On the relationship between optical variability, visual saliency, and eye fixations: A computational approach

Antón Garcia-Diaz Computer Vision Group, University of Santiago de Compostela, Galicia, Spain

Víctor Leborán Computer Vision Group, University of Santiago de Compostela, Galicia, Spain

Xosé R. Fdez-Vidal Computer Vision Group, University of Santiago de Compostela, Galicia, Spain

Xosé M. Pardo Computer Vision Group, University of Santiago de Compostela, Galicia, Spain

A hierarchical definition of optical variability is proposed that links physical magnitudes to visual saliency and yields a more reductionist interpretation than previous approaches. This definition is shown to be grounded on the classical efficient coding hypothesis. Moreover, we propose that a major goal of contextual adaptation mechanisms is to ensure the invariance of the behavior that the contribution of an image point to optical variability elicits in the visual system. This hypothesis and the necessary assumptions are tested through the comparison with human fixations and state-of-the-art approaches to saliency in three open access eye-tracking datasets, including one devoted to images with faces, as well as in a novel experiment using hyperspectral representations of surface reflectance. The results on faces yield a significant reduction of the potential strength of semantic influences compared to previous works. The results on hyperspectral images support the assumptions to estimate optical variability. As well, the proposed approach explains quantitative results related to a visual illusion observed for images of corners, which does not involve eye movements.

Keywords: optical variability, contextual adaptation, saliency, efficient coding, eye fixations, face saliency, hyperspectral

Citation: Garcia-Diaz, A., Leborán, V., Fdez-Vidal, X. R., & Pardo, X. M. (2012). On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of Vision*, 12(6):17, 1–22, http://www.journalofvision.org/content/12/6/17, doi:10.1167/12.6.17.

Introduction

Biological vision establishes unrivaled benchmarks in terms of efficiency, robustness, and general performance in active visual tasks. These capabilities demand an active and dramatic selection of information that poses a main cause for visual attention. Evidence indicates that bottom-up processing (plenty of adaptation mechanisms) and data-driven saliency play a central role in the control of human visual attention and determine visual priority in cooperation with top-down relevance. The terms saliency as a data-driven property of image points, relevance as a semantic property, and priority as the combination of both are used for the sake of clarity, similar to Fecteau & Munoz, 2006. Concerns on the understanding of the human visual system (HVS) as much as on the development of active vision systems have fostered an important and cross-disciplinary effort to improve the estimation of saliency and the efficiency of low level representations. As a result, the bio-inspired modeling and the applications of saliency have registered a steady increase of research activity.

However, there is a lack of computational models that address the relationship between the contextual data-driven adaptation observed in early visual coding and the perception of saliency. Most existing models decompose the image through its projection on a predefined and fixed basis of low level features. They leave all the adaptive work well in a rigid process of normalization and weighted summation of the initial responses, or well directly in a subsequent local measure of dissimilarity or improbability of the ensemble of features. Approaches to this problem are very interesting for both the understanding of the HVS and for computer vision applications, as far as they may yield improved models of adaptive low level features and saliency.

doi: 10.1167/12.6.17 Received March 7, 2011; published June 12, 2012 ISSN 1534-7362

Furthermore, most models of saliency are grounded either on a bio-inspired hierarchical approach of early visual processing (Itti, Koch, & Niebur, 1998; Le Meur, Le Callet, Barba, & Thoreau, 2006) or on an information theoretic foundation (N. D. Bruce & Tsotsos, 2009; Zhang, Tong, Marks, Shan, & Cottrell, 2008; Seo & Milanfar, 2009). The first group is very conditioned by the interpretation of psychophysical results by the feature integration theory (FIT) proposed by Treisman and collaborators. In an illustrative passage, at the beginning of a reference work related to this theory, Treisman & Gormican (1988, p. 1) state:

> Most theorists agree that the early description derives from spatial groupings of a small set of simple primitives that are registered in parallel across the visual field. These primitives, or functional features, need not correspond to simple physical dimensions like wavelength or intensity.

Models of the second group already point to a more reductionist approach and they ultimately claim to compute an efficient approximation of the inverse of the probability density of the low level content present in the image. However, both approaches lack specification of the physical sources involved, and more importantly, of the different ways in which they contribute to visual saliency.

Remarkably, there are a number of evidences pointing to an invariant behavior of the HVS in the way it manages low level content, and particularly saliency. B. W. Tatler, Baddeley, & Gilchrist (2005) showed that while consistency between subjects decreases over time even without forcing a common starting location, there is no evidence for variation in the discrimination between the saliency at fixated and nonfixated locations. They used a number of specifically modeled low level features to account for saliency. Recent results by Foulsham & Underwood (2008) agree with this observation. In the light of this finding Tatler and collaborators assessed four different hypotheses for the involvement of saliency in the course of time: a) saliency divergence with a relative drop of bottom-up influence in comparison to top-down one as proposed by Parkhurst, Law, & Niebur (2002), b) saliency rank, which means the selection of locations with basis only in saliency such as in the model of attention of Itti et al. (1998), c) random selection with distance weighting independent of bottom-up and top-down processes as proposed by Melcher & Kowler (2001), and d) strategic divergence, which as proposed by the authors means that top-down strategies chosen by observers are different, while the bottom-up frame of reference remains the same. This last possibility is the only one compatible with a decrease in the consistency between observers, even with free starting locations, and the constancy of low level content of fixations over time, both reported in the study. From comparison of eye fixations on natural images between patients with visual agnosia and healthy subjects, Mannan, Kennard, & Husain (2009) showed that consistency between observers in the very first fixations was equivalent for healthy and unhealthy subjects. However, for subsequent fixations, only unhealthy subjects (impaired to understand the image) maintained the consistency between fixation patterns. This result also points to a constant influence of saliency and an increasing and divergent influence of relevance in the spatial distribution of fixations in healthy subjects. All of this suggests invariance in the perception of visual saliency—strictly data-driven—that makes even more interesting the development of efficient computational approaches to yield an accurate estimation of the same.

In previous works we have shown that the decorrelation of local scale features is sufficient to explain a variety of psychophysical results and to predict human fixations at state-of-the-art performance (Garcia-Diaz, Fdez-Vidal, Pardo, & Dosil, 2009). As a generalization, in this paper we propose an estimation of optical variability that involves few magnitudes—intensity, spectral wavelengths, and spatial frequencies. This measure is shown to rely on a contextually adapted representation of the image arising from biologically plausible operations. Unlike most previous approaches to visual saliency that use a fixed basis of features to decompose the input image, the basis of components employed to compute optical variability is adapted for each specific scene. That is, our approach is rooted on a physical rather than only informational theoretic ground and links a parsimonious contextual adaptation of the low level representation to the computation of saliency. Therefore, a major contribution is the explicit proposal of the invariance of the HVS to cope with relative optical variability.

A concrete implementation of such a measure is able to outperform many other state-of-the-art models of saliency in the prediction of human fixations in natural scenes, using two open access eye-tracking datasets. The impact of several approximations is assessed as well and a high robustness in the management of scales and spectral sensitivities is demonstrated. Besides, results on an additional open access dataset of images with faces suggest a lower influence of face relevance than recently reported using a classical model of saliency. Beyond a variety of psychophysical results also reproduced by different previous computational models, the proposed measure is able to quantitatively explain the linearity of perceived saliency versus corner angle. To our knowledge, this behavior is not correctly reproduced by any other model. Overall, the proposed approach exhibits an improved performance and robustness in major benchmarks and yields new

insights in the perception of visual saliency, beyond the reach of existing models.

The paper starts with a presentation of some necessary background information, followed by the definition of optical variability and the proposed hypothesis of invariance of saliency. Next, implementation details, datasets, and evaluation procedures used are described. Then results for each of the four selected experiments are shown. Finally, a discussion of the results and their implications is given.

Background

To obtain the contribution of a particular sample from a set of samples to variability in a multidimensional space, a measure of generalized or statistical distance may be used. It yields the distance to the center of the distribution and thus a measure of sample distinctiveness. Otherwise, the statistical distance can be obtained from the norm of the vector associated to the sample in a decorrelated and whitened representation of the set of samples, that is, on a representation in which the feature basis has been adapted (through shifting, rotation, and scaling of axes) to the statistical structure of the samples, so that the covariance matrix becomes the unity matrix and the mean vector becomes zero.

That is, being $\mathbf{X} = (x_1, \ldots, x_M)$ the original representation (with M components), $\mathbf{Y} = (y_1, \ldots, y_M)$ the whitened representation, and the respective covariance matrices

$$\mathbf{C}_\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1M} \\ \vdots & \ddots & \vdots \\ x_{M1} & \cdots & x_{MM} \end{pmatrix} \quad (1)$$

with, in general $x_{ij} \neq 0$, while

$$\mathbf{C}_\mathbf{Y} = \begin{pmatrix} y_{11} & \cdots & y_{1M} \\ \vdots & \ddots & \vdots \\ y_{M1} & \cdots & y_{MM} \end{pmatrix} = (\delta_{ij}) \quad (2)$$

this whitening transformation can be expressed as a matrix product

$$\mathbf{Y} = \mathbf{W}\mathbf{X} \quad (3)$$

where W is usually referred to as the unmixing matrix.

Under these conditions, the multivariate variance contributed by a given sample can be taken as the squared norm of the Y vector associated to that sample. That is by $||\mathbf{Y}||^2 = \sum_i^M y_i^2$, being M the number of components.

In Fourier optics any image may be regarded as a wavefront piece and therefore approached as a superposition of ideal monochromatic plane waves, as shown in the appendix. The local contribution to such a superposition may be described in terms of the spatial power distributions of chromatic components—related to electromagnetic wavelength—and of the corresponding power distributions of magnitude and orientation of the spatial frequencies present for each of them—related to the wave number vector (i.e., the direction of propagation of the plane wave). Therefore, we could think of each point in the image as a sample with different components of radiant intensity associated to each combination of value of spectral wavelength and of value of 2D spatial frequency.

In a continuous domain the number of components would be infinite and the problem of whitening would be intractable. It is necessary to impose a discretization considering a finite number of possible spectral wavelengths M_λ, of possible spatial frequency radii M_ρ, and of possible spatial frequency angles M_α, with a certain bandwidth on each dimension. As well, only a finite number of image points acting as samples can be considered. Thereby, we assume the corresponding approximations and change integrals by sums in the equations drawn in the appendix.

Thus, we assign each pixel a feature vector whose components are the intensity values at its position for each combination of elementary intervals of spectral wavelength and of radius and angle of spatial frequency. The sum of its components is the total intensity at a given point. That is,

$$\mathbf{f} = (i_{\lambda_1\rho_1\alpha_1}, \cdots, i_{\lambda_{M_\lambda}\rho_{M_\rho}\alpha_{M_\alpha}}). \quad (4)$$

and

$$I = \sum_{j=1}^{M_\lambda}\sum_{k=1}^{M_\rho}\sum_{l=1}^{M_\alpha} i_{\lambda_j\rho_k\alpha_l} \quad (5)$$

This feature vector would play the role of the X original vector in the expression 3 and has $M = M_\lambda \times M_\rho \times M_\alpha$ components.

Using available sensors, it is easy to see that the number of spectral components in the visible range may be rather high. In a typical hyperspectral image like those shown below, this number is several tens. As well, several tens of wavelets, each with a different orientation and a different scale, is also a reasonable number. Therefore, an overall amount of over a thousand components that approximate plane waves is perfectly possible with off-the-shelf current sensors. That is, each vector $\mathbf{X}$ and $\mathbf{Y}$ in Equation 3 would have this number of components, making the rank of the unmixing matrix $\mathbf{W}$ also over a thousand. Typical whitening schemes have a complexity cubic or higher against the number of components while they are linear against the number of samples. That is, a feasible whitening scheme should trade off the number of components involved and the redundancy reduction

achieved to keep a low complexity and get a high performance.

Otherwise, an important issue in the estimation of variability is related to the domain considered. The term *window* is usually employed to refer a given limited portion of the electromagnetic spectrum. As well, it is widely used to refer spatial limits in works in optics and computer vision. Hence, it is used to denote limits in the transmission and reception of optical and visual information from a given domain. Here the term is extrapolated to apply it to the reception of information from the environment by the brain, through the capture and representation of images using the visual system. Therefore, it refers the limited domain of optical magnitudes that the HVS—or any other visual system—is able to sense due to different factors. These limits, discretizations, and thresholds imposed to those magnitudes will constrain any visual transfer function.

Optical variability from adaptive whitening

The approach adopted here to reduce complexity and alleviate computational loads in the estimation of variability consists in whitening separately groups of coordinates, specifically, chromatic and spatial components, and within these last only scales of each orientation. This approach agrees with the hierarchical processing of color and spatial frequencies in the HVS. Overall, we have observed that it even improves the capability of predicting fixations compared to joint whitening strategies. Of course, it is a particular definition of optical variability that assumes that enough reduction of redundancy is achieved by independently whitening chromatic and scale components.

