relationship to the viewer). For instance, if one has same texture patch in different locations in space, these are lumped together, regardless of compactness. Another alternative is the concept of Information

Bottleneck, [71], and our approach can be understood as a special case tailored to the statistics and invariance classes of interest, that are task-specific, sensor-specific, and control authority-specific. These ideas can be seen as seeds of a theory of *"Controlled Sensing"* that generalizes Active Vision to different modalities whereby the purpose of the control is to counteract the effect of nuisances. This is different than Active Sensing, that usually entails broadcasting a known or structured probing signal into the environment. Our work also relates to attempts to define a notion of information in statistics [51, 9], economics [54, 4] and in other areas of image analysis [44] and signal processing [32]. Our particular approach to defining the underlying representational structure relates to the work of Guillemin and Golubitsky [31]. Our work also relates to video coding/compression: As I have pointed out, poor man's versions of some of our constructions could be computed using standard operations from the video coding standards. However, I advocated structures that are adapted to the image data (superpixels, TAG, representational graph) rather than on fixed blocks. We can do this because, to achieve invariance to viewpoint, we have no need to encode deformation of these regions, just their correspondence.

With respect to Information Theory, relating our definitions of information in terms of coding length to the various notion of entropy customary in the trade would require establishing the relation between a distribution of coding length of invariant and sufficient statistics and the set of images that it represents. I hypothesize that one could quantify the average coding length (Actionable Information) against the probability distribution of allowable control actions to arrive at a lower bound on the Actionable Information Gap,

$$\int (\mathcal{I} - H(\phi^\wedge(I)|u))\, dP(u). \tag{22}$$

Consider two extreme cases: The Gibsonian explorer that has full control authority, so $dP(u)$ is uniform, and $\mathcal{G} \to 0$. Note that, at least in theory, occlusions, quantization and noise are invertible, for their effect can be reduced by moving around obstacles, moving closer to textured surfaces, and by measuring repeated images at a stand-still respectively. The explorer in Google's car, on the other hand, has no control authority whatsoever, so $dP(u)$ is a measure concentrated on the path that the Google car is actually following. Therefore, $\mathcal{G} > 0$. More in general, there will be a trade-off between control authority (the entropy of the measure $dP(u)$) and task-performance ($\mathcal{G}$ for exploration, recognition, or any other nuisance-invariant task), that could eventually extend traditional rate/distortion theory into a control-authority/task-performance theory for machine perception.

Last, but not least, our work relates to Active Vision [1, 10, 6], and to the "value of information" [54, 26, 33, 21, 18]. The specific illustration of the experiment to the sub-literature on next-best-view selection [61, 7]. Although this area was popular in the eighties and nineties, it has so far not yielded usable notions of information that can be transposed to other visual inference problems, such as recognition and 3D reconstruction.

Similarly to previously cited work [77, 11], [25] propose using the decrease of uncertainty as a criterion to select camera parameters, and [3] uses information-theoretic notions to evaluate the "informative content" of laser range measurements depending on their viewpoint.

Clearly, one can raise a number of objections to the concepts defined here, both on mathematical and on philosophical grounds. For start, if we define occlusion as a nuisance, then a sufficient statistic can never be known until we explore the entire world and beyond, for we cannot know what is "on the other side of the hill[31]". However, the sufficient statistics are defined by the task, and if the task is navigation, then a sufficient statistic is aggregated until all openings in a space have been explored. If the task is recognition of a particular object or class, partial occlusions can be resolved, and total occlusions (*i.e.* the absence of the object of interest in the visual field) requires active search to resolve, and will not end until the object is found. Also, the invariant sufficient statistic described in [68] assume that the image is a Morse function. While Morse functions are dense in $\mathcal{C}^2$, which is dense in $\mathcal{L}^2$, and therefore they can approximate any square-integrable function arbitrarily well, co-dimension one extrema (edges, ridges, valleys) are qualitatively different than elongated blobs. Nevertheless, one could extend the analysis to (multi-scale) edge and ridge detectors, for instance following the guidelines of [49, 16], and still have a thin set that encodes all the actionable information. This extension is the subject of future work.

The operational definition of information introduced, and the mechanisms by which it is computed, suggest some sort of *"manifesto of visual representation"* for the purpose of viewpoint- and illumination-independent tasks (Sect. 5).

1. (Hippocratic oath:) *First, do no harm:* I have shown that this is possible, by storing statistics that are invariant with respect to viewpoint and contrast, for all possible partitions at all possible scales of an image. Thus Rao & Blackwell's theorem does not stand in the way of developing efficient representation of visual scenes.

2. (Occlusions:) *No single segmentation is right or wrong.* The set of all possible segmentations *may be useful.* If it were not for occlusions, there would be no need to introduce a notion of segmentation in visual inference.

3. (Quantization:) *Textures/structures are present at multiple scales at the same location.* The causality principle of scalar-valued signal scale-space does not apply to images. A "texture" is defined by *two* regions, ω, Ω, and a statistic on ω, ψ_ω that is stationary for all $\omega \subset \Omega$. Without quantization there would be no need to introduce a notion of (stochastic, or ensemble) texture in visual inference.

4. (Illumination is hard:) General illumination models are intractable for the purpose of analysis. One can use multiple spectral bands and local contrast changes limited to each segment and scale. Assume illumination is constant at the time-scale of the exploration, lest $\delta\mathcal{G}(I, t)$ cannot be computed.

5. (The ART of Vision:) The representational structure $\mathcal{R}$ should be designed to be *stable* with respect to un-modeled phenomena n, in the sense that small changes in n yield small changes in $\mathcal{R}$, or
$$\|\frac{\partial \mathcal{R}}{\partial n}\| = \kappa \tag{23}$$

[31] Occlusions have long fascinated humans both for practical reasons (*e.g.* the Duke of Wellington's quote "All the business of war, and indeed all the business of life, is to endeavor to find out what you don't know by what you do; that's what I called 'guessing what was on the other side of the hill'.") and for aesthetic ones (*e.g.* Leopardi's "L'Infinito").

where $0 \leq \kappa \leq \epsilon < \infty$ is the (bounded-input, bounded-output, or BIBO) *gain* and ϵ is a small constant. This is trivial for additive noise models $I = h \circ \rho \circ w + n$, but in general $I = f(\rho, S; g, h, n)$ exhibits more complex dependencies, and can be studied in the context of a particular application or model.

Whether the representational structure $\mathcal{R}$ implied by the computation of Actionable Information will be useful for visual recognition will depend on the availability of efficient (hyper-)graph matching algorithms that can handle topological changes (missing nodes, links or faces).

Coming back to the preamble to this article, the results presented have implications on the "signal-to-symbol barrier" issue.

The first result presented in this paper that is epistemologically relevant, that follows directly from [68], is that, for the case of viewpoint and illumination, (a) it is possible to devise and compute invariant statistics, (b) that such invariant statistics are sufficient (*i.e.* they are equivalent to the image up to changes of viewpoint and illumination/contrast), and (c) such statistics are *discrete*.[32] This means that, while there is in principle no benefit in storing a discrete representation, there is no loss either. In other words, the sign in Rao and Blackwell's condition $R(u|I) \leq R(u|\phi(I))$ is actually "$=$" as in the definition of a sufficient statistic. The indirect benefit is that the nuisances are directly removed in the *representation*, rather than having to be marginalized (MAP) or extremized (ML) as part of the decision process.[33] But while this result calls for discrete representations, it does not necessarily call for data analysis. In fact, even in the ART [68], there is no notion of locality, or the need to partition the image domain in the representation.

