Saliency from hierarchical adaptation through decorrelation and variance normalization

Antón Garcia-Diaz, Xosé R. Fdez Vidal, Xosé M. Pardo, Raquel Dosil

Computer Vision Group, Dept. of Electronics and Computer Science, University of Santiago de Compostela

Abstract

This paper presents a novel approach to visual saliency that relies on a contextually adapted representation produced through adaptive whitening of color and scale features. Unlike previous models, the proposal is grounded on the specific adaptation of the basis of low level features to the statistical structure of the image. Adaptation is achieved through decorrelation and contrast normalization in several steps in a hierarchical approach, in compliance with coarse features described in biological visual systems. Saliency is simply computed as the square of the vector norm in the resulting representation. The performance of the model is compared with several state-of-the-art models, in predicting human fixations using three different eye-tracking datasets. Referring this measure to the performance of human priority maps, the model is proved to be the only one able to keep the same behavior through different datasets, showing free of biases. Moreover, it is able to predict a wide set of relevant psychophysical observations, to our knowledge, not reproduced together by any other model before.

Keywords: saliency, bottom-up, eye fixations, decorrelation, whitening, visual attention

1. Introduction

Research on the estimation of visual saliency has experienced an increasing activity in the last years from both computer vision and neuroscience perspectives, giving rise to a number of improved approaches. Furthermore, a wide diversity of applications based on saliency are being proposed that range from image retargeting [1] to human-like robot surveillance [2], object learning and recognition [3, 4, 5], objectness definition [6], image processing for retinal implants [7], and many others.

Existing approaches to visual saliency have adopted a number of quite different strategies. A first group, including many early models, is very influenced by psychophysical theories supporting a parallel processing of several feature dimensions. Models in this group are particularly concerned with biological plausibility in their formulation, and they resort to the modeling of visual functions. Outstanding examples can be found in [8] or in [9]. Most recent models are in a second group that broadly aims to estimate the inverse of the probability density of a set of low level features by different procedures. In this kind of models, low level features are usually obtained by an off-line process of statistical analysis of a large set of images, aiming to represent the set of natural images. Saliency is computed on these features through a particular estimation of improbability. Outstanding examples of these models are the approaches of [10, 11, 12, 13]. Other models, although without an explicit ground, can also be interpreted from an information theoretic perspective in terms of estimations of the inverse of the probability density. For instance, those models that seek distinctive spectral features in the domain of the spatial frequencies, like the models proposed in [14, 15], but also a more recent model that computes distances in a color space for different spatial frequency bands [16].

Approaches that are not strictly data-driven include the combination of saliency with semantic maps, trying to catch attractiveness of faces, persons, and other objects [17, 18], or the ad-hoc adaptation of saliency models to different datasets by learning weights that optimize the prediction of human fixations in those datasets [19]. Also, the adaptation of the spatio-chromatic representation to the specific dataset from learning specifi-

Email addresses: anton.garcia@usc.es (Antón Garcia-Diaz), xose.vidal@usc.es (Xosé R. Fdez Vidal), xose.pardo@usc.es (Xosé M. Pardo), raquel.dosil@usc.es (Raquel Dosil)

cally decorrelated coordinates [20] or independent components [21] has been proposed. These two last approaches already point to benefits from the adaptation of the feature basis. However these approaches are mostly ad-hoc, since they rely on an off-line computation applied to a specific dataset. They do not produce a representation adapted to each specific image.

1.1. Our approach

A natural approximation to sample distinctiveness can be done by computing the statistical distance in a representative coordinate system. This can be simply done through a vector norm computation if such system is statistically whitened. Thereby, it makes sense to think in the adaptation –through whitening– of the feature basis to the specific statistical structure of a particular image, considering pixels as samples. The resulting representation would yield a simple and straight measure of point saliency through a vector norm computation.

However, typical schemes of statistical whitening have a cubic or higher complexity on the number of coordinates, while linear on the number of samples. These facts prevent their use with representations of images involving three color components, several scales and several orientations. In this paper we generalize a preliminary approach [22] to overcome this problem by imposing a whitening transformation independently on reduced groups of feature components.

The proposed approach grounds on a classical hierarchical decomposition of images that first separates chromatic components, and next performs on each of them a multiscale and multioriented decomposition. Such approach is coarsely inspired in the image representation described in early stages of the visual pathway. Besides, different implementations and simplifications of the same can be found in a variety of early and recent models of computer vision with different purposes. Therefore, we propose to apply on-line whitening on chromatic components in a first stage. This operation is followed by a multiorientation and multiscale decomposition of the resulting whitened chromatic components. Next, further whitening is impossed to groups of oriented scales for each whitened chromatic component. This strategy allows to keep the number of components involved in whitening limited, overcoming problems of computational complexity.

As a result, an specifically adapted representation of the image arises. The resulting image components have zero mean and units of variance. As well, they are partly decorrelated. To obtain a saliency map, we simply compute point distinctiveness by taking, for each pixel, the squared vector norm in this representation divided by the sum of the same across all the pixels.

The proposed model is validated and compared with state-of-the-art approaches by measuring the predictive capability of human fixations in three open access datasets through state-of the-art procedures based on Receiver Operating Characteristic (ROC) analysis and Kullback-Leibler divergences (KLD). Additionally, the model will be shown to reproduce a wide set of relevant psychophysical results in which other models show failures.

The paper is organized as follows. Section 2 provides a detailed description of the AWS model for saliency computation. Section 3 evaluates the capability of the model in predicting eye-fixations. Section 4 shows the ability to reproduce a selection of psychophysical and perceptual observations. Finally, in Section 5 the main conclusions of the work are presented.

2. Model

The key point of the model of saliency proposed relies on the on-line adaptation of the basis used for representation to the specific statistical structure of the image. This implies a step beyond the adaptation to a given set –like the set of natural images– that is under the decomposition methods of most existing approaches to saliency. This adaptation uses pixels as statistical samples and seeks for a set of decorrelated and whitened coordinates, able to deal with the information present in the image and to also provide a reliable estimation of the statistical distance of each sample –pixel–to the center of the distribution. Therefore, the proposed model will be referred to as the adaptive whitening saliency (AWS) model.

2.1. Chromatic decomposition and whitening

Chromatic components undergo the first adaptative stage. Each pixel in the image has an associated vector of red (r), green (g) and blue (b) components. In general, the (r, g, b) coordinates are highly correlated in the ensemble of samples. Provided the covariance matrix in these coordinates is:

$$\mathbf{C}_{rgb} = \begin{pmatrix} \sigma_r^2 & \sigma_{rg} & \sigma_{rb} \\ \sigma_{rg} & \sigma_g^2 & \sigma_{gb} \\ \sigma_{rb} & \sigma_{gb} & \sigma_b^2 \end{pmatrix} \tag{1}$$

we typically have that all the elements are non-zero and non-negligible. Some color spaces (e.g. the Lab model) reduce this correlation between components by producing a representation that is decorrelated in the set of natural images, but not necessarily in specific images.

To decorrelate color information, the whitening procedure consisting in decorrelation and variance normalization (as described in the appendix) is simply applied to the r, g, b components of the image. Being $x_1 = r$, $x_2 = g$ and $x_3 = b$ the RGB coordinates of any pixel in the image, they are involved in the transformation

$$(x_1, x_2, x_3) \rightarrow (z_1, z_2, z_3) \tag{2}$$

Thereby, we get a $\mathbf{z} = (z_1^{chr}, z_2^{chr}, z_3^{chr})$ whitened representation with a new vector associated to each pixel. In this representation the covariance matrix is the indentity matrix, and thus each coordinate has units of variance. Indeed, the vector norm gives a measure of chromatic distinctiveness as the statistical distance of each color point to the average color. Such a simple measure is equivalent to the explanation proposed by [23] to color search asymmetry phenomena reported for humans in a set of simple synthetic images. However, to compute point saliency in cluttered natural scenes, spatial distinctiveness also must be taken into account.

Alternatively to a RGB color space, we have also tested the use of other color spaces like the Lab model. The conversion from RGB to a Lab model involves a non-linear transformation. Besides, the Lab model produces a representation that preserves, on average, perceptual distances. Therefore, differences in the resulting decorrelated components and even an advantage for the Lab model may be expected. However, we did not find a significant advantage for any of the alternative color spaces over RGB in our experimental evaluation. Thereby, there was no apparent reason to recode the RGB images to other color space before whitening.

2.2. Oriented multiscale decomposition and whitening

We represent the spatial structure by decomposing each of the whitened chromatic components (i.e. z_1^{chr}, z_2^{chr}, and z_3^{chr}) through a measure of local energy at different spatial frequency bands centered at different frequency modulus values (scales) and different orientations.

To obtain local energy, we use a bank of log-Gabor filters, since their real and imaginary parts in the spatial domain form a pair of filters in phase quadrature. These filters present several advantages over the Gabor filters. Namely, they have a zero DC component and a long tail towards high frequencies, approaching better the receptive fields of cortical cells [24]. The expression of these filters in the frequency domain is given by:

$$\log Gabor_{so}(\rho, \alpha) = \exp\left(-\frac{(\log(\rho/\rho_s))^2}{2\left(\log\left(\sigma_{\rho s}/\rho_s\right)\right)^2}\right) \cdot \exp\left(-\frac{(\alpha - \alpha_o)^2}{2(\sigma_{\alpha o})^2}\right) \tag{3}$$

being (ρ, α) the spatial frequency in polar coordinates, (ρ_s, α_o) the central frequency of the filter, s the scale index, and o the orientation index.

