Spatio-Chromatic Information available from different Neural Layers via Gaussianization

Jesús Malo

Abstract

How much visual information about the retinal images can be extracted from the different layers of the visual pathway? This question depends on the complexity of the visual input, the set of transforms applied to this multivariate input, and the noise of the sensors in the considered layer. Separate subsystems (e.g. opponent channels, spatial filters, nonlinearities of the texture sensors) have been suggested to be organized for optimal information transmission. However, the efficiency of these different layers has not been measured when they operate together on colorimetrically calibrated natural images and using multivariate information-theoretic units over the joint spatio-chromatic array of responses.

In this work we present a statistical tool to address this question in an appropriate (multivariate) way. Specifically, we propose an empirical estimate of the capacity of the system based on a recent Gaussianization technique that reduces the (challenging) multivariate PDF estimation problem to a set of (robust) univariate estimations. Total correlation measured using the proposed estimator is consistent with predictions based on the analytical Jacobian of a standard spatio-chromatic model of the retina-cortex pathway. Results show that (1) the capacity increases substantially along the layers of the model, and (2) the capacity of the cortical representation follows the PDF of natural scenes over the chromatic and achromatic dimensions of the stimulus space.

In theoretical neuroscience, the proposed capacity measure based on Gaussianization is useful to check the Efficient Coding Hypothesis and to assess the relative relevance of the chromatic and spatial elements of the models. From an applied perspective, in image compression, the proposed measure suggest which visual features should be addressed first; and in image quality, the proposed measure can generalize the visual information fidelity concept.

Keywords: Retina-cortex pathway; LMS; Chromatic adaptation; Opponent channels; Chromatic saturation; Texture sensors; Chromatic and achromatic CSFs; Divisive Normalization; Information capacity; Total Correlation; Mutual Information; Gaussianization

1 Introduction

Seeing is all about extracting information about a scene from the neural responses induced by the image of the scene. Therefore, both visual neuroscientists and image coding engineers are interested in similar information theoretic tools.

Neuroscience has a long tradition in using information theory both to quantify the performance of neurons [1] and to formulate principles that explain observed behavior using the so-called Efficient Coding Hypothesis [2, 3], and it is useful at many scales of sensory processing [4–6].
On the one hand, a large body of literature studies information transmission starting from a spiking neuron [7], including all sorts of additional constraints such as energy or size [8–10]. From these single-cell capacity estimates, different summation strategies [11] or independence assumptions [12] are considered to give global estimates of the amount of information transmitted by a set of sensors. This detailed low-level descriptions are certainly the basis for higher-level behavioral effects, but psychophysics is usually described with more abstract models. In fact, cascades of *linear+nonlinear* layers of more schematic neurons [13] describe a wide range of psychophysical phenomena, including color [14, 15], spatial texture [16], and motion [17].

On the other hand, these more schematic neural layers where inputs and outputs are arrays of continuous real values (not spike trains) are also studied according to the *Efficient Coding Hypothesis*. Information-theoretic goals are used to derive physiological behavior but also psychophysical effects. Examples include the derivation of chromatic opponent channels [18], the saturation of achromatic and chromatic channels [19–21], the emergence of linear texture sensors (not only achromatic [22, 23], but also spatio-chromatic [24], equipped with chromatic adaptation [25]), and linear sensors sensitive to motion [26]. Finally, the saturation of the responses of these spatio-temporal sensors has also been derived from information-theory arguments in the case of achromatic textures [27, 28] and motion [29].

In this work we will use these higher-level models which, being connected to physiology, are more related to the psychophysics of color and spatial texture. Information-theoretic study of models at this level, may address questions like: *what perceptual behavior is more relevant to encode images?*, *color constancy or contrast adaptation?*, *what mechanisms contribute to extract more information about the images?* the chromatic opponent channels or the texture filters?, *what kind of images are better represented by the visual system?* smooth achromatic shapes or sharp chromatic patterns?.

In order to give quantitative answers to the above questions, one needs unbiased estimators of transmitted information over the cascade of layers. And these estimators have to work on real-valued high-dimensional vectors.

In this work we investigate the transmission of spatio-chromatic information through a psychophysically inspired cascade of linear+nonlinear layers that address in turn the psychophysical phenomena described above. First we point out that while the analytical description of the model may be useful to describe reduction in *total correlation*, in specific situations (as done before in [30,31]), the estimation of the network *capacity* in general requires empirical estimators that operate on real-valued high-dimensional arrays to cope with arbitrary response PDFs and arbitrary noise schemes. And then, we address this issue using a recent estimator of *mutual information* between multivariate variables based on Gaussianization [32,33].

While the conventional way to check the Efficient Coding Hypothesis goes in the *statistics-to-psychophysics* direction by obtaining biologically plausible behaviors from information-theoretic arguments; here, following [34], we go in the *psychophysics-to-statistics* direction by showing that a biological system that uses no statistical knowledge has remarkably good information-theoretic behavior.

The proposed analysis and tool represent a contribution in visual neuroscience, but they have consequences in image compression as well because they may be used
to change the goal when optimizing current transform coders, which have the kind of architecture considered here [35,36].

2 Materials: illustrative vision model and calibrated images

The use of the Gaussianization tool presented in this work to measure the information capacity of neural channels is illustrated in a standard spatio-chromatic retina-cortex model fed with color-calibrated natural stimuli.

In this section we first review the elements of this standard model, and then, we show the distribution of input stimuli in terms of chromatic contrast, achromatic contrast and luminance. This distribution identifies the regions of the image space that should drive an ecologically-efficient system.

2.1 A standard spatio-chromatic neural pathway

The model considered here for illustrative purposes is a cascade of linear+nonlinear layers [13, 30]. In this setting, the i-th layer takes the array of responses coming from a previous layer \( x^{(i-1)} \), applies a set of linear receptive fields that lead to the responses, \( r^{(i)} \), and these outputs interact to lead to the saturated responses, \( x^{(i)} \):

\[
\ldots x^{(i-1)} \xrightarrow{\mathcal{L}^{(i)}} r^{(i)} \xrightarrow{\mathcal{N}^{(i)}} x^{(i)} \ldots
\]  

Specifically, the model used in the simulations below consist of three of such linear+nonlinear layers:

\[
x^{(0)} \xrightarrow{\text{LMS}} r^{(1)} \xrightarrow{\text{Adapt.}} x^{(1)} \xrightarrow{\text{ATD}} r^{(2)} \xrightarrow{\text{Satur.}} x^{(2)} \xrightarrow{\text{DCT-CSF}} r^{(3)} \xrightarrow{\text{Div. Norm.}} x^{(3)}
\]  

where we can identify the following elements (or processes in each layer):

**Linear spectral integration.** The spectral image, \( x^{(0)} \), is analyzed at each spatial location by linear photoreceptors tuned to Long, Medium and Short wavelengths, in particular we use the standard cone fundamentals LMS in [37]. The \( r^{(1)} \) array contains the LMS retinal images.

**Chromatic adaptation.** We use the simplest chromatic adaptation model: the Von Kries normalization [15], where the linear LMS signals are divided by an estimate of what is considered to be white in the specific scene and re-scaled to a canonical white. The \( x^{(1)} \) array contains the Von Kries-balanced LMS retinal images.

**Linear opponent color space.** The LMS signals at each spatial location are linearly recombined into an opponent representation with Achromatic (luminance), Tritanopic (red-green) and Deuteranopic (yellow-blue) sensors. Specifically, \( r^{(2)} \) contains the ATD images of Jameson & Hurvich opponent sensors [38–40].
**Weber-like saturation.** The linear ATD responses saturate \([41, 42]\) to give, brightness (nonlinear A), and nonlinear versions of images T and D. This saturation can be modeled in sophisticated ways with psychophysical \([15, 43, 44]\) or statistical grounds \([20,21]\), but in \(x^{(2)}\) we will use a simple dimension-wise nonlinearity using a \(\gamma < 1\) exponent with parabolic correction at the origin \([30]\) to avoid singularities in the Jacobian. This exponential with fixed \(\gamma\) is the simplest model for the Weber-like luminance-brightness relation and the observed saturation in chromatic opponent channels \([40]\).

Up to \(x^{(2)}\) the model consists of purely chromatic transforms that operate at each spatial location. In these initial layers spatial context is not considered except for the scarce use made in chromatic adaptation to estimate the white. The final linear+nonlinear layer addresses spatio-chromatic texture.

**Linear texture filters and Contrast Sensitivity.** In the simple model considered here spatial transforms are applied over each A, T, and D images in parallel. Neglecting the interactions between ATD images is consistent with the results found in analyzing the spatio-chromatic statistics of natural images: the chromatic variation of the receptive fields found closely follows Von Kries-corrected ATD directions regardless of the spatial distribution of the receptive field \([25]\). Here we use a crude local-DCT model for the local-oriented receptive fields in V1: the local oscillations applied to the A part of the image account for the achromatic texture sensors, and applied to T and D arrays account for the double-opponent cells \([45]\). The gain of these linear filters (receptive fields) is weighted according to their frequency using achromatic and chromatic Contrast Sensitivity Functions (CSFs) \([46, 47]\). The weights for the local-DCT functions are based on the CSFs of the Standard Spatial Observer \([48]\), defined for sinusoids, and a procedure to transfer the weights from one domain to another \([49]\). The bandwidth of the achromatic and chromatic CSFs is markedly different. The array \(r^{(3)}\) consists of the spatial transforms of the A patch, the T patch, and the D patch, frequency weighted and stacked one after the other.