The second epistemological implication of this manuscript is that it is the *combination of the ecological statistics* (occlusion of line-of-sight yielding highly kurtotic gradient distributions in the image) *and the ability to move* (to invert the occlusion and quantization processes) that calls for a representation that partitions the image domain into (multiple, possibly overlapping) local regions. Plants do not move,[34] and therefore they would have no benefit in developing an internal representation. Although they acquire plenty of sensory data (temperature, pressure, radiometry), make plenty of decisions (sprout, flower, drop leaves), and perform plenty of actions (grow, bend, turn towards the sun), they have not developed a central nervous system.

Note that, while I have argued that in the case of invertible nuisance there is no loss in constructing a decision function that is invariant, in general there is no gain either, unless one factors time-complexity into the picture. Consider the two alternative mechanisms to eliminate the nuisances in a decision problem. The first is *marginalization*. If our *task* defines a loss function λ, and

[32] Other researchers in the past had this intuition, for instance David Marr wrote *"Our view is that vision goes symbolic almost immediately, at the level of zero crossings, [and this is] probably accomplished without loss of information"*. While this statement is patently incorrect, zero crossings aside, if one adopts the classical definition of information, it is actually true in the case of actionable information, and the proof of it resides in [68]. Alan Turing had also explored the "signal-to-symbol" issue in his seminal paper [73] that popularized reaction-diffusion partial differential equations. But artificial systems are not necessarily governed by reaction-diffusion dynamics, so the symbolization process is not an accident due to biological hardware constraints, but it is instead imposed by the nature of nuisances affecting the visual data formation process.

[33] If one sees a cat in the woods and needs to decide whether it is his dinner or vice-versa, he better avoid marginalizing-out all possible viewpoints and illuminations.

[34] Plants do move, of course, but not in the time scale of evolutionarily relevant processes that occur in the environment. A plant does not run away from a fire.

therefore a conditional risk (Bayes discriminant) $R(\alpha_i|x) = \sum_{j\in\{0,1\}} \lambda(\alpha_i, \omega_j)P(\omega_j|x)$, and our nuisance acts on the data y in a distributed fashion, so that $\nu \perp \omega \mid x$, then in general the expected risk $R(\alpha) \doteq \int R(\alpha|x)dP(x)$ satisfies $R(\alpha \circ \phi) \geq R(\alpha)$, with the equal sign defining a sufficient statistic for that task. Then the likelihood can be written as $p(y|x, \omega) = \int k(y - x \circ \nu)dP(\nu|\omega)$, where $k(y - x \circ \nu) \doteq p(y|x, \nu)$, and the discriminant can be written as a function of the data y, instead of the “hidden” data x, as the ratio

$$\psi_{mar} = \frac{\int k(y - x \circ \nu)dP(\nu|\omega_0)dP_X(x|\omega_0)}{\int k(y - x \circ \nu)dP(\nu|\omega_1)dP_X(x|\omega_1)}. \tag{24}$$

When the prior $dP(\nu)$ is improper, we can compute the discriminant by *registration*, or maximum-likelihood, by solving an optimization problem

$$\psi_{reg} = \min_{\nu_0,\nu_1} \frac{\int k(y - x \circ \nu_0)dP_X(x|\omega_0)}{\int k(y - x \circ \nu_1)dP_X(x|\omega_1)}. \tag{25}$$

These discriminants do not require that the nuisance be invertible. They do, however, require that a prior is available (for marginalization) and that a complex averaging procedure (marginalization) or optimization (registration) be performed *at decision time*. If one encounters an animal in the woods and needs to decide whether it is his dinner or vice-versa, clearly marginalizing with respect to all possible nuisances (position, orientation, pose, illumination, occlusion etc.) is a losing proposition. When available, a *canonization* discriminant can be computed instantaneously via

$$\phi_{can} = \frac{p_X(y \circ \nu^{-1}(y)|\omega_0)}{p_X(y \circ \nu^{-1}(y)|\omega_1)} \tag{26}$$

and yields an equi-variant estimator. Therefore, at equal performance (conditional risk) one would prefer to make a decision based on a maximal invariant. Indeed, one may be willing to trade off discriminative power (hence conditional risk) for the benefit of time-complexity. Naturally, those nuisances that are not invertible given the control authority of the sensors must be eliminated at run-time by marginalization of registration. One example of these nuisances are intra-class variation for the problem of recognizing object categories.

A corollary of what is shown in this paper is that if we were not capable of mobility, if our sensory modalities were not subject to occlusion phenomena, and if we did not have time constraints, there would be no benefit in developing an internal representation that is essentially discrete, and there would be no benefit in data analysis. Also, while marginalization and registration can always be performed, canonization – that is decision based on an invariant feature – can only be performed without a loss if the nuisances are invertible. The lesson from Gibson is that all nuisances are invertible if we have control authority over the sensing process. This is the fundamental link that ties together sensing and control: Without control, no (time- and risk-optimal) sensing can be performed in the presence of line-of-sight and quantization phenomena, and clearly without sensing no (feedback) control could be performed.

Acknowledgments

I wish to thank Andrea Vedaldi, Joseph O’ Sullivan and Richard Wesel for suggestions leading to a reformulation of Sect. 4.3, Alan Yuille, Serge Belongie and Ying-Nian Wu for discussions, Tahee Lee for processing the Google StreetView dataset, Brian Fulkerson and Jason Meltzer for providing examples, Roger Brockett for feedback and references, and Jitendra Malik for the Tunicate example. This research is supported by FA9550-09-1-0427, ONR and ARO. A short version of this paper appears in the Proc. of the Intl. Conf. on Comp. Vision [66].

References

[1] J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, 1988. 21

[2] L. Alvarez, F. Guichard, P. L. Lions, and J. M. Morel. Axioms and fundamental equations of image processing. *Arch. Rational Mechanics*, 123, 1993. 5, 10

[3] T. Arbel and F.P. Ferrie. Informative views and sequential recognition. In *Conference on Computer Vision and Pattern Recognition*, 1995. 21

[4] K. J. Arrow. *Information and economic behavior*. Federation of Swedish Industries Stockholm, Sweden, 1973. 21

[5] A. Ayvaci, M. Raptis, and S. Soatto. Optical flow with occlusion detection as a convex optimization problem. March 2010. 13, 14

[6] R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988. 21

[7] R. Bajcsy and J. Maver. Occlusions as a guide for planning the next view. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(5), May 1993. 20, 21

[8] M. A. Batalin and G. S. Sukhatme. Efficient exploration without localization. In *IEEE International Conference on Robotics and Automation, 2003. Proceedings. ICRA'03*, volume 2, 2003. 20

[9] J. M. Bernardo. Expected information as expected utility. *Annals of Stat.*, 7(3):686–690, 1979. 21

[10] A. Blake and A. Yuille. *Active vision*. MIT Press Cambridge, MA, USA, 1993. 21

[11] F. Bourgault, A.A. Makarenko, S. Williams, B. Grocholsky, and H. Durrant-Whyte. Information based adaptive robotic exploration. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, volume 1, 2002. 20, 21

