Improving unsupervised saliency detection by migrating from RGB to multispectral images

Journal:	*Color Research and Application*
Manuscript ID	COL-19-044.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Martinez , Miguel Angel; Universidad de Granada Facultad de Ciencias, Óptica Etchebehere, Sergi; Hewlett Packard Enterprise Co Valero, Eva; University of Granada Nieves, Juan Luis; University of Granada
Keywords:	conspicuity, machine vision, spectral imaging, multispectral images, visual saliency

Improving unsupervised saliency detection by migrating from RGB to multispectral images

MIGUEL ÁNGEL MARTÍNEZ,[1,*] SERGI ETCHEBEHERE,[2] EVA M. VALERO,[1] AND JUAN LUIS NIEVES[1]

[1] *Department of Optics, Facultad de Ciencias, Universidad de Granada, Granada 18071, Spain.*
[2] *HP Barcelona. Camí de Can Graells, 1-21, 08174 Sant Cugat del Vallès, Barcelona, Spain.*
**martinezm@ugr.es*

Abstract: Saliency detection has been an important topic during the last decade. The main goal of saliency detection models is to detect the most relevant objects in a given scene. Most of these models use RGB images as an input because they mainly focus on applications where features (e.g. faces, textures, colors or human silhouettes) are extracted from color images and there are many labeled databases available for RGB-based saliency data. Nevertheless, the use of RGB inputs clearly limits the amount of information from where to extract the salient regions since spectral information is lost during the color image recording. On the contrary, multispectral systems are able to capture more than three bands in a single capture and can retrieve information from the full spectrum at a pixel. The main aim of this study is to investigate the advantages of using multispectral images instead of RGB images for saliency detection within the framework of unsupervised models. We compare the performance of several unsupervised saliency models with both RGB and multispectral images, using a specific dataset of multispectral images with ground-truth data extracted from observers' fixation patterns. Our results show a general improvement when multispectral information is taken into account. The saliency maps estimated by using the multispectral features are closer to the ground-truth data, with the simplest Graph-based (GBVS) and Boolean Map-based (BME) models showing good relative gain compared with other approaches.

1. Introduction

The human visual system is able to detect the relevant or important information out of all the amount of data that enters the eye. This cognitive process known as visual attention is complex and its complete understanding and simulation have been widely explored. Back in 1998, Itti et al., [1] proposed the first completely functional saliency model, which tried to simulate where the human visual system would focus its attention on a given RGB image. After Itti's revolutionary work, many other models were to come which attempted to improve the results. In order to extract salient information, most models utilize some specific features, from the more basic intensity, color and orientation to the more advanced features like: motion, optical flow, flicker, multiple superimposed orientations (crosses or corners), texture contrast [2].

All the previously cited features use trichromatic images as an input. These are the more common types of images (RGB color images), which try to simulate how the human visual system responds to light and extracts color information. The human eye, and therefore most conventional cameras, have three kinds of channels or photoreceptors, sensitive to different parts of the visible light spectrum. Consequently, when capturing an image, the incoming light recorded by the camera sensor is encoded with three numbers (R-, G-, and B-digital values or L-, M- and S-cone responses) and thus the spectral information is lost. Nevertheless, such spectral information might be useful for certain applications. In recent years there has been a growing interest in devices (more and more affordable) able to capture all this extra information, not only with a better spectral resolution in visible light but also being able to capture light in other areas of the spectrum such as the ultraviolet, infrared and thermal. The increase in the availability of these multispectral and hyperspectral cameras has facilitated

huge advances in fields such as robotics, remote sensing, satellite imaging, medicine, food control and even object detection [3-5].