**Nonlinear interactions between texture sensors.** Following \([13,50]\) the saturation of the sensors tuned to chromatic/achromatic textures has been modeled using a psychophysically-tuned Divisive Normalization \([30, 35]\). Note that Divisive Normalization has been found to be equivalent to neural field models such as the Wilson-Cowan equations \([51]\). The same parameters for Divisive Normalization have been used for the achromatic part and the chromatic parts of the response array. As in the linear case, no interaction has been considered between the A, T and D parts.

An implementation of this illustrative model using the Matlab libraries Colorlab and Vistalab \([52,53]\) is available at [http://isp.uv.es/code/visioncolor/vistamodels.html](http://isp.uv.es/code/visioncolor/vistamodels.html).

Intuition about the information-theoretic performance of the model may be obtained by visualizing the geometrical effect of the series of transforms on the manifold of natural images. This qualitative intuition is discussed in Appendix A.
2.2 Calibrated natural images

The IPL color image database [21, 25, 29] is well suited to study color adaptation because its controlled illumination under CIE A and CIE D65 allows straightforward application of Von Kries adaptation. With the knowledge of the illumination there is no need for extra approximations such as the gray-world assumption to estimate the white. Acquisition of the images in the controlled setting and resulting CIE xy data is illustrated in Fig. 1.

Alternative choices for the database could have been the database of spectral images [54,55] and the color calibrated Barcelona database [56], which also provide the illumination information from calibration gray spheres located in the scenes.

We extracted $19 \cdot 10^6$ image patches from the database (expressed in CIE XYZ tristimulus values) and transformed them into the linear LMS representation. In fact, in this work we will consider the behavior of the model from $r^{(1)}$ on. This amounts to considering that the input to the system is the set of linear LMS images. In order to keep the dimensionality small for a proper comparison of the empirical and theoretical estimates of information quantities done below, we kept the spatial extent small: only $3 \times 3$ pixels. As a result, the input stimuli, $r^{(1)}$, and the response arrays live in $27$-dimensional spaces.

In order to check whether the capacity of the system is adapted to the statistics of the natural input we consider the distribution of the considered samples over three relevant visual features: the average luminance, the achromatic contrast, and the chromatic contrast of the pattern in the image. To define these features LMS images were expressed in the Jameson & Hurvich ATD space using no chromatic adaptation. In this representation average luminance in $cd/m^2$ units is simply the average of the A image, the achromatic contrast was defined as the RMSE contrast of the A image (standard deviation divided by the mean luminance), and the chromatic contrast was defined as the mean of the RMSE contrasts of the T and D images (in
each case, the corresponding standard deviation divided by the norm of the average color in the patch).

We computed these three features for the $19 \cdot 10^6$ images and we found a markedly non-uniform density: image patches are mostly dark, and the variance of the deviations with regard to the mean is small (see PDF in Fig. 2).

This specific distribution highlights the regions of the image space that should be favoured by an efficient processor. In the Results section 4 we plot the capacity of the cortical sensors (at layer $x^{(3)}$ in Eq. 2) over these dimensions of the image space.

3 Methods: capacity from analytic Jacobian and Gaussianization

In a sensory system where the input undergoes certain deterministic transform, $S$, but the sensors are noisy:

$$S \rightarrow x = s(r) + n$$

the transmitted information about $r$ available from $x$ (i.e. the mutual information shared by input and output, $I(r, x)$) is also called the capacity of the system (or channel) $S$ [57].

In this work we are interested in comparing the performance of the system at different locations of the image space with the distribution of natural scenes over that space (e.g. comparing $I(r, x)$, or a convenient proxy, with $p(r)$ in Fig. 2). Moreover, this comparison should be done for different layers of the visual pathway to analyze their relative contribution to the information transmission. Ideally, we’d like to describe the trends of the performance from the analytical description of the response.

In this section, we first discuss up to which point the trends of the capacity over the image space can be predicted from the analytical response. We will see that while the analytical Jacobian may be enough to understand the behavior of the system at a single layer (as done in [30, 31]), the relative relevance of the different layers cannot be easily inferred from the corresponding derivatives. Therefore, an empirical tool is extremely useful in this context.

In order to address empirical the estimation, we describe how the recently proposed Rotation-Based Iterative Gaussianization (RBIG) [32] can be used to estimate the capacity of a channel from input and output samples [33].
3.1 Capacity over the image space and analytical Jacobian

First we make some general considerations on the capacity which, being common knowledge in the literature related to the Efficient Coding Hypothesis [6, 58], are listed here from explicit expressions for the convenience of the reader. More interestingly, we elaborate on these expressions to point out the knowledge about the performance that can be gained from the analytical description of the response.

3.1.1 General considerations on capacity

The amount of information transmitted by the sensory system can be written through two equations of interest in visual neuroscience because they (i) relate the capacity of the system with the neural response $S$, and (ii) identify the goals to be fulfilled by the neural system to optimize its capacity. These equations suggest how certain behavior of the system (transform or saturation) may impact the transmitted information, and they point out surrogates for the capacity to evaluate the performance of the system, such as measuring redundancy instead.

From the definitions of mutual information in terms of entropy and conditional entropy [57], $I(r, x) = h(r) + h(x) - h(r, x) = h(x) - h(x|r)$; and from the concept of total correlation, that describes the redundancy within a vector, i.e. the information shared by the univariate variables [59,60], $T(x) = \sum_i h(x_i) - h(x)$, it is easy to see (appendix B) that the following expressions hold:

$$I(r, x) = h(r) + E_r\left\{\log_2 |\nabla_r S|\right\} - h(n) + E_n\left\{D_{KL}\left(p(s(r))\mid\mid p(s(r) + n)\right)\right\}$$  
(4)

$$I(r, x) = \sum_i h(x_i) - T(x) - h(n)$$  
(5)

where $E_v\{\cdot\}$ stands for expected value over the random variable $v$, and $D_{KL}(p|q)$ stands for the Kullback-Leibler divergence between the probabilities $p$ and $q$.

On the one hand, Eq. 4 tells us about the ingredients that qualitatively determine the transmitted information: the input, the deterministic transform executed by the system (i.e. the response), and the uncertainty of the response (i.e. the noise). Of course the information at the output increases with the complexity of the input, $h(x)$. Regarding the transform, the determinant of the Jacobian with regard to the input, $|\nabla_r S(r)|$, describes how the system changes the volume element of the image space at a certain input $r$. While volume-preserving transforms ($|\nabla S| = 1$, e.g. locally orthogonal transforms) do not contribute to improvements in information transmission, expansions or contractions of regions of the stimulus space ($|\nabla S| > 1$ or $|\nabla S| < 1$ respectively) lead to increases or decreases of the transmitted information about those stimuli. The last two terms depend on the noise: the capacity obviously decreases with the entropy of the noise (which is related to the noise energy [57]), and the last term also depends on the noise in a less obvious way: in the low-noise limit the divergence tends to zero, it monotonically increases with the noise being bounded by the value corresponding to the high-noise limit, where the divergence tends to $D_{KL}(p(s(r))|p(n))$. Note that $E_n\{D_{KL}(p(s(r))|p(n))\} = h(n) - h(s(r))$. This means that the last term $E_n\{\cdot\} < h(n)$ for deterministic responses with positive differential entropy. In summary, when designing a sensory system, the entropy of the input stimulus is given
by the natural environment, and the intrinsic quality of the sensors (their uncertainty) is also given by the biological constraints, therefore, the way to improve the communication performance is choosing the best response function $S$ with the available sensors.

On the other hand, assuming that noise cannot be reduced, Eq. 5 points out what the system should do to improve the capacity: (i) it should increase the sum of entropies of the responses of the individual sensors, and (ii) it should reduce the redundancy (the total correlation, $T$) within the response array. Since increasing the entropies $h(x_i)$ via response amplification is limited by energy constraints, optimization of the first term should come from marginal equalization. These requirements determine the geometric goal of the response transform: the response vectors should follow a factorial code with uniform or Gaussian marginal PDFs.

### 3.1.2 Communication performance from the analytical Jacobian

Eq. 5 identifies univariate and multivariate strategies for information maximization. When trying to assess the performance of a sensory system, reduction of the multivariate total correlation, $T(x)$, seems the relevant term to look at because univariate entropy maximization can always be performed after joint PDF factorization through a set of (easy-to-do) univariate equalizations.

