Actionable Information in Vision

Stefano Soatto

University of California, Los Angeles
http://vision.ucla.edu; soatto@ucla.edu

Abstract

I propose a notion of visual information 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). I illustrate these ideas on visual exploration, whereby a "Shannonian Explorer" guided by the entropy of the data 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. The operational definition of visual information suggests desirable properties that a visual representation should possess to best accomplish vision-based decision and control tasks.*

1. 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. The influence of Shannon's 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? Despite its pervasive reach today, Shannon's notion of information had early critics, among who James J. Gibson. Already in the fifties he was convinced that *data* is not *information*, and the value of data should depend on what one can do with it, *i.e*. the task. Much of the complexity in an image is due to *nuisance factors*, such as illumination, viewpoint and clutter, that have little to do with the decision (perception) and control (action) task at hand.

The goal of this manuscript is to define an operational notion of "information" that is relevant to visual decision and control 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) of a maximal statistic that is invariant to the nuisances associated to a given task.[1] I show that, 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 is drifting along unaware of the structure of the environment, while one seeking to maximize Actionable Information 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. An expanded version of this manuscript, with more details and discussion in relation to existing literature, is available at [25].

[1] It is tempting to write off this idea on grounds that neither viewpoint nor illumination invariants exist [7, 8]. But this would be simplistic: [30] shows that non-trivial viewpoint invariants exist for Lambertian objects of general shape. Similarly, [8] considers complex illumination fields, but invariants can be constructed for simpler illumination models, such as contrast transformations [2], even though these are valid only locally.

2009 IEEE 12th International Conference on Computer Vision (ICCV)
978-1-4244-4419-9/09/$25.00 ©2009 IEEE

2. Preliminaries

I represent an image as a function $I : D \subset \mathbb{R}^2 \to \mathbb{R}^k_+;\ x \mapsto I(x)$ that is $\mathbb{L}^2$-integrable, but 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 $I(x,t),\ t \in \mathbb{Z}_+$ and a sequence by $\{I(x,t)\}_{t=1}^T$, or simply $\{I\}$. The image relates to the scene, which I represent as a shape $S \subset \mathbb{R}^3$, made of piecewise continuous surfaces parameterized by x, $S : D \to \mathbb{R}^3;\ x \mapsto S(x)$, and a reflectance $\rho : S \to \mathbb{R}^k$, which I also parameterize as $\rho(x) \doteq \rho(S(x))$. I model illumination changes by contrast transformations, *i.e.* monotonically increasing continuous functions $h : \mathbb{R}^+ \to \mathbb{R}^+$. Changes of viewpoint are represented by a translation vector $T \in \mathbb{R}^3$ and a rotation matrix $R \in SO(3)$, indicated by $g = (R, T)$. 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 $\pi : \mathbb{R}^3 \to \mathbb{P}^2;\ X \mapsto x = \lambda X$ 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 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 [26], changes of viewpoint are modeled with diffeomorphic domain deformations $w : D \subset \mathbb{R}^2 \to \mathbb{R}^2$. 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 the image formation model:

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

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, with an abuse of notation, as $I = f(\xi; \nu) + n$.

2.1. 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.* spatial changes in the reflectance ρ). Based on natural intensity and range statistics [22], I model occlusions as the "replacement" of f, in a portion of the domain $\Omega \subset D$, by another function β having the same statistics:

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

Digital images are quantized into a discrete lattice, with each element averaging the function I over a small region where the quantization error is lumped into the additive noise n.

2.2. 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, $\phi(I) \in \mathbb{R}^K$. A statistic is *invariant* if it 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 is a constant $\phi(I) = \mathrm{c}\ \forall\ I$. Among all invariant statistics, we are most interested in the *largest*[2], also called *maximal invariant* $\boxed{\hat{\phi}_\nu(I).}$ 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 from a policy based on such a statistic is the same as one based on the raw data, *i.e.* $R(u|I) = R(u|\phi(I))$. A trivial example is the identity $\phi(I) = I$. Among all sufficient statistics, of particular interest is the smallest[2], or *minimal*, one $\boxed{\phi^\vee(I).}$ 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).

3. Placing the Ecological Approach to Visual Perception onto Computational Grounds

3.1. Actionable Information

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

$$\boxed{\mathcal{H}(I) \doteq H(\hat{\phi}_\nu(I)).} \tag{4}$$

When the maximal invariant is also a sufficient statistic, we have *complete information* $\mathcal{H}(I) = H(\phi^\vee(I))$. 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 I discuss in Sect. 3.3, invariant and sufficient statistics are, in general, different sets, so we have an *"information gap."* In Sect. 4 I show how to compute Actionable Information, which for unknown environments requires spatial integration of the information gap.

