Signal Processing: *Image Communication* 28 (2013) 642–658

Contents lists available at SciVerse ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

RARE2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis

Nicolas Riche*, Matei Mancas, Matthieu Duvinage, Makiese Mibulumukini, Bernard Gosselin, Thierry Dutoit

UMONS, 31, Bld Dolez, Mons, Belgium

ARTICLE INFO

Article history:
Received 30 October 2012
Received in revised form
28 March 2013
Accepted 31 March 2013
Available online 6 April 2013

Keywords:
Bottom-up saliency
Comparative statistical analysis
Multi-scale rarity mechanism
Regions of interest
Saliency models evaluation
Visual attention

ABSTRACT

For the last decades, computer-based visual attention models aiming at automatically predicting human gaze on images or videos have exponentially increased. Even if several families of methods have been proposed and a lot of words like centre-surround difference, contrast, rarity, novelty, redundancy, irregularity, surprise or compressibility have been used to define those models, they are all based on the same and unique idea of information *innovation* in a given *context*.

In this paper, we propose a novel saliency prediction model, called RARE2012, which selects information worthy of attention based on multi-scale spatial rarity. RARE2012 is then evaluated using two complementary metrics, the Normalized Scanpath Saliency (NSS) and the Area Under the Receiver Operating Characteristic (AUROC) against 13 recently published saliency models. It is shown to be the best for NSS metric and second best for AUROC metric on three publicly available datasets (Toronto, Koostra and Jian Li).

Finally, based on an additional comparative statistical analysis and the effect-size Hedge' g^* measure, RARE2012 outperforms, at least slightly, the other models while considering both metrics on the three databases as a whole.

1. Introduction

There is no common definition of human attention, and it can differ depending on the domain (psychology, neuroscience or engineering) or the considered approach. But, in a general sense, human attention can be defined as the natural capacity to prioritize the incoming stimuli and selectively focus on part of them. The goal of the attentional process is to identify as quickly as possible those parts of our environment that are key to our survival. Humans but also all animals use this mechanism in their daily life and even during dreams when the rapid eye movements occur (REM stage), which are saccades and fixations on the dream scene.

The interest of attention prediction is more and more understood by the scientific community with an exponential number of papers dealing with saliency algorithms. Attention modeling has very wide applications such as machine vision, surveillance, data reduction and compression, human computer interfaces, advertising assessment or robotics. In this context, efficient attention models are of great importance for vision and signal processing algorithms improvements in the future.

In computer science, attention modeling is mainly based on the concept of "saliency maps", which provides, for each pixel, its probability to attract human attention. The idea is that the gaze of people will direct to areas which, in some way, stand out from the background. Saliency implies a competition between an objective "bottom-up" attention and a subjective "top-down" information. Bottom-up attention is a generic approach also known as stimulus-

* Corresponding author. Tel.: +32 65 37 47 25.
E-mail addresses: Nicolas.Riche@umons.ac.be (N. Riche), Matei.Mancas@umons.ac.be (M. Mancas), Matthieu.Duvinage@umons.ac.be (M. Duvinage), Bernard.Gosselin@umons.ac.be (B. Gosselin), Thierry.Dutoit@umons.ac.be (T. Dutoit).

0923-5965/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.image.2013.03.009

driven or exogenous attention. Furthermore, it relies on the information innovation that the features extracted from the image can bring in a given spatial context. The top-down component of attention, which is also known as task-driven or endogenous attention, integrates specific knowledge that the viewer could have in specific situations (tasks, models of the kind of scene, recognized objects, etc.). The eye movements are not a direct output of the algorithms, but they can be computed from the saliency map by using winner-take-all [1] or more dynamical algorithms [2].

In this paper we present a novel attention algorithm and we focus on a fair comparison with other state of the art attention models. The algorithm proposed which we will call "RARE2012" is purely bottom-up. This is an important point for model evaluation as top-down information can drastically increase a model performance. Indeed, several models use additional post-processing which provide top-down information like centred Gaussians which leads to an artificial increase of their results. Moreover, several saliency models have a lot of parameters, which make fair comparison very difficult. Some research, like Borji and Itti [3] or Judd et al. [4], attempts to provide a benchmark between bottom-up models using several similarity measures and sometimes several datasets of images. We based our validation on Borji and Itti approach and codes [3]. A complementary statistical evaluation has also been added. The codes of the model proposed in this paper are freely available online [5].

The paper is organized as follows. Section 2 contains an overview of recent saliency models and more specifically of methods used in our comparative study. In Section 3, the architecture of our method is described in detail. The results are presented in Section 4: after a qualitative evaluation on psychophysical observations and three databases, two metrics are used to quantify the prediction of the proposed method. Section 5 details an additional two-metric based statistical analysis of the results showing the overall effectiveness of RARE2012. Finally, Section 6 provides a discussion and conclusion.

2. Related work

It is very hard to find an optimal taxonomy, which classifies all the saliency approaches. The literature is very active concerning still images saliency models. While some years ago only some labs in the world were working on the topic, nowadays a hundred different models have been published. Those models have various implementations and technical approaches despite that they all derive from the same idea of information innovation in a given context.

Some attempts of taxonomies proposed an opposition between "biologically driven" and "mathematically based" methods. Unfortunately, the biological plausibility of the methods is a difficult point to judge. Another criterion is the computational time or the algorithmic complexity, but it is very difficult to make this comparison as all the existing models do not provide cues about their complexity. Moreover, the implementations can be found in several programming languages. Finally a classification of models based on centre-surround contrast compared to information theory methods do not include different approaches as spectral residual for example. Although several taxonomies can coexist, we propose an original context-based taxonomy. In this framework, there are three classes of models with different contexts which are mostly local, global and normality.

In this section, we define the proposed saliency models categories and provide a brief overview of the recent saliency models that are used for the evaluation in this study. We focus on the models used for our evaluation and do not intend to provide an overview of all existing saliency models. For this purpose, we selected most recently published models which are also available online and classify them using the proposed taxonomy. We also focused on models which use eye-tracking as gold standard and not the models which use manual segmentation as evaluation. Some models obviously use both local and global information. In this case, the classification is made on the primary considered context.

2.1. Local context: salient objects are contrasted compared to their surroundings

The first approach, called local context, is about pixels surroundings: here a pixel or patch is compared with its surroundings at one or several scales like in [6]. Five models from this context are proposed for the study and described in the following subsections.

2.1.1. AIM: Attention Based on Information Maximization (2005)

AIM was created by Bruce and Tsotsos in 2005 [7]. The principle of this bottom-up attention model aims at maximizing information sampled from a scene. Shannon's self-information measure is computed by using patches from the image and their surrounding patches projected on a new basis obtained by performing an ICA (Independent Component Analysis) on a large sample of 7×7 RGB patches drawn from natural images. Overall, this approach quantifies how unexpected the content in a local patch is based on its surrounding.

2.1.2. STB: Saliency ToolBox (2006)

This toolbox [8,9] is a partial reimplementation of the Neuromorphic Vision Toolkit (iNVT) from Laurent Itti [1]. His model is composed of three steps: feature extraction, centre-surround inhibition and feature maps fusion. First, three types of static visual features are selected (colours, intensity and orientations) at several scales. The second step is the centre-surround inhibition which will provide high response in case of high contrast, and low response in case of low contrast. The third step consists in an across-scale combination, followed by normalization to form "conspicuity" maps which are single multi-scale contrast maps for each feature. Finally, a linear combination is made to achieve inter-features fusion.

2.1.3. GBVS: Graph-Based Visual Saliency (2006)

Harel et al. introduced the Graph-Based Visual Saliency (GBVS) model [10]. In this model, they first extracted similar feature maps to Itti's maps (see previous subsection) leading to three multi-scale feature maps (intensity,

colour and orientation). Then, a fully connected graph over all grid locations of each features map is built and a weight is assigned between each nodes. This weight depends on the spatial distance and features of nodes. Finally, each graph is treated as Markov chains to build an activation map and all activation maps are merged into the final saliency map. Here only locally contrasted features are integrated all over the image, thus the model is mainly based on local context.

2.1.4. YINLI: Visual Saliency Based on Lossy Coding (2009)

In 2009, Yin Li proposed a new saliency model inspired by biological vision [11,12]. In this model, the approach is strictly local and based on conditional entropy, which is computed by the lossy coding length of multivariate Gaussian data. The final saliency map is generated by accumulating the coding length. Local information has a priority on the global information integration.

2.1.5. SEO: Saliency Detection by Self-resemblance (2009)

The bottom-up Saliency Detection by Self-resemblance (SDSR) model implemented by Seo and Milanfar consists of two parts [13,14]. First, they propose to use local regression kernels as features (matrix of local descriptors). The underlying hypothesis is that eye fixations are driven by local feature contrast. In a second step, they want to quantify the likeness of each pixel to its surroundings and use a non-parametric kernel density estimation for such features, which results in a saliency map consisting of local "self-resemblance" measure. Even if patches of the image are compared on a wider space than only surround, this is not computed on the entire image.

2.2. Global context: salient objects are different from all the others in the image

The second approach considers the entire image as a context and compares pixels or patches of pixels with any other pixels or patches from any location in the image. There are a lot of recent work in this category like [15,16] or [17], but in the following subsection we will further describe only the models used for our evaluation.

2.2.1. TORRALBA: Saliency Detection by using Local Features (2006)

Torralba stated that saliency is defined in terms of the probability of finding a set of local features within the image as derived from the Bayesian framework [18]. Local image features are salient when they are statistically distinguishable from the background on the rest of the image, i.e. the whole image is considered. In addition to the purely bottom-up approach, two parallel pathways are included in the model: one pathway computes local features (saliency) and the other computes global (scene-centred) features. The contextual guidance model of attention combines bottom-up saliency, scene context and top-down mechanisms at an early stage of visual processing, and predicts the image regions likely to be fixated by human observers performing natural search tasks in real world scenes. The model considered here for validation is the purely bottom-up model without the task scene priors.