Then, the reduction in redundancy, $\Delta T(r, x) = T(r) - T(x)$, is a possible measure of performance: the system is efficient in regions where $\Delta T$ is big.

Interestingly, this performance measure, $\Delta T$, can be written in terms of univariate quantities and the response model [60]:

$$\Delta T(r, x) = \sum_i h(r_i) - \sum_i h(x_i) + E_r \left\{ \log_2 |\nabla_r S| \right\}$$

$$= \Delta h_m(r, x) + E_r \left\{ \log_2 |\nabla_r S| \right\}$$

Eq. 6 is good for our purposes for two reasons: (1) in case the marginal difference, $\Delta h_m$, is approximately constant over the space of interest, the performance is totally driven by the Jacobian of the response, so it can be theoretically studied from the model, and (2) even if $\Delta h_m$ is not constant, the expression is still useful to get robust estimates of $\Delta T$ because the multivariate contribution may be get analytically from the Jacobian of the model and the rest reduces to a set of univariate entropy estimations (which do not involve multivariate PDF estimations). In the Results section 4, estimates of $\Delta T$ using Eq. 6 are referred to as theoretical estimation (as opposed to model-agnostic empirical estimates purely based on samples Gaussianization) because of this second reason.

In previous works, Eq. 6 has been used to describe the communication performance of Divisive Normalization [30] and Wilson-Cowan interaction [31] on achromatic scenes exclusively from the analytical expressions of the corresponding Jacobian. In both cases, these studies used Eq. 6 to analyze the performance at a single layer, and $\Delta h_m$ was explicitly shown to be constant over the considered domain. Therefore the considerations on the analytical Jacobian certainly explained the behavior of the system.

However, $\Delta h_m$ may not be constant in general, and hence, the trends obtained from the Jacobian of the model can be counteracted by the variation of $\Delta h_m$. 
In particular, the situation gets complicated if one wants to study the relative effect of different layers in the cascade. At a single layer (whose Jacobian accumulates the effect of all previous layers [30]) the marginal difference of entropies with regard to the input may be constant over the domain, and hence negligible. However, reasoning only with the Jacobians in the case of comparisons between multiple layers will only be valid if all the marginal differences of entropy between every layer are constant over the domain. This more strict condition is harder to fulfil in a specific network. For instance, in the illustrative cascade considered in Eq. 2 and used the experiments, this condition does not hold (as shown in the Results section 4.1).

Similar considerations can be made with Eq. 4: we also find this transform-dependent term whose behavior can be successfully analyzed over the considered image space [30, 31]; however, there is no guarantee that the other terms are constant and can be disregarded in the analysis, particularly dealing with comparisons between multiple layers. Moreover, the situation seems worse in Eq. 4 because the terms that should be constant are multivariate in nature, and hence in principle, more difficult to estimate.

Therefore, since the intuition from the analytical response is conclusive only in restricted situations, there is a need for empirical methods to estimate the capacity directly from sets of stimuli and the responses they elicit.

3.2 Channel capacity from Gaussianization

Estimation of $I(r, x)$ directly from samples and using the definitions based on the PDFs is not straightforward: it implies the estimation of multivariate PDFs and this challenging problem would introduce substantial bias in the results. This is also true for surrogates for performance such as $\Delta T(r, x)$.

Using equivalences of $I$ with other information-theoretic quantities the estimation can be reduced to entropy estimations or total correlation estimations. Definitions based on $h$ could use the entropy estimator of Kozachenko-Leonenko [61], as done in [62, 63] to compute the information shared by corresponding scenes under different illuminations.

In this work we solve the estimation of the capacity of the different layers of the visual pathway through the relation between $I$ and $T$. In particular, we use a novel estimator of $T$ which only relies on (easy-to-compute) univariate density estimations: the Rotation-Based Iterative Gaussianization (RBIG) [33].

The RBIG is a cascade of nonlinear-linear layers, each one made of marginal (i.e. easy) Gaussianizations followed by a rotation. This invertible architecture is able to transform any input PDF into a zero-mean unit-covariance multivariate Gaussian even if the chosen rotations are random [32]. As a result both elements of the transform are straightforward: univariate equalizations and random rotations. This ability to completely remove the structure of any PDF makes RBIG a member of the so-called class of density destructors [64]. Density destruction is useful to estimate $T(x)$: as the redundancy of a Gaussianized signal is zero, the redundancy of the original signal, $T(x)$, corresponds to the sum of the individual variations that take place along the layers of RBIG, while converting the original variable $x$, into the Gaussianizer variable $g(x)$. 
Interestingly, the individual variation in each RBIG layer only depends on (easy to compute) univariate negentropies, therefore, after the $L$ layers of RBIG, the total correlation is [32]:

$$T(x) = \sum_{l=1}^{L} \Delta T(x^{(l-1)}, x^{(l)}) = \sum_{l=1}^{L} J_m(x^{(l)})$$  (7)

where the marginal negentropy of a $d$-dimensional random vector is given by a set of $d$ univariate divergences $J_m(v) = \sum_{i=1}^{d} D_{KL}(p(v_i) | \mathcal{N}(0, 1))$. Therefore, using RBIG, the challenging problem of estimating one $d$-dimensional joint PDF to compute $T(x)$ reduces to solve $d \times L$ univariate problems. Moreover, as opposed to Eq. 6, RBIG estimation does not involve any averaging over the whole dataset.

In the density destructor framework, where $T$ is easy to compute using RBIG, $I$ can be obtained by reducing it to a single redundancy estimation. In this work $I$ is estimated using three density destructors as:

$$I(r, x) = T([g_r(r), g_x(x)])$$  (8)

where, $g_r$ and $g_x$ are different Gaussianization transforms applied to the input and the output, and then we compute the $T$ for the concatenated variable $[g_r(r), g_x(x)]$ through an extra destructor.

This procedure is possible because $I$ does not change under invertible transformations (as the density destructors) applied separately to each dataset [57]. Therefore, $I(r, x) = I(g_r(r), g_x(x))$. Since we removed $T$ within each individual dataset by applying individual density destructors, the only redundant information that remains in the concatenated vectors is the one shared by the original datasets, therefore $I(g_r(r), g_x(x)) = T([g_r(r), g_x(x)])$, and hence Eq. 8. See [33] for more elaborate proof.

Note that Eq. 8 involves three different RBIG transforms but only one computation of $T$, as opposed to three RBIGs and three computations of $T$ required if one applies the straightforward relation $I(r, x) = T(r, x) - T(r) - T(x)$. This is an advantage of Eq. 8 to avoid error accumulation.

4 Results

In Section 2 we reviewed the basic elements of conventional retina-cortex models as a cascade of linear+nonlinear neural layers. These different layers use a different representation of the spatio-chromatic content. The general question is: why the image representation is transformed in such a way?

In this Section we study how appropriate these intermediate representations are in information-theoretic terms. Appendix A illustrates the geometrical effect of the different layers on the manifold of natural scenes. That illustration is a qualitative confirmation of the positive effect of the perceptual transforms in multivariate redundancy reduction and marginal equalization, as discussed above on Eq. 5. Beyond the qualitative visualization of Appendix A, here we apply the proposed information measure based on Gaussianization, Eq. 8, to actually quantify the performance of the considered layers: what is the global capacity of certain layer?, how this capacity is distributed over the image space?, is this distribution similar to the PDF of natural scenes?
4.1 Accuracy of the performance estimates based on Gaussianization

Here we show that the estimates of multivariate information-theoretic quantities obtained from RBIG are accurate for the spatio-chromatic data that propagate through the considered neural path.

Computing $\Delta T(r, x)$ with RBIG through Eq. 7 involves training two Gaussianization transforms, one for $r$ and one for $x$, and subtracting two redundancy estimations, $T(r)$ and $T(x)$, each one affected by its own error.

We compare the RBIG estimation with the estimation obtained with Eq. 6, which will be referred to as theoretical because it is based on the analytical expression of the Jacobian, and (as discussed above) leads to a robust estimate of the redundancy reduction because it only involves (easy) univariate PDF estimations.

The comparison shown in Fig. 3 is done in the last Divisive Normalization layer, $x^{(3)}$, that models spatio-chromatic texture sensors at V1. This last layer is useful to check the correctness of the whole architecture because the Jacobian at the last layer includes all the Jacobian matrices of the previous layers.

The computation at each location of the stimulus space is based on the samples belonging to the corresponding bin of the histogram shown in Fig. 2. The results show the average of 10 estimations for each bin. We used 80% of the available samples in each realization. In any case, no more than $5 \cdot 10^5$ randomly chosen samples per bin were considered. Note that the population in the low-luminance/low-contrast bins may be substantially larger, but considering that many samples slows down the estimation. Bins with less than 500 samples (the high-luminance / high-contrast corners of the domain) were discarded in the estimation because results are not reliable. In those low-populated bins we plot a constant value from the boundary of bins with population bigger than 500, but this assignation is arbitrary and these (flat) regions should not be considered in the discussion.