The first step is chromatic whitening. Formally, being M_λ the number of discrete values of spectral wavelengths, $\mathbf{W}_c$ the chromatic whitening unmixing matrix, and $\lambda_i^{'}$ a given whitened spectral wavelength, the idea is to compute the transformation

$$(i_{\lambda_1'}, \ldots, i_{\lambda_{M_\lambda}'}) = \mathbf{W}_c(i_{\lambda_1}, \ldots i_{\lambda_{M_\lambda}}) \tag{6}$$

that is a coordinate transformation in the spectral domain from an original chromatic representation $\mathbf{f}^{'} = (i_{\lambda_1}, \cdots, i_{\lambda_{M_\lambda}})$ to a whitened one $\mathbf{f}'' = (i_{\lambda_1'}, \cdots, i_{\lambda_{M_\lambda}'})$. Besides, similar to Equation 22 of the appendix, we have that the image intensity at each point is the sum of chromatic intensities

$$I = \sum_{j=1}^{M_\lambda} i_{\lambda_j} \tag{7}$$

Thus, the vector $\mathbf{f}'$ may be regarded as a spectral decomposition of the image at a given point. Otherwise, the squared norm in the whitened representation is the statistical distance or T^2 of Hottelling, that is

$$T^2_{chromatic} = \sum_{j=1}^{M_\lambda} i^2_{\lambda_j'} = ||\mathbf{f}''||^2 \tag{8}$$

which is in fact a multivariate measure of variance. Since the samples are the pixel values, each point has a T^2 value that gives its contribution to variance through the ensemble of samples. It is hence a measure of the pixel contribution to variance of chromatic spectral components on the image plane.

Otherwise, the original monochromatic spectral components can be expressed by the discrete version of Equation 23 in the appendix, so that at a given point

$$i_{\lambda_j} = \sum_{k=1}^{M_\rho}\sum_{l=1}^{M_\alpha} i_{\lambda_j\rho_k\alpha_l} \tag{9}$$

As denoted in Equation 6, the whitened spectral components are linear combinations of the original spectral components. As a result, an expression equivalent to Equation 9 can be written for the whitened chromatic components that decomposes each of them as a combination of spatial frequency bands,

$$i_{\lambda_j'} = \sum_{k=1}^{M_\rho}\sum_{l=1}^{M_\alpha} i_{\lambda_j'\rho_k\alpha_l} \tag{10}$$

From this decomposition and for each whitened chromatic component at each pixel, we get a vector of $M_\rho \times M_\alpha$ components $\mathbf{f}_j''' = (i_{\lambda_j';\rho_1,\alpha_1}, \cdots, i_{\lambda_j';\rho_{M_\rho},\alpha_{M_\alpha}})$. Each of these representations of whitened components can be further whitened, using as original coordinates those of the spatial frequency bands. Instead of such an approach, a simplification is adopted here. Whitening is proposed for each set of spatial frequency bands at a given spatial frequency angle,

$$(i_{\lambda_j'\alpha_l\rho_1'}, \cdots, i_{\lambda_j'\alpha_l\rho_{M_\rho}'}) = \mathbf{W}_{jl}(i_{\lambda_j'\alpha_l\rho_1}, \cdots, i_{\lambda_j'\alpha_l\rho_{M_\rho}}) \tag{11}$$

which reduces the number of components involved in whitening to M_ρ (i.e., the number of scales). Therefore, the rank of every unmixing matrices $\mathbf{W}_{jl}$ is reduced to a maximum of M_ρ. Otherwise, we have as many transformations as the product of the number of chromatic components by the number of orientations, that is $M_\lambda \times M_\alpha$ parallel transformations.

As a result, we have a novel representation that assigns to each pixel a vector $\mathbf{f}''''$ of $M = M_\lambda \times M_\rho \times M_\alpha$ components that are partially whitened. We estimate the optical variability OV contributed by each point to the whole image as the squared norm of this vector.

That is,

$$OV = \sum_{j=1}^{M_c}\sum_{k=1}^{M_\rho}\sum_{l=1}^{M_\alpha} i^2_{\lambda'_j\rho'_k\alpha_l} = ||\mathbf{f}''''||^2 \quad (12)$$

Unlike for Expression 8 obtained from color whitening, this result is not the T^2 of Hotelling of the original components. It is an approximation that arises from the summation of the T^2 obtained for different subsets of original coordinates. It is worth noting that the approximations adopted did not reduce the effectiveness in explaining visual behavior in the experiments described below.

The saliency of a given point is computed as a relative measure of optical variability, that is, considering N points (pixels) in the image, the saliency of one point (p) is given by:

$$S_p = \frac{OV_p}{\sum_{p=1}^{N} OV_p} \quad (13)$$

Therefore, saliency may be interpreted as a measure of the probability density for a point to be attended or broadly as a measure of the strength of point distinctiveness.

It is worth remarking that the described approximations in the computation of variability are inspired in coarse features of the HVS, namely, the independent hierarchical processing of color and spatial information, as well as orientation specific contextual interactions. Indeed, the representation on which the measure of variability relies is whitened for the spectral and spatial structures of a specific scene. The basis of features is thus adapted for the ensemble of local feature values in the particular image. Therefore, it is a retinotopic representation adapted to the specific visual context.

The flowchart in Figure 1 summarizes the proposed procedure to compute optical variability.

Some further approximations

A simple characterization of an image closely related to its optical description in spatial frequencies can be formulated in terms of local energy components at different scales and orientations, thus different values of radius and angle of spatial frequencies for different spectral components. The relation 7 is not true for a nonorthogonal wavelet decomposition but it can be taken as a reasonable approximation. Besides, the accuracy in that relation is not essential in our analysis, but what is really important is the reliability of the resulting whitened components. We have observed that the computation of whitening through principle components analysis (PCA) and independent components analysis (ICA) in our scheme is barely affected by the overlapping between the original filters in the Fourier domain. Distinctiveness of a given point taken as a sample would be easily computed through the norm in the hierarchically whitened representation.

The only remarkable difference that would remain in comparison to a coarse visual processing scheme is the use of monochromatic spectral components rather than broadband overlapping trichromatic components (like LMS or RGB). Going a step further, an additional approximation consists in using the responses to such broadband spectral detectors rather than to narrow spectral bands, for instance, to use (r, g, b) components instead of the narrow spectral components $(\lambda_1, \cdots \lambda_{M\lambda})$. We can apply exactly the same whitening schemes proposed above and we can take the resulting norm at each point in the image as a measure of relative variability or distinctiveness. Otherwise, the implications of this approximation will be examined in an experiment involving hyperspectral images in the visible spectrum. There, results using narrow spectral components and responses to broad detectors will be compared and analyzed.

If we think of saliency as an objective measure captured by the HVS as a result of an adaptive neuro-optical transfer function, then saliency must be the same for different subjects with the same visual window when observing the same image. Indeed, in the approximations pointed above, broad sensitivities against chromatic wavelengths and spatial frequencies, discretizations, and separate dimensions for whitening can be seen as neural constraints acting on the definition of the visual window.

Connection to contextual adaptation

The classical receptive fields of early visual cortex cells are tuned to different scales and orientations, as described by Hubel & Wiesel (1968). Meanwhile, color opponencies are characteristic of responses to color along the early visual pathway. This classical scheme has been interpreted as coding the independent or sparse components of natural images (Olshausen & Field, 1996; Hoyer & Hyvärinen, 2000). Furthermore, color opponencies have been shown to emerge from efficient representations of hyperspectral images of natural scenes (Lee, Wachtler, & Sejnowski, 2002). That is, the classical receptive fields can be seen as whitened components of the set of natural images. Therefore, the long-term adaptation that produced such receptive fields can be related to decorrelation and interpreted as a mechanism for efficient coding of natural scenes, as early proposed by Barlow (1961) and Barlow & Foldiak (1989).

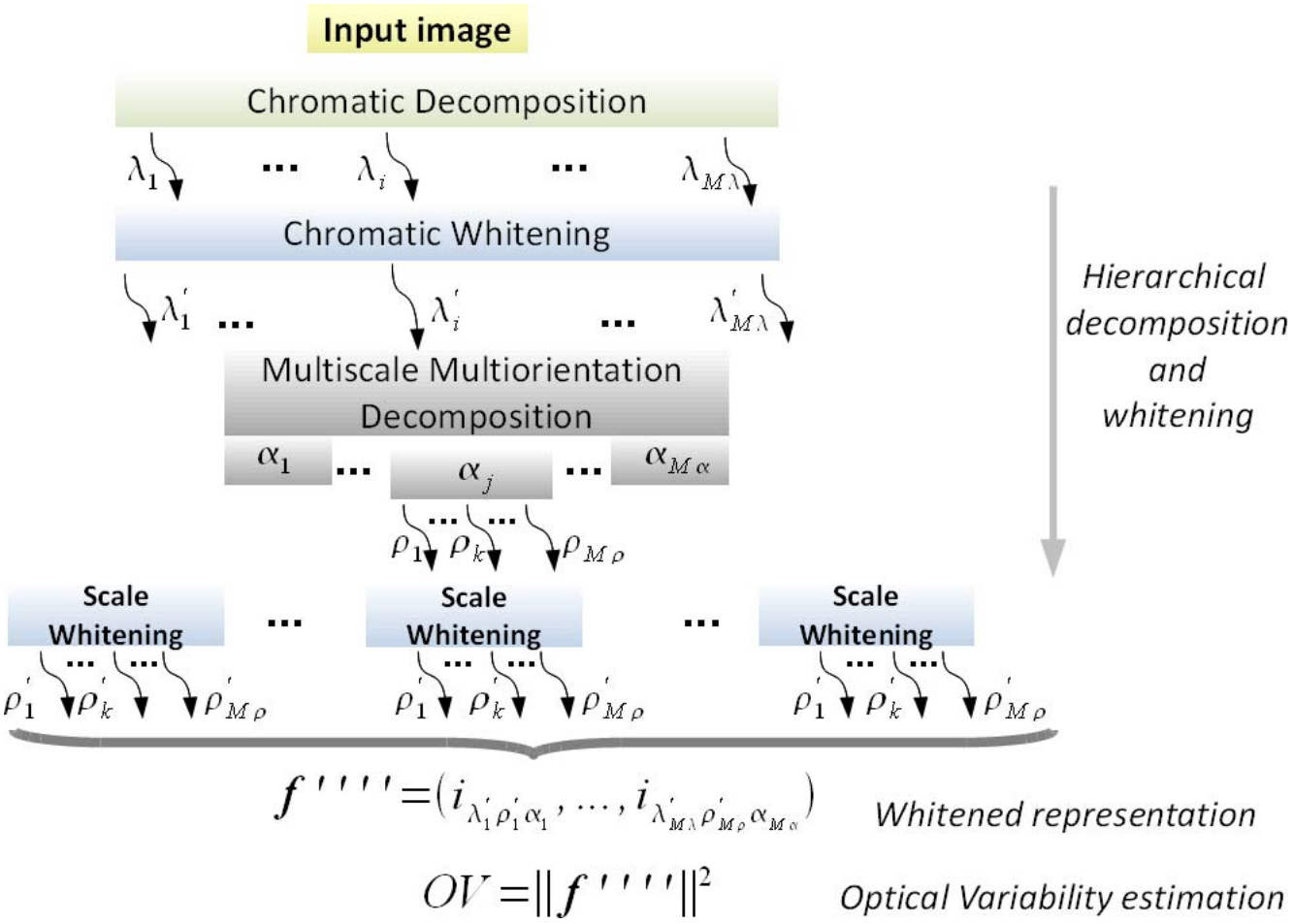

Figure 1. Estimation of optical variability through partial whitening.

The high redundancy in specific scenes due to the presence of objects with a particular color, texture, and shape is a classical observation of vision science (Attneave, 1954). Moreover, such redundancy, or in other words the remarkably restricted feature distributions that characterize specific images, constitutes a powerful stimulus for a short-term and contextual adaptation of neural coding (Barlow & Foldiak, 1989; Webster & Mollon, 1997).

A variety of neural mechanisms of temporal and contextual adaptation (e.g., the contextual adaptation to spatial frequencies outside the classical receptive fields) have been described all across the early visual pathway and beyond (Rieke & Rudd, 2009; Kohn, 2007; Clifford et al., 2007). One of the main functional benefits of this adaptation is thought to be the decorrelation of cell responses in order to improve representational efficiency. Moreover, adaptation effects are similar on a wide range of time scales with longer stimulation producing stronger effects (Kohn, 2007). Indeed, neural adaptation under natural stimulation has been shown to produce overall a decorrelation of neural responses (Vinje & Gallant, 2000; Ecker et al., 2010; Atick, Li, & Redlich, 1993).

As pointed by Schwartz, Hsu, & Dayan (2007), adaptation mechanisms pose a decoding ambiguity, since the receivers of a neuron response are supposed to be unaware of the adaptation operated at that neuron. They call this ambiguity the *coding catastrophe*. A correlate of this catastrophe can be found in the perceptual adaptation underlying a variety of visual illusions.