[12] R. Brooks. Visual map making for a mobile robot. In *1985 IEEE International Conference on Robotics and Automation. Proceedings*, volume 2, 1985. 20

[13] A. Buades, B. Coll, and J.M. Morel. A non-local algorithm for image denoising. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, volume 2, 2005. 20

[14] J. B. Burns, R. S. Weiss, and E. M. Riseman. The non-existence of general-case view-invariants. In *Geometric Invariance in Computer Vision*, pages 120–131, 1992. 5

[15] R. B. Butler. The effect of deprivation of visual incentives on visual exploration motivation in monkeys. *Journal of comparative and physiological psychology*, 50(2):177, 1957. 20

[16] E. J. Candès and D. L. Donoho. Curvelets, multiresolution representation, and scaling laws. In *Wavelet Applications in Signal and Image Processing*, 2000. 22

[17] V. Caselles, B. Coll, and J.-M. Morel. Topographic maps and local contrast changes in natural images. *Int. J. Comput. Vision*, 33(1):5–27, 1999. 10

[18] R. Castro, C. Kalish, R. Nowak, R. Qian, T. Rogers, and X. Zhu. Human active learning. In *Proc. of NIPS*, 2008. 21

[19] T. Chan and L. Vese. An active contours model without edges. In *Proceedings of Int. Conf. Scale-Space Theories in Computer Vision*, pages 141–151, 1999. 16

[20] H. F. Chen, P. N. Belhumeur, and D. W. Jacobs. In search of illumination invariants. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, 2000. 5

[21] K. Claxton, P.J. Neumann, S. Araki, and M.C. Weinstein. Bayesian value-of-information analysis. *International Journal of Technology Assessment in Health Care*, 17(01):38–55, 2001. 21

[22] D. Cremers and S. Soatto. Motion competition: a variational approach to piecewise parametric motion segmentation. *Intl. J. of Comp. Vision*, pages 249–265, May, 2005. 11

[23] A.L. Da Cunha, M.N. Do, and M. Vetterli. On the information rates of the plenoptic function. *ICIP, Atlanta, GA*, 2006. 12

[24] A. J. Davison and D. W. Murray. Simultaneous localization and map-building using active vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):865–880, 2002. 20

[25] J. Denzler and C.M. Brown. Information theoretic sensor data selection for active object recognition and state estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):145–157, 2002. 21

[26] E. Fogel and Y.F. Huang. Value of Information in System Identification-Bounded Noise Case. *Automatica*, 18(2):229–238, 1982. 21

[27] M.O. Franz, B. Schölkopf, H.A. Mallot, and H.H. Bülthoff. Learning view graphs for robot navigation. *Autonomous robots*, 5(1):111–125, 1998. 20

[28] J. J. Gibson. The theory of information pickup. *Contemp. Theory and Research in Visual Perception*, page 662, 1968. 12

[29] J. J. Gibson. The myths of passive perception. *Philosophy and phenomenological research*, 37(2):234–238, 1976. 4

[30] J. J. Gibson. *The ecological approach to visual perception*. LEA, 1984. 5

[31] M. Golubitsky and V. Guillemin. Stable mappings and their singularities. *Graduate texts in mathematics*, 14, 1974. 21

[32] I. J. Good and D. B. Osteyee. *Information, weight of evidence. The singularity between probability measures and signal detection*. Springer, 1974. 21

[33] J.P. Gould. Risk, stochastic preference, and the value of information. *Journal of Economic Theory*, 8(1):64–84, 1974. 21

[34] U. Grenander and M. I. Miller. Representation of knowledge in complex systems. *J. Roy. Statist. Soc. Ser. B*, 56:549–603, 1994. 14

[35] C. Guo, S. Zhu, and Y. N. Wu. Toward a mathematical theory of primal sketch and sketchability. In *Proc. 9^{th} Int. Conf. on Computer Vision*, 2003. 20

[36] J. Huang and D. Mumford. Statistics of natural images and models. In *Proc. CVPR*, pages 541–547, 1999. 8

[37] S.B. Hughes and M. Lewis. Task-driven camera operations for robotic exploration. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 35(4):513–522, 2005. 20

[38] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Rev. Neuroscience*, 2(3):194–203, 2001. 20

[39] K. James. On some possible characteristics of information in J. J. Gibson's ecological approach to visual perception. *Leonardo*, 13(2), 1980. 5

[40] N. Jojic, B. Frey, and A. Kannan. Epitomic analysis of apperance and shape. In *Proc. ICCV*, 2003. 20

[41] S.D. Jones, C. Andersen, and J.L. Crowley. Appearance based processes for visual navigation. In *Processings of the 5th International Symposium on Intelligent Robotic Systems (SIRS'97)*, pages 551–557, 1997. 20

[42] B. Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–97, 1981. 11

[43] I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, 1988. 17, 18

[44] K. C. Keeler. *Map representation and optimal encoding for image segmentation*. PhD dissertation, Harvard University, October 1990. 21

[45] H. Kunita. *Stochastic differential equations on manifolds*. Cambridge University Press, 1991. 17

[46] K.N. Kutulakos and C.R. Dyer. Global surface reconstruction by purposive control of observer motion. *Artificial Intelligence*, 78(1-2):147–177, 1995. 20

[47] K.N. Kutulakos and M. Jagersand. Exploring objects by invariant-based tangential viewpoint control. In *Computer Vision, 1995. Proceedings., International Symposium on*, pages 503–508, 1995. 20

[48] M. Li and P. Vitanyi. *An introduction to Kolmogorov complexity and its applications*. Springer, 1997. 5

[49] T. Lindeberg. *Edge Detection and Ridge Detection with Automatic Scale Selection*, volume 30. Cambridge University Press, 1998. 22

[50] T. Lindeberg. Principles for automatic scale selection. Technical report, KTH, Stockholm, CVAP, 1998. 11, 16, 33

[51] D. V. Lindley. On a measure of the information provided by an experiment. *Annals of Math. Stat.*, 27(4):986–1005, 1956. 21

[52] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An invitation to 3D vision, from images to geometric models*. Springer Verlag, 2003. 3, 7, 14, 17

[53] D. Marr. *Vision*. W.H.Freeman & Co., 1982. 20

[54] J. Marschak. Remarks on the economics of information. *Contributions to scientific research in management*, 1960. 4, 21

[55] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proceedings of the 7th European Conference on Computer Vision*, pages 128–142. Springer-Verlag, 2002. 16

[56] D. Mumford and B. Gidas. Stochastic models for generic images. *Quarterly of Applied Mathematics*, 54(1):85–111, 2001. 8, 10

[57] Y. Nesterov. A method for unconstrained convex minimization problem with rate of convergence O (1/k2). In *Doklady AN SSSR*, volume 269, pages 543–547, 1983. 14

[58] P. Newman and K. Ho. SLAM-loop closing with visually salient features. In *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, pages 635–642, 2005. 20

[59] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set : A strategy employed by V1? *Vision Research*, 1998. 20

[60] P. Peruch, J.-L. Vercher, and G. M. Gauthier. Acquisition of spatial knowledge through visual exploration of simulated environments. *Ecological Psychology*, 7(1):1–20, 1995. 20