The result shows that the empirical RBIG estimate (top row) closely follows the theoretical estimate (middle row) all over the stimulus space. Note also that the difference between the estimates is small (green surfaces in the plots of the theoretical result), and this difference is similar to the uncertainties both estimates combined (green surfaces in the plots of the RBIG result). The average values of the relative difference and relative standard deviation are 0.12 and 0.09 respectively. Similar agreement is obtained for all the previous layers.

Finally, the term of Eq. 6 based on the analytical Jacobian is shown in the bottom row. This result confirms the point made in Section 3.1.2: the analytical Jacobian not always represents the trends of $\Delta T$. In the specific example in Fig. 3 the analytical term roughly determines the behavior for high chromatic contrast, but it is not the case for low chromatic contrasts.

As stated in Section 3.1.2, in the case of the magnitude that we really want to estimate (the capacity) we do not have alternatives to RBIG to compare with. This is because, as seen above, the analytical Jacobian does not capture all the variability of the problem. Nevertheless, as $I$ can be written in terms of sums of total correlations, one may expect that, given the accuracy of the RBIG estimates for $\Delta T$, the RBIG results for the capacity shown in the next sections can be trusted all over the image space.
Figure 3: Agreement between estimations of redundancy reduction ($\Delta T$, in bits) between the LMS input and the spatio-chromatic representation in V1, estimated via RBIG (top row) and via the theoretical approach (Eq. 6, middle row). The bottom row shows the component of the theoretical $\Delta T$ that comes from the analytical Jacobian. Results are shown for every region of the achromatic contrast/luminance space for different chromatic contrast (different columns). The surfaces in the blue-yellow colormap are the average of 10 estimations in each case. The green surfaces at the plots of the middle row represent the absolute difference between the theoretical and the RBIG estimates. The green surfaces at the plots of the top row represent the combination of the uncertainties of the estimates ($\sigma_{RBIG}^2 + \sigma_{theor}^2$) / 2.
4.2 Redundancy reduction and mutual information through the network

Here we show the redundancy removed when the signal travels from the retina up to the different considered layers, and the amount of information transmitted from the retina up to the different layers.

Fig. 4 shows the redundancy reduction along the network in the achromatic contrast / luminance plane for two chromatic contrasts (the minimum and the maximum in our images). Despite the differences in evolution through the layers for the two chromatic contrasts shown at the figure we can identify general trends which also apply to the other values of chromatic contrast, not shown at the figure: (1) \( \Delta T \) is always positive, i.e. redundancy is effectively reduced by the system. (2) \( \Delta T \) always increases along the pathway, which suggests that inner representations are better than earlier representations in terms of the information captured, and finally, (3) the increments along the way mainly occur at the linear stages, i.e. the transform to opponent color representation, and the transform from the spatial domain to the local-frequency domain. Note that these two stages are the ones that rotate the representation similarly to PCA (see Appendix A).

Nevertheless, suggestions about efficiency given by a small redundancy in the representation have to be confirmed by the values of information about the scene actually available from each layer.

While \( \Delta T \) is computed either theoretically or with RBIG, as detailed in the previous section, estimation of \( I \) needs some assumption on the noise of the sensors in the different layers. The issue of the nature and the origin of uncertainty of the neural response is still under debate [65–67]. The advantage of the RBIG estimate is that it doesn’t rely on a specific analytical PDF of the noise, but only on the availability of noisy samples. It can handle responses corrupted with arbitrary noise sources.

In order to illustrate the capabilities of RBIG in the experiments involving \( I \) we only considered a crude noise model because the procedure would not change using more sophisticated uncertainties in the sensors.

Specifically, we considered the following situation. We assumed that the sensors at the different layers have equivalent signal-to-noise ratio, so we focus on the quality of the representations and not on the quality of the sensors. We assumed that the variance of the noise for each sensor is proportional to the energy of the response. And (for simplicity) we assumed white Gaussian noise with standard deviation given by the 5% of the response deviation, for every sensor and every layer.

Fig. 5 shows the information about the scene (in bits/sensor) available from the different neural layers assuming the same signal to noise ratio of the sensors. Capacity per volume (per bin) of the stimulus space shows the kind stimuli better represented by the different layers.

This computation used the same number of samples, realizations, and bin considerations that were applied in the experiment of Fig. 4 described above.

Only two chromatic contrast are shown in Fig. 5, but the following trends also hold of the other chromatic contrasts omitted in the figure: assuming sensors with the same SNR, (1) the cortical representation is substantially better than the retinal representation since it captures more information about the scene, (2) different intermediate representations are progressively better from retina to cortex, and
Figure 4 Redundancy reduction along the layers of the visual pathway. Results are shown over the achromatic contrast and luminance space for two fixed chromatic contrasts: the minimum (zero, on the top) and the maximum in out set (on the bottom).
Figure 5 Information about the scene available from different layers of the visual pathway. Results are shown over the achromatic contrast and luminance space for two fixed chromatic contrasts: the minimum (zero, on the top) and the maximum in our set (on the bottom). In each case, we assume sensors at the different representations have equivalent noise level (5% of the signal deviation). This result implies that using sensors of equivalent quality, the cortical representation is more appropriate because it captures more information from the input.
improvements of the representation come both from the linear and the nonlinear stages.

It is important to note the differences of this \( I \) result (the interesting measure) with regard to the \( \Delta T \) result (the conventional surrogate). First, while no substantial gain is obtained through the nonlinear stages in terms of redundancy reduction, the improvements in transmitted information due to the nonlinearities are more apparent. Second, the distribution of \( I \) and \( \Delta T \) over the stimulus space is substantially different, and the distribution of capacity seems more matched to the statistics of the natural scenes. The experiments below further elaborate on these points.

In order to investigate the first point further we computed \( \Delta T \) and \( I \) at the different layers for the whole manifold of natural scenes instead of considering images at specific locations of the stimulus space. Note that this computation is not just integrating the surfaces per volume of stimulus domain we have shown before. We computed \( \Delta T \) via the theoretical approach and via RBIG, and we computed \( I \) via RBIG with the same noise assumptions considered above. We randomly choose 0.5·10^6 for each realization from the dataset of 19·10^6 samples. Results in Fig. 6 show the average of 10 realizations with the corresponding standard deviation. Standard deviation is relatively smaller than in Fig. 3 because the number of samples in the estimations is substantially bigger.

Regarding \( \Delta T \) (Fig. 6, left), form a technical point of view, consistency of the RBIG estimate with the theoretical approach based on the analytical Jacobian matrices confirms the accuracy of the RBIG result.

More importantly, these results confirm the progressive improvement of the stimulus representation along the pathway. As useful references for \( T \) and \( I \), the total correlation at the LMS input is 4.08 ± 0.01 bits/sensor; and assuming that 0.01% deviation at LMS is negligible, the upper bound for mutual information would be
$1.90 \pm 0.01$ bits/sensor. Therefore, the cortical representation reduces $82.2\%$ of the redundancy present at the LMS input, and (with sensors with $5\%$ deviation) the V1 captures $60.4\%$ of the available information, as opposed to $29.7\%$ that would be captured by LMS sensors with the same accuracy.

This result confirms the statement made above: gains in redundancy reduction mainly come from the linear transforms: chromatic opponency and the application of local-frequency sensors. Nonlinear transforms seem not to contribute to the communication efficiency if one only looks at the redundancy reduction. However, considering the amount of transmitted information (Fig. 6, right) leads to a different view: nonlinear stages substantially contribute to the increase in capacity.

Spatial transforms definitely have a bigger contribution to the amount of transmitted information than chromatic transforms, despite the fact that tiny patches of $0.05$ degrees were considered in our computations.

4.3 Comparison with the PDF of natural images
Figs. 4 and 5 show that the distribution of $I$ over the stimulus space is substantially different. Take for example the redundancy reduction at the Divisive Normalization layer in V1 and the information transmitted up to that layer in the case of high chromatic contrast: bottom right plots in Figs. 4 and 5 respectively. Note that while both measures indicate better performance for low-luminance stimuli, the (more relevant) capacity measure is substantially more similar to the PDF of natural images (see last plot of Fig. 2).

Fig. 7 explicitly compares the transmitted information up to the Divisive Normalization cortical layer for stimuli at different locations of the image space (top) with the distribution of natural scenes over that space (bottom). This result shows that the considered psychophysically-tuned network (no statistical knowledge was used in this crude biological model) transmits more information in the more frequent regions of the image space: note that the peak of the capacity distribution shifts to higher achromatic contrasts for bigger chromatic contrasts (as the PDF of natural images), and the amount of transmitted information decreases for high chromatic contrasts (as the PDF).

5 Discussion and final remarks

**Efficiency of the network and hierarchy of transforms.** The measurements of transmitted information presented here imply that the cortical representation allows to capture substantially more information about the scene than the retinal representation with sensors of the same signal/noise quality.