3.2. 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

[2] In the sense of inclusion of sigma-algebras.

be *inverted* [26]. 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 [26]. 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.

When a nuisance is not a group, it 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 (ML), $\max_\nu R(u|f(\xi;\nu))$, or by marginalization (MAP) $\int R(u|f(\xi,\nu))dP(\nu)$, if a probability measure on the nuisance $dP(\nu)$ is available. In either case, an optimal decision cannot 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. 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.

Segmentation as Redundant Lossless Coding

An occlusion $\Omega \subset D$ is a region that exhibits the same (piece-wise spatially stationary [22]) statistics of the unoccluded scene (3). It can be multiply-connected, generally has piecewise smooth boundaries, and is possibly attached to the ground. 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. Occluders connecting to the ground where no occlusion boundary is present would have to be "completed", 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, as customary in vision, set 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). But while no single segmentation is "right" or "wrong," the set of *all possible segmentations is a sufficient statistic.* It does not reduce the complexity of the image, 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.

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,[3] 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. In between such critical scales, structures become part of aggregate statistics that we call *textures*. 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*:

$$\psi_\omega(I(x+v)) = \psi_\omega(I(x)),\ \ \forall\, v \mid x \in \omega \Rightarrow x + v \in \Omega$$
$$\bar{\Omega}\backslash\Omega \neq \emptyset \Rightarrow \psi_{\bar{\Omega}}(I) \neq \psi_\Omega(I). \quad (5)$$

The region ω, that defines the intrinsic scale $s = |\omega|$, is minimal in the sense of inclusion: If $\bar{\omega}$ satisfies the stationarity condition, then $\exists\, v \mid x \in \omega \Rightarrow x + v \in \bar{\omega}$. Note that, by definition, $\psi_\omega(I) = \psi_\Omega(I)$. A texture segmentation is thus defined, for every quantization scale s, as the solution of the following optimization with respect to the unknowns $\{\Omega_i\}_{i=1}^N, \{\omega_i\}_{i=1}^N, \{\psi_i\}_{i=1}^N, N(s)$

$$\min \sum_{i=1}^{N(s)} \int_{\Omega_i} \|\psi_{\omega_i}(I(x)) - \psi_i\|^2 dx + \Gamma(\Omega_i, \omega_i) \quad (6)$$

where Γ denotes a regularization functional.[4]

As described in Sect. 3.1, *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?

3.3. The Actionable Information Gap

As I have hinted at in Sect. 2.1, 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 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.[5] This simple observation is the key to Gibson's approach to ecological perception, whereby *"the occluded becomes unoccluded"* in the process of "Information Pickup" [12].

[3] Two-dimensional signals do not obey the "causality principle" of one-dimensional scale-space, whereby structure cannot be created with increasing scale [20].

[4] A bare-bone version pre-computes the statistics ψ_i on a fixed domain ω, and aggregates statistics using a mode-seeking algorithm [31] 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.

[5] 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.

To make this concrete, recall from Sect. 3.1 the definition of *complete information* and note that – because of the non-invertible action of the nuisances – it must now depend 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 $\boxed{\mathcal{I} \doteq H(\phi^{\vee}(\{I\}))}$. I define the Actionable Information Gap (AIG) as the difference between the Complete Information and the Actionable Information

$$\boxed{\mathcal{G}(I) \doteq \mathcal{I} - \mathcal{H}(I).} \tag{7}$$

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. Letting $dt \to 0$, given the assumptions implicit in the model (1), we have

$$\Omega(t) = \arg\min_{\Omega} \int_{D\setminus\Omega} \left(\nabla I w(x,t) + \frac{\partial I}{\partial t}(x,t) \right)^2 dx + \int_D \|\nabla w\|_{\ell^1} dx + \int_{\Omega} \|\nabla I\|^2 dx. \tag{8}$$

Once an incremental occlusion has been found, the Decrease in Actionable Information Gap (DAIG) can be measured by the Actionable Information it unveils:

$$\boxed{\delta\mathcal{G}(I,t) = \mathcal{H}(I(x,t)_{|_{x\in\Omega(t)}}).} \tag{9}$$

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

$$\int_0^T \delta\mathcal{G}(I,t)dt = -\mathcal{H}\left(\phi^{\vee}(\{I\}_{t=0}^T)\right) = \mathcal{H}\left(\{\hat{\phi}_{\nu}(I)\}_{t=0}^T\right) \tag{10}$$

and therefore $\mathcal{G}(I) = 0$. We have developed a variational technique for detecting occlusions based on the solution of a partial differential equation (PDE) that has the minimum of (8) as its fixed point. However, a hurried man's solution to (8), and the ensuing computation of (9), can be found by block matching followed by run-length encoding of the residual, as customary in MPEG. 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 (8) guarantees that no actionable information is lost, at least in theory, asymptotycally for $T \to \infty$.