[61] R. Pito, I.T. Co, and M.A. Boston. A solution to the next best view problem for automated surface acquisition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):1016–1030, 1999. 21

[62] C. P. Robert. *The Bayesian Choice*. Springer Verlag, New York, 2001. 10

[63] S. Se, D. Lowe, and J. Little. Vision-based mobile robot localization and mapping using scale-invariant features. In *In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2001. 20

[64] J. Shao. *Mathematical Statistics*. Springer Verlag, 1998. 2, 9

[65] R. Sim and G. Dudek. Effective exploration strategies for the construction of visual maps. In *2003 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2003.(IROS 2003). Proceedings*, volume 4, 2003. 20

[66] S. Soatto. Actionable information in vision. In *Proc. of the Intl. Conf. on Comp. Vision*, October 2009. 25

[67] C. Stachniss, G. Grisetti, and W. Burgard. Information gain-based exploration using rao-blackwellized particle filters. In *Proc. of RSS*, 2005. 20

[68] G. Sundaramoorthi, P. Petersen, and S. Soatto. On the set of images modulo viewpoint and contrast changes. *TR090005*, Submitted, JMIV 2009. 7, 9, 10, 16, 22, 23

[69] C.J. Taylor and D.J. Kriegman. Vision-based motion planning and exploration algorithms for mobile robots. *IEEE Trans. on Robotics and Automation*, 14(3):417–426, 1998. 20

[70] S. Thrun. Learning metric-topological maps for indoor mobile robot navigation. *Artificial Intelligence*, 99(1):21–71, 1998. 20

[71] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proc. of the Allerton Conf.*, 2000. 21

[72] S. Todorovic and N. Ahuja. Extracting subimages of an unknown category from a set of images. In *Proc. IEEE Conf. on Comp. Vis. and Patt. Recog.*, 2006. 20

[73] A. M. Turing. The chemical basis of morphogenesis. *Phil. Trans. of the Royal Society of London, ser. B*, 1952. 2, 23

[74] A. Vedaldi and S. Soatto. Features for recognition: viewpoint invariance for non-planar scenes. In *Proc. of the Intl. Conf. of Comp. Vision*, pages 1474–1481, October 2005. 5, 10

[75] A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. In *Proc. of the Eur. Conf. on Comp. Vis. (ECCV)*, October 2008. 16

[76] M. Wertheimer. *Laws of organization in perceptual forms*. W. D. Ellis, editor, A Sourcebook of Gestalt Psychology, pages 331–363. Harcourt, Brace and Company, 1939. 11

[77] P. Whaite and F.P. Ferrie. From uncertainty to visual exploration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):1038–1049, 1991. 20, 21

[78] N. Wiener. *Cybernetics, or Control and Communication in Men and Machines*. MIT Press, 1949. 4

[79] Y. N. Wu, C. Guo, and S. C. Zhu. From information scaling of natural images to regimes of statistical models. *Quarterly of Applied Mathematics*, 66:81–122, 2008. 20

[80] H. Zhang and J.P. Ostrowski. Visual motion planning for mobile robots. *IEEE Transactions on Robotics and Automation*, 18(2):199–208, 2002. 20

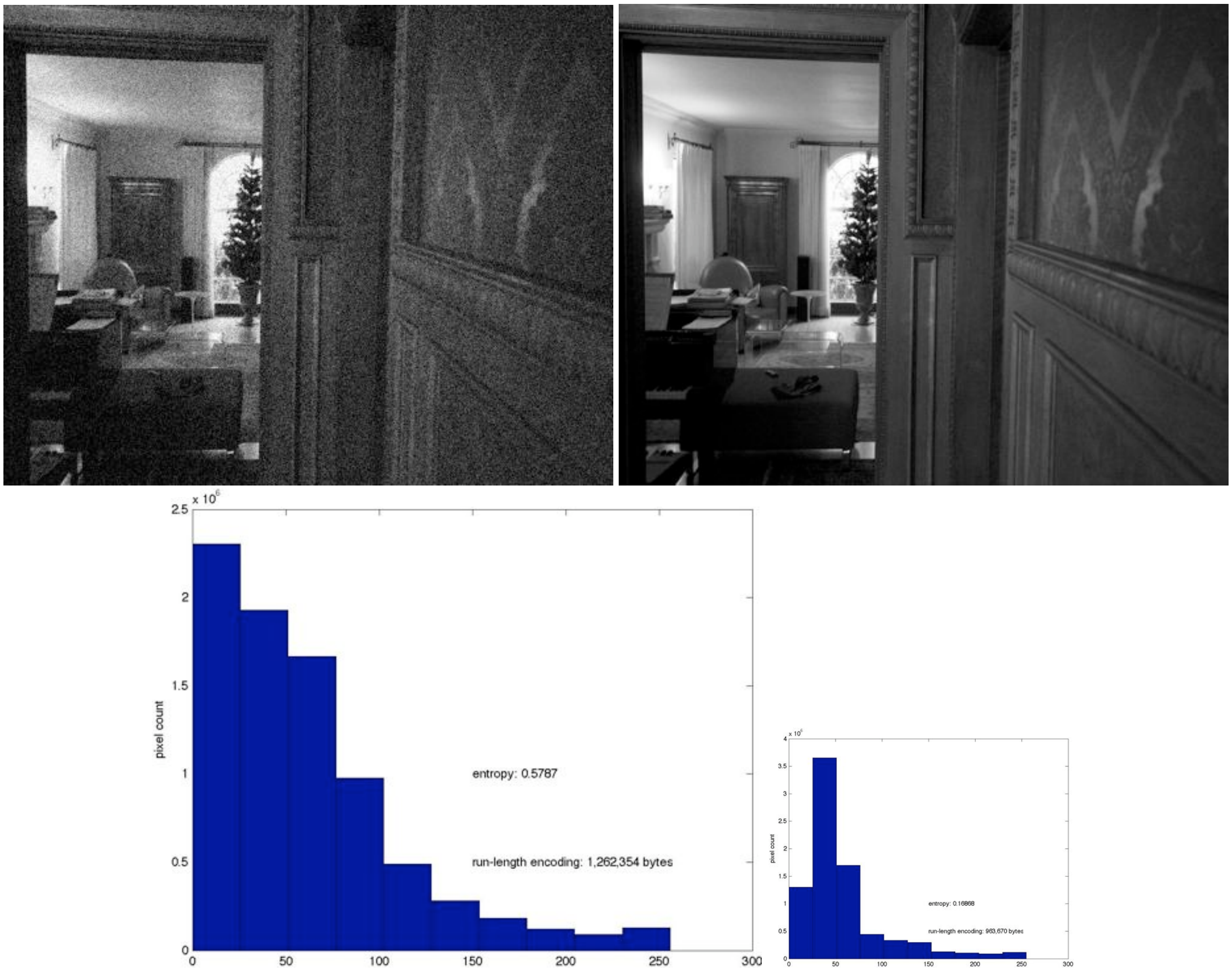

Figure 2: **What can you see in the left image that is not in the right?** *The left image "contains more information" if we measure information as entropy or coding length (bottom-left). Indeed, one pays more to* transmit *or* store *the image on the left. However, our goal is to* use *these images for a decision or control task that involves properties of the* scene. *We therefore need a novel notion of information that is not at odds with these tasks.*