In this study, we analyze the advantages of using multispectral images with the aim of salient object detection using some of the more known saliency models and adapting them to receive and take advantage of spectral information. The topic of saliency detection and prediction is described in general at an introductory level in the review/book by Li and Gao. [6]. The idea is to compare the original models developed for RGB images with their adapted multispectral versions by using the most common evaluation metrics and investigate whether there is an improvement or not. Although multispectral images go beyond what the human vision can perceive, multispectral saliency detection does not imply a perfect simulation of bottom-up visual attention, but rather a broader detection of objects that stand out spectrally from their neighbors, which can also be related to knowledge and task associated visual attention, the so called top-down visual attention. Specific Visual Attention Models (VAM) have been developed for spectral images [7-9] (see section 2), but in these studies the comparison between RGB and multispectral images was not addressed specifically. Our study aims to tackle this issue using the least favourable situation for multispectral images, which is using models that have been specifically developed with RGB images in mind. Two specific fields of application that can benefit from the results shown in this paper could be: surveillance and security field (to detect objects or events of interest in urban scenes using modified camera surveillance devices to make them multispectral), and a second one could be the active monitorization of the state of preservation of the elements present in urban scenes.

This paper is organized as follows: Section 2 reviews some of the more relevant related studies modeling visual attention; Section 3 describes the methodology and the framework of the research, Section 4 analyzes the results obtained, and the conclusions are in Section 5.

2. Visual attention modelling

During the last decade it has been of great interest to look for appropriate answers about what determines in the end where and why an observer aims their gaze to particular locations in a scene. When some areas in an image attract the visual attention and the point of gaze of an observer it is said that these regions show high saliency, (i.e. specific low-level visual features are attracting the observers' interest), and thus the saliency map is a biologically plausible model for bottom-up attention as proposed by Koch and Ullman (1985) [10]. Their definition of saliency relied on center-surround principles considering that points in the visual scene are salient if they differ from their neighbors. There are many features characterizing a visual scene, amongst which we could cite edges, contrast, luminance and color as the main visual features defined at different scales. Classical bottom-up visual models get relatively good results when they use these features to localize the highly salient features in a scene, both for natural and artificial images. Latterly, including task-dependent constraints within the saliency algorithms has been found to improve the derived salient maps [11]. These kinds of models, which operate at higher visual levels (i.e. top-down models) use a prior knowledge to get visual attention. Eye tracking systems are usually employed to record observers' gaze paths as they view a collection of images. After discarding saccade fixation locations, the corresponding fixation map can be obtained.

As explained in the previous section, the most influential attempt to create a complete saliency model was made by Itti et al. [1] inspired by the theoretical work of Treisman et al. [12] in the feature integration theory (FIT) where three basic features that influence to the visual attention were proposed: intensity, color and orientation. The Itti model proposes how to extract these three features from a digital color image based on bottom-up scene-based properties by selecting pre-attentively computed simple features and combining all of them into a conspicuity map for each channel. Doing this to different sizes of the same image through a Gaussian blur pyramid, center-surround difference at each feature simulates the

neuronal receptive fields found in the human visual system. Finally, after obtaining the relative saliency contribution of each feature, a linear combination, resulting in the final saliency map is produced. Moreover, as established by Tatler et al. [13] there are differences between visual features in attended and non-attended spatial locations in an image. To be more specific, these differences are determined by various contrasts, luminances, chromaticity, energy and orientation. Nevertheless, doubt on these findings has been cast by Braddely and Tatler [14] who found that a fixation map is dominated by high-frequency edges; the authors argue that contrast does not contribute to saliency and that the other features are "behaviorally irrelevant".

Later on, many models appeared improving different assets of this initial approach: the use of a log-spectrum in the input image [15], using the information theory to extract salient information [16], using high level features [17] and supervised learning trained by large eye-tracking datasets [18]. Recently the majority of leading benchmark models have been based on convolutional networks and deep learning techniques [19].

In this section we describe first the RGB-based models used in our study, and then some models developed specifically for multispectral images.

2.1. RGB-based saliency prediction

Out of all the existing models we have selected five and adapted them to receive multispectral images as input. This selection was done taking into account their impact, their accuracy, and the feasibility of adapting them to multispectral images.