In the implementation employed in this paper, four orientations $(0°, 45°, 90°, 135°)$ are used, seven scales for the first z-score (roughly equivalent to luminance), and only 5 scales for the remaining two components. This difference is justified by the observation that the finest and coarsest scales of these components barely showed any relevant information. Accordingly, while the minimum wavelength for the first z-score is 3 pixels, 6 pixels for color have been used instead. The use of orientations in color components has been observed to improve performance, compared to the use of isotropic responses. Besides, orientation selectivity of chromatic multiscale receptive fields has been shown to take place in V1 and is thought to influence saliency [25]. It has been also tried to include isotropic responses to luminance in addition to the oriented responses, but the results were practically the same. Consequently, they were considered redundant in the computation of saliency, and discarded for the sake of efficiency.

The bank of filters is applied on each of the whitened chromatic components previously obtained. From the complex response to the filter in a given frequency band we compute local energy as the modulus of the response [26][27]. That is:

$$\mathbf{e}_{cos} = \sqrt{(\mathbf{z}_c * f_{os})^2 + (\mathbf{z}_c * h_{os})^2} \tag{4}$$

where index c denotes a whitened chromatic component, $\mathbf{z}_c$ is a retinotopic representation of such component, and f and h denote respectively the even symmetric log-Gabor giving the real part of the response, and the odd symmetric log-Gabor giving the imaginary part of the response. They form indeed a pair of filters in phase quadrature.

Therefore, we obtain a representation of the image in terms of the local energy corresponding to different whitened color components, different scales and orientations.

The next step deals with the adaptation of this representation that already codes the spatial structure. To do so, we have chosen to decorrelate and whiten, independently and in parallel, each set of oriented scales

for each of the chromatic components. For a given whitened chromatic component and orientation, each pixel has a local energy value for each scale. Therefore, each scale s_i can be viewed as an original coordinate axis. The ensemble of scales determines a set of original axis in which each pixel is represented by a point with its coordinates determined by the local energy values for the corresponding scales. From the ensemble of samples (all the pixels), we can compute the covariance matrix in such scale coordinates. If the number of scales is M_s, then the covariance matrix is a $M_s \times M_s$ matrix.

$$\mathbf{C}_{co} = \begin{pmatrix} \sigma^2_{co;s_1} & \cdots & \sigma_{co;s_1 s_{Ms}} \\ \vdots & \ddots & \vdots \\ \sigma_{co;s_1 s_{Ms}} & \cdots & \sigma^2_{co;s_{Ms}} \end{pmatrix} \quad (5)$$

As well known, in natural images different scales are highly correlated, which in general makes all the matrix elements non-zero. Therefore, the whitening procedure based on decorrelation and variance normalization is applied again to achieve a new set of whitened scale coordinates z_i^{sc}. The resulting covariance matrix becomes the identity. This new whitened feature basis is composed of axis that are shifted, rotated and rescaled from the original scale axis. In sum, for each set of scales we have transformed the original scale coordinates of pixels to new coordinates that are decorrelated and with the variance as the norm.

2.3. Saliency

To compute saliency, we simply use the sum of the squared norm of the vectors in the obtained representation as an estimation of pixel (i.e. sample) distinctiveness and we normalize it to the sum across all the pixels.

That is, for each pixel i

$$\|\mathbf{z}_{ico}\|^2 = \mathbf{z}_{ico}^T \mathbf{z}_{ico} \quad (6)$$

where $\mathbf{z}_{ico}$ is the vector associated to the pixel for a color component c and an orientation o with as many components as whitened scales (M_s).

This provides a retinotopic measure of the local feature contrast. In this way, a measure of conspicuity is obtained for each orientation of each of the color components. The next steps involve a Gaussian smoothing and the addition of the maps corresponding to all of the orientations. That is, for a given color component $c = 1...M_c$ and pixel i, the corresponding saliency (S_{ic}) is calculated:

$$S_{ic} = \sum_{o=1}^{M_o} \|\mathbf{z}_{ico}\|^2 \quad (7)$$

Color components undergo the same summation step to get a final map of saliency. Additionally, to ease interpretation of this map as probability to receive attention, it is normalized by the integral of the saliency in the image domain (i.e. the total population activity). Hence, saliency of a pixel i (S_i) is given by:

$$S_i = \frac{\sum_{c=1}^{M_c} S_{ic}}{\sum_{i=1}^{N} \sum_{c=1}^{M_c} S_{ic}} \quad (8)$$

The values obtained for the ensemble of points, arranged in a 2D matrix deliver a map of saliency $\mathbf{S}$ of the same dimensions than the input image.

The figure 1 shows a graphic outline of the model.

It must be noticed that other approaches to integration different from the squared norm have been explored like the raw vector norm, higher power exponents of the vector norm, and even an exponential transformation of the different components followed by a summation.The use of the raw vector norm achieved close -but inferior- performance in predicting fixations, while all the other approaches behaved much worse. All the alternatives failed in several of the psychophysical experiments described in this paper. This observations agree with the view that saliency is related to the classical statistical distance of the feature vector associated to a point from the centre of the distribution of features present in the image. The use of the squared vector norm in our hierarchical whitening approach can be viewed as an efficient estimation of an overall T^2 of Hotelling in an original high-dimensional representation.

Regarding the computational complexity of this implementation, PCA implies a load that linearly grows with the number of pixels (N), and in a cubic manner with the number of components (M), specifically $O(M^3 + M^2N)$. Since we have kept the number of components (color components or scales) fixed and small, the asymptotic complexity depends on the number of pixels. This is determined by the use of the FFT in the filtering process, which is $O(Nlog(N))$. Most saliency models have a complexity which is $O(N^2)$ or higher.

3. Comparison with human fixations

In the last years, the most extended validation procedure for novel models of saliency has been the ability to predict fixations recorded from humans during the free-viewing of natural images, without any specific goal or task [10, 11, 13]. There are some alternative procedures, more difficult to interpret strictly in terms of saliency. For example, object segmentation or task-driven visual

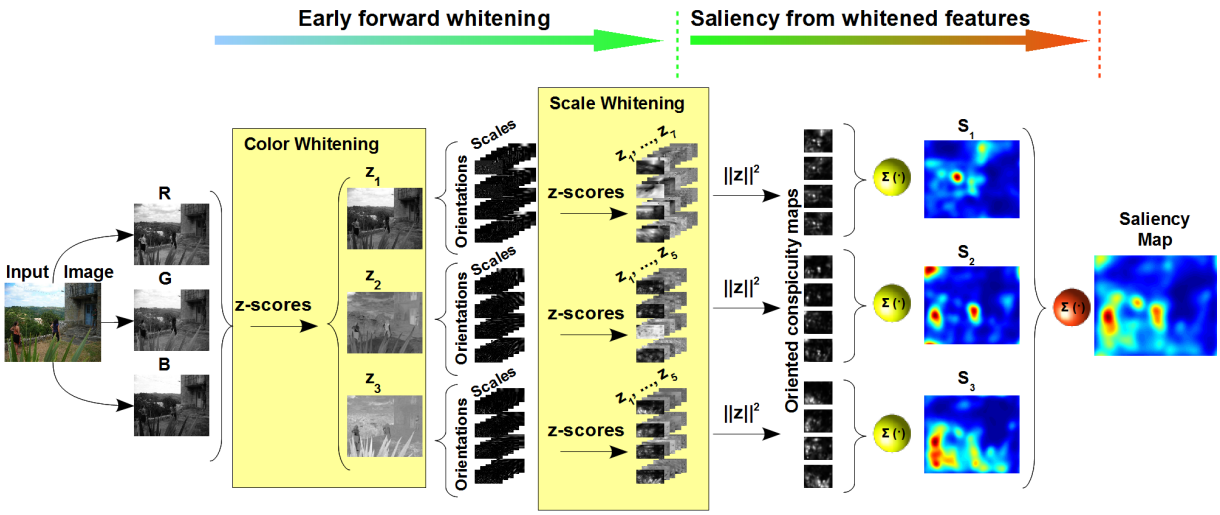

Figure 1: Adaptive whitening saliency model.

search. A very recent work analyses a number of models of saliency through the comparison with human ratings in a task of visual search of military vehicles in photographs, finding statistically significant correlation for most models [28]. However, the dataset employed is top-down biased and also has important feature biases (mostly open green landscapes and sky). The target is in most cases non salient due to its camouflage design, or takes up a large portion of the image –holding a number of salient and non salient parts. Thereby, the implications of such significance in correlation results raise important difficulties of interpretation. Is such correlation related to a general estimation of saliency or rather to an efficient detection –or even segmentation– of military vehicles in countryside scenes?. These concerns have not been set out for the tests based on the prediction of human fixations. This procedure does not rely on reflective decisions about what is *conpicuous*, but relies on a fast action when faced to an image. Moreover, it is precisely related to positions in the space, not to an object of undetermined and changeable area on the image.

Therefore, a major goal in the modeling of saliency is pushing this benchmark further on.

3.1. Datasets and models

Three open-access eye-tracking datasets of natural images have been used. In the three datasets the subjects did not receive any specific instruction. That is, they meet the requirement of free-viewing. The images have been shown in a ramdom order to each subject. The figure 2 shows three example images from each of the datasets.