2.2.2. AWS: Adaptive Whitening Saliency (2009)

This model of bottom-up saliency is based on the variability in local energy as a measure of salience [19,20]. To do this, first, RGB image is transformed into Lab colour space. In the next step, the luminance is transformed into a multi-oriented multi-resolution representation by using Gabor filters. Each representation is then decorrelated by using a principal component analysis (PCA) and a statistical distance is computed to the centre of the distribution. The final saliency map is obtained by summing the extracted maps. The decorrelation is a global operation which considers the whole image.

2.2.3. CASD: Context Aware Saliency Detection (2010)

In 2010, Goferman et al. have introduced context-aware saliency detection based on four principles [21]. First, local low-level considerations, including factors such as contrast and colour are used. Second, global considerations, which suppress frequently occurring features, while maintaining features that deviate from the norm are taken into account. Higher level information as visual organization rules, which state that visual forms may possess one or several centres of gravity about which the form is organized are then used. Finally, human faces detection is also integrated into the model which brings partly top-down information.

2.3. Normal context: salient objects imply differences to what the normal image would be

Finally, the third saliency category takes into account a context which is based on a model of what the normality should be like in [22]. In the following subsections, we briefly explain the five normality-based models which are used in our evaluation.

2.3.1. HZ: Spectral residual approach (2007)

The authors, Hou and Zhang, proposed a model that is independent of any feature in 2007 [23]. In this method, the first step is to compute the image Fourier spectrum (the amplitude and phase maps). Then, they computed the log-spectrum of the amplitude map. They also computed a filtering amplitude map by multiplying the log-spectrum map with a local average filter. The spectral residual map can be obtained by subtracting these last two maps. The saliency map is achieved through Fourier transform inversion. It should be noted that the phase spectrum is preserved during the process. The idea is that if the image log-spectrum is far from the $1/f$ of natural images (image filtered spectrum), there is something abnormal which deserves attention.

2.3.2. SUN: Saliency Using Natural Image Statistics (2008)

Another model is SUN (for Saliency Using Natural statistics) from Butko et al. (2008) that proposes a Bayesian framework to compute a saliency map [24,25]. In their paper, two methods are implemented. First, the features are calculated as responses of biologically plausible linear filters, such as DoG (Differences of Gaussians) filters. Second, the features are calculated as the responses to filters learned from natural images using independent

component analysis (ICA). SUN with ICA (Method 2) outperforms SUN with DoG filters (Method 1) but both methods predict well people fixations during free viewing. The self-information measure is not applied to the current image statistics but on statistics from a database of natural images (among which the current image is not present). Those images act like typical "normal" images and difference from the statistics of those images might attract attention.

2.3.3. FTSRD: Frequency-Tuned Salient Region Detection (2009)

In 2009, Achanta proposed a very simple model [26] based on local colour and luminance feature contrast. First, the input RGB image is transformed to Lab colour space. Second, the Lab image is blurred with a Gaussian kernel to eliminate noise and texture details from the original Lab image. Finally, the saliency map is computed by using an euclidean distance between the Gaussian-filtered and the original image. The Gaussian-filtered image eliminates small objects and provide an idea about how the image appears to the eyes at a first glance. Objects which are very different from this "normal" image will attract attention.

2.3.4. JIANLI: Frequency and Spatial Saliency (2011)

Jian Li proposes a saliency detection model based on the combination of the global information from the frequency domain analysis and local information from the spatial domain analysis [27]. By using the frequency domain, the spectral residual approach is applied. In the spatial domain analysis, Jian Li enhances those regions that are more informative by using a centre-surround mechanism. The final step is a merge of these two channels to produce the saliency map. This method is a hybrid combination between a local method and a normal context method, and could be located in both sections.

2.3.5. QDCT: Saliency Detection Using Quaternion DCT (2012)

This model applies a spectral saliency method to predict human gaze. More precisely, the authors integrate and evaluate quaternion DCT-based spectral saliency map [28]. They utilize weighted quaternion colour space components and multiple resolutions. Furthermore, they propose the use of the eigen-axes and eigen-angles for spectral saliency models that are based on the quaternion Fourier transform. As HZ, QDCT uses a model of what image should globally be.

3. RARE2012: our proposed saliency model

In this section, the architecture of our method (Fig. 1) is described in detail. There are three main steps. First, we extract low-level colour and medium-level orientation features. Afterwards, a multi-scale rarity mechanism is applied. Finally, we fuse rarity maps into a single final saliency map. A comparison is then made with the RARE algorithms family. In the proposed taxonomy of Section 3, RARE2012 is a part of the second category as it considers information at several scales but globally on the whole image.

3.1. Feature extraction

The first stage of RARE2012 assumes that features can be extracted and processed in parallel or sequentially depending on their complexity.

Contrary to RGB color space, some alternative colour spaces (like Lab, YCbCr, etc.) better uncorrelate colour information. Moreover, the nonlinear relations between their component are intended to mimic the nonlinear response of the eye. To obtain a maximum colour features decorrelation, we transform the RGB colour space into three linearly uncorrelated maps by using Principal Component Analysis (PCA) decomposition. Similar to the other spaces, the first map contains mainly information about the luminance while the two others contain information about the chrominance.

At this stage, the algorithm split in two pathways. The first one, mainly deals with colours (low-level features) while the second one with textures (medium-level features). While the first pathway directly uses the PCA-based colour transformation and computes its rarity, the second pathway extracts orientation features maps by using a set of Gabor filters. These filters were chosen because they are similar to simple cells of the visual cortex (V1) in the brain [29]. As you can see in Eq. (1), a Gabor filter (assumed to be centred at zero) is the product of a sinusoid and a Gaussian where ϕ is the phase offset, λ represents the wavelength of the sinusoidal factor, θ is the orientation (the angle of the normal to the sinusoid), γ is the spatial aspect ratio and σ is the sigma of the Gaussian envelope. In the implementation of these filters, 8 orientations (0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135° and 157.5°), are used at 3 different scales. As any convolution filters, Gabor filters induce side effects, so for each of the 24 resulting maps, a border attenuation is applied.

$$g(x,y;\lambda,\theta,\phi,\gamma)=\exp\left(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda}+\phi\right) \tag{1}$$

where

$$x'=(x\cos\theta+y\sin\theta) \quad \text{and} \quad y'=(-x\sin\theta+y\cos\theta).$$

The decomposition at several scales is recombined in a single map for each orientation. To combine data information for each channel, a selection algorithm is applied to the 8 orientation maps. The first step is to compute for each map an efficiency coefficient, EC_i, which is higher if the map has important peaks compared to its mean (see Eq. (2)). These coefficients let us sort the different maps (map_i) based on each map efficiency coefficient EC_i. Each map is than multiplied by a fixed weight defined as i/N where N is the number of maps to mix (here $N=8$) and i the rank of the sorted maps as shown in the first line of Eq. (3).

$$EC_i=(max_i-mean_i)^2 \tag{2}$$

$$\forall i\in[1\ \ N],\quad \begin{cases} IF\dfrac{EC_i}{EC_N}\geq T & M_i=\dfrac{i}{N}\times map_i \\ IF\dfrac{EC_i}{EC_N}<T & M_i=0 \end{cases} \tag{3}$$

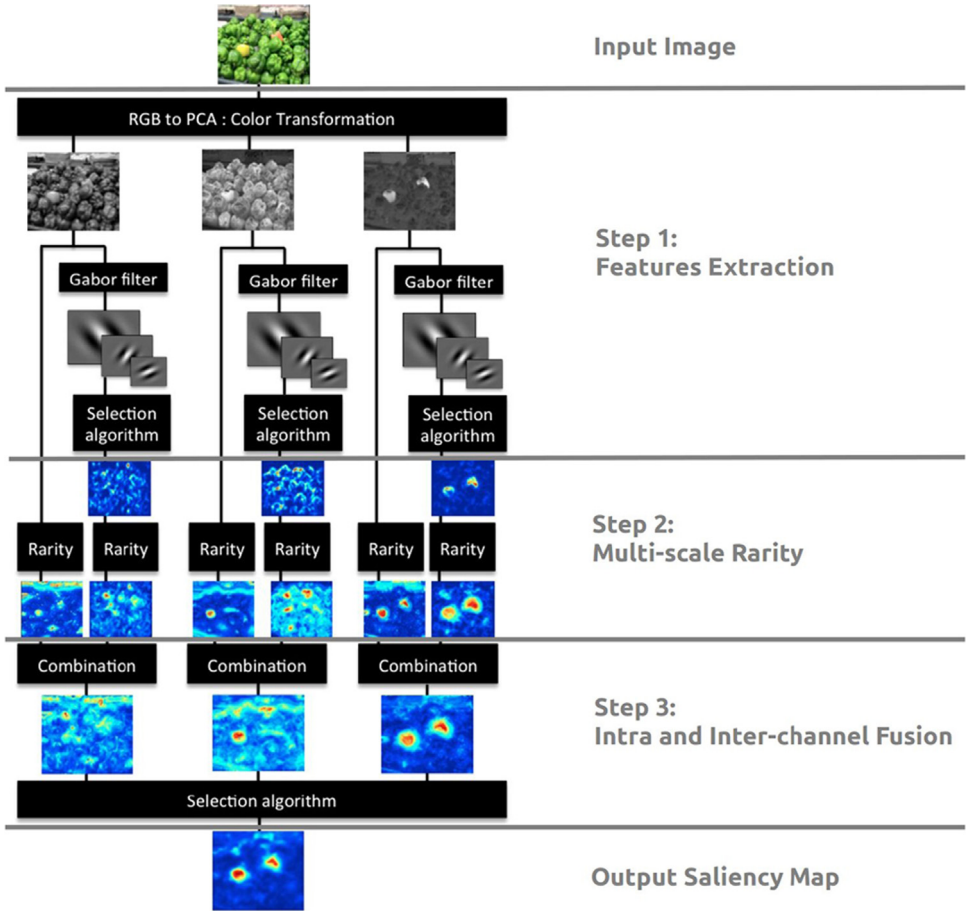

Fig. 1. Diagram of our proposed model. First, from the input image, colour and orientation features are extracted in parallel or sequentially. Then, for each feature, a multi-scale rarity mechanism is applied. Finally, two fusions (intra- and inter-channel) are made from the rarity maps to provide the final saliency map.

where map_N is the most efficient map and map_1 is the less efficient one. Finally the less efficient maps are fully eliminated if they are under an empirical threshold of $T=0.3$ as shown in the second line of Eq. (3).