(1) ITTI: Itti's model [1] has been selected due to its influence on salience detection research and the many times it has been used in previous studies. As we have explained, Itti uses center-surround differentiation over three main features: intensity, color and orientation.

(2) GBVS: Harel et al. [20], proposed the graph based visual saliency, a modification of Itti's model; whilst using the same feature extraction, it proposes new activation, normalization and combination steps based on graph computation. Activation and normalization are achieved by implementing a Markovian approach: a fully connected graph, with a weight assigned to each edge connecting one node of the feature map to all the other nodes except itself. Therefore, by adding these two graph-based approaches to the steps of activation and normalization, and using the feature extraction already proposed by Itti, and a linear concatenation of normalized activation maps, they were able to significantly improve both the performance and the accuracy of the other existing saliency methods.

(3) RARE: Published in 2012 by Riche et al. [21], it proposes finding salient information by looking at the rarity of the different features. Rarity is calculated by using co-occurrence matrices of a given pixel or region; giving high values to pixel that has values that are less frequent. It uses principal component analysis (PCA) over the RGB images in order to find higher discriminations; it also uses Gabor filters to analyze different orientations.

(4) BMS: Boolean map saliency, proposed by Zhang and Sclaroff in 2013 [22]. The idea is to binarize the different channels of the image by using random thresholding, and extract the salient information by analyzing their topological structure. This model is quite simple and using low-cost processing it reaches high scores when compared to other models.

(5) LDS: Continuing with the same idea as RARE, learning discriminative subspaces on random contrasts [23] tries to project the images into more discriminative sub-spaces that allow targets to pop out. It calculates the principal components using a big set of image patches and by maximizing the contrast between target and background it learns what sub-spaces are more suited to show this differentiation.

2.2. Spectral-based saliency prediction

Although the previously cited models are able to predict salient information with a high accuracy (although lower than supervised models), they extrapolate all the information from an RGB image. The idea of using multispectral or hypersepctral images in order to predict salient information is not new and there have been several attempts to create spectral-image-based saliency models. Most of them adapted Itti's model to receive different features such as:

- Using space transformation methods such as Principal Component Analysis (PCA) [8] or Non-negative Matrix Factorization (NMF) [9] in order to reduce the dimensionality of the multispectral images and get a higher contrast of the more distinguishable objects.
- Computing spectral differentiation metrics between the different pixels, to more easily compute spectral differences between center and surround [9, 24].
- Taking advantage of the higher spectral resolution to more accurately select the blue-yellow (BY), and red-green (RG) vectors extracted from the corresponding group of spectral bands [9].

All the above studies were presented as complete saliency models instead of an adaptation of previous models, so it was difficult to distinguish whether the performance of these models is related specifically to the usage of multispectral or hyperspectral information. In our case, we use models specifically developed for RGB images and adapt them to receive multispectral information as input. Our aim is to investigate if there is an improvement in the models' performance when they use a more complete source of information to obtain the saliency prediction.

We are aware that the selected models are not among the best performing since the advent of convolutional neural networks (CNN-based saliency prediction approaches [25-26]). However, supervised models would require a high amount of labeled spectral images to produce acceptable results, since they would per force have to be re-trained if spectral images are to be used as input. Currently, there are no labeled spectral images' databases of more than four channels for saliency detection. We think it is worth investigating if using spectral information can provide a significant improvement in unsupervised saliency prediction before tackling the huge task of capturing and labelling a sufficient amount of spectral data to test with supervised approaches for saliency prediction. Besides, finding efficient ways to adapt existing models to receive different input data has also an intrinsic interest.