We saw that the intermediate layers (or perceptual computations) in the considered retina-cortex pathway progressively improve the signal representation according to the information transmission criterion. Transforms acting on spatial content of the signal lead to bigger improvements in information transmission than the purely chromatic transforms carried out in the first layers. The biggest contribution to the increment in transmitted information is due to the analysis of ATD images through local-oriented filters and Divisive Normalization. In the considered setting, about $70\%$ of the increment in transmitted information comes from the spatial transforms as opposed to the $30\%$ that comes from the chromatic transforms.
Figure 7. Information available at the cortical representation after Divisive Normalization at different regions of the image space (top) compared to the PDF of natural images (bottom).
Interestingly, the cortical representation captures relatively more information (has bigger capacity) in the regions of the image space where natural images are more frequent.

**Capacity via Gaussianization.** From the technical point of view, we pointed out that the characterization of redundancy is not always a good surrogate for the capacity of the system: in fact, in the considered network, behavior of $\Delta T$ and $I$ over the image space are markedly different. Moreover, the trends of the information transmitted may not be inferred from the analytical response. As a result, empirical estimates, such as the proposed measure based on Gaussianization, are necessary. In this regard, Gaussianization estimates of total correlation are consistent with theoretical predictions based on the analytical Jacobian all over the stimulus space.

**Relation to previous work.** The analysis done here has a number of relevant differences with previous work that already analyzed the statistical performance of psychophysically-plausible transforms.

For instance, previous works were limited either because explored a limited range of models or because they used limited performance measures. In [68] the authors addressed the interesting study of the gain that can be obtained from different redundancy reduction transforms, basically using Eq. 6, but restricted the analysis to linear cases to neglect the term that depends on the Jacobian. In [34] authors do consider a more general (nonlinear) model, but their analysis is limited because they didn’t use a multivariate measure for the redundancy, but a set of mutual information measures between pairs of coefficients at the considered layer.

The communication efficiency analysis done here is very related from the technical point of view to the work of Foster et al. who are also concerned about the use of accurate information-theoretic measures to study human vision [63]. The similarity is that they use non-parametric measures of mutual information that operate directly on natural samples as we do. The main difference is their focus on color vision, and specifically on characterizing the performance of humans in different illumination conditions (e.g. determining the number of discriminable colors) [69–71]. This is related to the amount of color information in a scene that can be extracted from color measurements under other illumination [62, 63]. These problems are related to entropy and mutual-information measures (which is the same problem that we address with RBIG), but they do not quantify the information flow through the visual pathway (mutual information between layers and redundancy within layers) that we address here to identify the most relevant layers. As an example, in [70,72] the redundancy is considered only because of its impact on the available information for illumination compensation, not as a measure of information transmission in the visual pathway.

An example of the conceptual difference is that chromatic transforms actually are less important in the information transmission hierarchy that we found when considering spatio-chromatic aspects at the same time. Moreover, note that the crude Von-Kries color adaptation transform is actually the only transform that reduces the available information. Therefore color adaptation may be more related to manifold alignment to improve color-based classification (the kind of goal studied
by Foster et al.) than to improve information transmission (the goal we study here). Despite these differences, the interesting improvements of Kozachenko-Leonenko entropy estimator [61] proposed in [62] should be compared in the future with RBIG estimates because this may have impact in other techniques applicable to spiking neurons [73].

This work originated from the analytical results for total correlation developed for cascades of linear+nonlinear networks [30], and from the analysis of redundancy reduction in Wilson-Cowan networks [31]. In both cases the analysis was restricted to achromatic stimuli. In [30] the approach was totally analytical, while [31] included RBIG estimations for the first time. However, the main difference is that those works didn’t considered the transmitted information which is the actual parameter we want to measure. In fact, in this work we have shown that $\Delta T$ may not be a good descriptor for $I$.

**Consistency between different databases and models.** Despite [31] is purely achromatic and does not consider $I$ (which are crucial conceptual differences), comparison with those results is interesting for different reasons: (1) results of the redundancy at the input retinal representation are comparable (beyond the achromatic/chromatic difference) and some interesting consequences can be extracted. (2) the small gain in redundancy reduction at the Weber saturation and at the Divisive Normalization saturation (also obtained in [31]) has been better explained here.

First, it is interesting to note that this work and [31] use different databases: the colorimetrically calibrated IPL database [21,25], and the radiometrically calibrated database by Foster-Nascimento-Amano [54,55] respectively. Interestingly (for the users of the databases), the redundancy measures at the retinal input are comparable, which means that the statistics of both databases is similar. Specifically, for achromatic patches subtending 0.06 deg in the Foster et al. database the total correlation is about 3.8 bits/sensor [31]. Here, for color patches subtending a smaller angle (0.05 deg in the IPL database) the total correlation is 4.1 bits/sensor. This redundancy is a little bit bigger because it includes color, which is redundant, but not that big because the size is smaller and hence less spatial structure is present, which should increase redundancy too. Moreover, this suggest that consideration of color on top of spatial information increases the redundancy by a small amount, which is consistent with the fact that spatial operations are more relevant in removing redundancy and transmitting information.

Second, the models considered in both works have similar structure but they are not exactly the same: the one here is simpler because it doesn’t have a specific layer for contrast computation. However in both cases small gain in redundancy reduction is obtained at the Weber-like saturation stages and at the cortical Divisive Normalization stage. This consistency of behavior is a safety check for the models, but more interestingly, the analysis of $I$ proposed here and the point made about the advantage of using $I$ as descriptor of performance explains the benefits of these saturations even though they do not contribute to the reduction of redundancy.
Consequences in image quality metrics and image coding  The Visual Information Fidelity (VIF) \cite{74,75} is an original approach to characterize the distortion introduced in an image which is based in comparing the information about the scene that a human could extract from the distorted image wrt the information that he/she could extract from the original image.

The results presented here can be incorporated in that attractive framework in different ways. On the one hand, one may improve the perceptual model including nonlinearities and more sophisticated noise schemes with no restriction because the non-parametric RBIG estimation is insensitive to the complexity of the model. On the other hand, estimations of mutual information in the original VIF scheme made crude approximations on the PDF of the signals to apply analytical estimations, which may be too biased. Better measures of $I$ not subject to approximated models could certainly improve the results.

Following previous tradition of improvements of JPEG/MPEG compression based on Divisive Normalization \cite{35,76}, current state-of-the-art in image coding also uses this kind of perceptually inspired linear+nonlinear architectures \cite{36}. The difference is that current architectures are optimized through the powerful automatic differentiation tools refined for deep-learning \cite{77}. In this case, the encoding and decoding transforms are optimized to minimize simultaneously the bitrate and the perceptual distortion. Nowadays, these two magnitudes have different nature. However, with the considerations done here, VIF distortion, which is expressed in information-theoretic units, could have more meaningful values, and the rate in the image coder could be bounded or modulated by the capacity of the perceptual system, thus leading to a better optimization goal.

Appendix A: Manifold of natural images through the visual pathway

Here we present scatter plots of the spatio-chromatic samples at the different layers of the considered model and the marginal PDFs throughout the network. These transforms illustrate in practice the qualitative comments made in Section 3.1.1 on Eq. 5: in order to increase capacity, redundancy between the responses at each layer throughout the pathway should be reduced, and the marginal PDFs should be equalized.

Moreover, visualization of the marginal PDFs and the corresponding joint PDFs is interesting in case one wants to propose models for the marginals, as in \cite{17,34}, to make analytical estimations of the marginal entropies in Eqs. 5 or 6.

Fig. 8 shows the geometrical effect of the point-wise chromatic transforms that occur at the first layers of the network. We display selected projections of the 27-dimensional arrays in order to illustrate (1) the distribution of the chromatic information (response of color sensors, LMS or ATD, at a fixed spatial location, in the scatter plots of the first row), and (2) the distribution of the spatio-chromatic information (response of color sensors, A, T and D, tuned to different spatial locations, either close spatial neighbors -middle row-, or more distant neighbors -bottom row-).