The fusion is then the sum of all the weighted maps M_i:

$$M = \sum_{i=1}^{N} M_i \tag{4}$$

To conclude this first stage of the algorithm (step 1 from Fig. 1), we obtain six feature maps: three low-level (which are the colours from the first path) and three medium-level (the orientation and texture information coming from the Gabor filters).

3.2. Multi-scale rarity mechanism

The rarity mechanism is the key of RARE2012. Indeed, a feature is not necessary salient alone, but only in a specific context. The mechanism of multi-scale rarity allows to detect both locally contrasted and globally rare regions in the image. First, a Gaussian Pyramid decomposition provides six feature maps at four different scales. A second step consists, for each feature, to compute the cross-scale occurrence probability of each pixel. It is obtained by the normalization of the sum of the occurrence probabilities of the pixel at all scales as shown in Eq. (5) where the n_i is the occurrence value of the current pixel j computed using a histogram within the i th scale or resolution level. Then, the self-information is used to represent the attention score for the pixel. This mechanism provides higher scores for contrasted and rare regions. This idea is illustrated for a single scale in Fig. 2: the dark object from the initial feature (left image) is considered as rare and displayed on the right image with a high amplitude. The rarity range can be set between 0 (all the pixels are the same) and 1 (one pixel different form all the others).

$$Attention(I_j) = -\log\left(\frac{1}{S*|I_j|}\sum_{i=1}^{S} n_i\right) \tag{5}$$

The output of the multi-scale rarity step (step 2 in Fig. 1) consists in a set of six maps called rarity maps.

3.3. Fusion

After the multi-scale rarity mechanism, the 6 rarity maps are fused together into a single saliency map (step 3

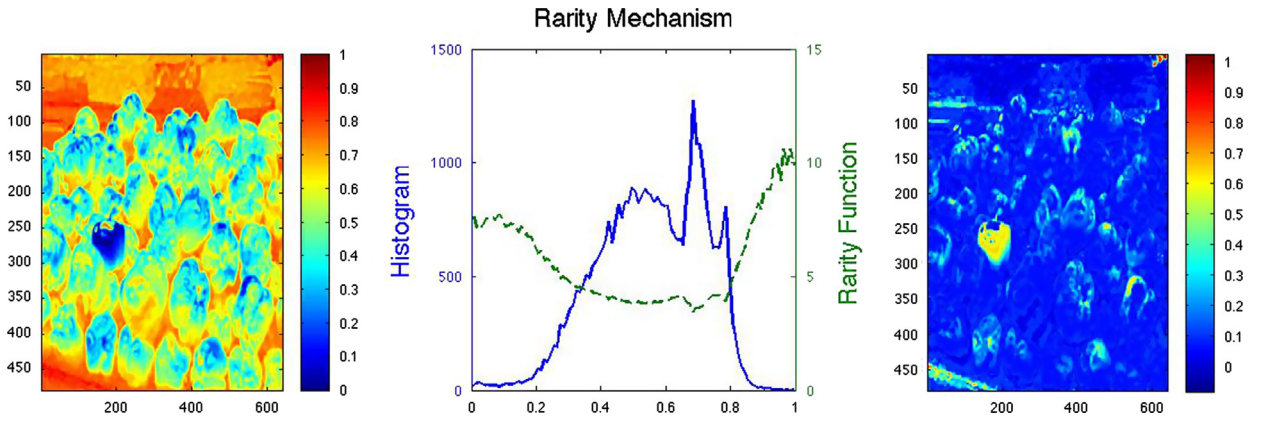

Fig. 2. Illustration of the rarity mechanism on a single scale. Rarity function (green curve in the middle graph) is computed from a histogram (blue curve) of a feature map (left image) to a given scale. This process is repeated at several scales. Output is a reconstruction of the map where high values are given for the most "rare" areas (right image).

in Fig. 1). This fusion is achieved in two main steps: an intra-channel fusion followed by an inter-channel one.

First, an *intra-channel* fusion (called combination in Fig. 1, step 3) is computed between colour and orientation rarity maps by using the fusion method provided by Itti et al. [30]. The idea is to provide a higher weight to the maps which have important peaks compared to their mean (Eq. (6))

$$S = \sum_{i=1}^{N} EC_i * map_i \tag{6}$$

where $N=2$ for each channel and EC_i is the efficiency coefficient computed as in Eq. (2). At the end of this first process, the model provides 3 channel saliency maps, one per colour channel.

In a second step, the final *inter-channel* fusion between those three maps (called selection algorithm in Fig. 1, step 3) is achieved to obtain the final saliency map. This final fusion uses exactly the same method as the one explained in Section 3.1 which uses Eqs. (2)–(4). The parameter $T=0.3$ is the same, but $N=3$ as there are 3 maps to fuse.

The output saliency map is now unique and of the same size that the original image provided to the algorithm.

3.4. The RARE family

RARE2012 is the latest development around the idea of multi-scale rarity-based saliency detection which begun with RARE2007 [31] and followed by RARE2011 [32]. Each one of these steps brought major changes in the algorithm pipeline and performance improvements.

RARE2007 extracts only colour information maps which are quantized into 11 classes of pixels each. A rarity mechanism is then applied. Finally, a *NS* (normalized and sum) fusion is made between rarity maps to output saliency map. Compared to RARE2007, RARE2011 mainly introduced the Gabor filtering to also extract the image orientation information. An optimal Otsu's quantization is then performed to separate each map into 8 classes of pixels. A rarity mechanism and a *NS* (normalized and sum) fusion are also applied. RARE2012 which is presented in this paper brought changes compared to RARE2011 (a) in the algorithm pipeline introducing serial and parallel features extraction which computes rarity on both color and texture instead of only texture, (b) in the rarity algorithm which is a new version which does not use any quantification, (c) in the colour space which is PCA-based and (d) in the fusion algorithm (Eq. (3)) which is modified to be more selective.

RARE2012 is also faster and provides better results than RARE2011 and RARE2007 as it will be shown in Section 4.3.

4. Saliency model evaluation

In this section, we compare our method with the 13 saliency models presented in related work on three datasets. After the dataset presentation, qualitative and quantitative results are detailed and explained.

4.1. Evaluation databases

Three recent eye-tracking datasets available online are used (Fig. 3). The first one, that we will call Toronto dataset, was made by Bruce and Tsotsos [7]; it includes 120 images with 20 viewers per image. The second, Kootstra dataset [33], provides 100 images with 31 viewers per image. Finally, we use Jian Li dataset [27], which provides eye-tracking data on 235 images with 19 viewers per image.

The Toronto dataset includes images of both outdoor and indoor scenes. A particularity of this database is that a large subset of images does not contain specific regions of interest like semantic objects or faces. Each image has been freely viewed by participants during 4 s.

In the Kootstra dataset, there are mainly images of outdoor scenes. This dataset is split into five different categories like images with animals, street scenes, buildings, natural symmetrical shapes, mainly flowers and plants and natural scenes. The viewing time was of 5 s for each image.

Finally, the last dataset used here has been published by Jian Li. One interesting property of this dataset is that it

Fig. 3. Sample representative images of the three datasets (first row: Toronto dataset, second row: Kootstra dataset, third row: Jian Li dataset) are displayed.

is organized in six different groups of images which are different from the ones of Kootstra (large objects, intermediate objects, small objects, cluttered backgrounds (CB), repeating distractors (RD) and large and small salient regions (LSSR)). The images were presented for a short time duration, sufficient for actually seeing the image.

4.2. Qualitative evaluation

Some qualitative results on synthetic patterns and selected images from the three datasets are presented here. The goal of this section is to visually show the results of RARE2012 on simple and more complex images.

4.2.1. Synthetic patterns

Psychophysical observations are synthetic stimuli showing a particular object (the target) among other objects (the distractors). All stimuli presented here have been widely used by the community [3,20]. Nevertheless, RARE2012 does not intend to fully explain human behaviour and the dataset shown here is not large enough and it has no eye-tracking data for an efficient comparison. The goal is to see if the global rarity and local contrast idea behind RARE2012 make sense compared with human behaviour which will fixate the pop-out target. There are two parts in this section. First, eight synthetic patterns are selected for the specificity of their targets which are linked to RARE2012 features: colour and orientation. In the second part, the selected targets are more complex. They are not necessarily directly linked to the features extracted by RARE2012.

In Fig. 4, RARE2012 suitably reproduces pop-out phenomena related to colour and orientation targets. Indeed, the saliency is high (in red) on the targets. These results are expected due to the nature of the targets. For the colour/luminance differences, they are well detected even if the colour difference is not very important. This is due to the nature of the proposed model which is based on global rarity. Even if an object has a low contrast, and there are no other high contrast objects, it will be well highlighted. Concerning the combination of colour and orientation