The measure of optical variability proposed in the previous section may be regarded as a coarse approximation to contextual adaptation through a hierarchical whitening of classical receptive fields. Indeed, it is a simple approach that aims to produce an efficient representation of the visual input. The starting whitening of color features proposed finds an early antecedent in the mechanistic model described by Atick et al. (1993) to explain perceptual results on color adaptation. Their model was in essence a neural network able to compute decorrelated components thanks to lateral feedback connections. Likewise, the independent whitening of scales for different orientations may be related to the lateral interactions required to handle with spatial contexts (Schwartz et al., 2007).

Hypothesis: invariance of saliency in bottom-up visual processing

Otherwise, a criticism to the efficient coding hypothesis relies in the fact that it *does not address why the coding catastrophe occurs because it lacks specification as to the computational goal beyond representation; rather, it embraces it without further question* (Schwartz et al., 2007).

Grounded in the proposed definition of optical variability and the link to contextual adaptation, we propose such a specification. A major goal underlying representational efficiency, and by extension the corresponding contribution to contextual adaptation, is to ensure the invariance of the bottom-up behavior elicited by optical variability in the image. Saliency as a constrained measure of relative optical variability in the visual window is hence hypothesized as an invariant in biological visual systems.

Some examples of the contextual adaptation of scales

To show the effect of scale whitening on low level features before saliency computation, we show a comparison of adapted versus classical responses in few illustrative examples. The responses (both before and after scale adaptation) shown here have been obtained using the implementation for RGB images described in the next section.

Overall, the simple adaptation scheme proposed has been observed to produce a kind of figure-ground separation in different components that may be linked to perceptual grouping as shown in Figures 2 and 3.

Furthermore, the contextual adaptation proposed with the goal of computing optical variability has been also observed to catch certain illusory contours that are not apparent in nonwhitened features. Two illustrative examples are given in Figure 4.

Materials and methods

We have tested the hypothesis formulated in the previous section in four different experiments that aim to tackle some major issues related to visual saliency. First, we compare the performance of the proposed approach in predicting fixations in two different open access eye-tracking datasets. As well, robustness against spatial resolution of the input image is studied. Next, a novel experiment is proposed that aims to catch the impact of the trichromatic approximation in computing optical variability. In third place, the approach is tested on a dataset of images containing human faces and the influence of relevance versus saliency is revised in the light of the results. Finally, a psychophysical result that does not involve eye movements but only perceptual comparisons is shown to be quantitatively explained by the proposed measure of saliency, for the first time to our knowledge.

Implementation details

In the specific implementation used in the following experiments we have performed whitening through the computation of principal components analysis and the normalization of the resulting components by the standard deviation. That is, if Z is a basis of features that results transforming the input X through PCA, the corresponding covariance matrix is diagonal

$$C_Z = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_M \end{pmatrix} \tag{14}$$

Each element of the trace is the eigenvalue of a corresponding eigenvector that also gives its variance. Then, normalizing the components of Z by the square root of the corresponding eigenvalue, that is by doing $y_i = z_i/\sqrt{\lambda_i}$, the elements of the diagonal become the unity. Consequently the covariance matrix for the resulting Y coordinates also becomes the unity matrix satisfying Expression 2. Thus, the overall variance of the ensemble of samples (i.e., all the pixels) is the unity for each of the transformed components. We have also tried ICA for whitening but the results were equivalent in the most favorable cases for ICA. Thereby, whitened principal components were chosen because of both their higher computational lightness and their slightly better performance.

The input color components were RGB components for all of the experiments; hyperspectral reflectance components were used as input in the second experiment.

For the spatial decomposition in spatial frequency bands (multiscale and multi-orientation) we have used a measure of local energy from the modulus of the complex responses to a bank of logGabor filters. These filters only have analytical expression in the frequency domain. Their transfer function is

$$\log Gabor_{so}(\rho,\alpha) = \exp\left(-\frac{\left(\log(\rho/\rho_s)\right)^2}{2\left(\log(\sigma_{\rho s}/\rho_s)\right)^2}\right) \cdot \exp\left(-\frac{(\alpha-\alpha_o)^2}{2(\sigma_{\alpha o})^2}\right) \tag{15}$$

The particular details of design have been thoroughly described in Garcia-Diaz et al. (2009). Figure 5

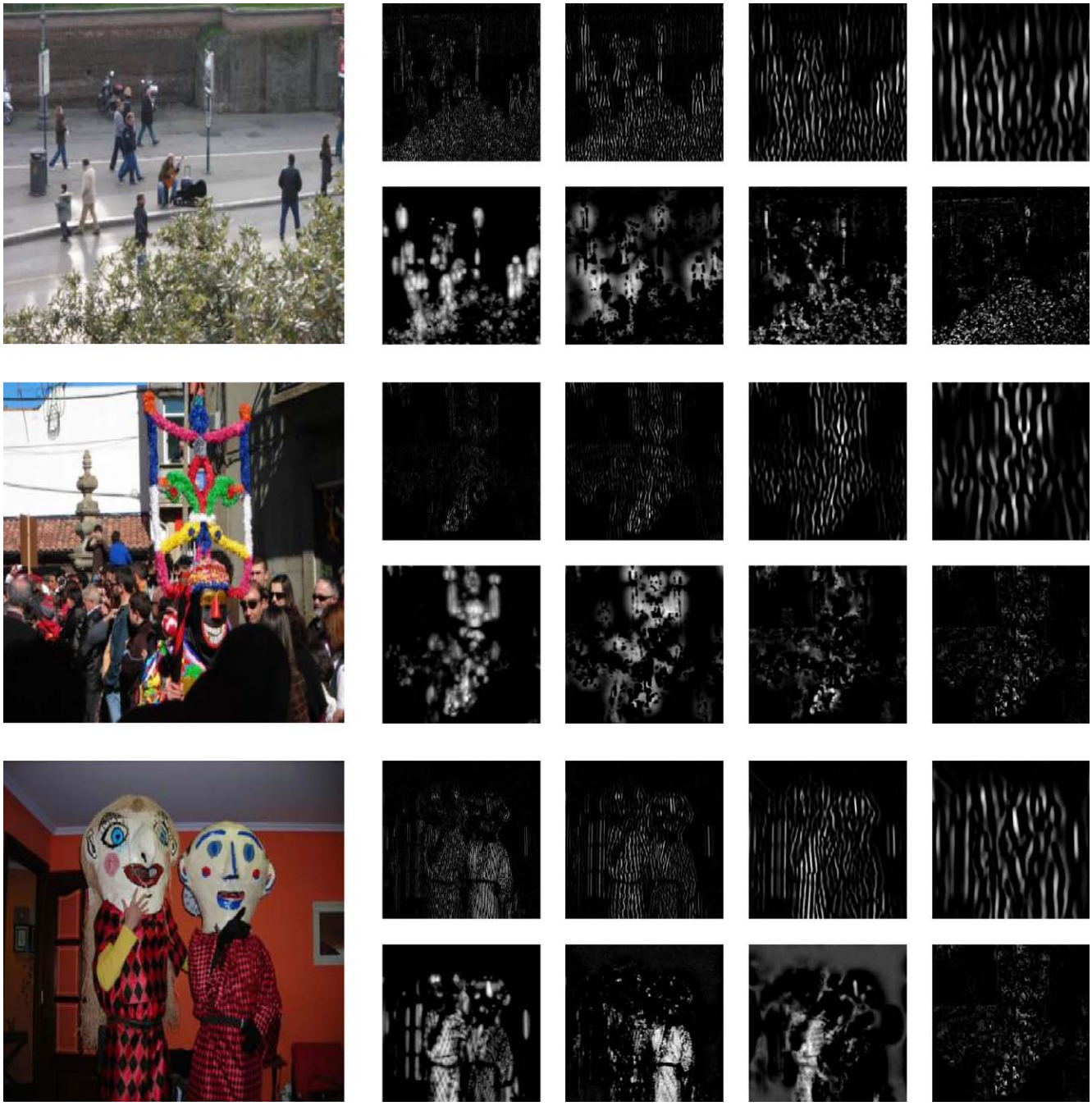

Figure 2. Examples on three natural images of classical low level features based on band-pass filters (top row) and the corresponding whitened (adapted) features (bottom row).

shows a flowchart of the specific implementation used with RGB images.

It must be noted that the same implementation without any specific tuning has been used in all the experiments. The whitening procedure and the bank of filters used for spatial decomposition were exactly the same. The only modified parameters were the size of the input image—varied as in the other models—in the first experiment and the number of whitened chromatic components in the second experiment that involved hyperspectral images. In the following the proposed estimation of optical variability is shortly referred to as AWS (adaptive whitening saliency).

Datasets

In the first experiment, two open access eye-tracking datasets of natural images were used. The first was published by Bruce and Tsotsos and has 120 images

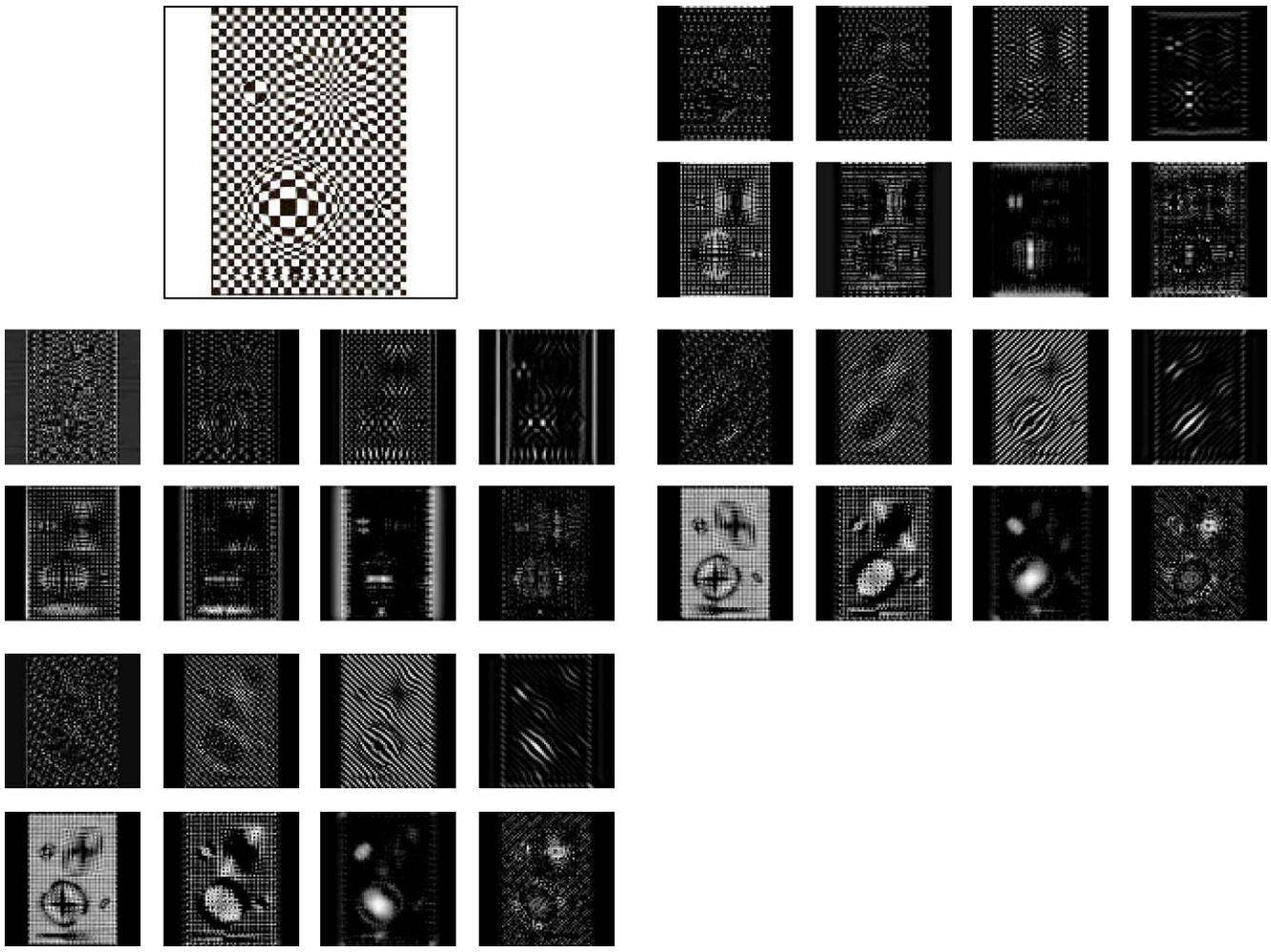

Figure 3. Example with an artistic Op-Art representation. Each of the four blocks of filter responses shows the classical low level features based on band-pass filters (top row) and the corresponding whitened (adapted) features (bottom row) for four different orientations.

and fixations from 20 subjects (Bruce & Tsotsos, 2006). This dataset has already been used to validate several state-of-the-art models of bottom-up saliency, for instance in Bruce & Tsotsos (2009); Gao, Mahadevan, & Vasconcelos (2008); Zhang et al. (2008); Seo & Milanfar (2009); and Hou & Zhang (2008). The second dataset was published by Kootstra et al. and consists of 99 images and the corresponding fixations of 31 subjects (Kootstra, Nederveen, & de Boer, 2008; Kootstra & Schomaker, 2009). The main purpose of the use of different datasets in this work is to assess the robustness and reliability of the evaluation procedure. Several example images of both datasets are shown in Figure 6.