The first dataset has been published by Bruce and Tsotsos and has 120 images and fixations from 20 subjects [29]. Each image has been viewed during 4 seconds. It has already been used to validate many state-of-the-art models of bottom-up saliency using different procedures [10, 11, 13]. Therefore, it provides a suitable reference for a fair assessment of a novel model in relation to existing approaches.

The second dataset has been published by Kootstra et al. and consists of 99 images and the corresponding fixations of 31 subjects [30]. The viewing time was 5 seconds for each image. One interesting property of this dataset is that it is organized in five different groups of images (12 images of animals, 12 of streets, 16 of buildings, 40 of nature, and 19 of flowers or natural symmetries). This feature may be expected to reveal possible biases in the models. Besides, it may help to analyze the causes of variability under the same experimental conditions.

Finally, the NUSEF dataset has been chosen because it is supposed to have a strong emotional burden [31]. This affective content may be expected to produce an increased human consistency not related to low level features, but to emotions related to abstract concepts strongly suggested by the images. Therefore, models of saliency may be expected to explain less amount of intersubject consistency than in the other datasets. It is composed of 758 images observed on average by 25.3 subjects. The viewing time was 5 seconds for each im-

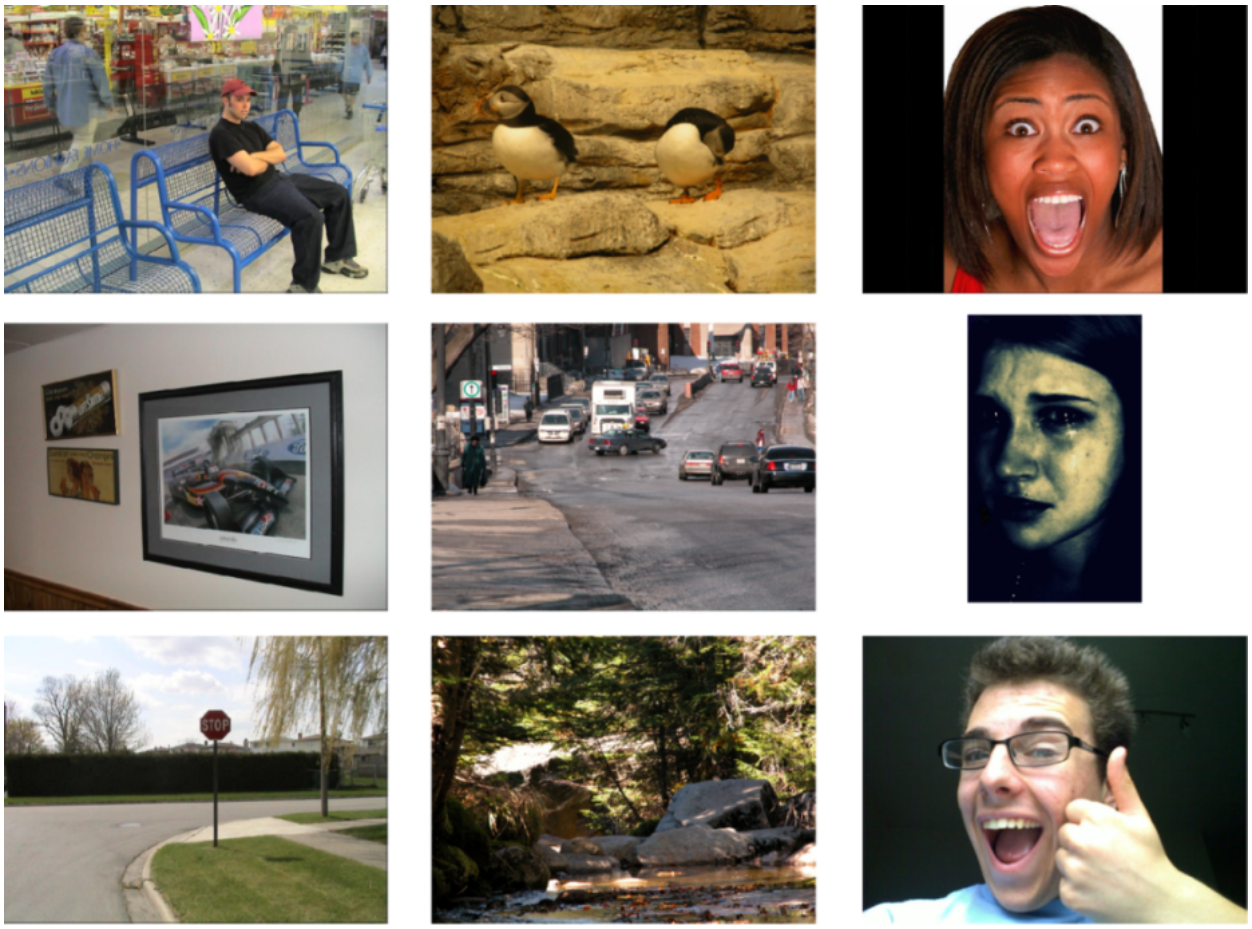

Figure 2: Examples of images from the three datasets used. Bruce and Tsotsos (left); Kotstra et al. (center); NUSEF (right)

age.

Otherwise, we compare the results of the proposed model with other 4 models. Namely: The model of Seo and Milanfar based on selfresemblance [13]; the SUN model [11] that adopts a bayessian approach based on previously learned image statistics; the AIM model that decomposes the image with independent components of natural images and uses self-information as a measure of distinctiveness [10]; and finally the classic model of saliency proposed by Itti et al. [8].

3.2. ROC analysis and KL divergence

To assess the usefulness of the saliency maps to discriminate between fixated and non fixated points, we have used the area under the curve (AUC), obtained from a receiver operating characteristic (ROC) analysis, and a Kullback-Leibler divergence (KLD) comparison. Both methods target the capability of the saliency maps to predict the spatial distribution of fixations through the comparison of the distributions of saliency in fixated versus non-fixated points.

To avoid center-bias, in each image, only points fixated in another image from the same dataset are used as non fixated points. As suggested in [32], standard error is computed through a bootstrap technique, shuffling the other images used to take the non fixated points, exactly like in [11] and in [13]. This last step should not be adopted if the goal is to assess a combination of saliency and a center-bias model. However, since saliency is data-driven and the center-bias is a spatial bias (working regardless of the specific data), they are different mechanisms. Thus, it makes sense to use different evaluations focusing on each of the components.

Furthermore, the use of the bootstrapping method yields a high sensitivity. As recently shown in [19], a ROC analysis as used by many authors (without a bootstrapping to prevent the influence of center-bias) raises problems of sensitivity. However, the use of the boot-

strapping procedure yields a standard error that is tipically below the 3% of the dynamic range spanned by the obtained values for the different models for both the ROC analysis and the KLD comparison.

Nevertheless, a double assessment through a ROC analysis and a KLD comparison is provided to ensure the reliability of the evaluation.

3.3. Results

The table 1 gathers the results obtained on the three datasets. The figures 3 to 5 provide examples of saliency maps for the best performing models that focus on specific features to support the discussion.

The values shown for the model of Itti et al. [8] on the dataset of Bruce and Tsotsos are higher than reported in previous works [11, 13] because, instead of using their saliency toolbox, the original implementation has been used, as made available for Matlab (`http://www.klab.caltech.edu/~harel/share/gbvs.php`). For the other models in this dataset the values that we have obtained are compatible with those published in [11] and [13], thus we have respected the reported values.

3.3.1. Discussion

Firstly, it is worth noting that the results with both measures, ROC analysis and KLD, yield an equivalent ranking and equivalent distances between models on each of the datasets. There are minor differences in two groups of the dataset of Kootstra but they do not give rise to any remarkable difference in the evaluation. Therefore, the comments that follow hold for both. Regarding the sensitivity of the measures, the standard error remains below the 3% of the spanned range of values.

Considering the three datasets, the ranking of models yields only a single change of positions between the model of Seo and Milanfar and the AIM model in the NUSEF dataset. Although with variable distances, the rest of models rank the same position. The AWS model holds clearly the first position in the three datasets.

However, looking at the five groups of the dataset of Kootstra et al., several changes of position involving different models occur. As a result, there is no more a clear coherent ranking. Even though, the AWS maintains the best performance with a distance on the next far beyond the standard error, except for the buildings group in which the model of Seo and Milanfar achieves a slightly better result, but within the uncertainty limits established.

Otherwise, the variations in performance across the datasets may be used to look for biases in the models. The strong advantage of AWS in the group of flowers and natural symmetries finds an explanation in the examples shown in the figure 3. The model of Seo and Milanfar and the AIM model miss completely the saliency of natural symmetries that catch a considerable amount of fixations in these images. In contrast, the AWS model manages to capture the saliency of symmetries in natural scenes.

Besides, other factors appear to contribute to the advantage of the AWS model. It shows a sensitivity to salient high frequency patterns like the striped pattern on the small head of a butterfly that is shown in the figure 4. For the model of Seo and Milanfar, the head seems to be just another part of an edge around the butterfly. Additionally, the behavior of our model when faced to color singletons appears to be more robust. The figure 5 shows a revealing example. The yellow and red peppers are among the most salient objects for both the AWS model and humans. In contrast, the AIM model and particularly the model of Seo and Milanfar find more salient the objects in the upper part of the image, thus showing a lack of sensitivity to color pop-out in this natural context.