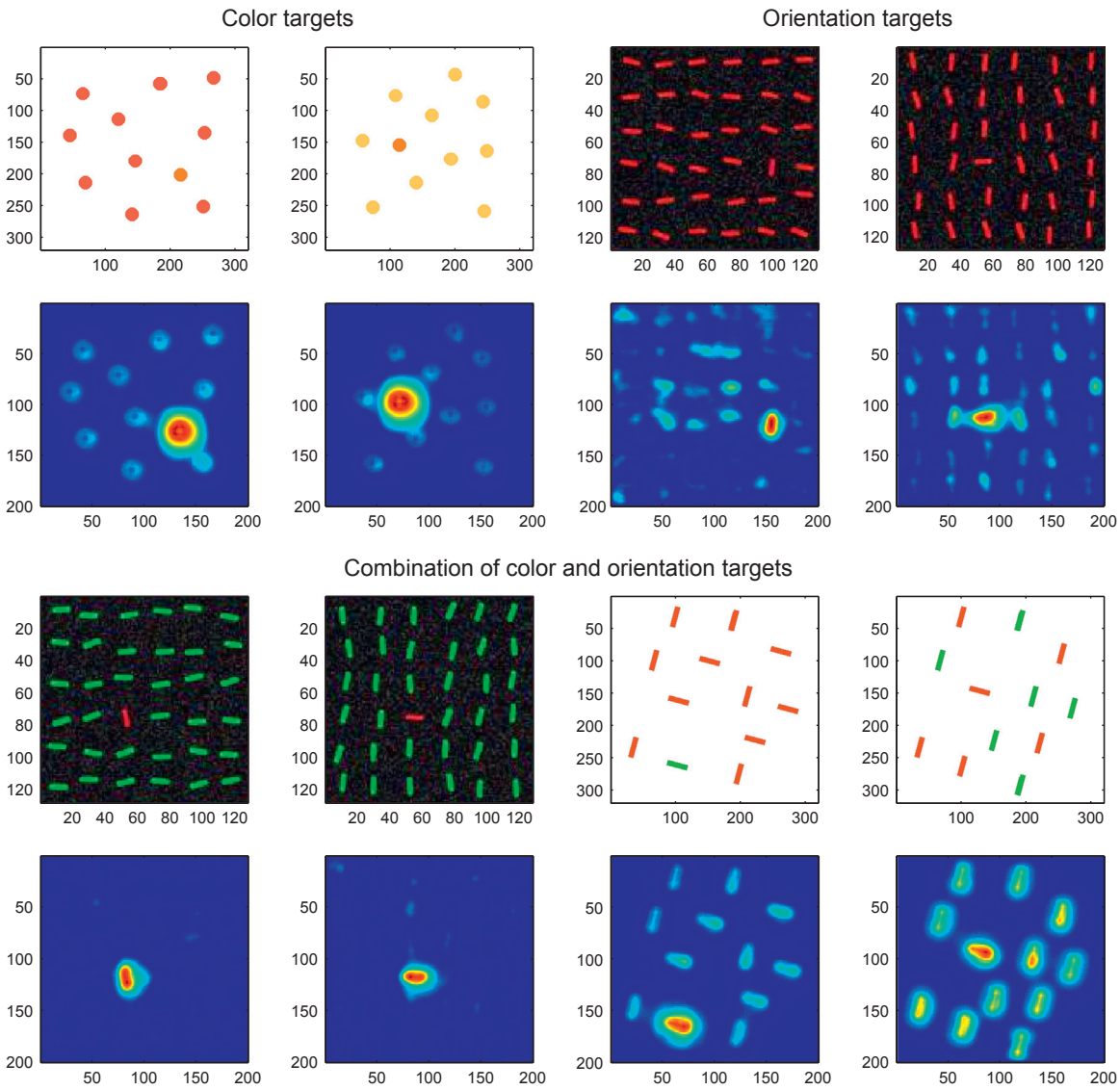

Fig. 4. Rows 1–2: Stimuli and RARE2012 saliency maps for colour and orientation targets presented separately. Rows 3–4: Stimuli and related saliency maps for colour and orientation mixed targets. Globally, RARE2012 works as desired.

targets, it is interesting to see the influence of mixed targets or the heterogeneity of distractors. Indeed, the more the distractors, the less selective the saliency map, even if the pop-out target is still detected as the maximum of the saliency map. This is again a consequence of the global rarity part of the algorithm.

In Fig. 5, our model points out all of the selected targets even if the features used here are more complex. The selection of targets includes: (1) luminance, (2) intersection and curvature, (3) density target and (4) visual search examples where all previous targets can be present. Our saliency maps are more noisy than in Fig. 4 but replicate the expected human behaviour.

4.2.2. Visualization result

In addition to synthetic patterns, Fig. 6 displays selected images from the three eye-tracking datasets. The eye-tracking results on these images which are superimposed to the images on rows one, three and five are compared to RARE2012 saliency maps on the same images on rows two, four and six. The peak value of the saliency maps fits to the maximum of the eye-tracking maps. This shows a subjective good prediction of the saliency maps of RARE2012.

Fig. 8 shows six images from different databases (first row). The second row contains the corresponding eye-tracking heatmaps. The following rows show the saliency maps of all the compared models beginning with RARE2012.

Fig. 7 shows six images from different databases (first row). The second row contains the corresponding eye-tracking heatmaps. The following 3 rows show the saliency maps of RARE2012, AWS and GBVS models. RARE2012 works well in the 3 first images and seems more selective than AWS and GBVS. The last three images

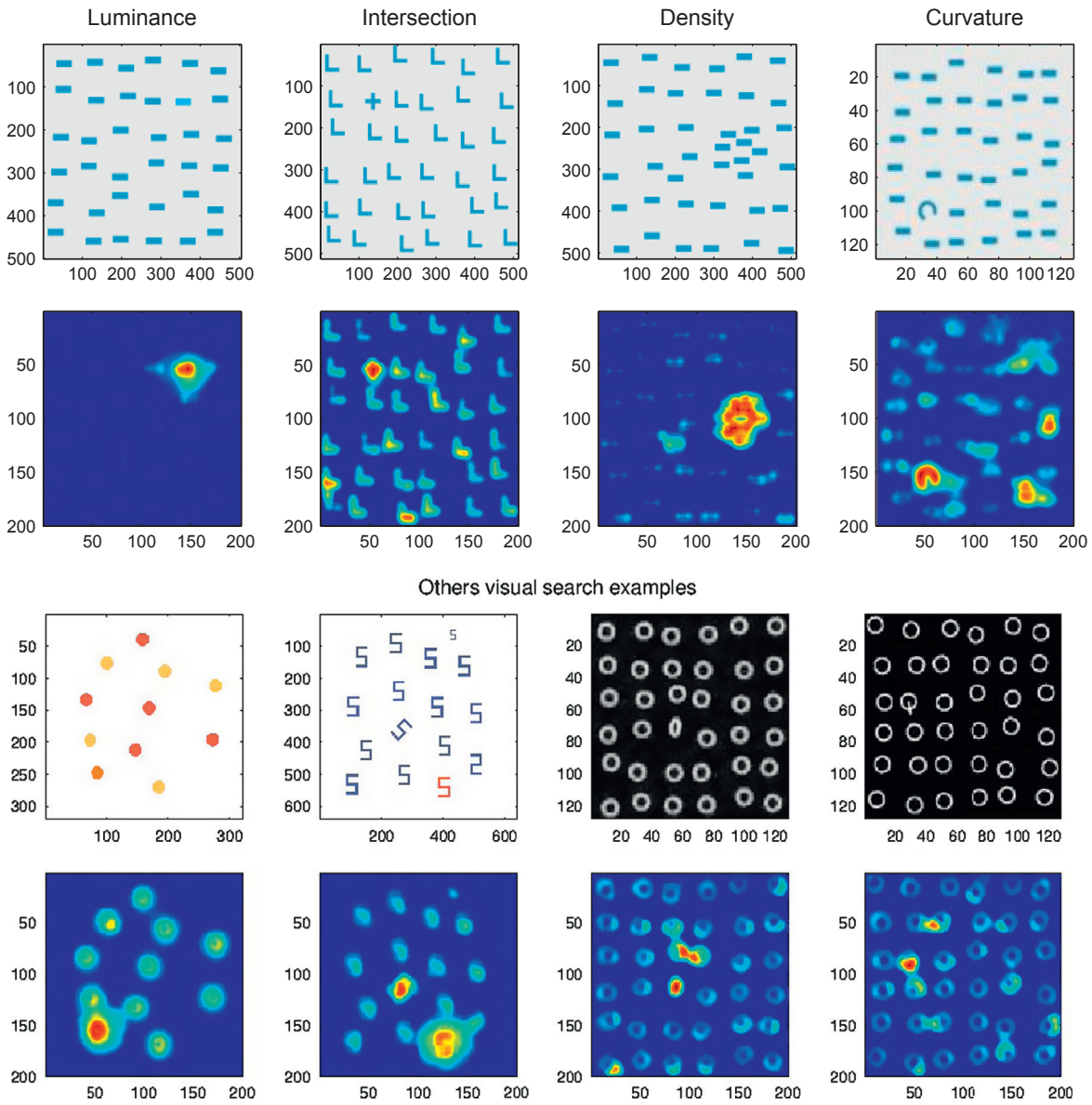

Fig. 5. Rows 1–2: Stimuli and RARE2012 saliency maps for targets with different specificities. Rows 3–4: Stimuli and related saliency maps for synthetic patterns come from visual search task. Overall, RARE2012 works slightly worse than in the first part. However, it also points out all targets.

are examples where RARE2012 fails to estimate people gaze position. AWS and GBVS do not seem to work especially better. This is mainly due to the fact that here the bottom-up cues do not match with top-down information (mainly faces). This example also shows that purely bottom-up models are nowadays good enough and they mainly fail when top-down information is present. Thus, future improvements will certainly come from more and more top-down information integration.

4.3. *Quantitative evaluation*

4.3.1. *Metric definition*

There are several similarity measures proposed in the literature to compare saliency and fixation maps. These measures include the Area Under Receiver Operating Characteristic (AUROC) [34], the Normalized Scanpath Saliency (NSS) metric [35] or the least square index [36]. Three other metrics namely the correlation-based measures as Correlation Coefficient (CC), dissimilarity measures like KL-divergence (KLD) and the string-edit distance are described in [37].

Among those similarity measures, two metrics have been chosen: the NSS and AUROC for their complementarity [38]. Indeed, contrary to the NSS measure which compares values or amplitudes of the maps, AUROC mainly measures the order and locations of the fixations. Moreover, the use of two complementary metrics ensures that the quantitative conclusions are much more robust from the choice of the metric.