In order to analyze the impact of the trichromatic approximation on the estimation of variability, we have conducted a novel eye-tracking experiment on a reduced set of open access hyperspectral images of close range, nonaerial natural scenes. The set of images employed comprises eight calibrated hyperspectral cubes with the surface reflectance of different scenes for 32–33 narrow spectral bands in the visible range. Details of the acquisition procedure are given in Foster, Nascimento, & Amano (2005). For each cube the authors have provided an RGB representation of the scene obtained from the simulation that arises from applying a natural illuminant (daylight at 4000 Kelvin) on the measured reflectances. The cubes and the corresponding images have been cut from the left to fit them to a maximum screen resolution of 1280×1024. This operation is intended to avoid any downsampling that would alter the original scales. The eight RGB images and hyperspectral cubes are shown in Figure 9. Seven subjects observed the eight RGB images shown in random order from a distance of 62 centimeters and their eye movements were recorded with a SMI eye-tracker. A 19-inch liquid crystal display (LCD) screen (Samsung 943B, LS19MYBESQ/EDC) was employed. Each image was shown for four seconds. Between each pair of scenes a dark blank screen with only a randomly positioned bar was shown for 1.5 seconds to remove possible aftereffects while keeping the gaze of observers

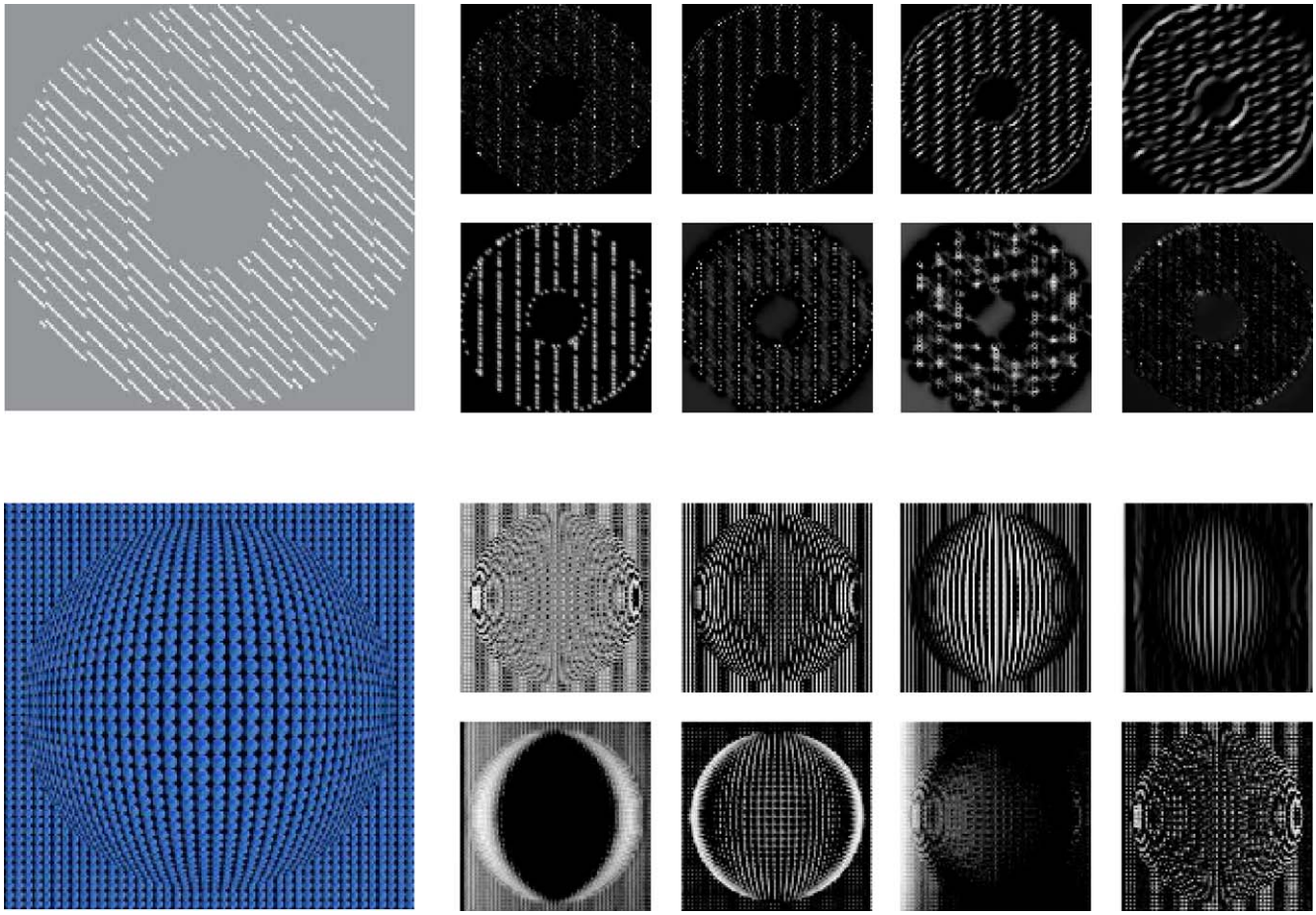

Figure 4. Examples of illusory contours on two images. For each image, the classical low level features based on band-pass filters (top row) and the corresponding whitened (adapted) features (bottom row), are shown.

on the screen. Fixations were extracted from recorded data through the analysis with the BeGaze software package using the default parameter values. Subjects were instructed to freely observe the shown scene.

In the third experiment we used the fixations in faces database (FIFAD) to revise the relative strength of saliency versus relevance in the light of the proposed approach to saliency computation. It comprises 200 images and the corresponding eye-tracking data for 8 subjects made available to the public by Cerf, Frady, & Koch (2009). In that work they described thoroughly the dataset and the results obtained in predicting fixations with a classical measure of saliency and semantic maps among other experimental observations. Therefore, the dataset includes semantic maps in which regions containing faces have been labeled by hand.

In the fourth experiment we used six images of corners in a grayscale gradient that present six different corner angles (30°, 45°, 75°, 105°, 135°, and 180°). These images have been used in a study on human subjects in which perceived saliency was estimated through perceptual comparisons and compared with the responses to difference of Gaussians filters (Troncoso, Macknik, & Martinez-Conde, 2005). Instead of such responses, we have kept the same procedure of comparison and used the saliency maps obtained for each of the images. The procedure involved taking the value of saliency at the central point of the grayscale gradient that makes the corner, where observers were instructed to look at for comparison with a standard stimulus stripe.

Measure of performance

The saliency maps have been compared with human fixations through the use of the area under the curve (AUC) obtained from a receiver operating characteristic (ROC) analysis, as proposed by Tatler et al. (2005). The method has been employed to validate a wide variety of state-of-the-art saliency models, providing a reliable measure for performance evaluation. In this procedure, one unique curve is drawn for a whole set of images. The area under this curve can be used to measure the capability of saliency to discriminate between fixated and nonfixated points. To avoid center-bias, in each image only points fixated in another image from the same dataset are used as

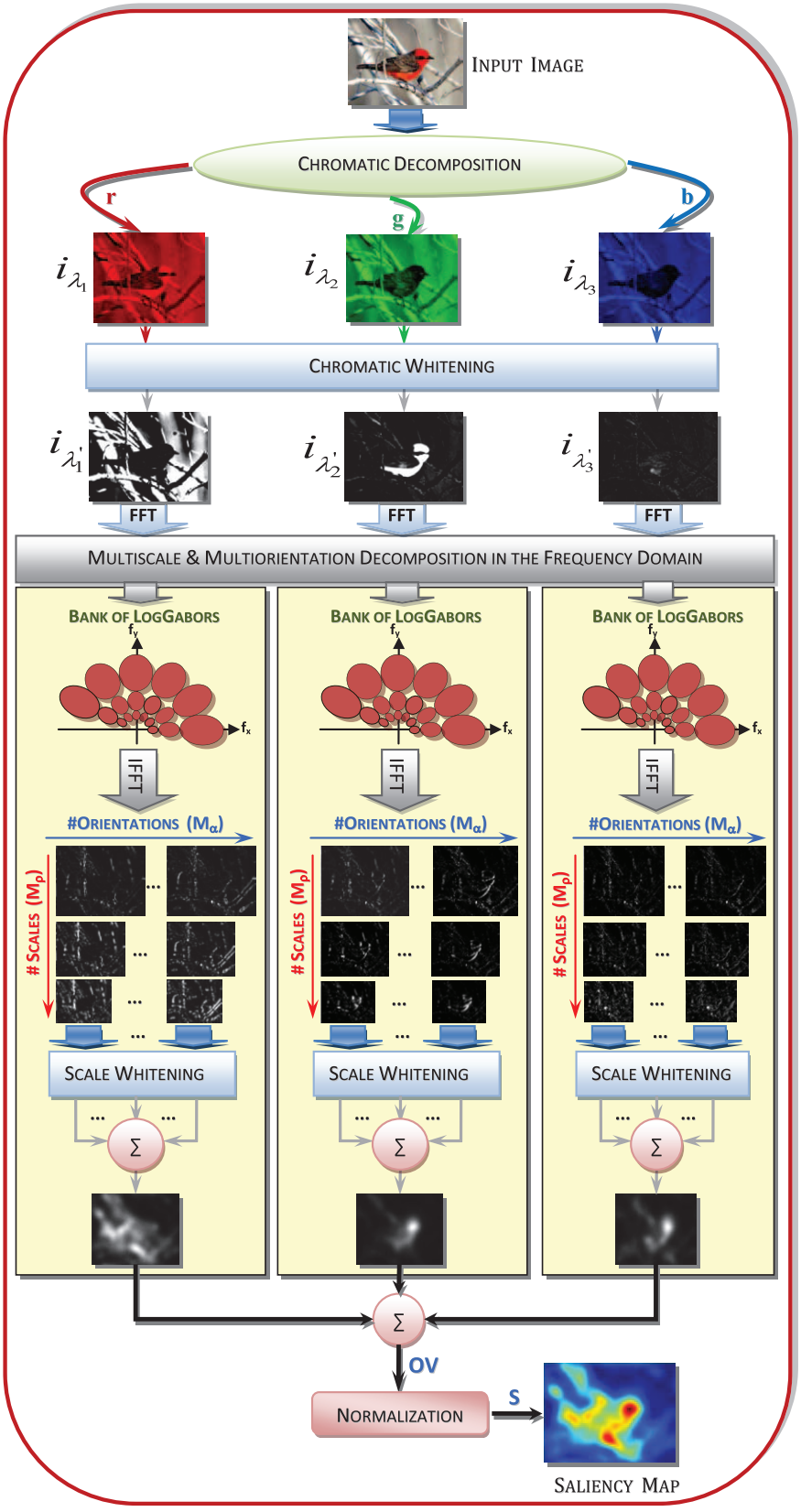

Figure 5. Optical variability and saliency estimation from RGB images. The chromatic components $(i_1, \cdots, i_M)$, are approximated by (*r, g, b*) components, being r,g, and b the red, green and blue (broadband) components.