In the input representation, LMS or $r^{(1)}$ vectors, the responses are strongly correlated, both between sensors of different spectral sensitivity and between sensors
COLOR responses at fixed spatial location  
SPATIO-CHROMATIC resp. at neighbor locations, $p$ and $p + \Delta p$

| $r^{(1)}$ | $x^{(1)}$ | $r^{(2)}$ | $x^{(2)}$ |
|-----------|-----------|-----------|-----------|
| Linear LMS | Nonlinear LMS | Linear ATD | Nonlinear ATD |
| Von-Kries adaptation | Opponency | Opponency |

| CLOSE | DISTANT |
|-------|---------|
| 4.10  | 4.74    |
| 3.30  | 3.25    |
| 2.50  | 2.45    |

| 0.13  | 0.08    |
| 2.91  | 3.42    |
| 2.43  | 2.76    |
| 2.49  | 2.88    |

Figure 8 Image manifold transformed through the chromatic layers. Different rows display projections that represent different aspects of the signal, and different columns show how these aspects transform over the series of layers of the network. In the first row the samples are colored according to the actual color they represent (appropriate rendering was done with Colorlab [81]). The scatter plots in the spatio-chromatic rows actually represent 4 dimensional data: the sample in a point in the 3d-axes represents the response of the corresponding chromatic sensor at the 3 corners of a square, and the color of the sample corresponds to the color of the 4th corner in the square. Correspondence of the color of the sample with the location in the space implies strong correlation between spatial neighbors. Numbers in blue display the 2nd order correlation measure proposed in [82]:

$$C = \frac{1}{2} \log_2 \left( \frac{1}{|\Sigma|} \right),$$

where $\Sigma$ is the covariance matrix and hence, $C$ is the difference between the entropy of the marginal distributions and the joint entropy assuming Gaussian approximations.
tuned to different spatial locations (see the alignment of the scatter plots and the numbers in blue representing 2nd order correlation). This represents the spectral [78, 79] and spatial [80] smoothness of natural scenes. As no chromatic adaptation has been applied yet, the scenes under reddish illumination lead to a distinctly separated blob in the color manifold (top plot of the first column). Note also the lower length of this cluster due to the lower luminance of the CIE A illumination in the experimental setting. In the spatio-chromatic rows (still first column) note that alignment of the scatter plot corresponding to sensors that are closer in space is bigger, consistently with a bigger correlation measure. Not surprisingly, spatial smoothness decays with distance.

Divisive normalization by the white (Von-Kries adaptation, 2nd column) aligns the distinct clusters corresponding to the CIE D65 and CIE A in the input representation (see scatter plot and increased correlation measure), but it does not introduce qualitative changes in the distributions with spatial information.

Transform to opponent channels (3rd column) substantially reduces the correlation between color sensors at a fixed location. It is remarkable how Jameson and Hurvich opponent transform rotates the manifold as a sort of PCA even though it is not based on any statistical consideration. This is consistent with efficient coding interpretations of this stage [18]. Nevertheless, responses of chromatic sensors T and D are still strongly correlated to their spatial neighbors. In fact the spatial correlation of the chromatic sensors is similar to the achromatic counterpart (in previous columns). In the case of chromatic sensors spatial correlation also decays with distance.

Nonlinear saturation of the response at each chromatic sensor (4th column) reduces the correlation even further. Note that here we just picked a crude exponential saturation with fixed exponent following the Weber-like curves of the brightness and opponent mechanisms [41, 42]. However, this psychophysically inspired choice reduced the correlation again, consistently with efficient coding interpretations of this nonlinearity [20, 21]. However, as in previous layers, this chromatic transform also has small effect on the spatial interaction between the sensors (mid and bottom rows of the 4th column). Finally, it is important to mention the effect of this saturation in the joint PDF: note the four-leaf-clover shape of the color manifold projected onto the nonlinear T-D plane (top-right scatter plot). This four-leaf-clover shape is a characteristic consequence of the saturation since it also appears in the spatial transforms based on Divisive Normalization, and has consequences on the bimodal nature of the marginal PDFs (see Figs. 9-10), that had been reported before [34].

Fig. 9 illustrates the effect of the spatial transforms: texture sensors with local oriented receptive fields tuned to different frequencies weighted by achromatic and chromatic CSFs (linear responses \( r^{(3)} \)), and Divisive Normalization (nonlinear responses \( x^{(3)} \)).

In each case, linear and nonlinear, on the left and right panels respectively, we display samples in 3d spaces corresponding to sensors tuned to zero frequency and to two low-frequency components of vertical and horizontal orientation. We also represent samples in 2d projections for a better assessment of the shape of the joint PDF. This is done for the three kinds of chromatic sensors: achromatic, red-green and yellow-blue (on the top, med, and bottom rows respectively).
Figure 9 Transform of the image manifold through the spatial layers. In the 3d scatter plots the stimuli (image patches) are displayed at the locations given by the responses they elicit in the three considered sensors. This illustrates the texture-chromatic meaning of each region of the spaces. On the contrary, 2d projections only display dots of the average color of the corresponding patches so that the elliptically symmetric shape or the four-leaf-clover shape of the densities can be seen.
In this case, the 2nd-order correlation measure, $C$, given in the previous figure (numbers in blue) has not been included because here differences are more subtle: $C$ is basically negligible in all cases and differences are in the level of the estimation error. Actually, differences between these representations have to be described by more appropriate higher-order (multivariate) measures given in the next section: the total correlation and the capacity.

The scatter plots in the first column illustrate how the texture sensors reorient the PDFs of the spatial neighbors of previous layers. In the bottom rows of the previous figure the PDFs were systematically correlated regardless of the chromatic manipulations. The spatial transforms applied to the A, T and D parts of the array $x^{(2)}$ virtually remove the 2nd order correlation. This is consistent with the efficient coding interpretation of this linear filter bank [22,23].

The top left scatter plot shows that stimuli with specific visual features actually give the expected response because it is located in the corresponding region of the space: note for instance the horizontal/vertical patterns of high contrast and different polarity aligned along their corresponding axis, and located along the zero-frequency axis according to their brightness.

As always throughout this Appendix, equivalent plots represent the same sets of samples. Therefore the scatter plots of the second and third rows (still left column) represent the very same set of images as the top-left scatter plot. However, the location of these images in the chromatic texture axes is markedly different because of two reasons: (1) chromatic contrast is smaller than achromatic contrast in natural images (note that images with high chromatic contrast in Fig. 2 are less frequent), and on top of that, (2) the chromatic CSFs has lower bandwidth than the achromatic filter, so the high frequency color oscillations are strongly attenuated by the system. As a result, the images are clustered along the DC color axis. This difference in amplitude of the oscillations can be seen in the extent of the samples over the 2d scatter plots on the left panel. More interestingly, in these 2d plots we can see that the elliptically symmetric distribution that had been reported before for natural achromatic images [83], also happens in the T and D channels.

The right panel shows the result of a Divisive Normalization transform applied to the responses of the linear sensors. Divisive Normalization implies saturation in each dimension but (as opposed to the crude dimension-wise saturation done in the chromatic channels) here the saturation of the response of a sensor for certain stimulus depends on the activity of the sensors tuned to other textures [16, 35]. Saturation is apparent in the 3d scatter plots because the higher contrast patterns in the achromatic plot have been pushed towards the DC axis and the low-brightness region has been expanded. In the chromatic cases saturation is also apparent since the low-saturation region has been expanded.

The 2d scatter plots show how, due to Divisive Normalization, the elliptically symmetric distribution characteristic of local-frequency representations change to this four-leaf-clover shape, as in the case of the color distribution after the saturation (previous figure, top-right scatter plot).

Fig. 10 shows well known results on marginal PDFs [17,19,84,85], together with generalizations to chromatic sensors of bimodal PDFs that had been reported only for achromatic texture sensors [34].
Regarding well known results, Fig. 10 (top left) shows how the first layers contribute to equalization of the PDF of the responses associated to brightness: Von-Kries adaptation expands the range of the responses to the stimuli under darker illumination, and the saturation contributes to reduce the peak at the dark region [19, 84]. In the chromatic case (bottom left), before Von-Kries adaptation the tail in the reddish region is heavier (half of the samples are under the reddish CIE A illumination). Then, adaptation reduces the bias in the PDF, and saturation leads to a bimodal PDF of reduced support in the red-green sensors (yellow-blue sensors behave similarly). The column at the center displays heavy tailed PDFs in the case of linear texture sensors, either achromatic (center top) or chromatic (center bottom) with decreasing variance depending on the frequency [17, 85].

On the other hand, the right column and bottom-left plot of Fig. 10 show how the joint distributions with a mode in each quadrant (four-leaf-clover shapes) in Figs. 8 and 9, lead to the bimodal marginal PDFs. These marginal PDFs are not specific of the considered dataset: they appear in the Van Hateren radiance calibrated dataset and in the Foster-Nascimento-Amano dataset [54, 55], as has been reported in [31, 34] respectively. This shape appears when applying Divisive Normalization saturations with the appropriate exponent: in [34] we proposed a functional form for this marginal related to the parameters of the Divisive Normalization, and this kind of four-leaf-clover joint PDFs and resulting bimodal marginal also appear when optimizing Divisive Normalization for image coding [36] (personal communication from the authors).

According to the discussion done in Section 3.1 analytical modelling of these marginal PDFs can be helpful to characterize the performance. However, in this illustrative section we just want to show new evidences (now for chromatic sensors) about the fact that the family of possible PDFs after Divisive Normalization is not restricted to Gaussian, as sometimes assumed [86].
From a more qualitative perspective, Figs. 8-10 suggest that the considered network does reduce the redundancy between the responses at each layer throughout the pathway and equalizes the marginal PDFs consistently with the comments made in Section 3.1.1 on Eq. 5.

The suggestions made in this Appendix are actually quantified using the proposed Gaussianization tool in the Results section 4.

Appendix B: Derivation of Eqs. 4 and 5

Equation 4. In the noisy channel context of Eq. 3, the capacity of the channel is defined as [57] \( I(r, x) = h(x) - h(x|r) \), so, as the uncertainty of the response given the input is due to the noise, it holds:

\[
I(r, x) = h(x) - h(n) = h(s + n) - h(n)
\]

where we are omitting the input in \( s(r) \) for clarity.