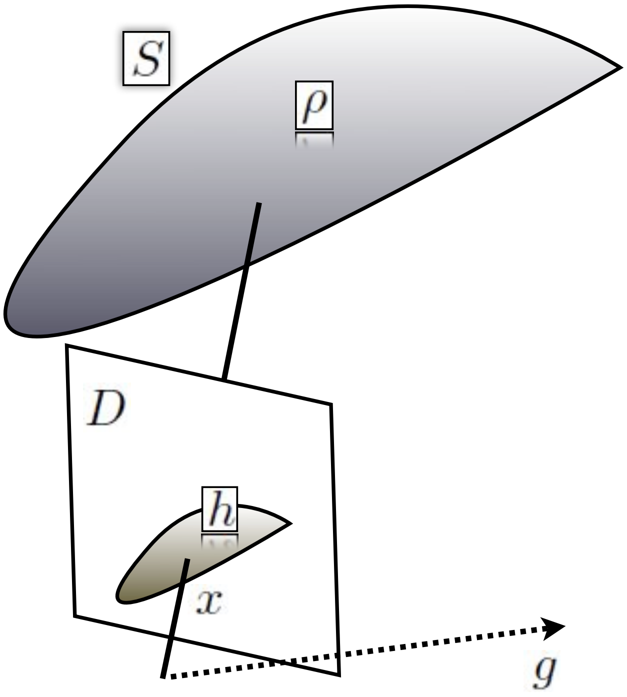

Figure 3:

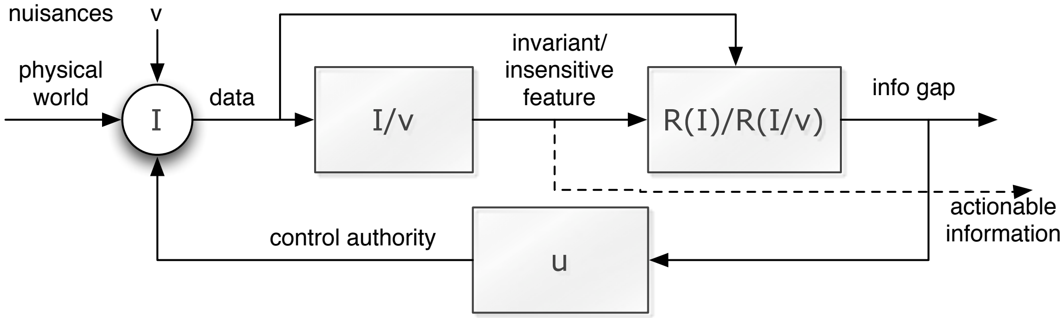

Figure 4: *For a given task, represented by a risk functional* R*, and for a given nuisance* ν*, one can in general compute invariant statistics. Their value is determined by the information gap: When it reaches zero, the invariant is also sufficient for the task. When it is larger, control authority can be exercised to minimize it; thence the invariant/insensitive feature is also sufficient (dashed line).*

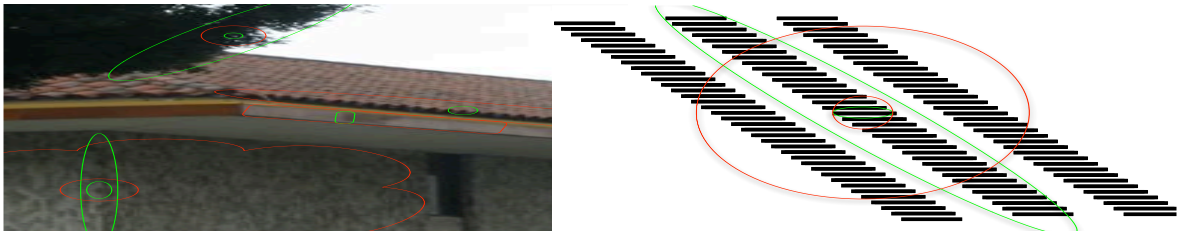

Figure 5: *The same point on an image can be represented, depending on scale, as "structure" (extrema and discontinuities, such as edges and ridges), then "texture" (spatially stationary, or cyclostationary, statistics), then again structure (green), and again texture (red) etc. All interpretations must be retained in the representation, rather than selecting one particular scale. One-dimensional signals obey a "causality principle" whereby structure can only be lost, but not created, with increasing scale [50]. This is not the case with two-dimensional images.*