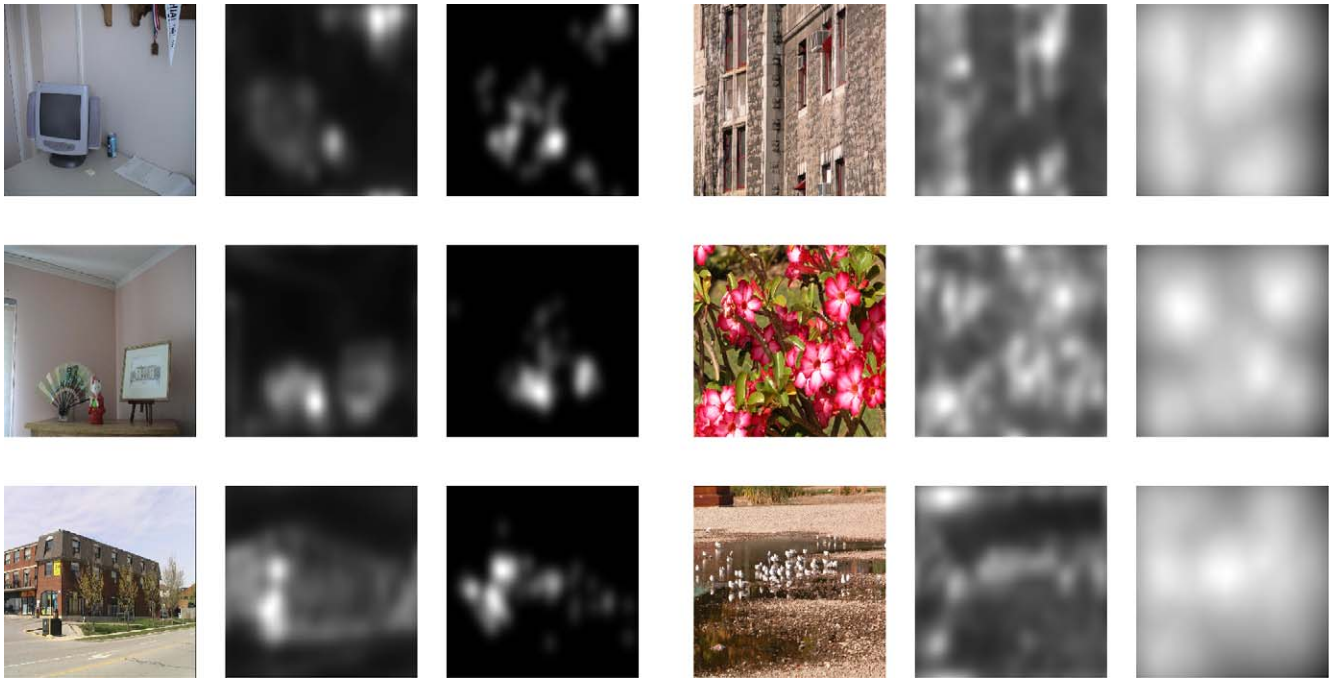

Figure 6. Three examples of images from each of the datasets of Bruce and Tsotsos (right) and Kootstra et al. (left). For each image the saliency map using AWS (center) and the fixations density map (left) provided by authors are shown. Note that fixations density maps have been computed from fixation locations through Gaussian kernels by Bruce & Tsotsos (2009) while Kootstra & Schomaker (2009) have used instead a distance transform, explaining the noticeable differences in aspect.

nonfixated points. As suggested by Tatler et al. (2005), standard error is computed through a bootstrap technique, shuffling the other images used to take the nonfixated points, exactly like in Zhang et al. (2008) and in Seo & Milanfar (2009), that is, the particular implementation by Zhang et al. (2008) following the method proposed by Tatler et al. (2005) has been adopted.

This choice is motivated by two main reasons. First, it has been recently used to assess several state-of-the-art models both by Zhang et al. (2008) and by Seo & Milanfar (2009). This fact clearly facilitates comparison in a fair fashion with existing approaches. Second, it is robust against tricks like border suppression used in many models.

Results

Robustness of performance against spatial resolution

An interesting aspect related to the performance of a measure of saliency is the impact of spatial resolution on its capability of predicting fixations. Therefore, we have measured the AUC values for several models for different spatial resolutions of the input image, expressed in pixels by degree of visual angle, in terms of the visual field observed by subjects in the specific eye-tracking experiments. All the models used for comparison are state-of-the-art models with the code made available by the authors. Specifically, these models are the graph-based visual saliency (GBVS) by Harel, Koch, & Perona (2007), the AIM model by N. D. Bruce & Tsotsos (2009), the model of saliency using natural images statistics (SUN) by Zhang et al. (2008), the model of saliency from self-resemblance (Sfr) by Seo & Milanfar (2009), and finally the classic model of saliency by Itti et al. (1998). Some of these models have recently reported the best results in predicting human eye fixations in natural images using open access datasets. As well, the ensemble yields a representative selection of the existing paradigms under the different bio-inspired approaches to the computation of visual saliency. We have used the code as it is, without altering the default values, except the image downsampling that has been varied in a wide range of values to test robustness against input resolution.

The results obtained are shown in Figures 7 and 8. Clearly, the AWS model presents not only the highest maximum performance on both datasets but also an unrivaled robustness against the spatial resolution of the input image. This fact reveals that, unlike other state-of-the-art approaches, the AWS model is not biased to deal with certain scales that are most often involved in the determination of saliency. As expected

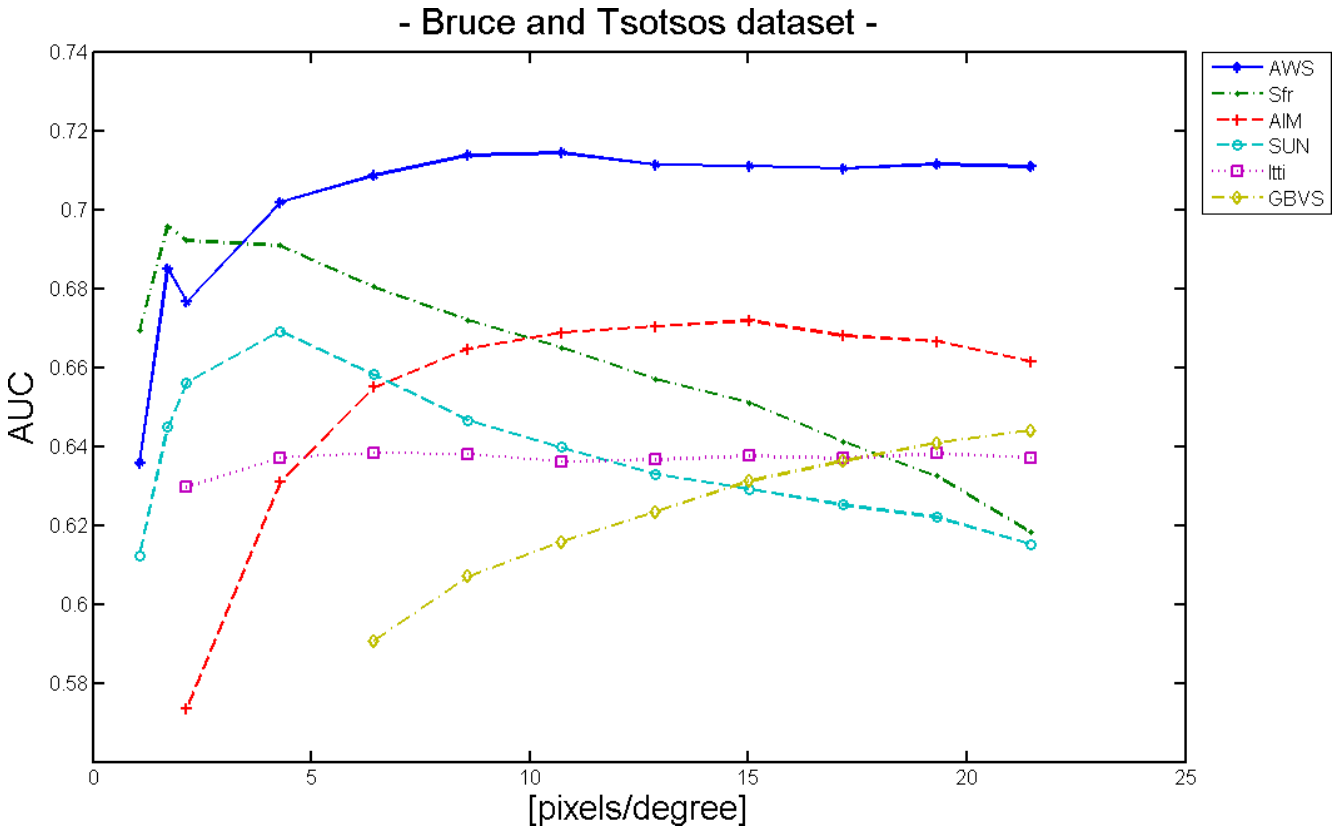

Figure 7. Comparison of the capability of different models to predict human fixations—measured through AUC values from ROC analysis, as explained in the first section—against the spatial resolution retained in the input image. Spatial resolution is expressed as pixels by degree of visual field for subjects. Results are shown for the dataset of Bruce & Tsotsos (2009). The standard errors are in the range 0.0005–0.0008 for all model and all spatial resolution values.

from its adaptive nature to the specific context, the AWS model is able to deal with a wider range of scales (i.e., a wider spectrum of spatial frequencies) and hence not only with the scales that are most frequently the salient ones.

In the dataset of Bruce and Tsotsos, there are several models—for instance the Sfr and the SUN models—that increase monotonically their performance as the spatial resolution decreases up to a given value from which they quickly decay. For low spatial resolutions the model of Seo and Milanfar manages to outperform the AWS (at the same spatial resolution). The reason is that it is optimized for a fixed (small) value of the size of the input image of 64×64 pixels, that is, for a low spatial resolution. We have checked that the AWS can also be tuned to outperform these results at such low spatial resolution. Moreover it must be noticed that the maximum AUC value achieved by their model for such a low spatial resolution is still clearly under the maximum achieved by AWS. Otherwise, a tuned version of the AWS for these low resolution values does not achieve the general maximum value either. This points to an amount of relevant saliency present in smaller scales that is lost with such a drastic down-sampling.

A very similar behavior is observed in the dataset of Kootstra et al. In this case none of the models outperforms the AWS, even at the lowest resolution values. This fact points to an additional bias in the Sfr model since it appears to work better when using only the middle and large scales in the dataset of Bruce and Tsotsos, used by the authors for validation, but not in a different dataset like the dataset of Kootstra et al. The relative distances between models and their behavior versus spatial resolution are very similar, in spite of the lower ability of saliency to predict fixations reflected in the shift to lower AUC values. This shift can be probably explained with basis on a higher clutter in the images of Kootstra et al. and a corresponding lower concentration of saliency. To check this point, we have derived for both datasets priority maps for each observer following the procedure based on the distance transform described in Kootstra & Schomaker (2009). We have taken the maximum AUC delivered by subjects as an indication of human consistency. The result yields $AUC = 0.7156$ for the dataset of Bruce and Tsotsos and $AUC = 0.6462$ for the dataset of Kootstra

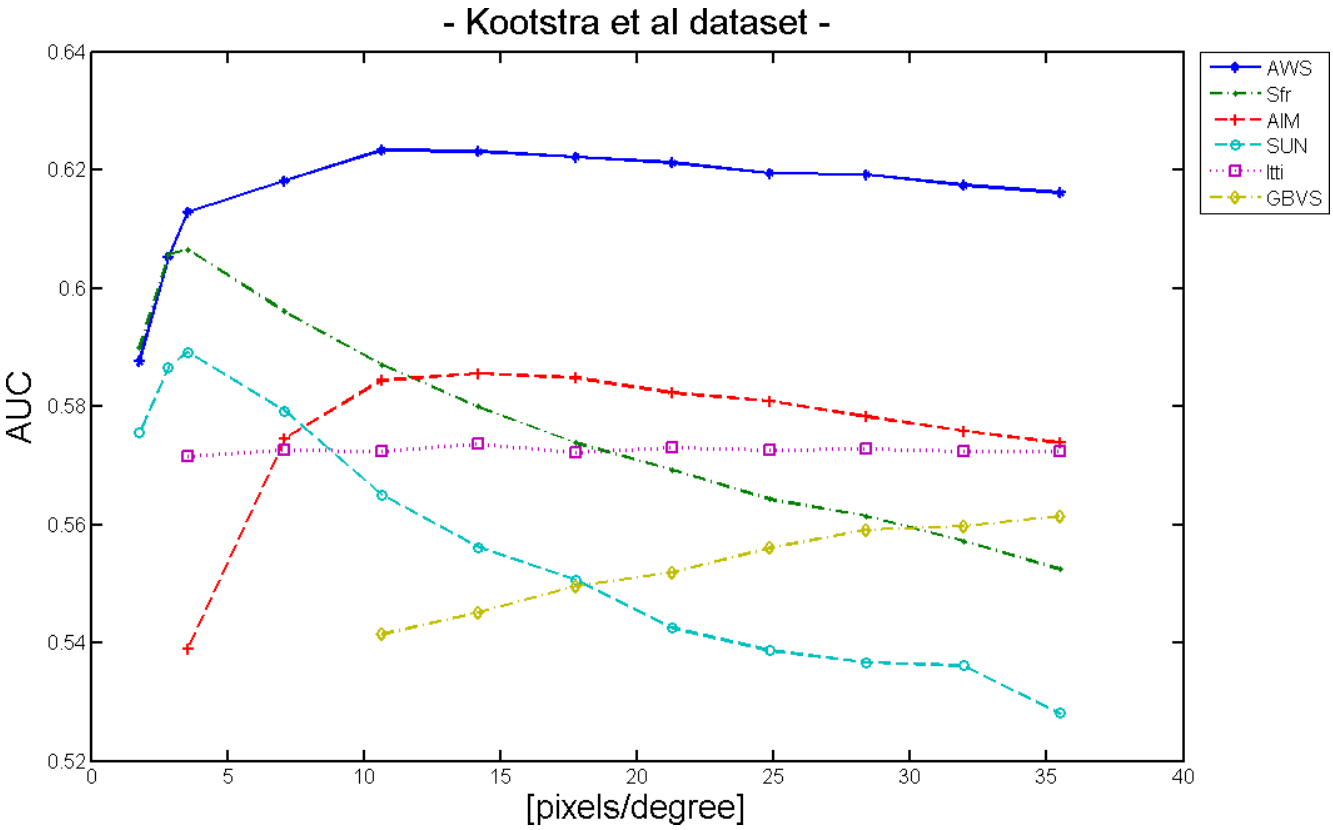

Figure 8. Comparison of the capability of different models to predict human fixations—measured through AUC values from ROC analysis, as explained in the first section—against the spatial resolution retained in the input image. Spatial resolution is expressed as pixels by degree of visual field observed by subjects. Results are shown for the dataset of Kootstra & Schomaker (2009). The standard errors are in the range 0.0005–0.0008 for all model and all spatial resolution values.

et al., again with standard error of 0.0008. Therefore, the overall shift observed in the results for models appears to reflect an equivalent shift in human consistency.