3. Methods

3.1 Image dataset and ground-truth data

We have used a set of nine multispectral images of urban scenes and their corresponding RGB version, three of them containing people, to test RGB vs Multispectral images saliency prediction performance. The results of this study are applicable to saliency detection in any framework, though the scenes captured in this work have only urban content (buildings, vehicles, urban furniture, people, plants…). The images were recorded using the PixelTeq (Halma, UK) SpectroCam VIS camera [27] which is composed of a monochrome silicon sensor with spatial resolution of 2456 x 2058 pixels, sensitive to wavelengths between 370 and 1100 nm (see Fig. 1 left). We are aware that more advanced sensors like InGaAs-based ones are sensitive to spectral regions beyond this range (i.e. up to 1700 or 2500 nm). This could certainly yield interesting results since we could explore different spectral regions with maybe interesting information for the saliency detection task. However, this would highly increase the cost of the imaging systems. In this regard, silicon-based sensors offer an affordable and easy to find solution yet presenting good performance for saliency detection. A filter wheel with 8 slots is placed between the lens and the sensor, and rotated to sequentially

capture the images corresponding to each band. The exposure time for each channel was determined independently to ensure that the scene was correctly exposed for the corresponding band.

We selected a range of filters with specific transmittances to cover the whole visible and near infrared (NIR) regions of the spectrum. In Fig. 1 right, the spectral responsivities of the channels are shown. Channels 1 to 5 have their responsivities within the visible range (roughly from 400 to 750 nm), channels 7 and 8 are sensitive in the NIR range (from 750 to 1000 nm), and channel 6 is both sensitive in the visible and NIR ranges. Of course, a higher number of channels with spectrally narrower sensitivities could help improving the saliency detection task by offering a higher amount of data. However, this would also increase the cost and complexity of the imaging system and the image data processing. For specific applications, an optimized filter selection could be carried out [28]. However, in this study, the available filters were meant for the general spectral imaging task, thus covering the whole visible and NIR range with certain overlap.

Each image has a resolution of 2456 x 2058 pixels x 8 different channels corresponding to the transmittance of each filter to the scene. In order to generate the RGB images from our multispectral data, we just selected three filters which were reasonably close to the standard peak wavelengths of R, G and B channels in a conventional RGB camera, and used them as the three channels of the RGB image. These filters were those corresponding to channels 5 (680 nm), 3 (555 nm) and 1 (450 nm), respectively. At this point one might think that the comparison is not fair since the RGB images only cover the visible range, and the multispectral system used in this article also cover the NIR range up to 1000 nm. However, this advantage is a part of the potential assets of multispectral systems. Not only offering a higher number of spectral channels within the same spectral range, but also extending its spectral range. Specifically, the sensor used in this study is a silicon-based sensor like the ones used in common RGB imaging systems. Therefore, we could extend the potential of any silicon sensor by removing the IR cut-off filter and adding the same color filters with the filter-wheel. There is no need to use a more complex and costly InGaAs-based sensor for it [29].

We used the RGB images to generate the ground-truth data for testing our hypothesis. A total number of 6 different observers composed of 4 women and 2 men, with mean age of 24 years were asked to look freely at the RGB version of the images while their eye movements were being recorded with an Eye-tracker device (Tobii II, from Tobii company, Sweden) [30]. The images were presented during 6 s. The objects with the highest number of fixations in each image (accumulating more than 70% of the fixation time), were marked as ground-truth salient objects and manually segmented from the images to generate the ground-truth data (see Fig. 2).

3.2 Features analyzed

In this subsection, we describe the features extracted from the spectral images and later fed as input for the adapted version of the visual attention models. We have used a range of features that can be divided into three main groups:

(a) **CIELab**: in general, color information is used in most of the saliency models, which use color or RGB images as input. The raw RGB color information can be used directly by the model, or else the RGB can be transformed into a different color space which better emulates human perception. The CIELab color space [31] is quite widely used for this purpose. The information conveyed by the three channels of the Lab feature (L*, a* and b*) is then fed to the

adapted models as a 3-dimensional image. Therefore, the model processes each channel independently and the corresponding activation maps are concatenated.