3.4. Comparison with human priority

Some questions in relation to the assessment procedure arise. Is it suitable the statistical significance to compare the models or is it too tight in practice?. It may occur that differences in model performance are similar to the variability shown by humans themselves, while being statistically significant. Otherwise, what are the reasons for the high variation in the absolute values across the datasets? Is there any means to create a dataset that provides reliable and definitive results in ranking the models?.

It is clear that the explanation is not a different experimental setup since the largest variation is found across the groups of the Kootstra dataset, all under the same setup. Any explanation should be related to differences in the image content, differences in the associated human behavior, and different biases in the models of saliency.

In order to explore answers to the raised questions, we propose to compare the performance of models with the peformance of single subjects. To assess the predictive performance of a single subject we resort to priority maps derived from the fixations of the subject on each image. From this measure, we may compute an estimation of the average subject performance and an estimation of human performance variability.

Table 1: AUC values obtained with different models of saliency for both of the datasets of Bruce and Tsotsos and Kootstra et al. Standard errors, obtained like in [11], range 0.0004-0.0008. For the groups of the Kootstra et al. dataset, standard errors range 0.0010-0.0018. (* Results reported by [11]; ** Results reported by the authors).

Model	Bruce and Tsotsos dataset	NUSEF dataset	Kootstra et al. dataset					
			Whole dataset	Buildings	Nature	Animals	Flowers	Street
AWS	0.7106	0.6035	0.6205	0.6105	0.5815	0.6565	0.6374	0.7020
Seo and Mil.	0.6896**	0.5802	0.5933	0.6136	0.5530	0.6445	0.5602	0.6907
AIM	0.6727*	0.5902	0.5842	0.5766	0.5628	0.5953	0.5881	0.6393
SUN	0.6682*	0.5782	0.5705	0.5514	0.5484	0.5401	0.6100	0.6458
Itti et al.	0.6456	0.5655	0.5702	0.5814	0.5478	0.6200	0.5217	0.6509
Gao et al.	0.6395*		–	–	–	–	–	–

Table 2: KL divergenge values obtained with different models of saliency for both of the datasets of Bruce and Tsotsos and Kootstra et al. Standard errors, obtained like in [11], range 0.001-0.002 for both of the datasets and the groups. (* Results reported by [11]; ** Results reported by the authors).

Model	Bruce and Tsotsos dataset	NUSEF dataset	Kootstra et al. dataset					
			Whole dataset	Buildings	Nature	Animals	Flowers	Street
AWS	0.321	0.071	0.099	0.109	0.058	0.188	0.142	0.307
Seo and Mil.	0.278**	0.048	0.071	0.110	0.049	0.175	0.057	0.281
AIM	0.203*	0.055	0.055	0.070	0.045	0.105	0.085	0.197
SUN	0.210*	0.043	0.039	0.046	0.033	0.049	0.097	0.173
Itti et al.	0.175	0.033	0.038	0.069	0.032	0.109	0.025	0.200

3.4.1. Human priority performance

To implement this measure, priority maps derived from fixations have been used, following the method described by [30]. This method lies in the subtraction of the distance between each point and its nearest fixation from the maximum possible distance in the image. As a result, fixated points have the maximum value and non fixated points have a value that decreases linearly with the distance to the nearest fixation. The resulting maps can be used as probability distributions of subjects fixations (priority maps), and can be considered as subjective measures of saliency. At least with few fixations per subject, as it is the case, this method yields better predictive results than the approach to compute priority maps based on filtering of fixations with Gaussians kernels [29]. This last approach tipically assigns zero or decimal priority to points beyond 2.3° of visual angle, since usually the width of the Gaussian is fixed to 1° and the amplitude is fixed to 255, which is the maximum of the dynamic range used in the ROC analysis. Consequently, with few fixations by subject, the priority maps present values below 1 for positions that can be close to fixations. Therefore, in a ROC analysis all these locations are equally considered zero priority points. Exactly the same as points much further from any fixation.

Furthermore, the linear distance-based method is parameter free. Thereby, we do not need to make assumptions on the range of points that may have influenced a given fixation. Of course, it can be argued that it is not justified to assume that priority drops linearly with distance to fixations. Nevertheless, it seems actually reasonable to assume that priority drops monotonically with distance to the nearest fixation. If the method to compare and evaluate maps is invariant to monotonic transformations, as ROC analysis is, then there is no issue with using linear, or any other monotonic maps. Hence, through a ROC analysis, the same one employed to evaluate models of saliency, the capability of these maps to predict the fixations of the set of subjects can be assessed, without concerns on the dynamic range of the ROC analysis. It must be noticed that we have not used the averaged priority maps shown in the figures 3 to 5, but maps computed specifically for each of the subjects using the procedure described above.

The previous evaluation for each subject has been done, only for those with fixations for all of the images. One individual has been excluded of the dataset of Bruce and Tsotsos, whose deviation from the average of humans was larger than twice the standard deviation, and who also had just one fixation in many images. This yields priority maps from 9 subjects for the dataset of Bruce and Tsotsos, and 25 subjects for the dataset of Kootstra. On the NUSEF dataset our approach to derive

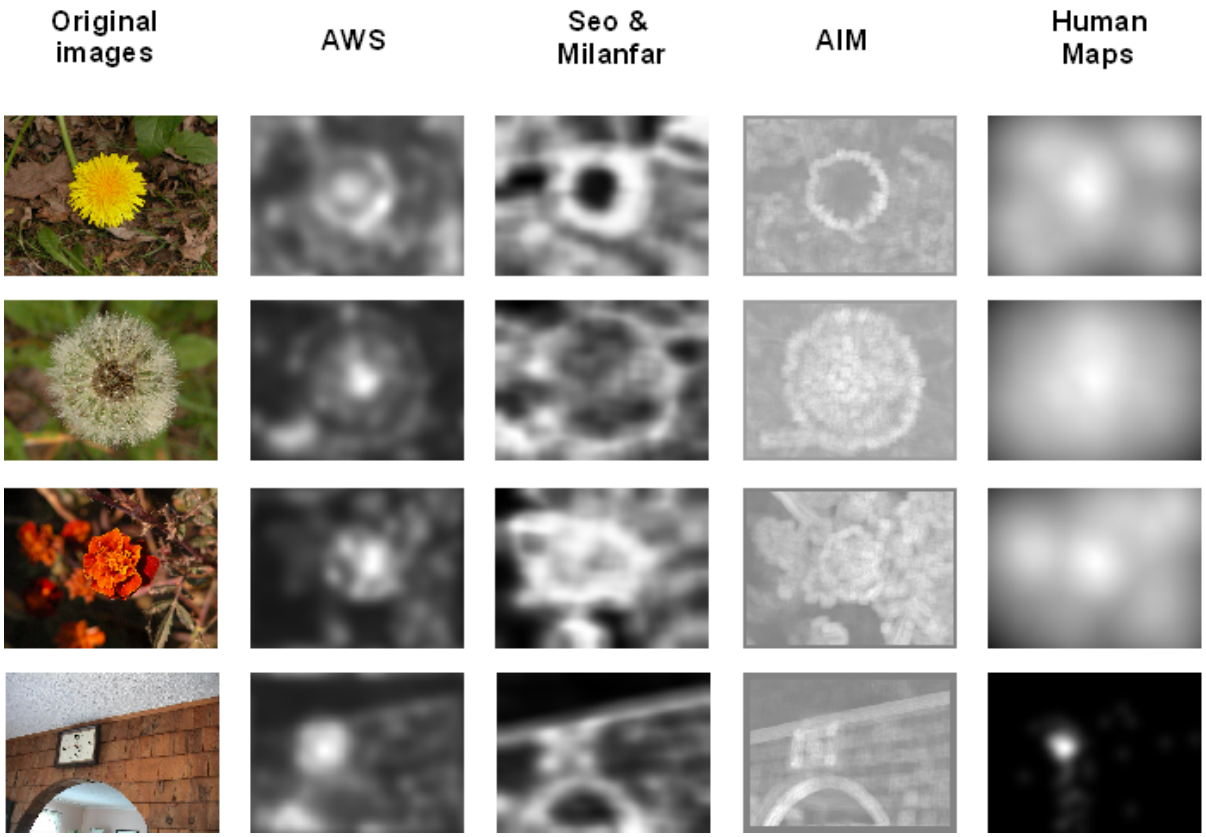

Figure 3: Examples of results with 4 images with a dominant symmetric point. For comparison, human priority maps provided by the authors are shown. Bruce and Tsotsos derived priority from fixations using Gaussian kernels [29], while Kootstra et al. used a distance-to-fixation transform [30]. This explains the noticeable differences in the dynamic range of the priority maps.

subject priority maps finds a problem: no subject has observed all the images. Nevertheless, all the images have been observed by at least 13 subjects. Therefore, we have built 13 pseudo-subjects gathering the priority maps of the first 13 observers that viewed each of the images, following the order of subjects provided by the authors.

The performance of the priority maps associated to a given subject was obtained through the assessment with fixations of other subjects in the dataset. Computing the average, we have the average performance of priority maps associated to different subjects. Besides, the double of the standard deviation provides an estimation of the range of predictive performance for the 95% of humans, under the assumption of a normal distribution for AUC priority values. This was true for the datasets and groups studied, with a kurtosis value very close to 3. Moreover, this interval of variability between subjects can be also used as a measure of the minimum relevant distance between two models. Differences lower than such variability may be regarded as producing no practical effect on performance. The results are given in the table 3.