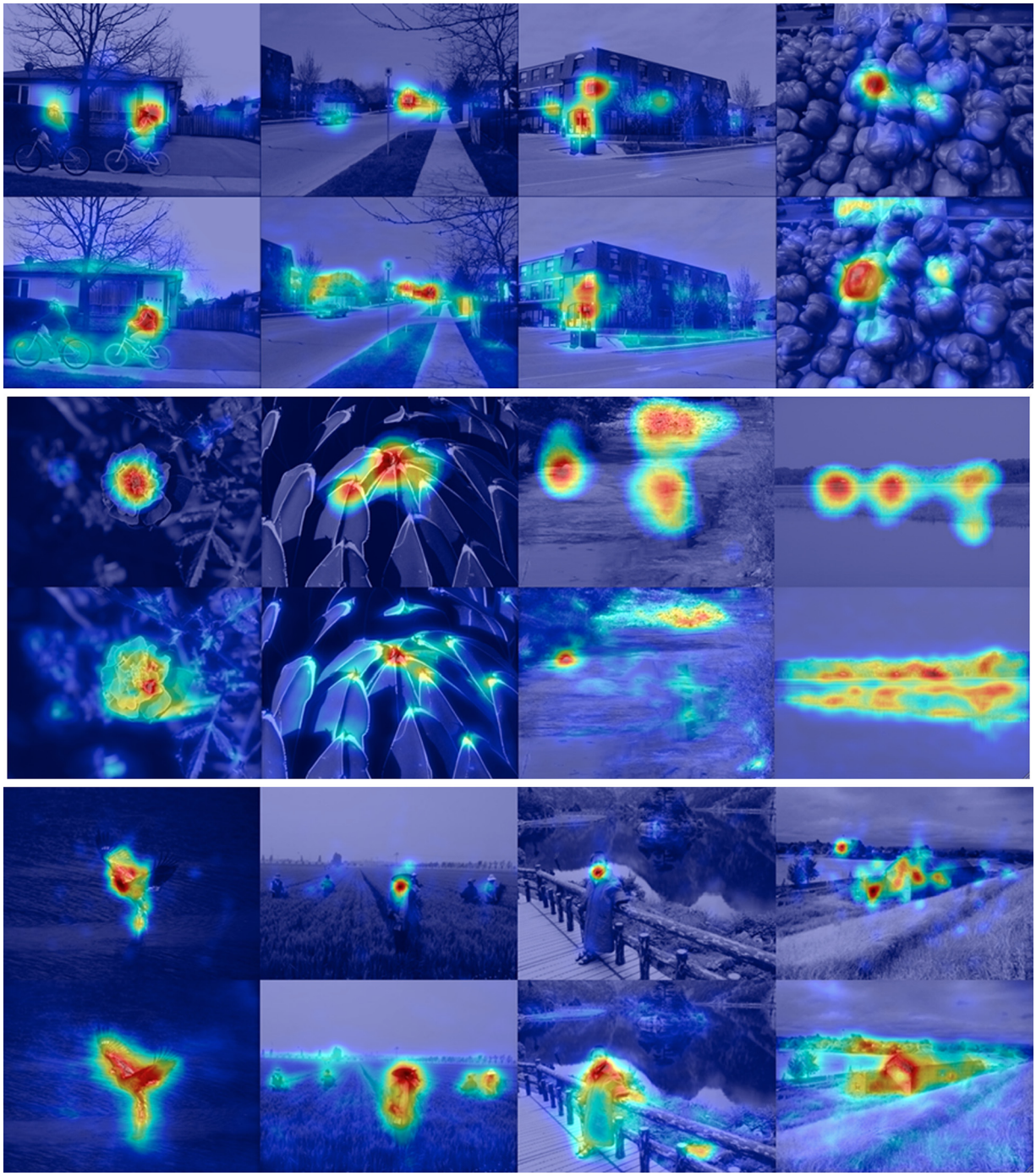

Fig. 6. Qualitative results from the three databases show that RARE2012 can reliably predict where people look in images. Rows 1–2: Eye-tracking and RARE2012 saliency map from Toronto dataset. Rows 3–4: from Kootstra dataset. Rows 5–6: from Li dataset.

The NSS metric represents the average of the response values at human eye positions in a model's saliency map. For this purpose, the saliency maps are normalized to have zero mean and unit standard deviation. The fixation map can then be thresholded to become a binary mask. The second evaluation is made by using the area under the receiver operating characteristics curve (AUROC). The first step is a normalization of the saliency map. Then, this map is thresholded to create binary masks that separate the positive samples called True Positive Rate (TPR) from the negatives called False Positive Rate (FPR). By thresholding the saliency map with several different thresholds and

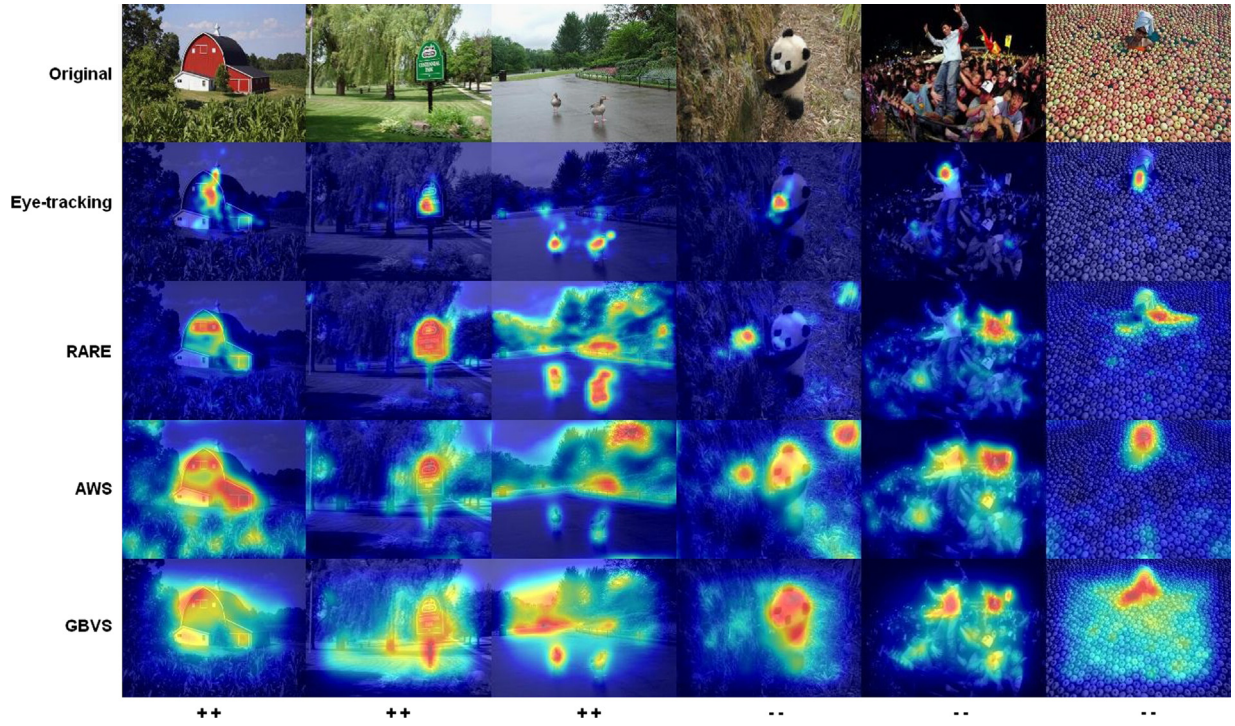

Fig. 7. Qualitative results for six images (first row) of RARE2012 and the two other best models (AWS and GBVS). RARE2012 works well in the first three images and less well in the three last images.

plotting TPR versus FPR, a ROC curve is computed. Once the curve is computed, the area under this curve can be quantified.

Two well-known problems for fair comparisons are the centre-bias and border effect. Centre-bias means that a lot of fixations from natural images databases are located near the image centre because when taking pictures, the amateur photographer often places salient objects in the image centre. The computational saliency models which include a centred Gaussian use the prior knowledge of working on natural images which is not a general assumption, and therefore can be considered as top-down information. Indeed, other categories of images (as advertisements or websites have very different behaviour as demonstrated in [31]). As the three datasets taken into account here provide only natural images, the use of centred weight within the saliency models will artificially improve their results if using the classical AUROC measures. Moreover, Zhang et al. [25] showed that AUROC scores are also corrupted by edge effects. If we remove edges of an image, AUROC scores increase as well. Indeed, human eye fixations are rarely near the edges of test images.

In [39], Borji suggests three possible remedies: (1) add a centred Gaussian to the output of every saliency model to obtain a fair comparison; (2) make the quantitative comparison on a dataset with no centre-bias; (3) design a suitable evaluation metrics. It is not fair to add centred Gaussians to some models while those weights are already included with other parameters in others. Getting datasets with no centre biases is almost impossible as all of them contain natural images. In this paper, the third solution is adopted by using the shuffled AUROC metric proposed by Zhang et al. [25]. In AUROC, saliency values and random points from the image are taken into account to create a binary mask. In the shuffled AUROC metric, saliency values and fixations from another image (instead of random) of the dataset are taken into account. We used the freely available Matlab implementation of Borji [40] for our quantitative evaluation.

Unfortunately, the shuffle AUROC idea cannot be applied to the NSS metric which suffers from the centre-bias issue. To quantify this issue, we introduce baseline model in our comparison called “Gauss”. This model is a centered Gaussian and can be seen like a baseline algorithm to have an idea of the impact of centre-bias. The codes used for NSS computation are also the freely available Matlab implementation of Borji [40].

4.3.2. *Performance evaluation*

Fig. 9 displays the results for the three datasets and two metrics of RARE2012 compared to the other 13 saliency models. For all models including RARE2012, we use the default parameters. The mean results are shown along with their standard deviations and RARE2012 is shown in purple. For the AUROC metric, RARE2012 performs second best for the three datasets after the AWS model and followed by QDCT for the Toronto dataset, SEO for Kootstra dataset and CASD for Jian Li dataset.

For the NSS metric, RARE2012 is the best followed by GBVS and AWS for Toronto dataset, the best followed by GBVS and YINLI for Koostra dataset and the best followed by GBVS and AWS for Jian Li dataset.