(b) **PCA:** in spectral images, usually the information contained in each pixel is high-dimensional. Our hypothesis is that this extra amount of information can be useful in the prediction of saliency. Nevertheless, many spectra are smooth functions and this means that there will be some amount of correlation between adjacent spectral bands. One way to exploit this correlation and try to keep the most relevant and distinctive features of the spectra is to use a dimensionality reduction technique such as PCA, which finds the best set of orthogonal components to represent the data while capturing the highest amount of their inner variance. PCA is also used as a feature in previously developed VAM for spectral images [24]. The principal component basis vectors are usually ranked by variance accounted for (VAF), and the number of principal component vectors used to represent the data is selected using a threshold criterion for accumulated VAF, usually ranking from 95 to 99%. For our data, we have checked that using three principal components we are able to account for at least 95% of the variance, so we have decided to use the projections of our image data onto the first three principal components as an additional feature for the saliency models. The three projected images are fed independently to the models, and the activation maps are computed and then concatenated. We have computed the PCA decomposition individually for each single image, to preserve its distinctive characteristics as much as possible and since the images were of a size that produced a sufficiently high number of pixels to allow for this approach.

(c) **SAM-SID**: When comparing spectral data to analyze differences between them, it is good practice to not only compare them channel by channel, but to also consider the spectrum as a whole. In the case of spectral images, since each pixel has N spectral components, the image can be considered as an array of signals and each pixel can be compared with the mean signal in the image, which could be a way to identify which are the most distinctive regions. There are different metrics used to numerically discriminate spectral signals, for instance Root Mean Square Error (RMSE) computes the square root of the mean of the channel-wise differences to the square, or Goodness-of-Fit Coefficient (GFC) which is the cosine of the angle between two spectral signals (considering them as vectors on a Hilbert space [32]). In our case we use the so-called SAM-SID distance [33] which is a combination of both the spectral angle mapper (SAM) and the spectral information divergence (SID). SAM is defined as the angle between two spectral signatures s and s', (and so the $\cos^{-1}$ of the GFC value) as expressed in the following formula:

$$SAM(s,s') = cos^{-1}\left(\frac{< s,s' >}{\| s \| \cdot \| s' \|}\right) \quad (1)$$

Meanwhile, SID is the discrepancy between the uncertainty of two spectral signatures, s and s', which is computed using their respective probability density distributions p and q:

$$D(s \parallel s') = \sum_{j=1}^{L} p_j log\left(\frac{p_j}{q_j}\right) \quad (2)$$

$$D(s' \parallel s) = \sum_{j=1}^{L} q_j log\left(\frac{q_j}{p_j}\right) \quad (3)$$

$$SID(s,s') = D(s \parallel s') + D(s' \parallel s) \quad (4)$$

Then the combination of both SAM and SID is done as the sinus of the angle by the information divergence:

$$D_{SAM-SID}(c,s) = \sin[SAM(c,s)] \times SID(c,s) \quad (5)$$

The advantage of this metric is that it combines sensitivity to differences in spectral amplitude distribution (SID) with sensitivity to differences in spectral shape (SAM). By taking the product of these two measures, the spectral discriminability of the SID-SAM mixed metric is increased because it makes two similar spectral signatures even more similar and two dissimilar spectral signatures more distinct [31]. Therefore, a one-channel feature is introduced as input to the saliency models, showing the SAM-SID difference between each pixel and the mean spectra of the scene. This feature is activated by the model and saliency is predicted.

Both PCA and SAM-SID are features that can only be extracted from multi-spectral images; nevertheless, the color information is already used as a feature in most saliency models. In Fig. 3, we show one scene (original scene), its segmentation ground truth, and the corresponding feature images (PCA, SAM-SID and Lab). The salient objects tend to have high intensity in some of the feature images, which can be useful for improving the performance of the VAM.

3.3 Model Adaptations

In this section we explain the adaptations carried out on the existing visual saliency models to enable them to receive spectral features as inputs. Since the different models have a completely different architecture, we have designed different ways of adapting them to accept the spectral features as inputs.

Fig. 4, summarizes the work-flow of the experiment performed for each of the models selected, with the aim of establishing if the use of spectral features as input produces an increase in the performance of the models. We first used RGB images as input for the model, and obtained the corresponding saliency map. Then, we used the adapted version of the model with the spectral feature images as input, and obtained the spectral-based saliency map. Finally, we used the ground truth and the set of metrics described in section 3.4 to compare the performance of the model in the two situations (RGB or spectral features as input).