3.4.2. Saliency versus priority

At a first look, we can see how the performance of priority varies across datasets similarly to saliency maps. The variability of subject performance in a given dataset is well an order of magnitude higher than the statistical significance of the measure of performance, except for the NUSEF that shows much less variability.

Remarkably, the performance of the AWS model is compatible with the estimated human priority performance for the three datasets and all the groups. The model of Seo and Milanfar is also compatible with the average human for the dataset of Bruce and Tsotsos, and for two of the five groups of Kootstra et al.. However this compatibility does not hold for either the whole dataset of Kootstra et al. or the NUSEF dataset. The model by Bruce and Tsotsos is only marginally compatible with the average human with their own dataset.

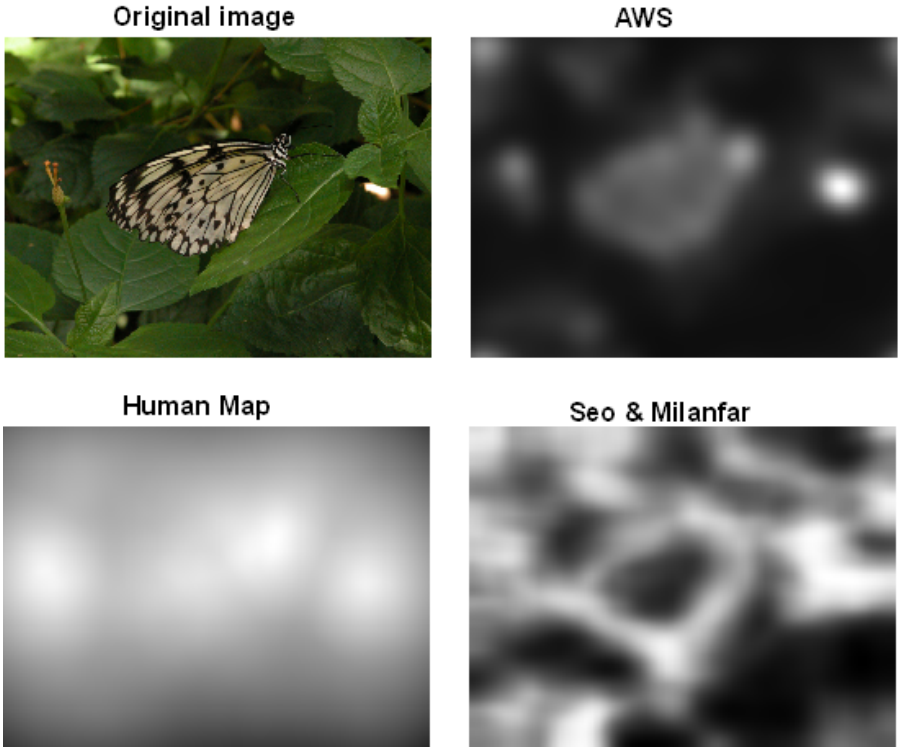

Figure 4: Example of a salient high spatial frequency pattern in the small head of a butterfly.

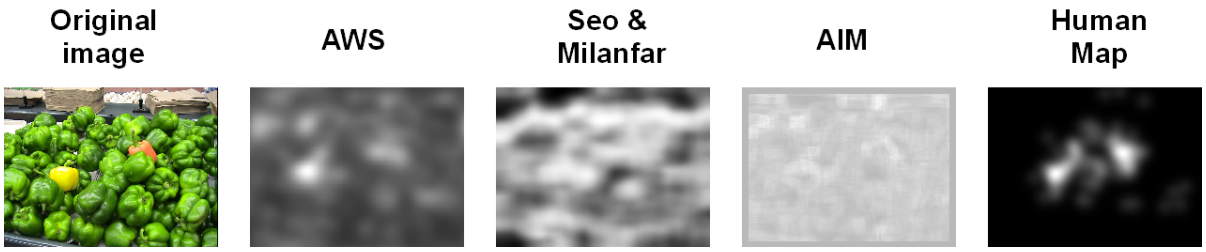

Figure 5: Example of dominant color saliency.

Furthermore, AWS is the only one that outperforms several subjects in all the cases, representing nearly half of the observers of Kootstra et al. dataset, and more than half of them in the dataset of Bruce and Tostsos and NUSEF. In this last two datasets, our model performs even slightly over the average value obtained with priority maps, as well as in the street group of Kootstra et al.

3.4.3. Discussion of results

The table 4 shows the difference between the results of each model and the average performance of human priority maps. Positive values imply higher predictive capability of the model and negative values imply higher predictive capability of the priority maps. As advanced, the interval of the 95% of subjects is provided as relevant difference, since AUC standard errors are comparatively negligible.

The results achieved by the AWS model in the three datasets and the groups of Kootstra become highly consistent when referred to the performance on the priority maps, not only in ranking position but also in the absolute value. The uncertainty limits cover without problems the compatibility between any pair of values. The proposed model seems thus to be remarkably robust against scene change. That is, the AWS model does not show any kind of scene bias. It does not seem to be specially fitted to manage better particular kinds

Table 3: Average predictive capability of humans using distance-to-fixation priority maps.

Bruce and Tsotsos dataset		NUSEF dataset		Kootstra et al. dataset							
				Whole dataset		Buildings		Animals		Street	
Mean	2σ	Mean	2σ	Mean	2σ	Mean	2σ	Mean	2σ	Mean	2σ
0.6946	0.0248	0.6010	0.0070	0.6254	0.0224	0.6154	0.0330	0.6672	0.0356	0.6923	0.0402
						Nature		Flowers			
Max	Min	Max	Min	Max	Min	Mean	2σ	Mean	2σ		
0.7156	0.6805	0.6069	0.5972	0.6462	0.6056	0.5874	0.0194	0.6419	0.0245		

Table 4: Results of comparing predictive capabilities of saliency models, subtracting the average predictive capability of humans. Positive sign means better, and negative sign means worse, than the average human. (All results derived from tables 1 and 2).

Model	Bruce and Tsotsos dataset	NUSEF dataset	Kootstra et al. dataset					
			Whole dataset	Buildings	Nature	Animals	Flowers	Street
95% of humans	± 0.025	± 0.007	± 0.022	± 0.033	± 0.019	± 0.036	± 0.025	± 0.040
AWS	0.016	0.002	-0.005	-0.005	-0.006	-0.011	-0.004	0.010
Seo and Mil.	-0.005	-0.021	-0.032	-0.002	-0.034	-0.023	-0.082	-0.002
AIM	-0.022	-0.011	-0.041	-0.039	-0.025	-0.072	-0.054	-0.053
SUN	-0.026	-0.023	-0.055	-0.064	-0.039	-0.127	-0.032	-0.047
Itti et al.	-0.049	-0.036	-0.055	-0.034	-0.040	-0.047	-0.120	-0.041

of saliency, present in different scenes. Also, it clearly exhibits the highest performance among the analyzed models that constitute a representative sample of the state of the art.

The model of Itti et al also presents consistency across datasets –with the lowest performance in the three cases. The remaining three models (Seo and Milanfar, AIM and SUN) are not able to keep a consistent behavior across the three datasets, when compared to performance of human priority.

Furthermore, AWS is the only model that maintains consistency when the five groups of Kootstra are considered. This points out to scene or feature biases in the different models, and to a difficulty to catch certain salient features that are present in natural images. Some examples related to symmetries, high frequency patterns and color have been already shown.

However, the fact that AWS completely matches the predictive capability of the used human priority maps deserves further comments. The maps of the proposed model show equivalent performance to an average human priority map. It seems that the model is able to explain the shared factor that drives human fixations, during the free-viewing of natural images. Therefore, there do not seem to be shared top-down effects driving fixations up to increase the predictive capability of humans to a level that saliency is not able to explain. From our viewpoint, this fact reinforces the importance of bottom-up saliency in the explanation of the behavior shared by different subjects. It also questions the real implications of results like those provided by [33] or by [34], involving top-down influences. Of course, it can be objected that the proposed approach to estimate human priority may miss some of the consistency between subjects and that it is susceptible of improvements.

Even if the mesure of human priority is seen as missing consistency, there is no reason to think that there is a special failure in a particular dataset, like for instance the NUSEF. The fact that the relative performance of priority versus saliency does not increase in this dataset deserves a comment. Indeed, the relative performance of saliency versus priority in NUSEF even increases for all the models considered when compared to the dataset of Kootstra. Therefore, there is no sign of increased consistency between subjects caused by the emotional burden of the images.

It seems that even with affective images, the top-down factors that might influence the selection of fixation points by subjects in a free-viewing task, produce divergent results across subjects. This observation is in agreement with the results of a study of Tatler et al. on the factors that drive early fixations [32].

It is worth noting that the performance of 95% of modeled subjects is always among 0.02-0.04 around the average value, except for the NUSEF dataset. In this dataset, priority maps seem to be more consistent. The variability in the performance of human priority drops to the third part of the variability observed in the other two datasets. Two differences may be expected to explain this fact: the much higher number of images in NUSEF; or the fact that priority maps in this dataset are derived from a collage of actual subjects, thus fading

any component of personal strategy in driving fixations. We have assessed the priority maps of NUSEF using only the first 100 images and the variability becomes 0.009, a bit higher but still well below the variability exhibited in the other two datasets with respectively 99 and 120 images. Therefore, the higher number of images does not seem to be the reason of the drop in variability registered in the NUSEF.