A hyperspectral analysis of eye movements

We have investigated the effect on the computation of optical variability of reducing the number of

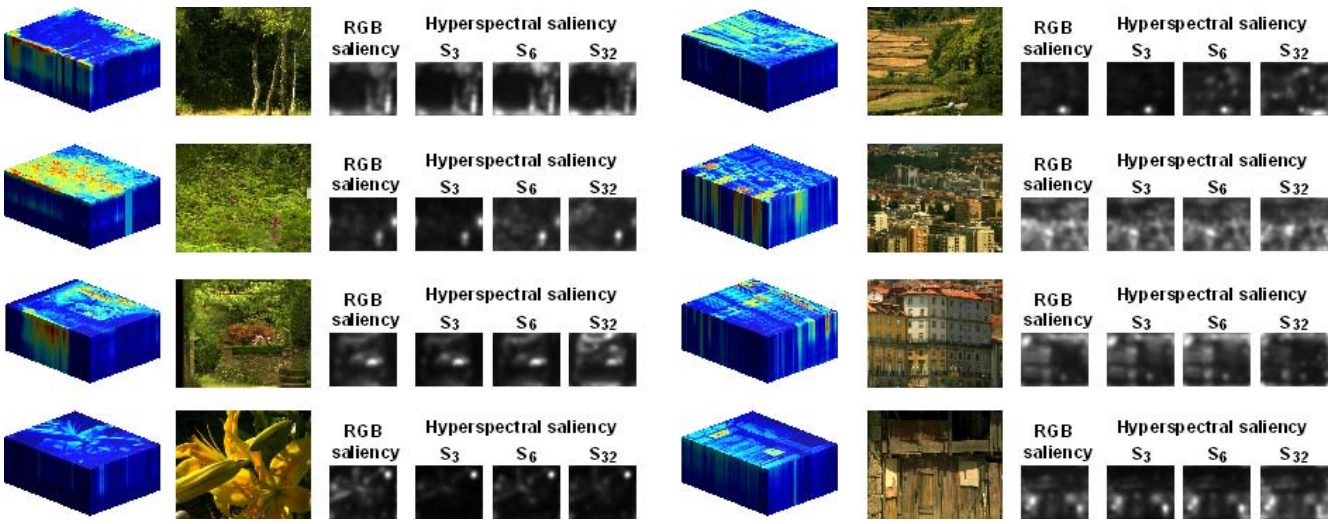

Figure 9. The hyperspectral reflectances cubes and the corresponding RGB images obtained from Foster et al. (2005). The resulting saliency maps are shown for the RGB image and for the hyperspectral cubes when using three, six, and 32 whitened spectral components for further spatial analysis.

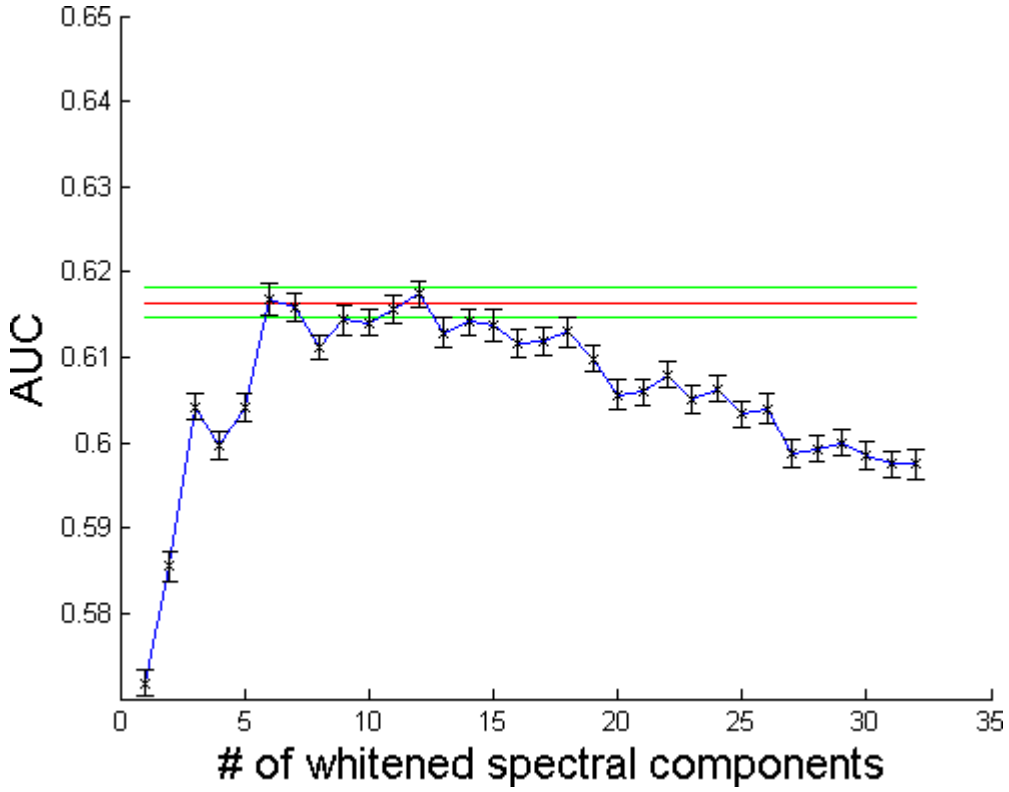

Figure 10. AUC values and the corresponding uncertainties for the saliency maps obtained from different numbers of whitened spectral components (black points) are shown. As well, the AUC value obtained with the saliency map from the RGB image is given by the red line and the uncertainty limits are given by the green lines.

chromatic components to only three and broadening their spectra. As well we have studied the effect of such approximation on the capability of predicting fixations.

To this end, we have applied the measure of saliency first on the RGB image, such as in the previous sections, and after on the hyperspectral cube using a variable number of whitened spectral components for further spatial analysis (this number ranges from 1 to 32).

Using the same procedure of ROC analysis employed in the previous sections we have assessed the performance of the different saliency maps computed in the prediction of fixations. In Figure 10 we show the obtained AUC values versus the number of whitened components involved. Additionally, the AUC value obtained using the saliency map from the RGB image is shown in the same figure. Due to the low number of images employed in this experiment, the standard errors are now clearly higher. However, they are still tight enough and they do not prevent us from extracting some conclusions.

It is remarkable that the saliency maps from both the RGB images and from the whitened spectral components are very similar and they show an equivalent capability of predicting fixations. A somewhat unexpected result arises however from the use of whitened spectral components of surface reflectance: The predictive capability shows a sensible increase with the number of components up to a maximum performance from 6 to 12 components from which it decays again, but not too much. The point is that the maximum does not occur for three whitened spectral components as could be expected. Therefore, the variability retained by the RGB components seems to be equivalent to the variability existing in a higher number of whitened narrow spectral components of surface reflectance, showing an equivalent capability of predicting fixations. In other words the use of a trichromatic representation would not mean a loss of perception of optical variability in natural scenes. Otherwise this fact agrees with the observation that a RGB image allows recovery of the spectrum of the illuminant and thereby allows us to estimate the spectral components of surface reflectances in a scene (Nieves, Plata, Valero, & Romero, 2008).

Early fixations and faces

We have tested the reach of the proposed hypothesis in a third open access eye-tracking dataset with many images with one or several faces. Since faces could

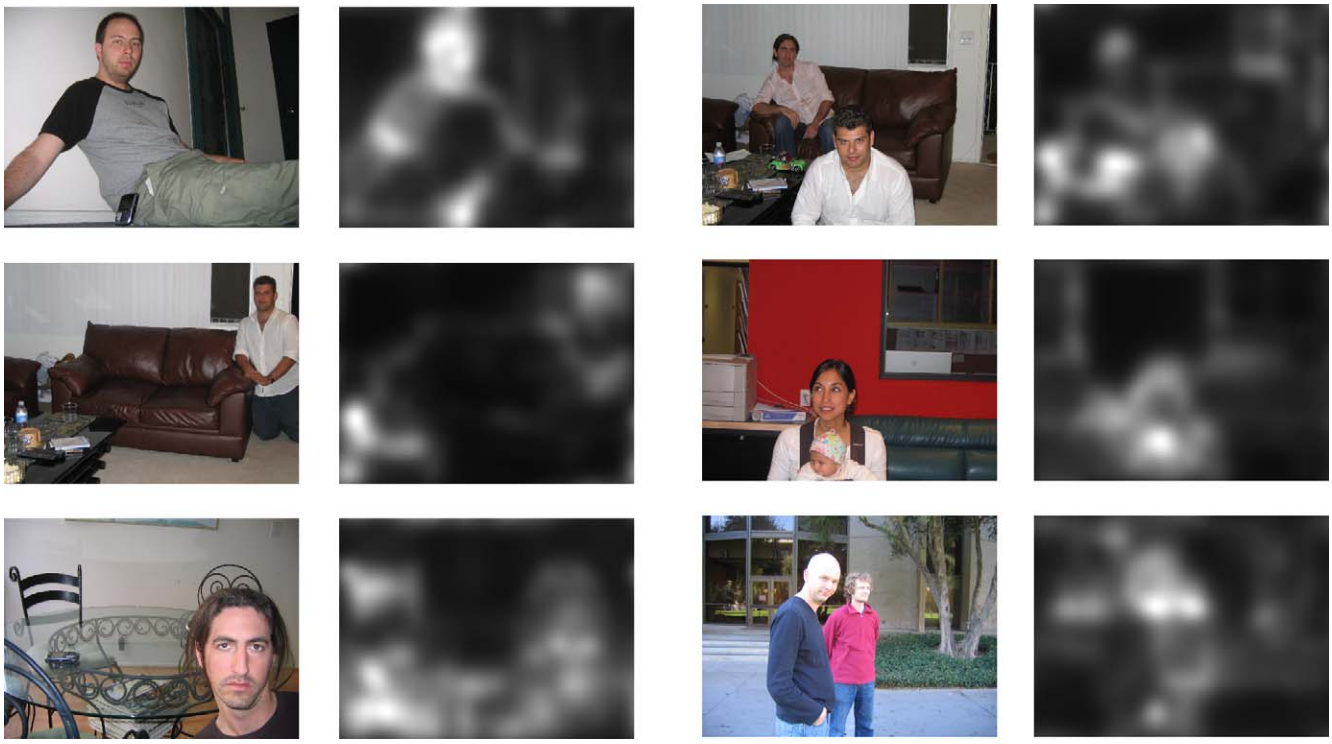

Figure 11. Six example images from the FIFAD dataset and the corresponding saliency maps using the AWS model.

introduce a top-down bias towards them, the ability of a measure of visual saliency to explain the spatial distribution of fixations in such images has a special interest.

In a detailed study, Cerf et al. (2009) reported a number of evidences of face saliency or attractiveness as guiding gaze from the very first fixations. Besides, they found that the addition of semantic maps to the saliency maps of Itti et al. (1998) produced a remarkable improvement in predicting the spatial distribution of fixations. They suggested that such result pointed to an attractiveness of faces because they are interesting for humans, since much of the salience of faces could not *be explained by their low-level features alone*. In other words, they suggested that faces introduced a strong influence of relevance able to drive early fixations.

Here we have compared the results obtained by the approaches studied by Cerf and collaborators with those provided by our measure of visual saliency as well as its combination with the semantic maps through the same weighting scheme used by them (75% saliency + 25% semantic map). Instead of the implementation made by the authors we have used the procedure of ROC analysis based on bootstrapping that was employed in the previous experiments for the sake of clarity. This choice also allows a more straight assessment of the predictive capability of saliency in comparison to that observed in the other datasets. The results are shown in Table 1 and also a graphical representation of the same is given in Figure 12.

In a first look, our approach clearly once again outperforms the model of Itti et al. (1998) and performs even better than its combination with semantic maps. Besides, the combination of our maps with semantic ones yields a relative improvement that is clearly lower, reducing the relative gain by more than the 30%.

Corner saliency versus corner angle

A preliminary version applied on scales of the approach proposed has been shown to reproduce a variety of psychophysical results that have been used to validate previous models (Garcia-Diaz et al., 2009). The generalized approach proposed here retains the ability to reproduce those experiments. Instead of repeating them, we just concentrate on a psychophys-

Model	AUC	SE
Itti	0.6522	0.0007
Itti + faces	0.7051	0.0005
AWS	0.7188	0.0006
AWS + faces	0.7568	0.0005

Table 1. AUC values and standard errors (SE) for models on the FIFAD dataset.