3.4. 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," *i.e.* a *viewpoint* $g(t)$ moving under the action of a control, which I assume can specify the instantaneous velocity $u(t) \in \mathbb{R}^6$ so that $\dot{g}(t) = \widehat{u}(t)g(t)$ starting from some initial position, which for convenience I assume to be the origin $g(0) = e$. [6] The agent measures an image at each instant of time, $I(x,t)$:

$$\boxed{\begin{cases} \dot{g}(t) = \widehat{u}(t)g(t) \quad g(0) = e \\ I(x,t) = f\left(\rho(x), S(x); g(t), h(t), n(x,t)\right). \end{cases}} \tag{11}$$

A myopic control would simply maximize the DAIG:

$$\hat{u}(t) = \arg\max_{u} \delta\mathcal{G}(I,t) \quad \text{subject to (11)} \tag{12}$$

and quickly converge to local minima of the Actionable Information Gap: The agent would stop at "interesting places" forever. To release it, one can introduce a "boredom function" that increases with the time spent at any given location. Still, the agent can get trapped by his own trajectories, as soon as he 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:

$$\{\hat{u}(t)\}_{t=1}^T = \arg\max_{u(\cdot)} \int_0^T \delta\mathcal{G}(I,t)dt \quad \text{subject to (11)} \tag{13}$$

until (10) is satisfied. This would require knowledge of the complete information $\mathcal{I}$. Our goal is to study instantaneous control strategies (12) that converge to $\mathcal{I}$ in an efficient manner (a dumb observer with only contact sensors can explore the space eventually, Sect. 5).

If the 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 structure of the nuisances, and instead treats as "information" the complexity of the raw data. I test this hypothesis in the experimental section 5.

4. The Representational Structures

I now describe the computation of Actionable Information and the representation it suggests. For each image, I perform texture segmentation at all scales: Starting from a 5-dimensional vector of color channels and positions, I use Quick Shift [31] to construct in one shot the tree of all possible segmentations (Fig. 1 top). I then consider

[6] Here $\widehat{u}$ indicates the operator that transforms a linear and angular velocity vector u into a *twist* $\widehat{u} \in se(3)$ [21].

the finest partition (a.k.a. "superpixels"), 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. 2 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, (Fig. 1 top-right and Fig. 2 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. 1 bottom), as faces of the TAG (Fig. 2 bottom). This structure, computed at all critical scales, is the *Representational Graph*, $\mathcal{R}$, whose coding length measures Actionable Information.

I compute the DAIG (9) by solving, at each time instant, (8) starting from a generic initialization, 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.

5. Experiments

In the first indoor experiment (Fig. 3), 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 AIG at the current position as well as at immediately neighboring ones. The robot is also able of computing the DAIG. Under these conditions, the agent reduces to a point $g = (Id, T)$ where $T = [T_X, T_y, 0]^T$. I indicate the vantage point with $X = [T_X, T_Y]^T$, and the control with $V = [u_X,\ u_Y]^T$.

In the second outdoor experiment (Fig. 5), the robot is Google's StretView car, over which we have no control authority. Instead, we assume that it has an intelligent (Gibsonian) driver aboard, who has selected a path close to the optimal one. 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, standard tools from multiple-view geometry have been used to reconstruct the trajectory of the vehicle and its relation with the 3D structure of the environment (Fig. 6).