On the other hand, does the emotional burden produce individual and divergent reactions of subjects, influencing fixations in NUSEF?. Is this the explanation to the equivalent drop in priority and saliency performance registered?. Further research, beyond the scope of this paper, is needed to deal with these questions on the lack of human consistency.

4. Reproduction of psychophysical and perceptual results

A frequent complement to the use of eye-tracking data is to evaluate the ability to reproduce psychophysical results. In this respect, the AWS model is able to reproduce a series of results, usually associated to saliency. The selection of cases includes: 1) perceptual comparisons that do not involve eye movements (orientation contrast); 2) quantitative behavior that is not invariant under monotonic transformations of the measure of saliency (Weber's law); 3) visual search asymmetries reproduced by other models of saliency; 4) efficiency and inneffeciency in visual search under typical and well described arrangements; and 5) figure-ground segmentation on natural scenes.

4.1. Non-linearity against orientation contrast

The saliency of a target, perceived by humans as a function of orientation contrast, has been observed to increase in a non-linear manner [35]. Saliency was measured through subjective comparisons with targets of different luminance contrast. It increases rapidly at the beginning, between $20 - 50°$, up to a nearly constant saturation value. The figure 6 shows how both the AWS model and the model of Seo and Milanfar match well the described behavior. This result has also been reproduced by [36]. However, other models fail to do it, at least with the setups made public by the authors [10, 8]. The same figure shows how the AIM model fails to reproduce this behavior and reaches the saturation with a minimal orientation contrast.

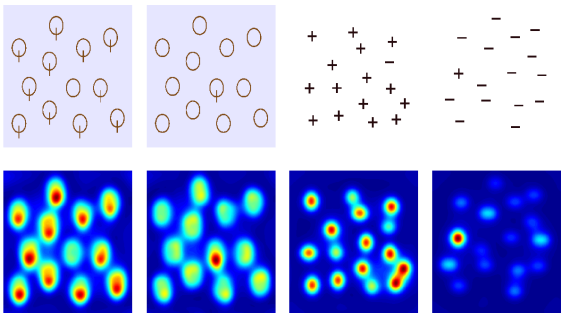

Figure 8: Left: Saliency against relative variation of length reproduces the Weber's law observed in humans. Right: two examples of the so called presence-absence asymmetry.

4.2. Weber's law and search asymmetries

A classical psychophysical study conducted by Treisman and Gormican, showed that certain features characterizing stimuli lead to an asymmetric behavior in visual search tasks [37]. One remarkable example is the presence-absence asymmetry, observed for a pair of stimuli differing only in the presence or absence of a given simple feature. While a target presenting that feature, surrounded by distractors lacking it causes a clear pop-out phenomenon, the reverse target-distractor distribution does no pop-out. As can be seen in the figure 8, AWS reproduces this behavior well, in two typical examples: the plus and minus symbols, and a circle with and without a bar.

In the same work, Treisman and Gormican report another result, closely related to this asymmetry. Saliency of a given stimulus satisfies the Weber's law, so that it grows linearly with a relative enlargement in one dimension. As can be seen also in the figure 7, AWS fulfills this behavior as well. The model of Seo and Milanfar fails in reproducing this result, while the AIM reproduces the behavior with an oscillation around the correct straight line, perhaps related to the definition of fixed sizes of the patches of independent components of natural images used to decompose the image.

Search asymmetries also affect color. Rosenholtz et al. have studied these asymmetries in depth, as well as the influence of background on their deployment [23]. Again, given a pair of stimuli, now with the same luminance and differing only in one color coordinate (in a MacLeod and Boynton color space), they exhibit different search times depending on which is the target and which is the distractor. Nevertheless, the background influences this effect, to the point to reverse it. For example, with a gray background, a redder stimulus is more salient than a less red one. However, if the background is red, then the redder stimulus becomes less salient.

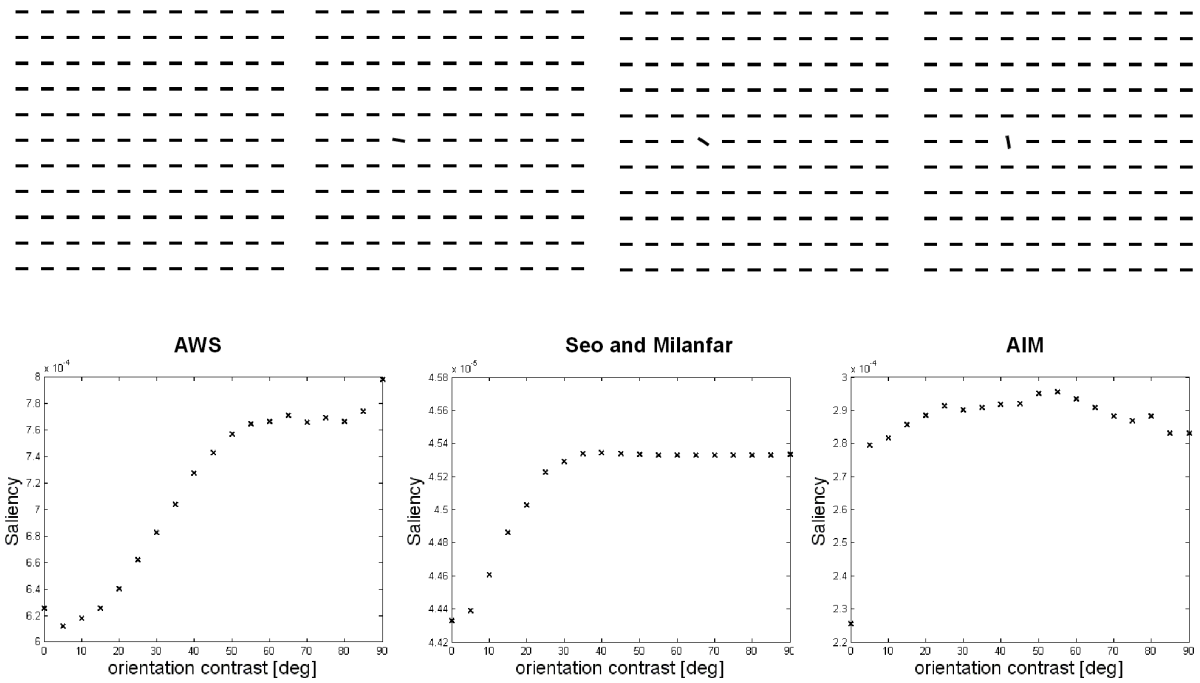

Figure 6: Obtained saliency against orientation contrast of the target and four examples of the images used (Images adapted from [35]).

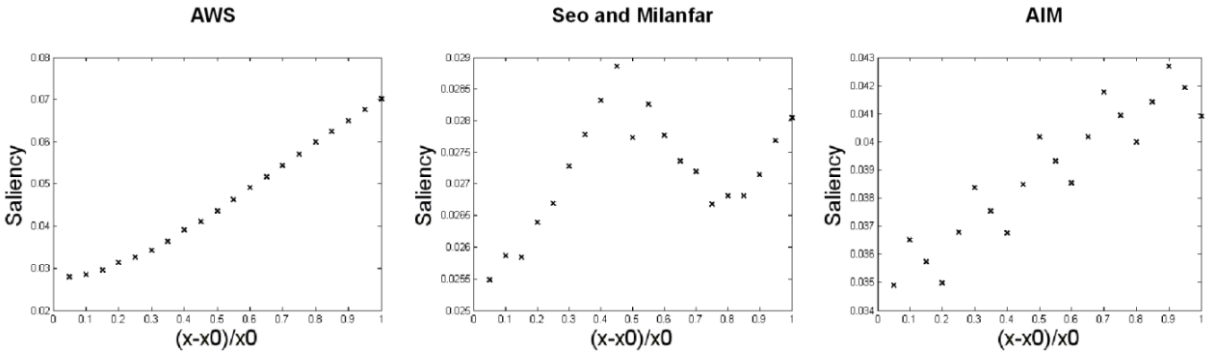

Figure 7: Left: Saliency against relative variation of length reproduces the Weber's law observed in humans. Right: two examples of the so called presence-absence asymmetry.

The figure 9 shows four images reproducing those used by Rosenholtz et al in their experiments. As well, the saliency maps obtained with the AWS model and the model of Seo and Milanfar are shown. Both the asymmetry and its reversal by a change of background are correctly reproduced by the AWS model. The singleton appears always salient, but it is more salient when it presents a higher color contrast against the background. On the other hand, the model of Seo and Milanfar fails even to detect the color singleton in any of the images and seems to suffer from a dominant behavior of spatial interactions between stimuli over color interactions. This result is in agreement with a behavior observed in some natural images as illustrated in the figure 5.

4.3. Efficient and inefficient visual search phenomena

The model also suitably reproduces pop-out phenomena related to orientation, color, size or texture, widely reported in literature [38]. In the figure 10 it is shown how singletons of color, orientation, and size, clearly pop-out with the AWS model. But also how a unique closed circle surrounded by randomly oriented open curves, or a cross surrounded by similarly oriented intersections, does not pop-out and undergoes an inefficient search.