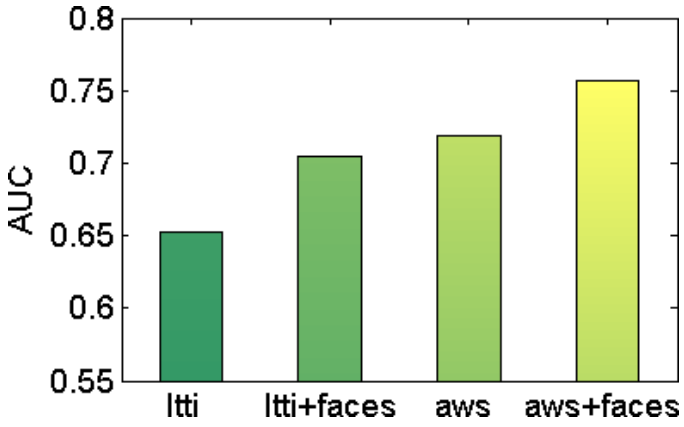

Figure 12. AUC values obtained for the different models compared on the FIFAD dataset.

ical result directly related to saliency that to our knowledge has not been explained by any other model before.

Inspired by a series of Vasarely's op-art works devoted to nested squares and with the aim to characterize and explain the illusion of higher luminance in their diagonals, as well as other related visual illusions, Troncoso et al. (2005) have studied the saliency of a corner in a gray scale gradient.

They measured saliency as a function of corner angle. To do it, they used seven images of different angle value, with the middle point of the gradient within the corner, always with the same luminance. Six of those images can be seen in Figure 13. They asked observers to compare the intensity at that central point, with a standard stimulus made of a vertical stripe with 55 segments of different luminance value. The order of segments was varied so that any had the same probability to appear at the same height than the central point of the corner. Given that the physical luminance of that point was the same for all of the corners, differences in the luminance chosen in the standard stripe were attributed to an illusory enhancement due to a different magnitude of saliency. The results obtained revealed that perceived saliency decreases linearly with corner angle.

The authors tried an explanation of such behavior with a basis on center-surround differences. They measured the responses of a difference of Gaussians (DoG) filter for all of the corners in the central point evaluated by observers. They succeeded in explaining the trend to decrease of saliency but not the linearity observed. They stated that the results pointed to a kind of center-surround competition and hypothesized two possibilities to explain the linear behavior obtained, namely, a nonlinear component in filtering or the intervention of mechanisms other than center-surround differences.

We have compared the relative saliency for several models in the central point of each corner. For our approach saliency is indeed the relative saliency. For a fair comparison, to assess the other models we have taken the saliency value at that point normalized by the overall saliency: $S = S_{center}/\Sigma S_{image}$. Taking raw values without normalizing did not yield better results in any case.

In Figure 13 the results obtained with the AWS model are shown. The saliency measured by the model decreases with corner angle for six corner angles (30°, 45°, 75°, 105°, 135°, and 180°). This result is in fair agreement with the reported linear behavior of humans. Saliency for an additional corner of 15° used by Troncoso et al. (2005) was clearly underestimated by the model and has not been used for linear fitting. Other models we tested have not been able to reproduce such behavior and to our knowledge this is the first model that claims to do it. Three illustrative examples of the failure of other models are also given in Figure 14.

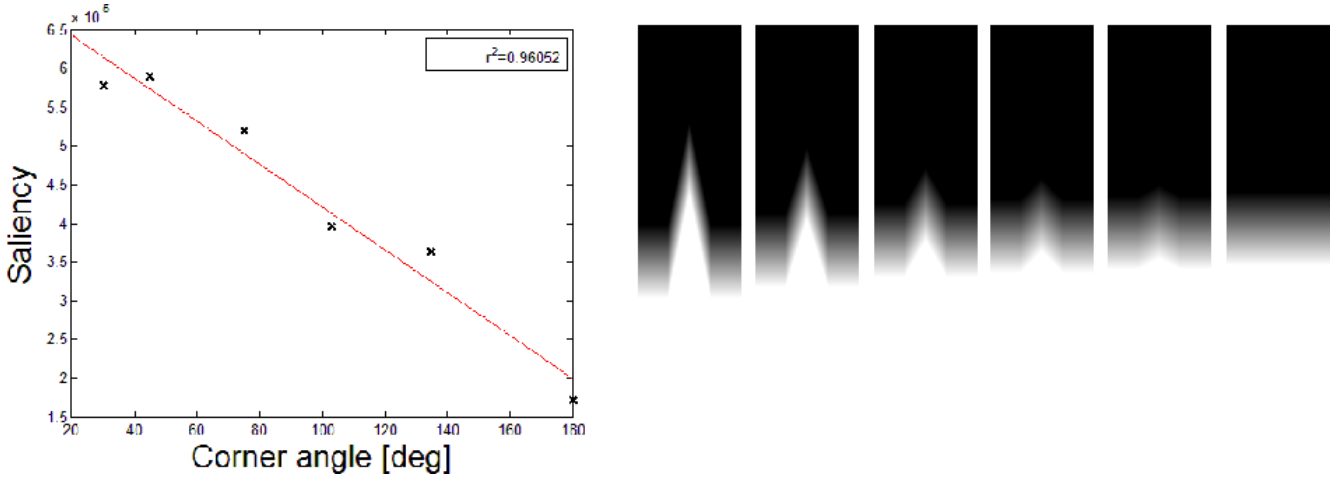

Figure 13. Saliency (using AWS) against corner angle and the six images used, obtained from Troncoso et al. (2005).

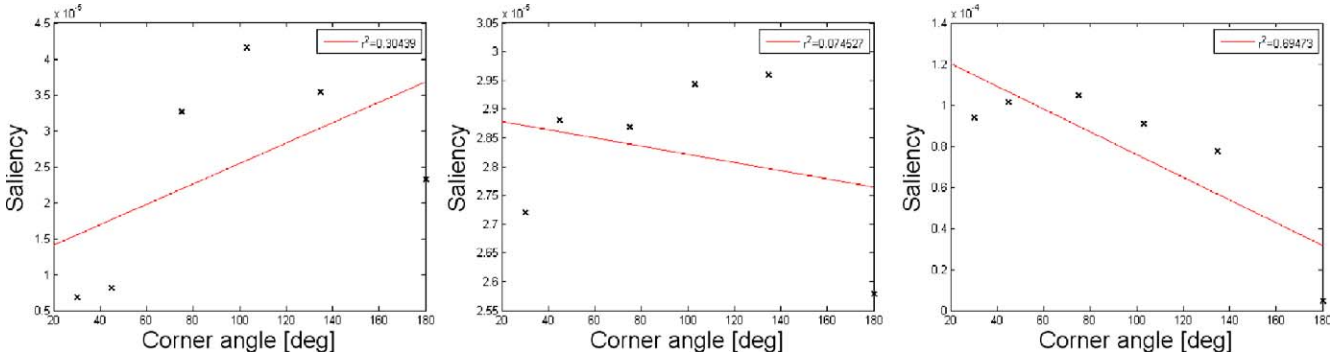

Figure 14. Saliency against corner angle for three other models. Right: Itti et al. (1998); Center: Bruce & Tsotsos (2009); Right: Seo & Milanfar (2009).

Discussion

We have shown how an analysis of the behavior of measures of saliency versus spatial resolution provides a simple tool useful to reveal existing biases in models. Such biases are due to design choices such as the definition of fixed sizes (in pixels) for the receptive fields and their surround, or the definition of fixed ranges of spatial frequencies to compute saliency, without considering the real dimensions of the visual field and its sampling rate. As well, the use of the dataset of Bruce and Tsotsos as a benchmark seems to have also contributed to certain amount of bias in some models.

Typically, the function call in available code implementations of models of saliency has a default value of downsampling factor, or even of the dimensions of the input image. But this poses a problem: They are working on different real scales, since the maximum spatial resolution available in the original images is different for different eye-tracking experiments. Otherwise, the proposed model based on optical variability does not force the image size and is able to deal with different data. Indeed, it achieves the maximum performance in both tested datasets for nearly the same value of spatial resolution, using different downsampling factors. It seems very reasonable to take spatial resolution (in pixels by degree of visual angle of observers) instead of the image size (in pixels), like the relevant parameter to compare and analyze results with different datasets.

Otherwise, as shown in Figures 7 and 8, the maximum predictive capability achieved by the proposed approach does not occur for the maximum spatial resolution available, but for a clearly lower value of about 10 pixels/degree of visual field for both datasets. This fact takes place in spite of the different values of maximum resolution used in the experiments conducted to obtain each of the datasets. It suggests the existence of a threshold of the visual acuity that is able to affect saliency perception and by extension inter-subject consistency in the spatial distribution of fixations. In other words, subjects with a loss of visual acuity that do not cross such threshold will exhibit the same consistency present among healthy subjects with normal visual acuity. Therefore, the hypothesis of hierarchical adaptive whitening in the HVS seems to predict a sustained consistency between subjects in spite of important variations of visual acuity. However further analysis is needed in this respect to determine the existence and value of such a threshold or whether the result merely arises from a shared bias in the datasets (or in the model). In particular, it would be worth using a selection of biased images in which saliency is expected to be driven by small scales to find how small a scale can be to affect the spatial pattern of human fixations in free surveillance.

It has been shown that the proposed measure of optical variability can be directly obtained with very similar results from the spectral components of surface reflectance of the scene, which are independent of the illuminant. That is, since the hyperspectral image is a calibrated and normalized representation of surface reflectance rather than measured luminance, the obtained results point to a direct relation between saliency and the optical properties of the objects in the scene. This yields a simple explanation to the robustness of visual saliency—when measured by optical variability—against important changes in illumination. It can be also seen as a major reason for an invariant management of optical variability: It provides a suitable reference directly related to object external properties and structure that is stable against variations of illumination conditions.

B. Tatler, Hayhoe, Land, & Ballard (2011) have recently revised a number of important concerns on the actual value of existing computational models for the explanation of eye-movements in daily visual tasks. One of the main underlying ideas is that saliency models show a rather low performance and that such performance may be explained because important objects have in average higher saliency. Thereby, if

relevance is the driving factor then fixated locations will be on average more salient than nonfixated locations. Therefore, they hold that probably reward explains better where we look.

However, the results achieved with an open access dataset of fixations on images with faces point to an interpretation of face *attractiveness* mostly supported by the structural and chromatic singularity of faces in a natural environment with low need of attractiveness for humans or any other kind of relevance, in agreement with previous psychophysical results that reported that *the ability to rapidly saccade to faces in natural scenes depends, at least in part, on low-level information* (Honey, Kirchner, & VanRullen, 2008). It is worth noting that such previous results do not take into account chromatic features but only content of spatial frequencies in an achromatic representation.

Thereby, the advantage previously reported by Cerf et al. (2009) of introducing semantic maps has been remarkably reduced by a 30% with an improved measure of saliency, purely data-driven and physically based. This fact leaves less room for face relevance in driving early fixations. It may be expected that further improvements in the measure of saliency (i.e., in the estimation of the optical variability within the visual window) will reduce even more the relative gain of using semantic maps, to turn it in an even weaker influence or to remove it completely.

Of course, we do not question the existence of strong spatial biases like the center bias or an orientation (horizontal) bias. Indeed, we have used a well-established evaluation method that discounts the effect they may have. This seems reasonable since spatial biases are supposed to be independent of the image content. In this situation, we show that a simple model of saliency is able to perform well over the model of Itti et al. Important works that question the actual influence of saliency on visual behavior like Rothkopf, Ballard, & Hayhoe (2007) use the version described in Itti & Koch (2000) with an even poorer performance than the original version in the prediction of human fixations. Compared to that measure of visual saliency our model doubles the gain versus random selection of fixations. Other models also have achieved remarkable improvements on that estimation. Different measures of saliency exhibit differences in behavior and resort to different assumptions on visual processing. An example of the importance of what is the measure of saliency chosen can be found in Verma & McOwan (2010). They show how a number of results claiming for top-down influences in change detection were saliency biased by using the model of Itti et al. (1998) instead of coarse approaches to saliency using measures of low-level richness with a poor validation.

The proposed approach is able to quantitatively explain the linearity of perceived saliency versus corner angle that—to our knowledge—has not been correctly reproduced before by any other model. This result has several interesting characteristics; it arises from perceptual comparisons without the involvement of eye movements and it poses a quantitative, not only qualitative, challenge. Moreover, corner saliency is supposed to underlie several visual illusions that are a form of contextual adaptation. Indeed, redundancy reduction has been pointed as a possible explanation of the visual behavior against corners (Troncoso, Macknik, & Martinez-Conde, 2011). Besides, unlike a ROC analysis of the capability of predicting fixations, this result is not invariant to monotonic transformations. Therefore, it represents a worthy complement to validate a model of saliency that, like ours, claims for both an improved performance and an increased explanatory capability over previous models.