5.1. Exploration via Information Pickup

The "ground truth" Entropy Map (Fig. 3 top) and Complete Information Map (Fig. 3 middle) are computed from (real) images regularly sampled on a 20cm grid and upsampled/interpolated to a 40×110 mesh (sample views are shown in Fig. 3 overlaid on the map of the room). The

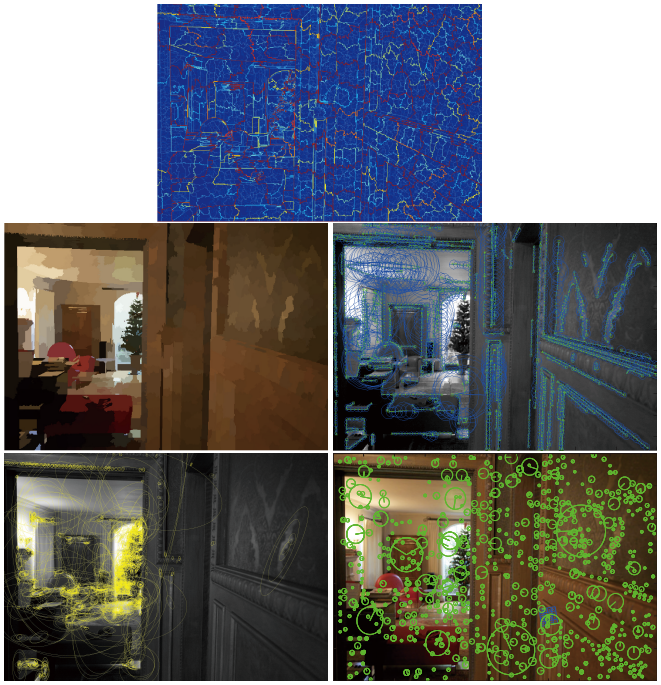

Figure 1. **Representational structures:** *Superpixel tree (top), dimension-two structures (color/texture regions), dimension-one structures (edges, ridges), dimension-zero structures (Harris junctions, Difference-of-Gaussian blobs). Structures are computed at all scales, and a representative subset 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).*

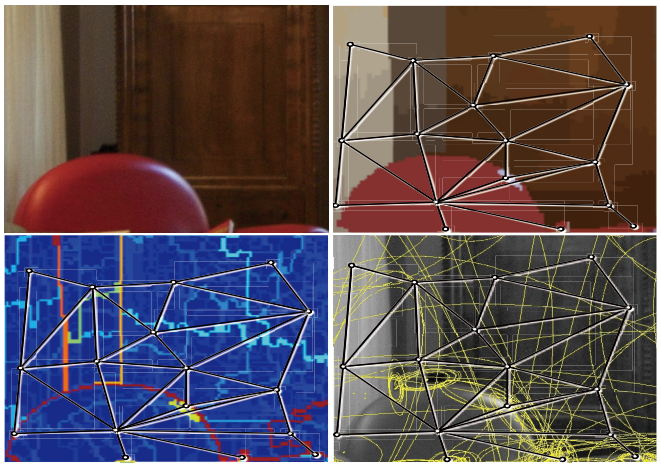

Figure 2. **Representational Graph** *(detail, top-left) Texture Adjacency Graph (TAG, top-right);* **nodes** *encode (two-dimensional) region statistics; pairs of nodes, represented by* **graph edges**, *encode the likelihood computed by a multi-scale (one-dimensional) edge/ridge detector between two regions; pairs of edges and their closure (***graph faces***) represent (zero-dimensional) attributed points (junctions, blobs).*

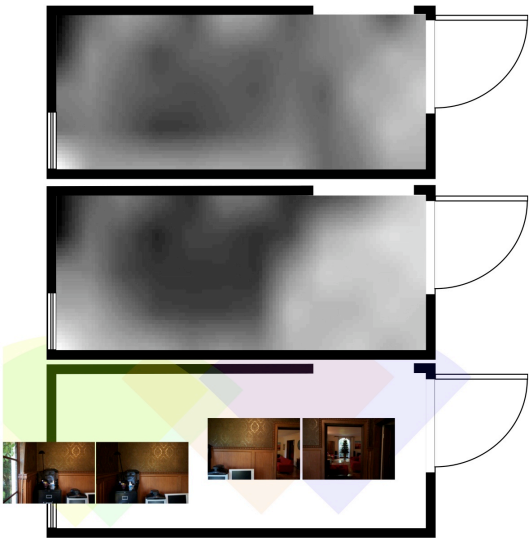

Figure 3. **Entropy vs. Actionable Information** *(first and second from the top) displayed as a function of position for a mobile agent with constant heading and 90^o field-of-view (bright = high; dark = low). Entropy relates to the structure of the image, without regard to the three-dimensional structure 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.*

first agent I consider is a **Brownian Explorer**, that follows a random walk governed by

$$\begin{cases} X(t+dt) = X(t) + V(t)dt; & X(0) \sim \mathcal{U}(S \subset \mathbb{R}^2) \\ V(t+dt) = V(t) + W(t)dt; & W(t) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I) \end{cases} \tag{14}$$

with $V(0) = 0$. The trajectory charted is shown in Fig. 4 (top). Clearly, one can do better with vision. For the **Shannonian Explorer** I consider 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 $H(X,t) \doteq [H(X,0) - \mathcal{B}(X,t)]_+$

$$\begin{cases} X(t+dt) = X(t) + \nabla H(X(t),t)dt \\ \mathcal{B}(X,t+dt) = \alpha \mathcal{B}(X,t) + \beta \mathcal{N}(X|X(t),\sigma^2)dt \end{cases} \tag{15}$$