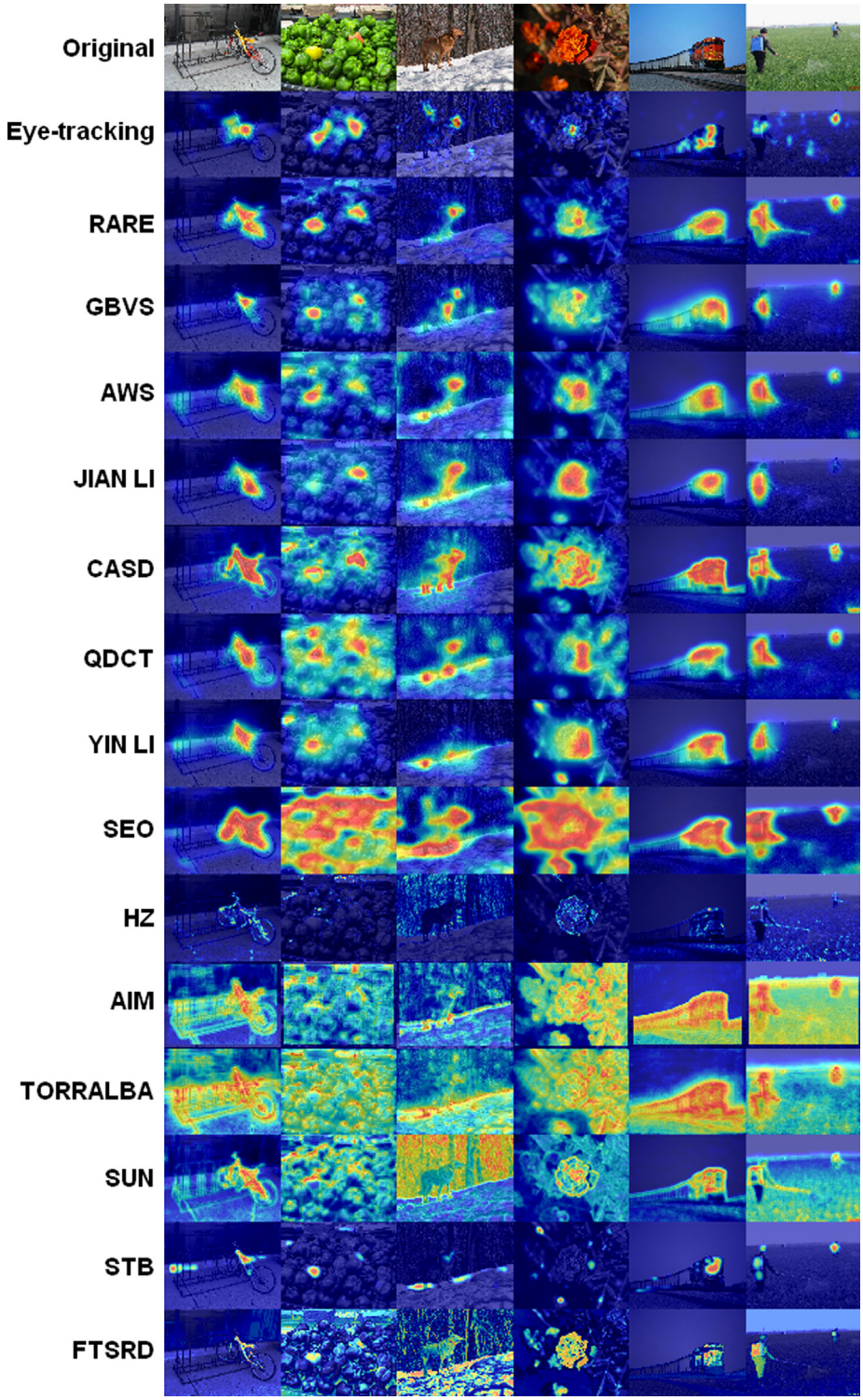

Fig. 8. Qualitative comparison of thirteen models and our RARE algorithm for 6 images (columns) compared with the eye-tracking ground truth (second row).

Concerning the other models from the RARE family [31,32], RARE2012 is faster (Fig. 10). It needs 1.97 s while RARE2007 needs (as a mean) 10.95 s to process an image and RARE2011 needs 29.48 s on the same platform and with Matlab implementations. The speed of the algorithm poorly depends on the initial image size as this one is anyway resized within the algorithm.

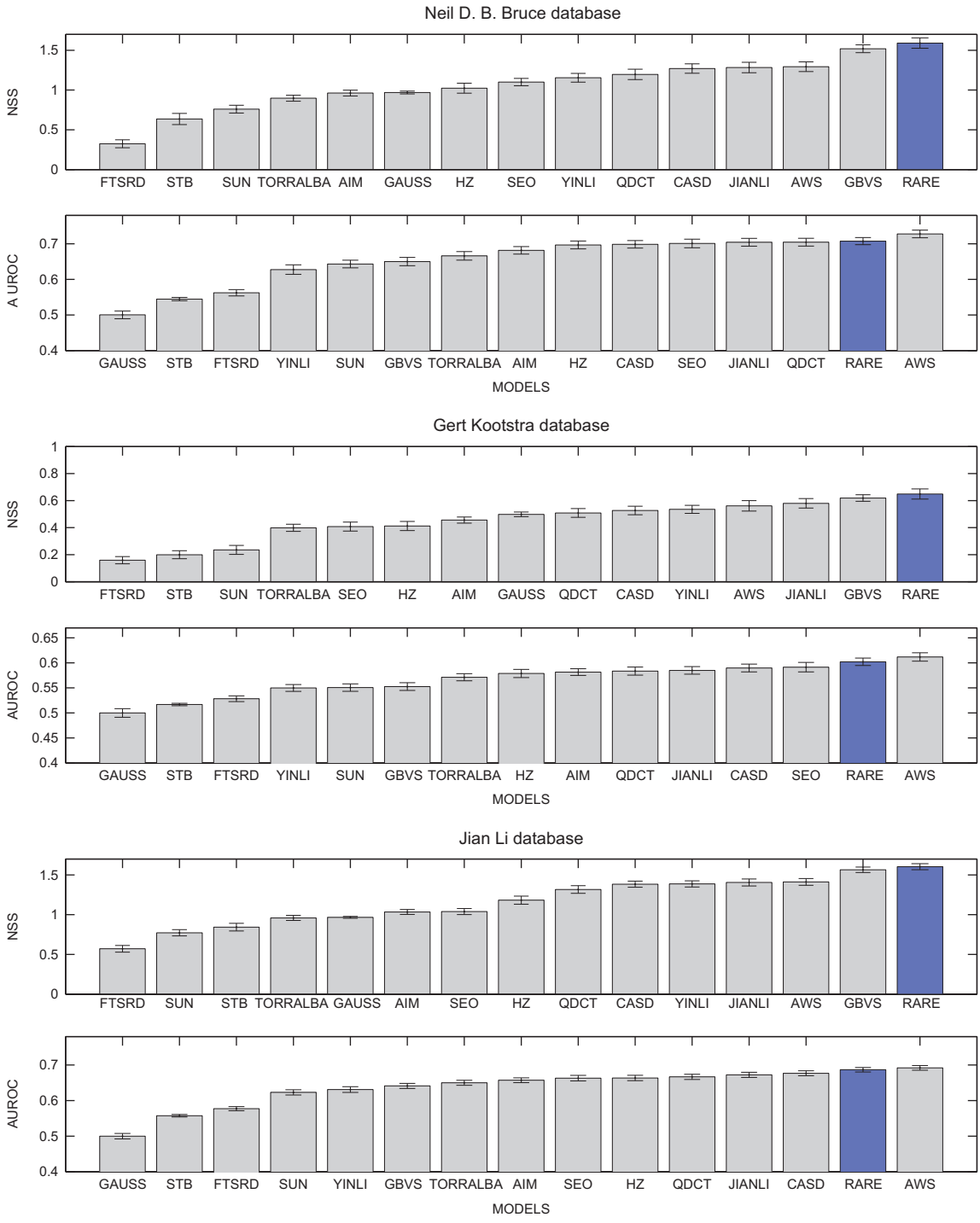

Fig. 9. Comparison of our model with 13 state-of-the-art saliency models and a centred Gaussian. The first two graphs display the NSS and AUROC metrics for the Toronto dataset. Graphs 3–4: for Kootstra dataset. Graphs 5–6: for the Jian Li dataset. RARE2012 outperforms the other models for the NSS metric and is the second best for AUROC metric.

Moreover, regarding the models results, for the three datasets Fig. 11 displays a comparison of the three models. For each metric RARE2012 provides better results than RARE2011 and RARE2007.

5. Additional statistical validation

In this section, the statistical validation is studied. First, the statistical approach is carefully detailed. Its interest compared

to a standard ANOVA test is exposed. The importance of effect sizes is highlighted. Then, the results are exposed and explained.

5.1. Statistical framework

In order to detect whether our model is competitive with respect to other state-of-the-art models, our statistical assessment is only focusing on comparisons including RARE2012. Indeed, the null hypothesis H_0 assumes that our model is the best one. Then, we are looking for evidence of rejecting it meaning that at least one other model is likely to outperform ours.

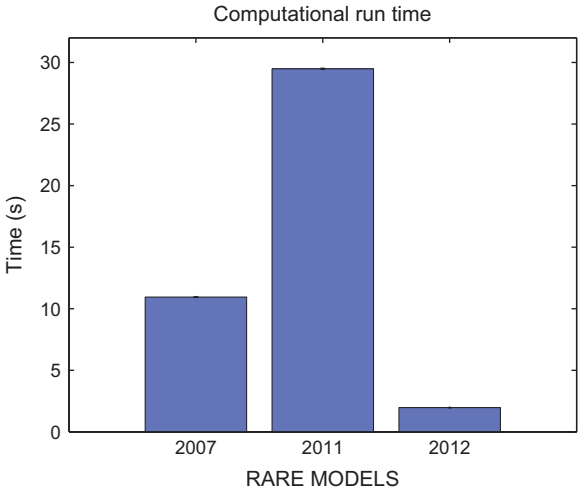

Fig. 10. A new rarity mechanism implementation as well as an optimization of the code led to a much faster algorithm in the case of RARE2012 (mean of 2 s per image).

Although our design follows a repeated measure analysis of variance (ANOVA), i.e. the structure of the collected data and the analyzed factors, for each performance measure [41,42], the ANOVA analysis and its omnibus F test were not performed in our procedure. Contrary to what a lot of researchers usually think, the omnibus ANOVA F test is not a necessary condition to control Family-Wise Error Rate (FWER) whatever the applied *post-hoc* tests [42,43]. If computed, the degrees of freedom are wasted for somehow useless statistical tests. Even worse, with this approach, an overall decrease of power could be observed. More precisely, omnibus F test might show no significativity while some of the underlying t-tests are significant. In this procedure, we did not compute power *a priori* because (1) there was a limited amount of images and if an insufficient number of images was detected, we were unable to increase their number, and (2) given their huge number, the observed power should be high enough.