With independence of mechanistic considerations—beyond the scope of this work—our approach to visual saliency like an estimation of optical variability from hierarchical whitening can be seen as one of the most parsimonious from a computational view but also as one of the closest to biological plausibility. Indeed, it relies on a decomposition of the image using the standard model of V1 combined with an adaptive whitening implemented through decorrelation and contrast normalization, two operations that may be related to mechanisms ubiquitous in the early visual pathway (Kohn, 2007).

We have proposed a coarse scheme linked to contextual (and data-driven) adaptation mechanisms that appears to produce a coarse figure-ground separation as well as illusory contours. Since adaptation mechanisms are thought to be similar for different time-scales we may expect that the proposed approach admits a coherent generalization to longer time scales, that is to a dynamic implementation that includes temporal adaptation. Indeed, the statistical interpretation provided here is compatible with existing statistical interpretations from dynamical free-viewing in terms of surprise (Itti & Baldi, 2009). Nevertheless, since we have resorted to optical magnitudes to root our approach without a loss of simplicity, we think that it contributes to clarify the concept and role of saliency in addition to the reported gains in performance.

Acknowledgments

Commercial relationships: None.
Corresponding author: Antón Garcia-Diaz.
Email: anton.garcia@usc.es.
Address: Computer Vision Group, University of Santiago de Compostela, Galicia, Spain.

References

Atick, J. J., Li, Z., & Redlich, A. N. (1993). What does post-adaptation color appearance reveal about cortical color representation? *Vision Research, 33*(1), 123–129.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review, 61*(3), 183–193.

Barlow, H. B. (1961). *Possible principles underlying the transformation of sensory messages sensory communication*. Cambridge, MA: MIT Press.

Barlow, H. B., & Foldiak, P. (1989). Adaptation and decorrelation in the cortex. In Richard Durbin, Christopher Miall, Graeme Mitchison, King's College (University of Cambridge) (Eds.) *The Computing Neuron* (pp. 54–72). Boston, MA: Addison-Wesley.

Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. In Y. Weiss, B. Schölkopf, and J. Platt (Eds.) *Advances in Neural Information Processing Systems: Vol. 18. Conference on Neural Information Processing Systems* (p. 155). Cambridge, MA: MIT Press.

Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, 9*(3):5, 1–24, http://www.journalofvision.org/9/3/5, doi:10.1167/9.3.5. [PubMed] [Article].

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision, 9*(12):10, 1–15, http://www.journalofvision.org/9/12/10, doi:10.1167/9.12.10. [PubMed] [Article].

Clifford, C. W., Webster, M. A., Stanley, G. B., Stocker, A. A., Kohn, A., Sharpee, T. O., et al. (2007). Visual adaptation: Neural, psychological and computational aspects. *Vision Research, 47*(25), 3125–3131.

Ecker, A. S., Berens, P., Keliris, G. A., Bethge, M., Logothetis, N. K., & Tolias, A. S. (2010). Decorrelated neuronal firing in cortical microcircuits. *Science, 327*(5965), 584.

Fecteau, J. H., & Munoz, D. P. (2006). Salience, relevance, and firing: A priority map for target selection. *Trends in Cognitive Sciences, 10*(8), 382–390.

Foster, D. H., Nascimento, S. M., & Amano, K. (2005). Information limits on neural identification of colored surfaces in natural scenes. *Visual Neuroscience, 21*(3), 331–336.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision, 8*(2):6, 1–17, http://www.journalofvision.org/8/2/6, doi:10.1167/8.2.6. [PubMed] [Article].

Gao, D., Mahadevan, V., & Vasconcelos, N. (2008). On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision, 8*(7):13, 1–18, http://www.journalofvision.org/content/8/7/13, doi:10.1167/8.7.13. [PubMed] [Article].

Garcia-Diaz, A., Fdez-Vidal, X., Pardo, X., & Dosil, R. (2009). Decorrelation and distinctiveness provide with human-like saliency. In Jacques Blanc-Talon, Wilfried Philips, Dan Popescu, Paul Scheunders (Eds.) *Lecture Notes in Computer Science: Vol. 5807. Advanced Concepts for Intelligent Vision Systems* (pp. 343–354).

Goodman, J. W. (2005). *Introduction to Fourier optics*. Greenwood Village, CO: Roberts & Company Publishers.

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. In Bernhard Schölkopf, John Platt, Thomas Hofmann (Eds.) *Advances in Neural Information Processing Systems: Vol. 19. Conference on Neural Information Processing Systems* (p. 545). Cambridge, MA: MIT Press.

Honey, C., Kirchner, H., & VanRullen, R. (2008). Faces in the cloud: Fourier power spectrum biases ultrarapid face detection. *Journal of Vision, 8*(12):9, 1–13, http://www.journalofvision.org/8/12/9, doi:10.1167/8.12.9. [PubMed] [Article].

Hou, X., & Zhang, L. (2008). Dynamic visual attention: Searching for coding length increments. In D. Koller (Ed.) *Advances in Neural Information Processing Systems: Vol. 21. Conference on Neural Information Processing Systems* (pp. 681–688). Red Hook, NY: Curran Associates, Inc.

Hoyer, P. O., & Hyvärinen, A. (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems, 11*(3), 191–210.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology, 195*(1), 215.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research, 49*(10), 1295–1306.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*(10–12), 1489–1506.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.

Kohn, A. (2007). Visual adaptation: Physiology,

mechanisms, and functional benefits. *Journal of Neurophysiology, 97*(5), 3155.

Kootstra, G., Nederveen, A., & de Boer, B. (2008). Paying attention to symmetry. In M. Everingham, C. J. Needham, R. Fraile (Eds.) *Proceedings of the British Machine Vision Conference* (pp. 1115–1125).

Kootstra, G., & Schomaker, L. R. (2009). Prediction of human eye fixations using symmetry. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society (CogSci09)*, July 29-August 1, 2009. Amsterdam, the Netherlands.

Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(5), 802–817.

Lee, T. W., Wachtler, T., & Sejnowski, T. J. (2002). Color opponency is an efficient representation of spectral properties in natural scenes. *Vision Research, 42*(17), 2095–2103.

Mannan, S. K., Kennard, C., & Husain, M. (2009). The role of visual salience in directing eye movements in visual object agnosia. *Current Biology, 19*(6), R247–R248.

Melcher, D., & Kowler, E. (2001). Visual scene memory and the guidance of saccadic eye movements. *Vision Research, 41*(25–26), 3597–3611.

Nieves, J. L., Plata, C., Valero, E. M., & Romero, J. (2008). Unsupervised illuminant estimation from natural scenes: An RGB digital camera suffices. *Applied Optics, 47*(20), 3574–3584.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*(6583), 607–609.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*(1), 107–123.

Rieke, F., & Rudd, M. E. (2009). The challenges natural images pose for visual adaptation. *Neuron, 64*(5), 605–616.

Rothkopf, C., Ballard, D., & Hayhoe, M. (2007). Task and context determine where you look. *Journal of Vision, 7*(14):16, 1–20, http://www.journalofvision.org/7/14/16, doi:10.1167/7.14.16. [PubMed] [Article].

Saleh, E. A. B., & Teich, M. C. (1991). *Fundamentals of photonics*. Hoboken, NJ: John Wiley & Sons.

Schwartz, O., Hsu, A., & Dayan, P. (2007). Space and time in visual context. *Nature Reviews Neuroscience, 8*(7), 522–535.

Seo, H. J., & Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Journal of Vision, 9*(12):15, 1–27, http://www.journalofvision.org/9/12/15/, doi:10.1167/9.12.15. [PubMed] [Article].

Tatler, B., Hayhoe, M., Land, M., & Ballard, D. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision, 11*(5):5, 1–23, http://www.journalofvision.org/11/5/5/, doi:10.1167/11.5.5.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45*(5), 643–659.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review, 95*(1), 15–48.

Troncoso, X., Macknik, S., & Martinez-Conde, S. (2011). *Vision's first steps: Anatomy, physiology, and perception in the retina, lateral geniculate nucleus, and early visual cortical areas*. In G. Dagnelie (Ed.), *Visual prosthetics*: *Physiology, bioengineering and rehabilitation* (p. 23). New York: Springer Verlag.

Troncoso, X. G., Macknik, S. L., & Martinez-Conde, S. (2005). Novel visual illusions related to Vasarely's nested squares show that corner salience varies with corner angle. *Perception, 34*:409–420.

Verma, M., & McOwan, P. W. (2010). A semi-automated approach to balancing of bottom-up salience for predicting change detection performance. *Journal of Vision, 10*(6):3, 1–17, http://www.journalofvision.org/10/6/3, doi:10.1167/10.6.3. [PubMed] [Article].

Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science, 287*(5456), 1273.

Webster, M. A., & Mollon, J. D. (1997). Adaptation and the color statistics of natural images. *Vision Research, 37*(23), 3283–3298.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A bayesian framework for saliency using natural statistics. *Journal of Vision, 8*(7):32, 1–20, http://www.journalofvision.org/8/7/32, doi:10.1167/8.7.32. [PubMed] [Article].

Appendix

Image representation

In Fourier optics any image can be considered as a wavefront piece and approached as a superposition of ideal monochromatic plane waves (Saleh & Teich,

1991; Goodman, 2005). A monochromatic plane wave can be characterized by means of its amplitude A, its spectral wavelength λ and its wave number vector $\mathbf{k}$ (i.e., its direction of propagation).

$$E(x,y,\lambda,\mathbf{k}) = A(\lambda,\mathbf{k})exp\Big(\mathbf{k}\cdot\mathbf{r} - i(c/\lambda)t\Big); \quad (16)$$

with $\mathbf{k}$ being a vector of free orientation and with norm $k = 2\pi/\lambda$, and being c the speed of light.

The visual system is only sensible to light intensity, which means to the squared norm of the different plane waves, and not to the ultrafast phase of light wavefronts. Besides, natural images are in general illuminated by diffuse or extended sources (e.g., the sun), hence the eye can be assumed to be an incoherent system which is linear in intensity. Consequently, the image intensity can be described by the expression:

$$I(x,y,\lambda,\mathbf{k}) = EE^{*} = A^{2}(\lambda,\mathbf{k})exp(2\mathbf{k}\cdot\mathbf{r}); \quad (17)$$

Hence, being u and v the rectangular components of the two dimensional spatial frequencies on an image plane parallel to the x–y plane, they are related to the wave number vector through the expression

$$\mathbf{k} = 2\pi f_x\mathbf{i} + 2\pi f_y\mathbf{j} + k_z\mathbf{k}; \quad (18)$$

so that the spatial frequencies contributed by a given plane wave depend on the projection of its wave number vector on the x–y plane. That means they can be derived from both the angle with the image plane and its spectral wavelength, so that:

$$\begin{aligned} f_x &= (1/\lambda)\sin\ \theta_x \approx (1/\lambda)\theta_x \\ f_y &= (1/\lambda)\sin\ \theta_y \approx (1/\lambda)\theta_y \end{aligned} \quad (19)$$

where θ_x and θ_y are the angles that the wave number vector makes with the planes y–z and x–z, respectively, and the sine becomes the angle in the paraxial approximation (for small angles).

That said, the spectral value determines the chromatic properties of the plane wave, while both the spectral value and the angle between the wave number vector and the image plane determine the spatial frequency contributed by the plane wave (Saleh & Teich, 1991). Besides on an image plane, the plane wave can be represented by an intensity value at each point. From the previous argument, it follows that the intensity of an image can be obtained from the integral of the light intensities in the continuum of plane waves, that is:

$$\begin{aligned} I(x,y) &= \int_{\lambda} I(x,y;\lambda)d\lambda \\ &= \int_{\lambda}\int_{f_x}\int_{f_y} I(x,y;\lambda;f_x,f_y)d\lambda df_x df_y \end{aligned} \quad (20)$$

where

$$I(x,y;\lambda) = \int_{f_x}\int_{f_y} I(x,y;\lambda;f_x,f_y)df_x df_y \quad (21)$$

Since spatial information is coded in the spatial frequencies, a given point can be referred by a single unidimensional index $(x,\ y) \rightarrow p$. Using more conveniently polar instead of rectangular coordinates to represent spatial frequencies, an image can be formalized by the expressions:

$$I(p) = \int_{\lambda} I(p;\lambda)d\lambda \quad (22)$$

and

$$I(p;\lambda) = \int_{\rho}\int_{\alpha} I(p;\lambda;\rho,\alpha)d\rho d\alpha \quad (23)$$

where ρ and α are the radius and the angle of the spatial frequency in polar coordinates, respectively.

The local contribution to such a superposition of monochromatic plane waves can be described in terms of the spatial power distributions of chromatic components—related to electromagnetic wavelength—and of the corresponding power distributions of magnitude and orientation of the spatial frequencies present for each of them, related to the wave number vector. The spectral power distribution is given by the left side of Equation 23, while the power distribution of spatial frequencies for a fixed λ can be represented by the argument of the integral in the right side of the same equation.

Reproducibility

A Matlab file with the implementation of the proposed model is available for download at http://www.gva.dec.usc.es/persoal/xose.vidal/research/aws/AWSmodel.html for reproducibility purposes.