Now, lets address the entropy of the sum. As \( h(a, b) = h(a) + h(a|b) \) [57], it holds,

\[
\begin{align*}
 h(s, s + n) &= h(s) + h(s + n) \\
 h(s, s + n) &= h(s) + h(s + n|s) = h(s) + h(n)
\end{align*}
\]

From the above identities we have, \( h(s + n) = h(s) + h(n) - h(s|s + n) \).

Now, considering the change of entropy under smooth differentiable transforms,

\[
I(r, x) = h(r) + E_r \left( \log_2 |\nabla r S| \right) - h(s|s + n)
\]

Now lets address the conditional entropy term. Note that the definition of the conditional entropy is an average using the joint entropy, therefore:

\[
h(s|s + n) = - \int p(s, s + n) \log_2 p(s|s + n) \ ds \ ds
\]

Now, given the definition of conditional entropy, and as the uncertainty given the input is due to the noise: \( p(s, s + n) = p(s + n|s)p(s) = p(n)p(s) \). And using Bayes on the other probability, \( p(s|s + n) = \frac{p(s + n|s)p(s)}{p(s + n)} = \frac{p(n)p(s)}{p(s + n)} \), we have,

\[
h(s|s + n) = - \int p(n)p(s) \log_2 \frac{p(n)p(s)}{p(s + n)} \ ds \ ds
\]

which is:

\[
h(s|s + n) = \int p(s) \left[ - \int p(n) \log_2 p(n) \ ds \ ds \right] ds - \int p(n) \left[ \int p(s) \log_2 \frac{p(s)}{p(s + n)} \ ds \right] ds
\]

Now, the bracket in the first term is the entropy of the noise (note that \( s \) is fixed), and considering that \( p(s) \) integrates to 1, the first term is just \( h(n) \). The bracket in the second term is \( D_{KL}(p(s)|p(s + n)) \) for a fixed realization of the noise, and the integral is just the expected value over all realizations of the noise,

\[
h(s|s + n) = h(n) - E_n \left( D_{KL}(p(s)|p(s + n)) \right)
\]

and from here we get Eq. 4. ■
**Equation 5** As the joint entropy of a vector is the sum of marginal entropies minus the shared information by all the components, \( h(x) = \sum_i h(x_i) - T(x) \), by substituting \( h(x) \) in Eq. 9, we get Eq. 5. ■

**Declarations**

**Availability of data and materials:** Color-calibrated image database is available at http://isp.uv.es/data_calibrated.html

Code for the visual model is available at http://isp.uv.es/code/visioncolor/vistamodels.html

Code for the RBIG estimator is available at http://isp.uv.es/rbig.html

**Acknowledgements:** I thank the attendees to my lectures on Information theory for Visual Neuroscience and particularly to Vir for encouraging me to write this with the promise of a travel.

**Funding:** This work was partially funded by the Spanish Government through the MINECO grant DPI2017-89867-C2-2-R and by the Generalitat Valenciana through the grant GrisoliaP/2019/035.

**Authors’ contributions:** The corresponding (and only) author completed all the work.

**Consent for publication:** The author supports the publication.

**Competing interests:** The author declares that he has no competing interests.

**Ethics approval and consent to participate:** Not applicable.

**References**

1. MacKay, D.M., McCulloch, W.S.: The limiting capacity of a neuronal link. Bull.Math.Phys. 14, 127 (1952)

2. Barlow, H.B.: Sensory mechanisms, the reduction of redundancy, and intelligence. Proc. of the Nat. Phys. Lab. Symposium on the Mechanization of Thought Process (10), 535–539 (1959)

3. Barlow, H.: Redundancy reduction revisited. Network: Comp. Neur. Syst. 12(3), 241–253 (2001)

4. Dimitrov, A., Lazar, A., Victor, J.: Information theory in neuroscience. J. Comput. Neurosci. 30(1), 1–5 (2011)

5. Friston, K.: The free-energy principle: a rough guide to the brain? Trends Cognit. Sci. 13(7), 293–301 (2009)

6. Tkacik, G., Bialek, W.: Information processing in living systems. Annu. Rev. Condens. Matter Phys. 7, 89–117 (2016)

7. Strong, S.P., Koberle, R., de Ruyter van Steveninck, R.R., Bialek, W.: Entropy and information in neural spike trains. Phys. Rev. Lett. 80, 197–200 (1998)

8. Perge, J., Koch, K., Miller, R., Sterling, P., Balasubramanian, V.: How the optic nerve allocates space, energy capacity, and information. J. Neurosci. 29(24), 7917–7928 (2009)

9. Sengupta, B., Faisal, A., Laughlin, S., Niven, J.: The effect of cell size and channel density on neuronal information encoding and energy efficiency. J. Cereb. Blood Flow Metabol. 33, 1465–1473 (2013)

10. Harris, J., Jolivet, R., Engl, E., Attwell, D.: Energy-efficient information transfer by visual pathway synapses. Current Biology 25(24), 3151–3160 (2015)

11. Borghuis, B., Ratliff, C., Smith, R., Sterling, P., Balasubramanian, V.: Design of a neuronal array. J. Neurosci. 28(12), 3178–3189 (2008)

12. Koch, K., McLean, J., Segev, R., Freed, M., Berry, M., Balasubramanian, V., Sterling, P.: How much the eye tells the brain. Curr. Biol. 16(14), 1428–1434 (2006)

13. Carandini, M., Heeger, D.J.: Normalization as a canonical neural computation. Nature Rev. Neurosci. 13(1), 51–62 (2012)

14. Abrams, A.B., Hillis, J.M., Brainard, D.H.: The relation between color discrimination and color constancy: When is optimal adaptation task dependent? Neural Computation 19(10), 2610–2637 (2007)

15. Fairchild, M.D.: Color Appearance Models. The Wiley-IS&T Series in Imaging Science and Technology. Wiley, Sussex, UK (2013)

16. Watson, A.B., Solomon, J.A.: Model of visual contrast gain control and pattern masking. JOSA A 14(9), 2379–2391 (1997)

17. Simoncelli, E.: Statistical models for images: Compression, restoration and synthesis. In: Farques, M.P., Hennepenstiel, R.D. (eds.) IEEE Asilomar Conf. Sign. Syst. Comp., vol. 1, pp. 673–678. IEEE Comp Soc., Asilomar, CA, USA (1998)

18. Buchsbaum, G., Gottschalk, A.: Trichromacy, opponent colours coding and optimum colour information transmission in the retina. Proc. Roy. Soc. Lond. B Biol. Sci. 220(1218), 89–113 (1983)

19. Laughlin, S.B.: Matching coding to scenes to enhance efficiency. In: In Braddick, O.J. & Sleigh, A.C. (Eds.) Physical and Biological Processing of Images, pp. 42–52. Springer, Berlin, Germany (1983)

20. MacLeod, D., von der Twer, T.: The pleistochrome: optimal opponent codes for natural colors. In: Heyer, D., Mausfeld, R. (eds.) Color Perception: From Light to Object. Oxford Univ. Press, Oxford, UK (2003)

21. Laparra, V., Jimenez, S., Camps-Valls, G., Malo, J.: Nonlinearities and adaptation of color vision from sequential principal curves analysis. Neural Computation 24(10), 2751–2788 (2012)

22. Hancock, P.J.B., Baddeley, R.J., Smith, L.S.: The principal components of natural images. Network 3, 61–70 (1991)
23. Olshausen, B., Field, D.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 281, 607–609 (1996). doi:10.1038/381607a0

24. Doi, E., Inui, T., Lee, T., Wachtler, T., Sejnowski, T.: Spatiochromatic receptive field properties derived from information-theoretic analyses of cone mosaic responses to natural scenes. Neural Comput. 15(2), 397–417 (2003)

25. Gutmann, M.U., Laparra, V., Hyvärinen, A., Malo, J.: Spatio-chromatic adaptation via higher-order canonical correlation analysis of natural images. PloS ONE 9(2), e86481 (2014)

26. Hyvärinen, A., Hurri, J., Hoyer, P.O.: Natural Image Statistics: A Probabilistic Approach to Early Computational Vision. Springer, Heidelberg, Germany

27. Schwartz, O., Simoncelli, E.P.: Natural signal statistics and sensory gain control. Nature Neurosci. 4(8), 819–825 (2001)

28. Malo, J., Gutiérrez, J.: V1 non-linear properties emerge from local-to-global non-linear ICA. Network: Computation in Neural Systems 17(1), 85–102 (2006)

29. Laparra, V., Malo, J.: Visual aftereffects and sensory nonlinearities from a single statistical framework. Frontiers in Human Neuroscience 9, 557 (2015). doi:10.3389/fnhum.2015.00557