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 $\mathcal{H}(I(x,t)|g(t) = X) \doteq G(X,t)$, trading off boredom and forgetfulness in $G(X,t) \doteq [G(X,0) - \mathcal{B}(X,t)]_+$

$$\begin{cases} X(t+dt) = X(t) + \nabla G(X,t)dt \\ \mathcal{B}(X,t+dt) = \text{as in (15)}. \end{cases} \tag{16}$$

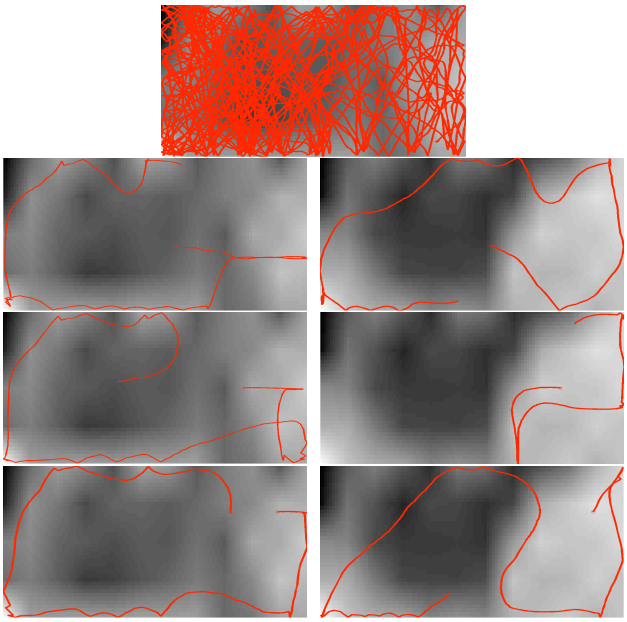

Figure 4. **Brownian (top), Shannonian (left) and Gibsonian (right) Information Pickup** *Representative samples of exploration runs are shown. The Shannonian Explorer (left column) is attracted by wallpaper (top edge of each plot) and the foliage outside the window (bottom-left corner of each plot). 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.*

Representative sample explorations are shown in Fig. 4 (left and right column respectively).

5.2. Exploration via Minimization of the DAIG

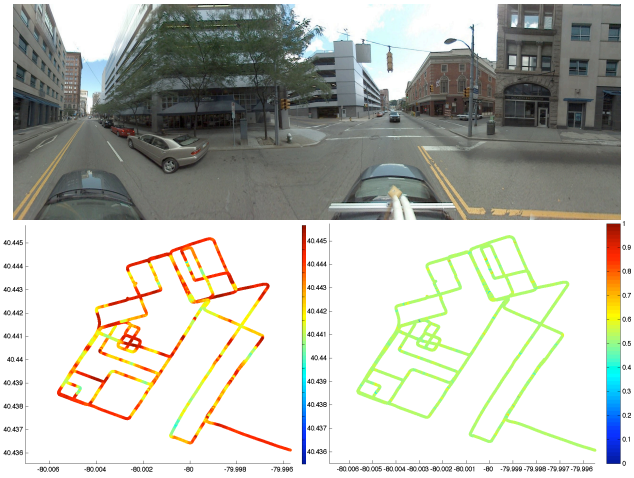

Figure 5. **Google StreetView Dataset** *Linear panorama, $2,560 \times 905$ pixels RGB. Entropy (left) and Entropy gradient along the path is shown color-coded at the bottom. Neither bear any relation to the geometry of the scene.*

It is patent in Fig. 5 that neither Entropy nor its gradient bear any relation to the structure of the scene. In Fig. 6, I show the top view of a 250 frame-long detail. 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. 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. 5). Similar considerations hold for highly specular objects such as cars and glass windows.