We thus defined only a limited amount of *a priori* comparisons by applying the prescription of [42,43]. First, we defined all the pairwise comparisons with our model before data collection in a similar way to the Dunnet test, which compares a control group to other alternatives to detect whether they significantly compete this control group. Obviously, training data that were used for optimization purposes do not have to be fed in the model assessment process to avoid data snooping bias. This bias typically overestimates p-values and leads to spurious conclusions. Second, we performed the standard paired t-tests, whose single assumption is data normality, with a standard alpha level of 2.5% for each performance measure (computing a simple Bonferonni adjustment). Given that those comparisons equal the degrees of freedom, we make

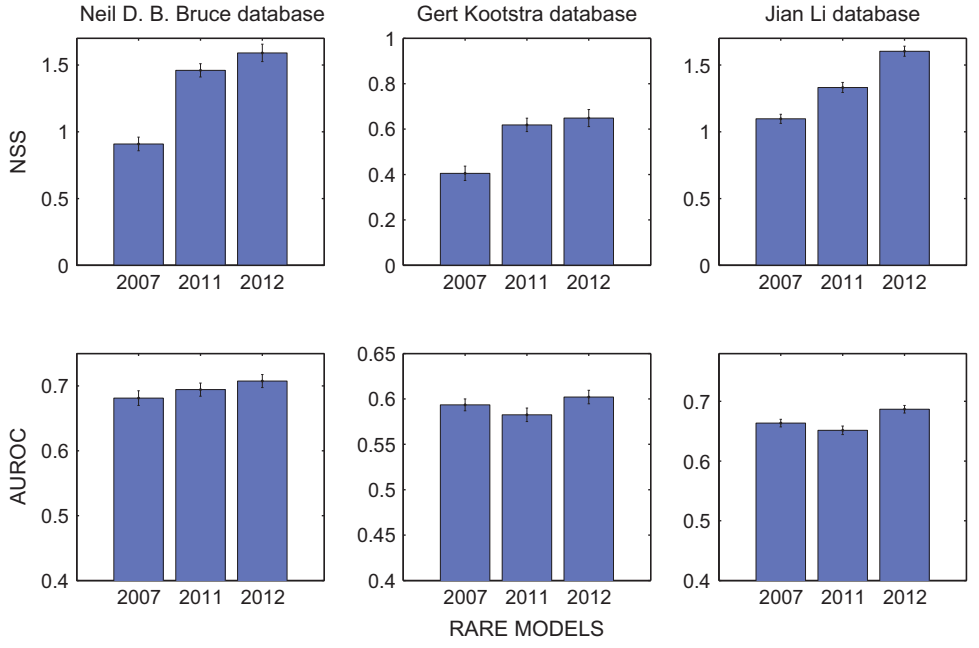

Fig. 11. For NSS metric, a permanent improvement can be observed from RARE 2007 to RARE2012 while for AUROC metric, RARE2012 outperforms RARE2007.

sure to control FWER inflation without any further adjustments, which leads to a much more powerful test. In our case, given that the observed correlation between performance measures is around $\rho = 0.57$, the multivariate Hotelling T^2 test could be useful to study performance model as a whole. Nevertheless, a gain in power is theoretically not always ensured, especially when *post-hoc* analysis needs to be computed to detect the incriminated measure of a significant effect. Furthermore, we intentionally want to decouple both performance measures in the analysis to get specific conclusions on both of them while being able to combine these conclusions [43].

However, significant results are not enough and effect-size is at least as important [42,44]. Nevertheless, although effect size has become a major interest in biomedical/psychology/medicine studies, in computer sciences, this importance has not really been pointed out. Significativity only assesses if there is enough evidence to determine whether there is a likely effect between two or more groups. It does not provide information about the size of this effect. If the difference is significant but trivial in terms of practical differences, the best method is not really outperforming the other ones.

The normalized unbiased Hedge' g^* effect-size measure somehow tackles this problem. Basically, the Hedge' g value is computed as

$$g = \frac{\overline{X}_1 - \overline{X}_2}{s^*} \tag{7}$$

where X_1, X_2 are sample distributions and the pooled standard deviation is defined as $s_* = \sqrt{(n_1-1)s_1^2 + (n_2-1)s_2^2/(n_1+n_2-2)}$ with standard deviations s_1, s_2. The precise correction for paired data of the unbiased Hedge' g^* is not given as it does not modify the interpretation.

As depicted in Fig. 12, this measure allows to evaluate this effect in a standard way. Furthermore, it provides some rules of thumb of how big the effect-size is. For instance, an absolute Hedge' g^* value around 0.1 ($\leq$0.16), 0.2 (0.17–0.32), 0.5 (0.33–0.55) or 0.8 (0.56–1.2) respectively mean a trivial, small, medium or a large effect

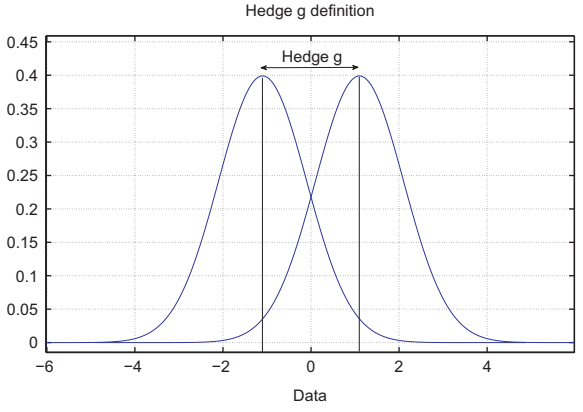

Fig. 12. Given two standard normal distributions, with means m_1 of 0 and m_2 of 2.2, and a standard deviation of 1 (identical for both distributions). Accordingly, in this example, Hedge' $g^* = \frac{(1.1-(-1.1)}{1} = 2.2$ [44].

Table 1
The AUROC distribution is quasi-normal. However, the NSS curve fits at best a log-normal distribution. Therefore, before statistical analysis, a logarithmic transformation was performed.

Statistics	AUROC	NSS
Mean	0.63	0.98
Mediane	0.62	0.93
Mode	0.2	−1.16
Kurtosis	2.73	4.11
Skewness	−0.005	0.71

according to [45,46]. To help interpretation, small, medium and large size-effects can be found in the American girl population height differences between 15 and 16, 14 and 18 and 13 and 18 respectively [45].

Furthermore, a single value effect size is not sufficient and a 95% interval should be studied. This helps to provide information about how the current effect size is a good estimation of the underlying one. Obviously, if more data are used, a more precise interval is provided allowing more reliable conclusions. To compute all these values, Matlab and a famous neuroscience toolbox were used [44].

5.2. Statistical results

Preliminary to the analysis, we checked by visual inspection if the normality assumption was met. Indeed, this was ensured for the AUROC measure (although also close to a log-normal curve) but not for the NSS indicator. Hopefully, the later distribution fits a log-normal curve and we considered a simple logarithmic transformation to compute our statistical analysis [43]. Detailed statistics are depicted in Table 1.

Regarding the p-values, results do not really provide additional information than a mean based-comparison as shown in Tables 2 and 3. Relying on the AUROC measure, there is sufficient evidence only for the AWS model to likely outperform our model. Based on the NSS measure, none of the alternative models show a significant better performance. Given the observed p-values, the power of the statistical test appears to be very high.

What concerns the Hedge' g^*, results show that our model is likely to have the best overall behaviour. The AUROC outperforming AWS model has only a trivial better performance than our model, whose tiny underperformance is observable in the confidence interval depicted in Table 2. However, RARE2012 is likely to only slightly outperform the CASD, JIANLI, QDCT, HZ and SEO models with a trivial to small effect-size. Basically, given the observed confidence interval, one can say that all those competitor models, including the AWS, depict similar performance whose differences could be explained, at least partially, by statistical fluctuations.

On the other hand, the NSS results are more successful as shown in Table 3. The GBVS model is the only one to compete our model. However, given the confidence interval, one can conclude that both models provide quasi-identical performance. The following other close performers are the AWS and JIANLI models. But, the corresponding effect-sizes at least indicate a small effect size. Except the

Table 2
Our model is slightly under-performing the AWS model with a trivial effect-size using the AUROC performance measure. Other competitors are close to our model with a trivial to small effect-size.

AUROC				
Models	***p*-values**	**Hedge *g***		
		Lower bound	**Mean**	**Upper bound**
AIM	1.00	0.19	0.25	0.32
AWS	**0.00**	**−0.14**	**−0.09**	**−0.05**
CASD	1.00	**0.03**	**0.09**	**0.15**
FTSRD	1.00	1.01	1.14	1.26
GAUSS	1.00	1.43	1.59	1.75
GBVS	1.00	0.38	0.45	0.52
HZ	1.00	**0.11**	**0.18**	**0.25**
JIANLI	1.00	**0.05**	**0.11**	**0.17**
QDCT	1.00	**0.08**	**0.14**	**0.20**
SEO	1.00	**0.08**	**0.14**	**0.20**
STB	1.00	1.42	1.56	1.71
SUN	1.00	0.48	0.57	0.66
TORRALBA	1.00	0.26	0.34	0.41
YINLI	1.00	0.44	0.53	0.62

Table 3
Based on the NSS performance measure, our model equals the GBVS model. Amongst the AUROC competitors, the closest one is the AWS model with a small effect-size under-performance.

NSS				
Models	***p*-values**	**Hedge *g***		
		Lower bound	**Mean**	**Upper bound**
AIM	1.00	0.74	0.82	0.90
AWS	1.00	**0.24**	**0.29**	**0.34**
CASD	1.00	0.27	0.33	0.38
FTSRD	1.00	1.40	1.54	1.69
GAUSS	1.00	0.84	0.94	1.05
GBVS	0.97	**0.00**	**0.05**	**0.10**
HZ	1.00	0.54	0.62	0.70
JIANLI	1.00	**0.23**	**0.30**	**0.36**
QDCT	1.00	0.36	0.42	0.49
SEO	1.00	0.65	0.72	0.80
STB	1.00	1.04	1.15	1.26
SUN	1.00	1.04	1.16	1.27
TORRALBA	1.00	0.83	0.93	1.02
YINLI	1.00	0.29	0.36	0.44

AWS model, none of the AUROC-based competitors is performing well enough compared to our model.