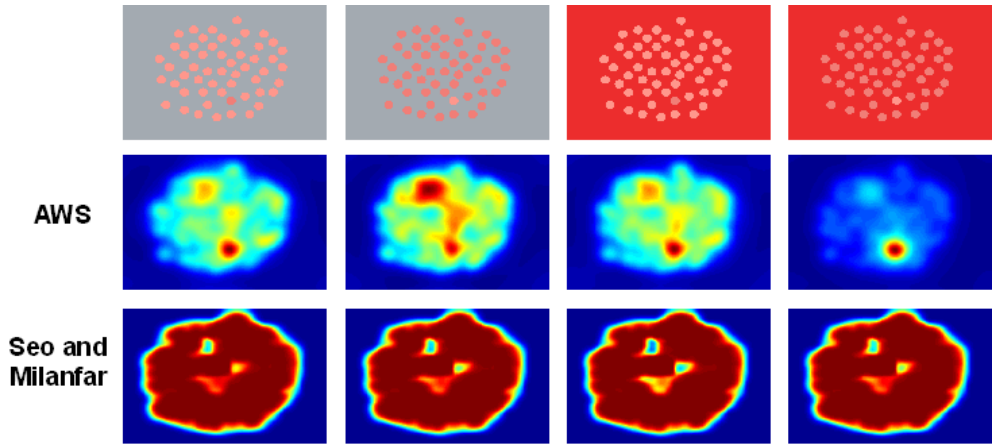

Figure 9: Color search asymmetry and its reversal by a change in the background color. Images adapted from [23]

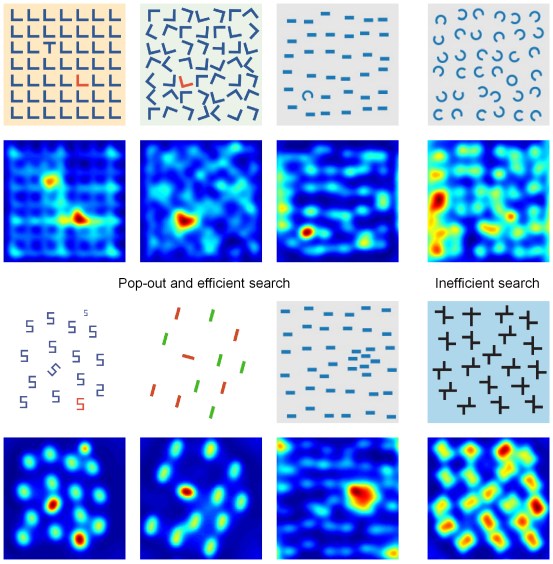

Figure 10: Typical examples of pop-out, efficient and inefficient search observed in humans, and reproduced by the AWS. Images adapted from [38, 10, 14].

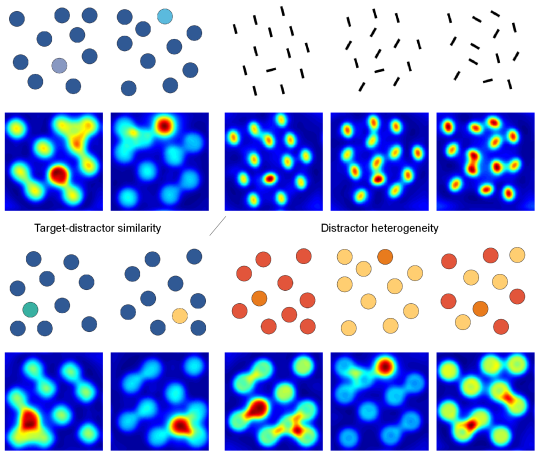

Figure 11: AWS matches human behavior against target-distractor similarity and distractor heterogeneity. Images adapted from [38] and [10].

Saliency maps catch well the non-linear influence of target-distractor color similarity, and from a given difference between target and distractors, saliency does not increase any more. This result is equivalent to the non-linearity against orientation contrast previously shown, but with color. It has been included here because the images used in the figure 11 are related to visual search experiments, not based on perceptual comparisons.

Finally, it must be pointed out how distractor heterogeneity, another important factor that affects the saliency of a color or orientation singleton in human observers, gives place to a similar behavior by the AWS model.

4.4. Figure-ground segmentation in natural scenes

The AWS model allows the extraction of proto-objects, in a similar manner to that used with previous models of saliency. This ability is very interesting, since it can be useful to reduce the search space in many visual operations, such as object detection and recognition, unsupervised image classification, or natural landmark generation and detection.

To segment the saliency maps, we have used a naive implementation of the watershed algorithm. The watershed approach has the advantage of being parameter free, which eases comparison with other pre-processing

approaches. The particular implementation employed here uses local maxima of the saliency map as sources.

To show the quality of these proto-objects, some results are provided in the figure 12 on 14 images with different degrees of clutter and lighting (luminance and color) conditions, as well as different relevant scales and spatial structure. For each image, the regions containing the six highest local maxima of saliency have been selected, which delivers at most 6 proto-objects.

As can be seen, in general the model extracts some proto-objects that correspond to meaningful objects, or to identifiable parts of them. Also, some salient textures are caught as proto-objects. Further valuable information can be found in partial saliencies and oriented conspicuities for a more refined approach. These results point to the suitability of the model for its application in machine vision solutions to the detection and analysis of unknown objects, as already done with other measures of saliency.

To summarize, it has been shown that the AWS is able to reproduce a wide and representative set of psychophysical phenomena, to our knowledge not reproduced together by any other model before. It is important to remark that, unlike other models that have tuned their setup or forced the object scale [13, 12], here, the exact implementation described in section 3 has been used, without any kind of additional bias. Furthermore, in the segmentation examples, it has been used a simple and parameter-free procedure that delivers results without any kind of special tuning or adaptation.

5. Conclussions

In this paper we have proposed a novel approach to visual saliency that arises as a simple norm vector computation from the adaptation of the basis of low level features used to decompose the image. Such adaptation has been simply achieved through hierarchical whitening of a classical representation of the image. This strategy extends the adaptation to the statistical structure of the set of natural images (off-line adaptation through modeling or statistical analysis) present in most models, to the statistical structure of specific images (on-line approach).

The particular implementation described is simple and light. It outperforms other important models of the state of the art. It improves their results in the prediction of human eye fixations, while keeping a low computational complexity. Besides, it reproduces a wide variety of relevant psychophysical results.

The use of biased groups of natural images from the dataset of Kootstra et al., combined with the analysis of specific examples has shown useful to reveal biases in the existing models of saliency. We think that such detection of biases is very important for the improvement of the existing approaches, since it provides guidelines to follow in further developments.

Regarding the comparison with human fixations, we point out a clear incompatibility between the huge variation in the results depending on the used dataset and the very tight values of uncertainty delivered by the procedure of Tatler et al. [32]. To overcome the problem, a comparison is proposed with the predictive capability shown by priority maps derived from the fixations of single subjects. Hence, results with two datasets, originally very different, become compatible. With this procedure, the AWS approach shows the same predictive capability than an average single-subject priority map. Moreover, it still clearly outperforms all other models that show an evident lack of robustness against a number of features like salient symmetries, high frequency patterns, and certain salient color arrangements. Such lack of robustness points to the existence of constraining design biases in these models.

Under the proposed approach to single-subject priority, bottom-up saliency explains the observed intersubject consistency without the need of any top-down mechanism. The equivalent results achieved on NUSEF, supposed to present a strong emotional burden, reinforce the idea of saliency as the major factor underlying the observed intersubject consistency in the spatial distribution of fixations in free-viewing tasks.

Appendix A. Whitening procedure

Regarding color information, it has been observed that results are barely affected by the choice of the whitening procedure, by testing several approaches based on PCA and ICA [39, 40]. The results are totally independent of the whitening method employed with scale information, since they only differ in a rotation that will not alter the subsequent computation of vector norm. Therefore, decorrelation is done through PCA, since it is a first order procedure that provides an ordered decomposition of the data. Its lower computational complexity is a clear advantage against higher order methods, like the diverse ICA algorithms. Thus, the principal components are obtained, and then normalized by their variance. This last step delivers a whitened, and still ordered, representation.

Let $\mathbf{x}$ be the representation of the image in the original –color or scale– space, $\mathbf{y}$ the corresponding representation in principal components, and $\mathbf{z}$ (z-scores) the corresponding representation in the whitened coordinates.