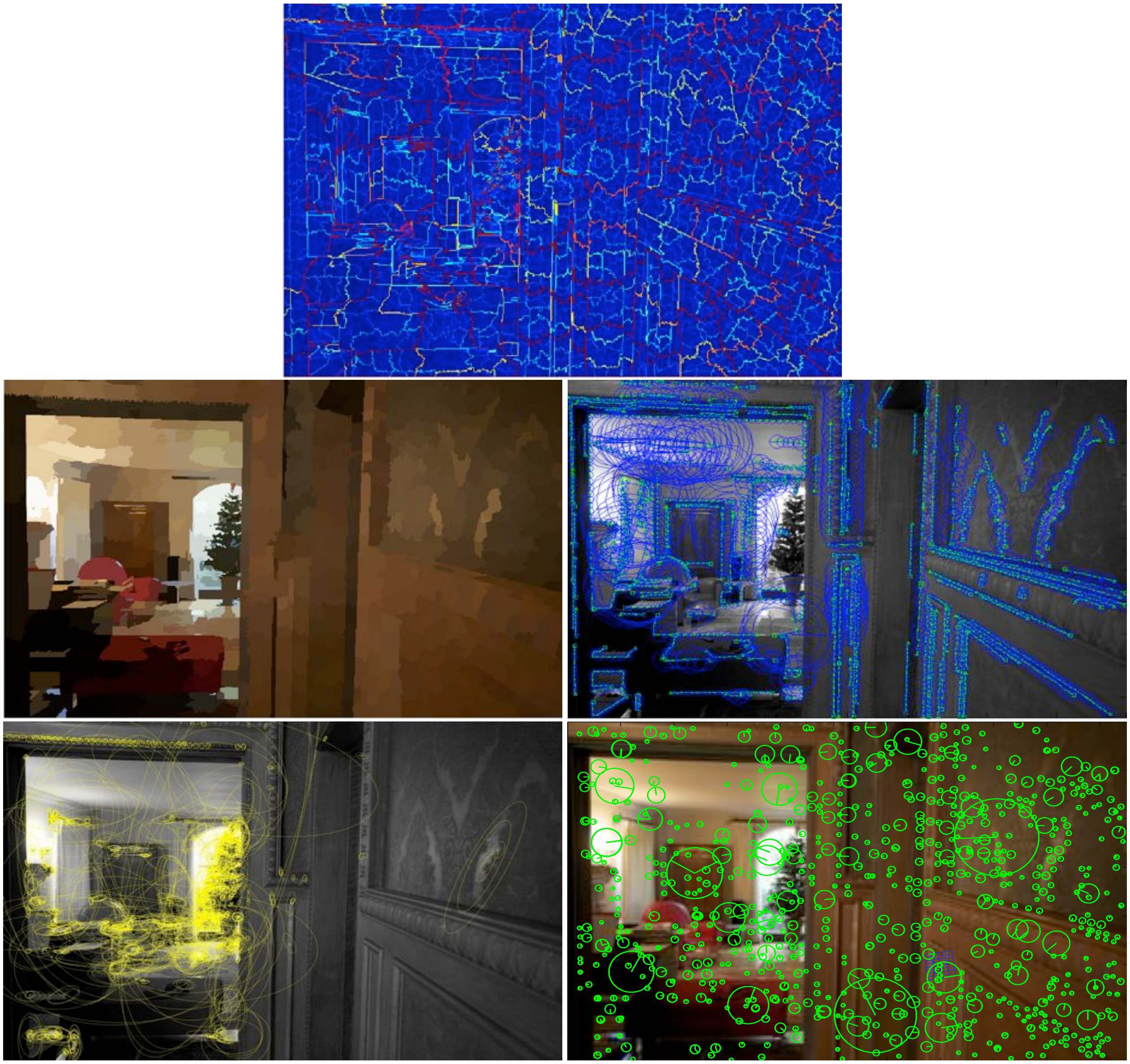

Figure 6: **Representational structures:** *Superpixel tree (top), dimension-two structures (color/texture regions), dimension-one structures (edges, ridges), dimension-zero structures (Harris junctions, Difference-of-Gaussian blobs). Structures are computed at all scales, and a representative subset of (multiple) scales are selected based on the local extrema of their respective detector operators (scale is color-coded in the top figure, red=coarse, blue=fine). Only a fraction of the structures detected are visualized, for clarity purposes. All structures are supported on the Representational Graph, described in the next figure.*

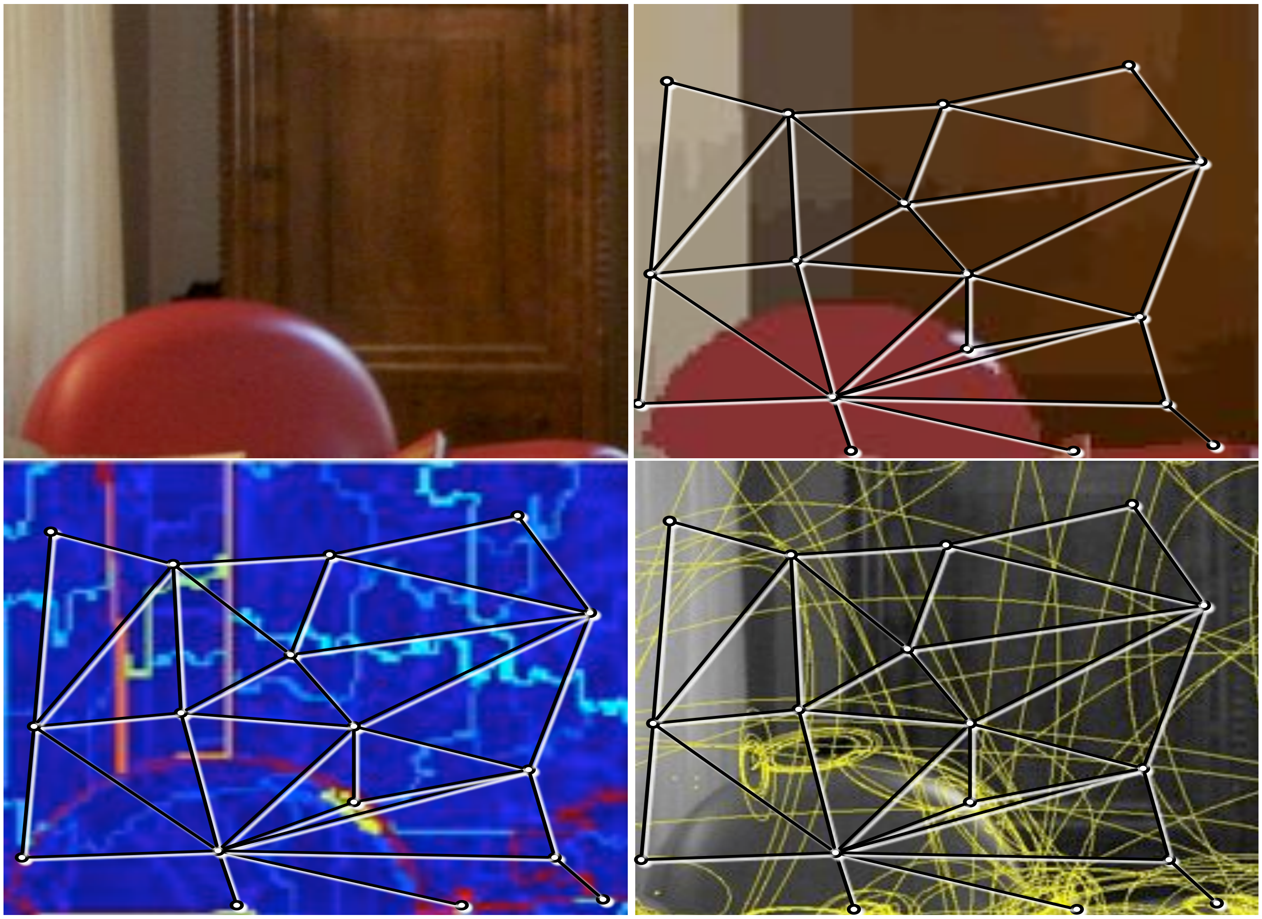

Figure 7: **Representational Graph** *(detail, top-left) Texture Adjacency Graph (TAG, top-right);* **nodes** *encode (two-dimensional) region statistics (vector-quantized filter-response histograms, or the ART of chromaticity within the region); pairs of nodes, represented by* **graph edges***, encode the likelihood computed by a multi-scale (one-dimensional) edge/ridge detector between two regions; pairs of edges and their closure (***graph faces***) represent (zero-dimensional) attributed points (junctions, blobs). For visualization purposes, the nodes are located at the centroid of the regions, and as a result the attributed point corresponding to a face may actually lie outside the face as visualized in the figure. This bears no consequence, as geometric information such as the location of point features is discounted in a viewpoint-invariant statistic.*