30. Martinez-Garcia, M., Cyriac, P., Batard, T., Bertalmío, M., Malo, J.: Derivatives and inverse of cascaded linear-i-nonlinear neural models. PLOS ONE 13(10), 1–49 (2018). doi:10.1771/journal.pone.0201326

31. Gomez-Villa, A., Bertalmio, M., Malo, J.: Visual information flow in wilson-cowan networks. ArXiv: Quant. Biol. https://arxiv.org/abs/1907.13046 (2019)

32. Laparra, V., Camps-Valls, G., Malo, J.: Iterative gausianization: from ICA to random rotations. IEEE Trans. Neural Networks 22(4), 537–549 (2011)

33. Johnson, J.E., Laparra, V., Santos, R., Camps, G., Malo, J.: Information theory in density destructors. In: 7th Int. Conf. Mach. Learn., ICM 2019, Workshop on Invertible Normalization Flows (2019)

34. Malo, J., Laparra, V.: Psychophysically tuned divisive normalization approximately factorizes the pdf of natural images. Neural computation 22(12), 3179–3206 (2010)

35. Malo, J., Epifanio, I., Navarro, R., Simoncelli, E.P.: Nonlinear image representation for efficient perceptual coding. IEEE Transactions on Image Processing 15(1), 68–80 (2006)

36. Ballé, J., Laparra, V., Simoncelli, E.P.: End-to-end optimized image compression. In: 5th Int. Conf. Learn. Repers., ICLR 2017 (2017)

37. Stockman, A., Sharpe, L.T.: The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype. Vision Research 40(13), 1711–1737 (2000)

38. Hurvich, L.M., Jameson, D.: An opponent-process theory of color vision. Psychol. Rev. 64(6), 384–404 (1957)

39. Capilla, P., Malo, J., Luque, M., Artigas, J.M.: Colour representation spaces at different physiological levels: a comparative analysis. Journal of optics 29(5), 324 (1998)

40. Stockman, A., Brainard, D.H.: Color vision mechanisms. In: Bass, M. (ed.) OSA Handbook of Optics (3rd. Ed.), pp. 147–152. McGraw-Hill, NY (2010)

41. Krauskopf, J., Gegenfurtner, K.: Color discrimination and adaptation. Vision Res. 32(11), 2165–2175 (1992)

42. Romero, J., Garcia, J.A., Jimenez del Barco, L., Hita, E.: Evaluation of color-discrimination ellipsoids in two-color spaces. J. Opt. Soc. Am. A 10(5), 827–837 (1993). doi:10.1364/JOSAA.10.000827

43. CIE-Commission: Colorimetry - Part 4: CIE 1976 L*a*b* colour space. Technical Report ISO/CIE 11664-4:2019, CIE (1976)

44. Luo, M., Hunt, R.: The structure of the CIE 1997 colour appearance model (CIECAM97s). Color Res. Appl. 22, 138–146 (1998)

45. Shapley, R., Hawken, M.: Color in the cortex: single- and double-opponent cells. Vision Res. 51(7), 701–711 (2011)

46. Campbell, F.W., Robson, J.G.: Application of Fourier analysis to the visibility of gratings. Journal of Physiology 197, 551–566 (1968)

47. Mullen, K.T.: The CSF of human colour vision to red-green and yellow-blue chromatic gratings. J. Physiol. 359, 381–400 (1985)

48. Watson, A.B., Malo, J.: V1 non-linear properties emerge from local-to-global non-linear ICA. Network: Computation in Neural Systems 17(1), 85–102 (2006)

49. Watson, A.B., Malo, J.: Video quality measures based on the standard spatial observer. In: Image Processing. 2002. Proceedings. 2002 International Conference On, vol. 3, p. 41 (2002). IEEE

50. Carandini, M., Heeger, D.: Summation and division by neurons in visual cortex. Science 258(5077), 336–340 (1992). doi:10.1126/science.1553418

51. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd Edition, 2nd edn. Wiley-Interscience, Hoboken, NJ, USA (2006)

52. Timme, N., Lapish, C.: A tutorial for information theory in neuroscience. eNeuro 5(3), 0052–182018 (2018)

53. Watanabe, S.: Information theoretical analysis of multivariate correlation. IBM Journal of research and development 4(1), 66–82 (1960)
60. Studeny, M., Vejnarova, J.: In: Jordan, M.I. (ed.) The Multi-information function as a tool for measuring stochastic dependence. pp. 261–298. Kluwer, Overijse, Belgium (1998)

61. Kozachenko, L.F., Leonenko, N.N.: Sample estimate of the entropy of a random vector. Probl. Inf. Trans. 23, 95–101 (1987)

62. Marin-Franch, I., Foster, D.H.: Estimating information from image colors: An application to digital cameras and natural scenes. IEEE Trans. Pattern Anal. Mach. Intell. 35(1), 78–91 (2013)

63. Foster, D.H.: The verriest lecture: Color vision in an uncertain world. JOSA A 35(4), 192–201 (2018)

64. Inouye, D.I., Ravikumar, P.: Deep density destructors. In: ICML (2018)

65. Goris, L., Movshon, J., Simoncelli, E.: Partitioning neuronal variability. Nat. Neurosci. 17(6), 858–865 (2014)

66. Moreno-Bote, R., Beck, J., Kanitscheider, I., Pitkow, X., Latham, P., Pouget, A.: Information-limiting correlations. Nat. Neurosci. 17(10), 1410–1417 (2014)

67. Kanitscheider, I., Coen-Cagli, R., Pouget, A.: Origin of information-limiting noise correlations. Proceedings of the National Academy of Sciences 112(50), 6973–6982 (2015). doi:10.1073/pnas.1508738112. https://www.pnas.org/content/112/50/E6973.full.pdf

68. Bethge, M.: Factorial coding of natural images: how effective are linear models in removing higher-order dependencies? JOSA A 23(6), 1253–1268 (2006)

69. Foster, D.H., Nascimento, S.M.C., Amano, K.: Information limits on neural identification of colored surfaces in natural scenes. Visual Neuroscience 21(3), 331–336 (2004). doi:10.1017/S0952523804213335

70. Foster, D.H., Marin-Franch, I., Amano, K., Nascimento, S.M.C.: Approaching ideal observer efficiency in using color to retrieve information from natural scenes. J. Opt. Soc. Am. A 26(11), 14–24 (2009)

71. Marin-Franch, I., Foster, D.H.: Number of perceptually distinct surface colors in natural scenes. Journal of Vision 10(9), 9–9 (2010)

72. Foster, D.H., Marin-Franch, I., Nascimento, S.M.C.: Coding Efficiency of CIE Color Spaces. In: Proc. 16th Color Imag. Conf., pp. 285–288 (2008). Soc. Imag. Sci. Tech.

73. Victor, J.: Binless strategies for estimation of information from neural data. Phys. Rev. E: Stat. Nonlin. Soft. Matter. Phys. 66(5), 051903 (2002)

74. Sheikh, H.R., Bovik, A.C., de Veciana, G.: An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Trans. Im. Proc. 14(12), 2117–2128 (2005)

75. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Trans. Im. Proc. 15(2), 430–444 (2006)

76. Malo, J., Gutiérrez, J., Epifanio, I., Ferri, F.J., Artigas, J.M.: Perceptual feedback in multigrid motion estimation using an improved dct quantization. IEEE Trans. Im. Proc. 10(10), 1411–1427 (2001)

77. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning, MIT Press, Cambridge, MA, USA (2016). http://www.deeplearningbook.org

78. Maloney, L.: In: Gegenfurtner, K.R., Sharpe, L.T. (eds.) Physics-based approaches to modeling surface color perception, pp. 387–422. Cambridge University Press, Cambridge, UK (1999)

79. Jimenez, S., Malo, J.: The role of spatial information in disentangling the irradiance–reflectance–transmittance ambiguity. Geoscience and Remote Sensing, IEEE Transactions on 52(6), 4881–4894 (2014)

80. Simoncelli, E.P., Olshausen, B.A.: Natural image statistics and neural representation. Annual Review of Neuroscience 24(1), 1193–1216 (2001)

81. Malo, J., Luque, M.J.: ColorLab: A Matlab Toolbox for Color Science and Calibrated Color Image Processing. http://isp.uv.es/code/visioncolor/colorlab.html (2002)

82. Cardoso, J.: Dependence, correlation and gaussianity in independent component analysis. J. Mach. Learn. Res. 4, 1177–1203 (2003)

83. Lyu, S., Simoncelli, E.P.: Nonlinear extraction of independent components of natural images using radial gaussianization. Neural Comp. 21(6), 1485–1519 (2009)

84. Huang, J., Mumford, D.: Statistics of natural images and models. In: IEEE CVPR, vol. 1, pp. 541–547 (1999)

85. Malo, J., Ferri, F., Albert, J., Soret, J., Artigas, J.M.: The role of perceptual contrast non-linearities in image transform quantization. Image and Vision Computing 18(3), 233–246 (2000)

86. Moorthy, A.K., Bovik, A.C.: Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE Transactions on Image Processing 20(12), 3350–3364 (2011)