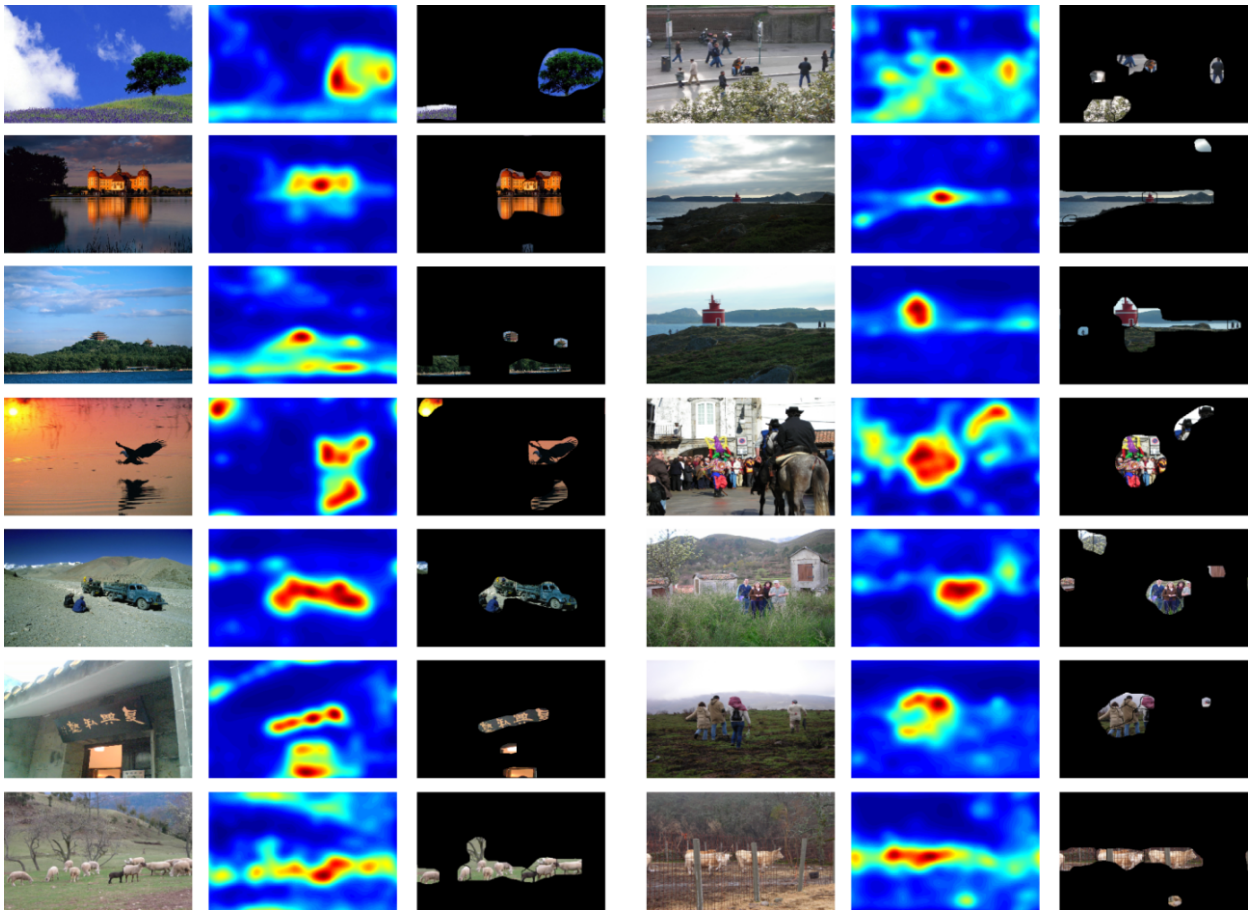

Figure 12: Examples of saliency-based segmentation: original image (lefts), saliency maps (center), and proto-objects (rigth) arraganged in two vertical blocks. Six of the images have been obtained from [12], the rest are ours.

That is,

$$\mathbf{x} = \left(x_j\right) \rightarrow \mathbf{y} = \left(y_j\right) \rightarrow \mathbf{z} = \left(z_j\right) \qquad \text{(A.1)}$$

with $j = 1...M$, where M is the number of components.

The whitening procedure can be summarized in two steps. First, as well known, principal components result from diagonalization of the covariance matrix, ordering eigenvalues (l_j) from higher to lower. To compute the covariance matrix there are N samples, as many as the number of pixels in the input image. The whitened $\mathbf{z}$ representation is then obtained through normalization by variance, given by the eigenvalues. This means that for each principal component:

$$z_j = \frac{y_j}{\sqrt{l_j}} \; ; \; j \in [1, M] \qquad \text{(A.2)}$$

These z-scores yield a whitened representation, with the covariance matrix being the unity matrix. The squared norm of a vector in these coordinates is in fact the statistical distance in the original $\mathbf{x}$ coordinates.

Appendix B. Reproducibility

A Matlab p-code file to reproduce the experimental results reported in this paper as well as all the saliency maps computed with the AWS model for the three eye-tracking datasets are available on the web page `http://www-gva.dec.usc.es/persoal/xose.vidal/research/aws/AWSmodel.html`.

References

[1] Z. Liu, H. Yan, L. Shen, K. N. Ngan, Z. Zhang, Adaptive image retargeting using saliency-based continuous seam carving, Optical Engineering 49 (2010) 1–10.

[2] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, R. Pfeifer, Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub, in: Int. Conf. on Robotics and Automation (ICRA), pp. 962–967.

[3] C. Kanan, G. Cottrell, Robust classification of objects, faces, and flowers using natural image statistics, in: IEEE int. Conf. on Computer Vision and Pattern Recognition (CVPR).

[4] J. Harel, C. Koch, On the optimality of spatial attention for object detection, in: Attention in Cognitive Systems 2009, pp. 1–14.

[5] D. Gao, S. Han, N. Vasconcelos, Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 989.

[6] B. Alexe, T. Deselaers, V. Ferrari, What is an object?, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 73–80.

[7] N. Parikh, L. Itti, J. Weiland, Saliency-based image processing for retinal prostheses, Journal of Neural Engineering 7 (2010) 016006.

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

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

[10] N. D. Bruce, J. K. Tsotsos, Saliency, attention, and visual search: An information theoretic approach, Journal of Vision 9 (2009) 5.

[11] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, G. W. Cottrell, SUN: a bayesian framework for saliency using natural statistics, Journal of Vision 8 (2008) 32.

[12] X. Hou, L. Zhang, Dynamic visual attention: Searching for coding length increments, in: Advances in Neural Information Processing Systems (NIPS), volume 21, pp. 681–688.

[13] H. J. Seo, P. Milanfar, Static and space-time visual saliency detection by self-resemblance, Journal of Vision 9 (2009) 12–15.

[14] X. Hou, L. Zhang, Thumbnail generation based on global saliency, in: Advances in Cognitive Neurodynamics (ICCN), pp. 999–1003.

[15] C. Guo, Q. Ma, L. Zhang, Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform, in: IEEE Conf on Computer Vision and Pattern Recognition (CVPR).

[16] R. Achanta, S. Hemami, F. Estrada, S. Ssstrunk, Frequency-tuned salient region detection, in: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).

[17] M. Cerf, E. P. Frady, C. Koch, Faces and text attract gaze independent of the task: Experimental data and computer model, Journal of vision 9 (2009).

[18] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: IEEE 12th Int.l Conf. on Computer Vision, IEEE, pp. 2106–2113.

[19] Q. Zhao, C. Koch, Learning a saliency map using fixated locations in natural scenes, Journal of vision 11 (2011).

[20] J. van de Weijer, T. Gevers, A. D. Bagdanov, Boosting color saliency in image feature detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (2006) 150–156.

[21] N. D. Bruce, P. Kornprobst, On the role of context in probabilistic models of visual saliency, in: IEEE International conference on image processing (ICIP), p. 30893092.

[22] A. Garcia-Diaz, X. Fdez-Vidal, X. Pardo, R. Dosil, Decorrelation and distinctiveness provide with human-like saliency, in: Advanced Concepts for Intelligent Vision Systems, pp. 343–354.

[23] R. Rosenholtz, A. L. Nagy, N. R. Bell, The effect of background color on asymmetries in color search, Journal of Vision 4 (2004) 224–240.

[24] D. J. Field, Relations between the statistics of natural images and the response properties of cortical cells, Journal of the Optical Society of America A 4 (1987) 2379–2394.

[25] L. Zhaoping, R. J. Snowden, A theory of a saliency map in primary visual cortex (V1) tested by psychophysics of colouror-ientation interference in texture segmentation, Visual Cognition 14 (2006) 911–933.

[26] P. Kovesi, Invariant measures of image features from phase information, Ph.D. thesis, Department of Psychology, University of Western Australia, 1996.

[27] M. C. Morrone, D. C. Burr, Feature detection in human vision: A phase-dependent energy model 1998, in: Proc. of the Royal Society of London. Series B, Biological Sciences, pp. 221–245.

[28] A. Toet, Computational versus psychophysical image saliency: A comparative evaluation study, IEEE Transactions on Pattern Analysis and Machine Intelligence (preprint) (2011).

[29] N. Bruce, J. Tsotsos, Saliency based on information maximization, in: Advances in Neural Information Processing Systems (NIPS), volume 18, p. 155.

[30] G. Kootstra, A. Nederveen, B. de Boer, Paying attention to symmetry, in: Proc. of the British Machine Vision Conference (BMVC), pp. 1115–1125.

[31] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, T. S. Chua, An eye fixation database for saliency detection in images, in: European Conf. on Computer Vision (ECCV), pp. 30–43.

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

[33] W. Einhuser, M. Spain, P. Perona, Objects predict fixations better than early saliency, Journal of Vision 8 (2008) 18.

[34] E. Birmingham, W. F. Bischof, A. Kingstone, Saliency does not account for fixations to eyes within social scenes, Vision Research 49 (2009) 2992–3000.

[35] H. C. Nothdurft, The conspicuousness of orientation and motion contrast, Spatial Vision 7 (1993) 341–363.

[36] D. Gao, V. Mahadevan, N. Vasconcelos, On the plausibility of the discriminant center-surround hypothesis for visual saliency, Journal of Vision 8 (2008) 13.

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

[38] J. M. Wolfe, T. S. Horowitz, What attributes guide the deployment of visual attention and how do they do it?, Nature Reviews Neuroscience 5 (2004) 495–501.

[39] A. Hyvrinen, E. Oja, A fast fixed-point algorithm for independent component analysis, Neural Computation 9 (1997) 1483–1492.

[40] J. F. Cardoso, A. Souloumiac, T. Paris, Blind beamforming for non-Gaussian signals, in: IEE Proceedings on Radar and Signal Processing 1993, volume 140, pp. 362–370.