6. Discussion

The goal of this work is to define and characterize a notion of visual information tailored to decision and control tasks, as opposed to transmission and storage tasks. It relates to visual navigation or robotic localization and planning [14]. In particular, [32, 6] propose "information-based" strategies, although by "information" they mean localization and mapping uncertainty based on range data. There is a significant literature on vision-based navigation [27, 24, 11, 9, 23, 17]. However, 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. The navigation experiments also relate to Saliency and Visual Attention [15], although there the focus is on navigating the *image*, whereas we are interested in navigating the *scene*, based on image data.

Although the problem of visual recognition is not tackled directly, the computation of actionable information implicitly suggests a representational structure that integrates image features of various dimensions into a unified representation that can, in principle, be exploited for recognition. In this sense, it presents an alternative to [13, 29], that could also be used to compute Actionable Information. This work also relates to the vast literature on segmentation, particularly texture-structure transitions [33]. It also relates to Active Learning and Active Vision [1, 5, 4, 18], also [10, 3, 19].

This work also relates to other attempts to formalize "information" including [16, 28], and can be understood as a special case tailored to the statistics and invariance classes

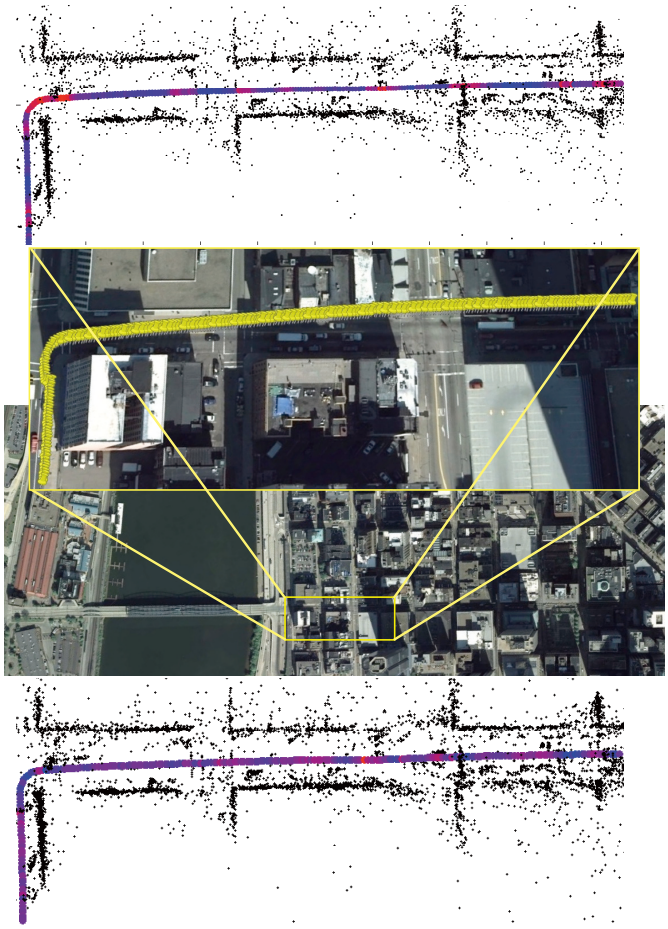

Figure 6. **Navigation via Minimization of the Actionable Information Gap** *Actionable Information gap (top) vs. Data Entropy gradient (bottom) color-coded (blue=small, red=large) for a 250-frame long detail of the Google Street View dataset, overlaid with the top-view of the point-wise 3D reconstruction 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).*

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.

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. 4),

discussed in detail in [25]. 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).

Acknowledgments

Thanks to T. Lee for processing StreetView, B. Fulkerson and J. Meltzer for examples, and AFOSR FA9550-09-1-0427 and ONR N00014-08-1-0414 for support.

References

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

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

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

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

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

[6] F. Bourgault, 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.

[7] 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.

[8] 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.

[9] 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.

[10] J. Denzler and C. 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.

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

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

[13] 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.

[14] S. 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.

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

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

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

[18] K. Kutulakos and C. Dyer. Occluding contour detection using affine invariants and purposive viewpoint control. In *Proceedings CVPR'94.*, pages 323–330, 1994.

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

[20] T. Lindeberg. Principles for automatic scale selection. Technical report, KTH, Stockholm, CVAP, 1998.

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

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

[23] 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.

[24] 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.

[25] S. Soatto. Actionable information in vision. *Technical Report UCLA-CSD-090007*, March 10, 2009.

[26] G. Sundaramoorthi, P. Petersen, V. S. Varadarajan, and S. Soatto. On the set of images modulo viewpoint and contrast changes. In *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, June 2009.

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

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

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.