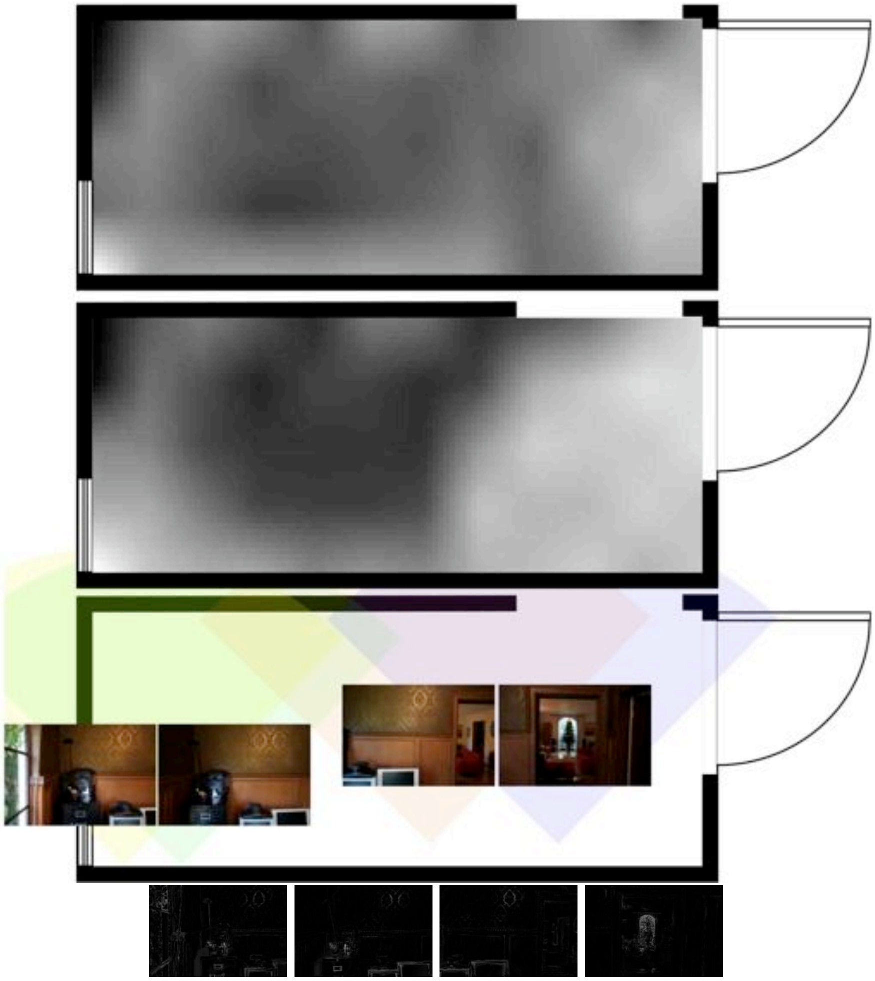

Figure 8: **Entropy vs. Actionable Information** *(first and second from the top) displayed as a function of position for a mobile agent with constant heading and* 90^o *field-of-view (bright = high; dark = low). Entropy relates to the structure of the image, without regard to the three-dimensional structure of the environment: It is high in the presence of complex textures (wallpaper and wood wainscoting) in the near field as well as complex scenes in the distance. Actionable Information, on the other hand, discounts periodic and stochastic textures, and prefers apertures (doors and windows), as well as specular highlights. Note the region on the right-hand side shows high levels of Actionable Information, proportional to the percentage of the field of view that intercepts the door aperture. Four representative images have been selected, corresponding to a field of view indicated by a colored cone (yellow, green, orange, and blue). Their coding residual is shown below. Note that, except for specular reflections, the complex wallpaper and wood grain does not trigger a high residual, but the opening behind the windows (yellow and blue viewing cone) does. The representational structures computed for every image collected (an approximation of the light field) constitutes the Complete Information, that is not available to the explorer beforehand.*

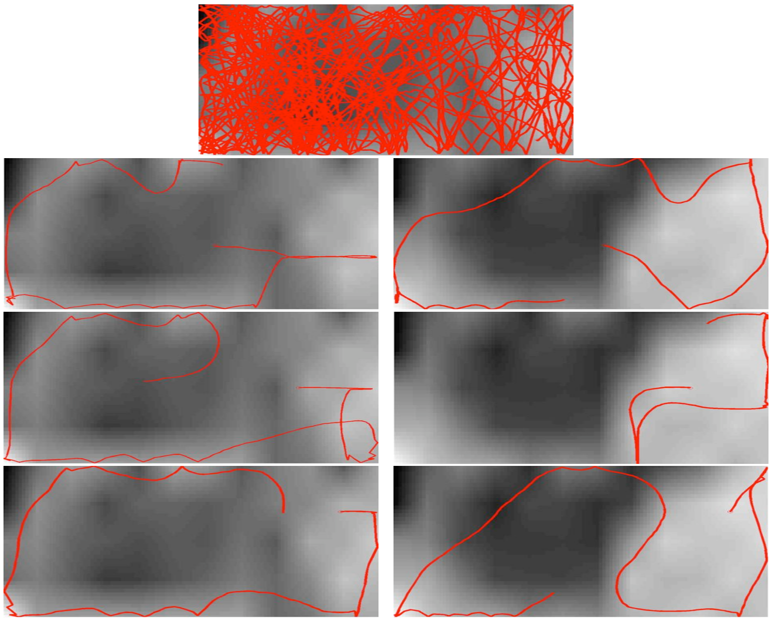

Figure 9: **Brownian (top), Shannonian (left) and Gibsonian (right) Information Pickup** *Representative samples of exploration runs are shown. The Shannonian Explorer (left column) is attracted by wallpaper (top edge of each plot) and the foliage outside the window (bottom-left corner of each plot). The Gibsonian Explorer (right column), aims for the window (bottom-left corner of the room) or the door (top-right corner of the room) like a trapped fly, and is similarly repelled by the control law that prohibits escape.*