Itti and GBVS use intensity, color and orientation as the main features and then the activation maps are computed. For these two models, we have substituted intensity and color for the CIELab features, and we have used the L* image to compute the orientation maps. Then, we added both PCAs (sequentially for each PCA component) and SAM-SID as extra features, leaving the models with a total of 4 feature global classes to be activated and combined. We have merged the activation maps with equal weights for all the features. Both RARE and LDS use PCAs to find a space which increases the differences between the objects. In this case we substitute the 3-dimensional input image by a 7-dimensional one, composed of the 3 CIELab channels, the first 3 principal components, and the SAM-SID image. We then run the model with the corresponding space transformations, and the final saliency map is obtained. The BMS model applies random thresholding to the different channels of the input image. In this case, instead of applying threshold to 3 different channels (RGB), we have used the random thresholding for the 7 different maps (CIELab + PCA + SAM-SID).

3.4 Validation

Once a model detects the main salient regions in an image, it is necessary to validate its performance over ground-truth data. There are several metrics commonly used in this field and standardized so different models can be compared, although consistent results cannot always be obtained [34]. Depending on the application and the kind of data used for validation, some metrics can be more appropriate than others. We decided to use the following three metrics for our experiment:

(a) **Area under curve (AUC):** this is computed from the receiving operator characteristic (ROC) curve. For different values of threshold in the saliency map produced by the model, true positives and false positives are computed by using the ground-truth data. Two main implementations of the AUC metrics are used: AUC-Borji [35] and AUC-Judd [18]. Another version of this metric was created in order to compensate the well-known center bias, the shuffled AUC [24], which was the one we used to validate our data. The main drawback of AUC metric is that low-valued false positives are not penalized [36]. This means that if the saliency map is predicting objects as salient that are not truly salient according to the ground truth, it could still reach high values of AUC. In other words, diffuse saliency maps in which many areas are highlighted with not very extreme values of saliency are not considered as poor quality.

(b) **Normalized Scan-path Saliency (NSS):** this is computed as the averaged normalized saliency at the ground-truth location. Chance level is assigned zero value, and a positive value would mean above chance results. This method solves the issue of not penalizing low-valued false positives, by assigning the highest score to a map that would detect all the pixels in the ground-truth salient regions as salient, and would have zero values in all the rest of the pixels in the image. [37]

(c) **Information Gain (IG):** this is a metric designed to compare two saliency maps taking into account the similarity of the probabilistic distribution with the ground-truth data [38]. Therefore, this metric is well suited to directly compare between two different saliency methods, computing the gain or loss in information with respect to the ground-truth data for the two maps that are compared.

Although there are many more different metrics for saliency benchmarking, most of them can be highly correlated with one of the three metrics that we have chosen; these three metrics are good representatives of different strategies in the definition of quality of saliency prediction.

4. Results

As we explained in the previous section, for each of the 9 multispectral images we calculated their saliency maps predicted by the 5 different models when using both the original features and the spectral ones. An example of these saliency maps can be seen in Fig. 4.

For each of the saliency maps the scores of the three different metrics described in section 3.4 were calculated. Table 1 shows the average and standard deviation over the 9 images for each of the models using both original and spectral features and each of the metrics, and also the relative difference between both inputs' scores. In the case of AUC and NSS, the difference between the original and the spectral features is shown, meaning positive a better score of the spectral features. The relative gain for the use of spectral features with respect to RGB features is also shown in the table. In the case of IG, since it already compares

the two maps, only the average over the images is shown; a positive result shows better accuracy of the spectral features over the original RGB-based features.