In conclusion, it appears that several alternative models are strong competitors for our model based on the AUROC measure. Nevertheless, a global assessment should consider several measures and, considering two complementary ones (AUROC-NSS), RARE2012 appears to be slightly better than the AWS model, which is the best performing alternative model.

6. Conclusion and future work

This paper presents a novel multi-scale rarity-based saliency model for still images called RARE2012. An extensive evaluation and statistical analysis are carried out to compare this model with other important models of the state of the art.

RARE2012 presents major changes compared to the previously published models of the RARE family [31,32]. First, the colour and orientation features are extracted in parallel or sequentially depending of their complexity. Colours are based on a PCA analysis to optimize information decorrelation. A new version of the rarity algorithm has been implemented and the inter- and intra-maps fusion has been changed to provide more reliable weights. These algorithmic innovation significantly improved the results on NSS and AUROC metrics along with computational efficiency.

After a qualitative evaluation of the model on both synthetic and real-life images, where RARE2012 predicts well the human gaze as desired, a quantitative study including three images datasets of a total of 455 images, 13 other saliency models and two comparison metrics (AUROC and NSS) is performed. RARE2012 outperforms all the other models for the NSS metric and is the second best for the AUROC metric. An assessment with these metrics used alone is not fair and complete enough to efficiently compare saliency models. Indeed, each metric will focus on a particular view of the comparison like the amplitude for the NSS and the order of the fixations for the AUROC. In addition, each metric can be influenced or not by top-down information like centered Gaussian that some models integrate and others not.

Finally, to determine the robustness of ranking with respect to statistical fluctuation, an additional analysis is applied. Therein, RARE2012 at least slightly outperforms the other models while considering both metrics and the three databases as a whole. Relying on the effect-size Hedge' g* measure, the AUCROC outperforming AWS model has only a trivial better performance than RARE2012 while our model outperforms all other models at least with a small effect size observed for the GBVS model.

While bottom-up models are now well established and their results convincing, the best way to go further in eye motion modeling is to use top-down information to tune the models to specific classes of images or to add information about semantic objects. A perspective for RARE2012 is to use top-down information to modulate the salient regions with context-driven attention.

Acknowledgements

N. Riche is supported by the “Fonds pour la formation a la Recherche dans Industrie et dans Agriculture” (FRIA). N. Riche and M. Mancas contributed equally to this work. M. Duvinage is a FNRS (Fonds National de la Recherche Scientifique) Research Fellow and the corresponding author for statistical analysis. Thierry Dutoit is member of EURASIP. This work is also funded by the Belgian Walloon region NumediArt project.

This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its author(s).

Part of the study was funded by LinkedTV EU FP7 project.

References

[1] 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 (11) (1998) 1254–1259.

[2] M. Mancas, F. Pirri, M. Pizzoli, From saliency to eye gaze: embodied visual selection for a pan-tilt-based robotic head, Advances in Visual Computing (2011) 135–146.

[3] Ali Borji, Laurent Itti, State-of-the-art in visual attention modeling, IEEE Transactions on Pattern Analysis and Machine Intelligence, in press, http://ilab.usc.edu/publications/doc/Borji_Itti12pami.pdf.

[4] T. Judd, F. Durand, A. Torralba, A benchmark of computational models of saliency to predict human fixations, IEEE Transactions on Pattern Analysis and Machine Intelligence (2012).

[5] Matei Mancas, Nicolas Riche, Computational attention website, 2012.

[6] Y.F. Ma, H.J. Zhang, Contrast-based image attention analysis by using fuzzy growing, in: International Multimedia Conference: Proceedings of the Eleventh ACM International Conference on Multimedia, vol. 2, 2003, pp. 374–381.

[7] N. Bruce, J. Tsotsos, Saliency based on information maximization, in: Advances in Neural Information Processing Systems, vol. 18, 2006, pp. 155–162.

[8] D. Walther, Interactions of Visual Attention and Object Recognition: Computational Modeling, Algorithms, and Psychophysics, PhD Thesis, California Institute of Technology, 2006.

[9] D. Walther, C. Koch, Modeling attention to salient proto-objects, Neural Networks 19 (9) (2006) 1395–1407.

[10] C. Koch J. Harel, P. Perona, Graph-based visual saliency, in: Proceedings of Neural Information Processing Systems (NIPS), 2006.

[11] Lei Xu Yin Li, Yue Zhou, Xiaochao Yang, Incremental sparse saliency detection, in: IEEE International Conference on Image Processing (ICIP), 2009.

[12] Junchi Yan Yin Li, Yue Zhou, Visual saliency based on conditional entropy, in: The Asian Conference on Computer Vision (ACCV), 2009.

[13] Hae Jong Seo, Peyman Milanfar, Nonparametric bottom-up saliency detection by self-resemblance, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1st International Workshop on Visual Scene Understanding (ViSU), June 2009.

[14] Hae Jong Seo, Peyman Milanfar, Static and space-time visual saliency detection by self-resemblance, Jounal of Vision 9 (12–15) (2009) 1–27.

[15] M. Cheng, et al., Global contrast based salient region detection, in: IEEE Conference on Computer Vision and Pattern Recognition, June 2011, pp. 409–416.

[16] Z. Liu, et al., Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut, IEEE Transactions on Multimedia 14 (August (4)) (2012) 1275–1289.

[17] F. Perazzil et al., Saliency filters: contrast based filtering for salient region detection, in: IEEE Conference on Computer Vision and Pattern Recognition, June 2012, pp. 1–8.

[18] Antonio Torralba, Aude Oliva, Monica Castelhano, John Henderson, Contextual guidance of eye movements and attention in real-world scenes: the role of global features on object search, Psychological Review 113 (October (4)) (2006) 766–786.

[19] A. Garcia-Diaz et al., Saliency based on decorrelation and distinctiveness of local responses, in: Proceedings of 13th International Conference on Computer Analysis of Images and Patterns, 2009, pp. 261–268.

[20] A. Garcia-Diaz, et al., Decorrelation and distinctiveness provide with human-like saliency, in: J. Blanc-Talon, et al., (Eds.), Proceedings of 11th International Conference on Advanced Concepts for Intelligent Vision Systems, 2009, pp. 343–354.

[21] Ayellet Tal Stas Goferman, Lihi Zelnik-Manor, Context-aware saliency detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[22] B. Schauerte, G.A. Fink, Focusing computational visual attention in multi-modal human-robot interaction, in: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, ACM, 2010, p. 6.

[23] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007.

[24] C. Kanan, et al., Sun: top-down saliency using natural statistics, Visual Cognition 17 (6/7) (2009) 979–1003.

[25] L. Zhang, et al., Sun: A Bayesian framework for saliency using natural statistics, Journal of Vision 8 (7) (2008) 1–20.

[26] F. Estrada R. Achanta, S. Hemami, S. Susstrunk, Frequency-tuned salient region detection, in: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[27] Jian Li, et al., Saliency detection based on frequency and spatial domain analyses, in: Jesse Hoey, Stephen McKenna, Emanuele Trucco (Eds.), Proceedings of the British Machine Vision Conference, 2011, pp. 86.1–86.11.

[28] B. Schauerte, R. Stiefelhagen, Predicting human gaze using quaternion dct image signature saliency and face detection, in: Proceedings of the 12th IEEE Workshop on the Applications of Computer Vision (WACV)/IEEE Winter Vision Meetings, Breckenridge, January 2012, pp. 9–11.

[29] J.G. Daugman, Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters, Journal of the Optical Society of America 2 (July (7)) (1985) 1160–1169.

[30] L. Itti, C. Koch, Comparaison of feature combination strategies for saliency-based visual attention systems, SPIE Human Vision and Electronic Imaging (HVEI'99) 3644 (1999) 473–482.

[31] M. Mancas, Relative influence of bottom-up and top-down attention, Attention in Cognitive Systems (2009) 212–226.

[32] N. Riche, M. Mancas, B. Gosselin, T. Dutoit, Rare: a new bottom-up saliency model, in: Proceedings of the IEEE International Conference of Image Processing (ICIP), 2012.

[33] Gert Kootstra, Bart de Boer, Lambert Schomaker, Predicting eye fixations on complex visual stimuli using local symmetry, Cognitive Computation 3 (1) (2011) 223–240.

[34] D. Green, J. Swets, Signal Detection Theory and Psychophysics, John Wiley, New York, 1966.

[35] L. Itti, R.J. Peters, A. Iyer, C. Koch, Components of bottom-up gaze allocation in natural images, Vision Research 45 (18) (2005) 2397–2416.

[36] J.M. Henderson, et al., Visual saliency does not account for eye movements during visual search in real-world scenes, Eye Movements: A Window on Mind and Brain (2007) 537–562.

[37] Olivier Le Meur, Thierry Baccino, Methods for comparing scanpaths and saliency maps: strengths and weaknesses, Behavior Research Methods (2012) 1–16.

[38] Qi Zhao, Christof Koch, Learning a saliency map using fixated locations in natural scenes, Journal of Vision 11 (2011) 1–15.

[39] Ali Borji, et al., Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study, in: IEEE Transactions on Image Processing, in press.

[40] Ali Borji, Evaluation measures for saliency maps: AUROC and NSS. ⟨https://sites.google.com/site/saliencyevaluation/evaluation-measures⟩.

[41] F. Sawyer Steven, Analysis of variance: The fundamental concepts, The Journal of Manual and Manipulative Therapy, 17 (2) (2009) 27E–38E.

[42] David C. Howell, Statistical Methods for Psychology (Psy 613 Qualitative Research and Analysis in Psychology), Wadsworth Publishing, 2012.

[43] Barbara G. Tabachnick, Linda S. Fidell, Using Multivariate Statistics, 6th ed. Prentice Hall, 2012.

[44] H. Hentschke, M.C. Stüttgen, Computation of measures of effect size for neuroscience data sets, European Journal of Neuroscience, 2011.

[45] Jacob Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Routledge Academic, 1988.

[46] Mark W. Lipsey, Design Sensitivity: Statistical Power for Experimental Research (Applied Social Research Methods), Sage Publications Inc., 1989.