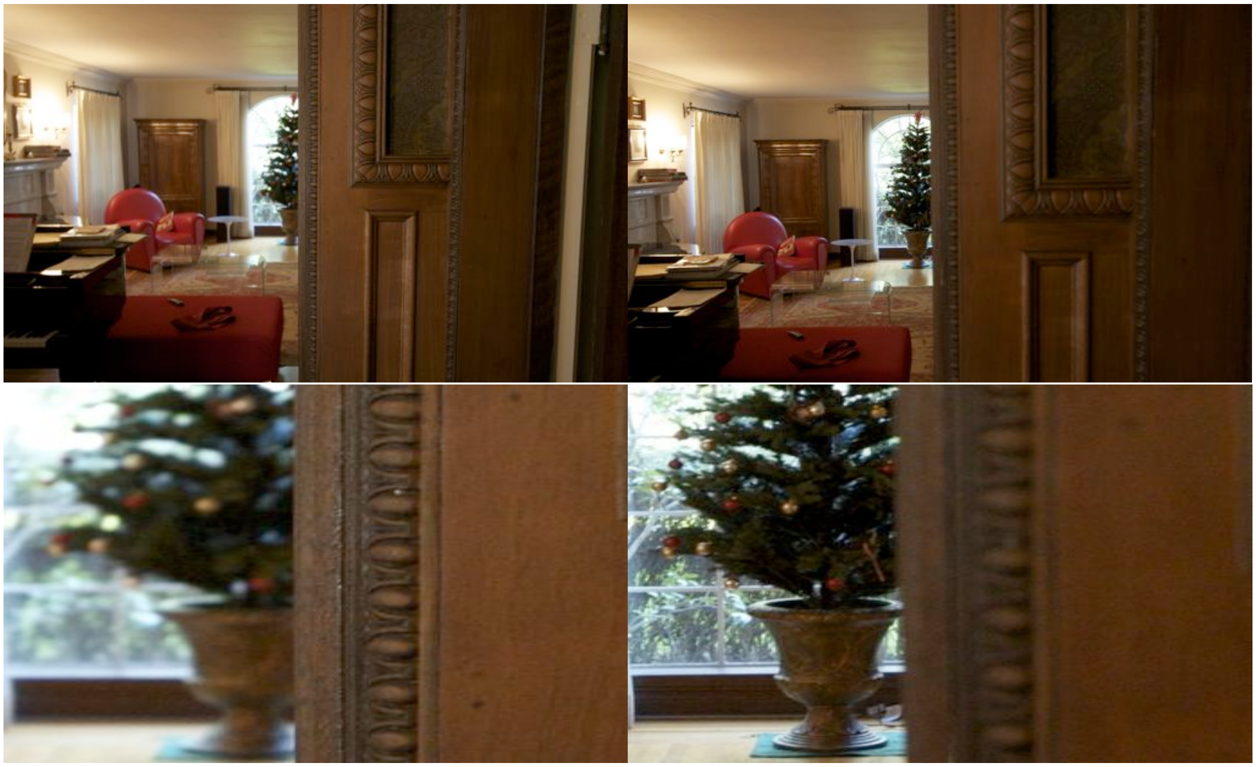

Figure 10: **Effects of Accommodation***: The same scene (top, detail at the bottom) viewed from similar vantage points while focusing in the near (left) and far field (right). Entropy is virtually identical (right is 4% lower,* 7.3414*nats vs.* 7.0451*nats), but the complexity on the left is due to the foreground texture, whereas on the right it is due to the structure of the background. Coding length is different, which reflects the self-similarity of the foreground texture (right is 47% higher,* $94,375$*Bytes vs.* $138,638$*Bytes). Actionable Information captures this fact as well (right is 52% higher,* $10,939$*bits vs* $16,608$*bits.) If accommodation is actively controlled, one can easily distiguish nearby texture from far away structure from the feedback signal of the accommodation control.*

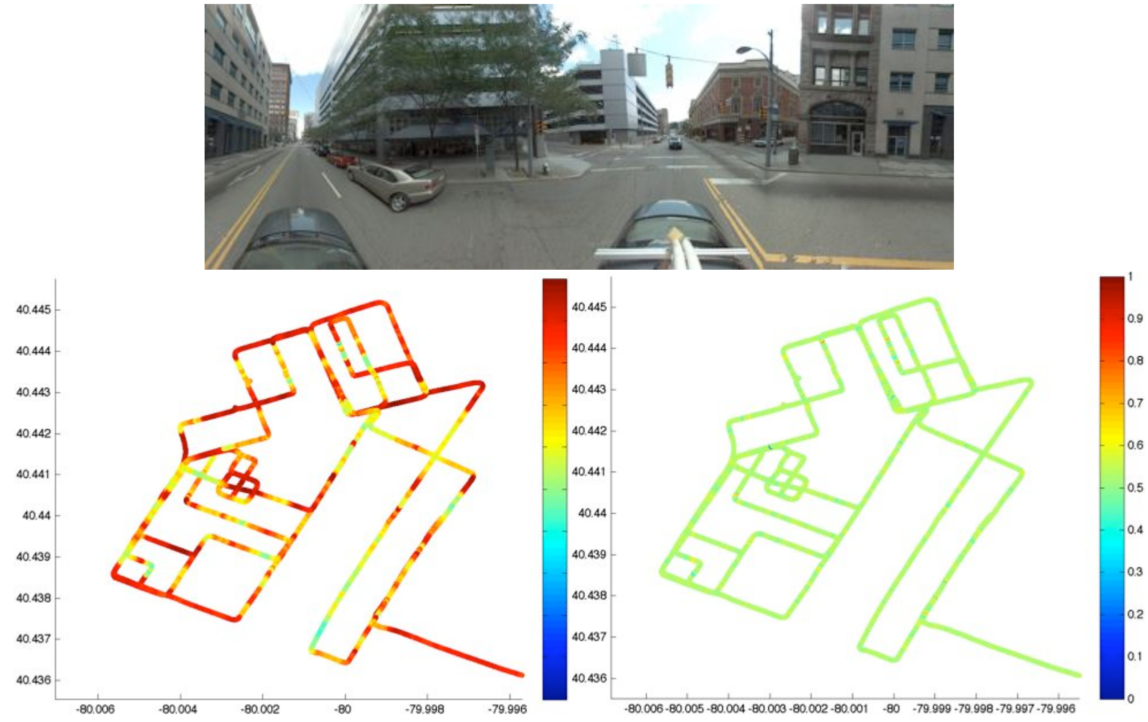

Figure 11: **Google StreetView Dataset** *Linear panorama,* $2,560 \times 905$ *pixels RGB. Entropy (left) and Entropy gradient along the path is shown color-coded at the bottom throughout the 12,000 frame-long sequence. Neither bear any relation to the geometry of the scene.*

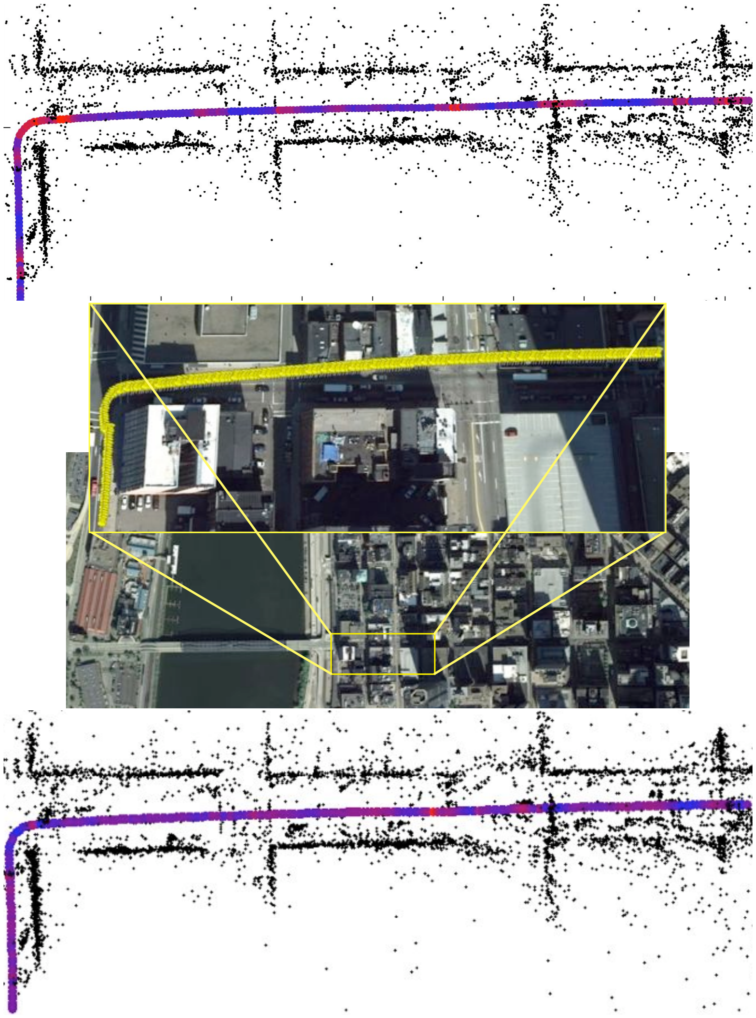

Figure 12: **Navigation via Minimization of the Actionable Information Gap** *Actionable Information gap (top) vs. Data Entropy gradient (bottom) color-coded (blue=small, red=large) for a 250-frame long detail of the Google Street View dataset, overlaid with the top-view of the point-wise 3D reconstruction computed using standard multiple-view geometry. For reference, the top-view from Google Earth is shown, together with push-pins corresponding to "ground truth" coordinates. The Entropy gradient (bottom) shows no relation with the 3D structure of the scene. Actionable Information (top), on the other hand, has peaks at turns and intersections, when large portions of the scene become visible (getting into the intersection) and thence disappear (getting out of the intersection).*