Analyzing the results in Table 1 we can observe some differences between the different models and also between the different metrics. We can see how both ITTI and GBVS models have one of the highest scores in AUC whilst the NSS score found is below the average across the models. One of the reasons of this noticeable difference between AUC and NSS in the ITTI and GBVS models might be the large amount of high (or salient) values in the maps; having a lot of false positives is penalized by NSS but not by AUC. Now looking at the RARE results, this is the model scoring the highest in AUC and second highest in NSS. We can appreciate in Fig. 5 how the resulting maps tend to contain high values in the salient object regions, and generally low values for non-salient regions. The BMS and LDS models are amongst the worst performing overall, having relatively low AUC and NSS scores both for RGB and multispectral images.

Now, we analyze the models' performance when we use the spectral features as input, which is the main aim of our experiment. Except for the RARE model in the IG metric, we have found that there is an improvement in the models' performance when used with spectral features. This improvement is much more marked for the NSS and IG metrics than for the AUC metric. For the Itti and GBVS models, there is a clear improvement in NSS values, which reach a level comparable to other models for the spectral features, while the performance is much poorer if we use the RGB image as input. For the RARE model, we can see the least improvement in AUC, the second smallest in NSS and even a decrease in the accuracy in IG.

The RARE model looks for rarity instead of center-surround difference for computing the saliency map, so its strategy is markedly different from the first two models analyzed. The model is already performing quite well (compared with the others) when using the RGB image as input, and the adaptations we have introduced might not be able to add enough value to the spectral features. Regarding BMS and LDS, they both improve the accuracy when spectral information is used: around 0.6 in NSS and 0.5 in IG, with BMS reaching the highest IG score. This considerable improvement in performance might be due to a more successful adaptation strategy when introducing the spectral features. The average relative gain for all five models is 9.2% for AUC and 61.2% for NSS. Finding precisely the factors that result in the observed improvement when using multispectral scenes as input for the visual attention models tested is not a straightforward task. One factor is related to the new features introduced (PCA and SAM-SID), that in some instances clearly highlight the salient objects, as can be seen in Figure 3 and also in Figure 6. The remaining factors are linked to the specific way each model uses the input features to extract the saliency maps, and a detailed discussion would be excessively long considering the number and diversity of the models presented here, and the fact that for some of them it is not easy to sequentially analyze each step and its relationship with the final saliency map delivered by the model.

5. Conclusions and future work

We have used AUC, NSS and IG metrics to assess the performance of five well-known visual attention models with multispectral and conventional RGB color images. Our results suggest that the saliency maps produced by using the multispectral features are closer to the ground-truth data. The higher gain for NSS is quite significant since this metric has

advantages over AUC. In fact, NSS will be adopted as gold standard quite soon in the VAM most popular benchmarks [39].

Saliency prediction performance has improved dramatically during the last three years after the outbreaks of the deep learning algorithms. Our promising results point out the fact that a CNN-based model, adequately trained using our specific spectral features, will improve the detection of the salient regions. A potential CNN-based Spectral saliency detection method will carry out a prediction of the salient regions analyzing in parallel all the spectral bands of an input image. This higher amount of information compared to RGB images, will allow the CNNs finding more complex features to detect saliency. Typically, we would need over 1,000 images to get a decent accuracy in image classification on the cross-validation set (or even more if a transfer learning on an already-trained model is not used). However, in the absence of such number of multispectral images adapted for a saliency task it would be difficult to hazard even a guess at the final spectral performance. After the results found in this study, a new multispectral image database is being built, together with its ground truth data. It is a matter for further studies to implement a CNN-based spectral saliency model, adequately trained with this labelled multispectral image dataset.

Acknowledgements

This research was supported by a joint agreement (reference number C-3368-00) between Tecnalia company and the Business-UGR Foundation, and through the Ministry of Economy and Competitiveness of Spain under research grant DPI2015-64571-R. We also thank Angela Tate for reviewing this text.

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, 1254–1259 (1998).
2. A. Borji, L. Itti. "State-of-the-art in visual attention modeling." IEEE transactions on pattern analysis and machine intelligence 35.1, 185-207 (2013).
3. L. M. Dale, A. Thewis, C. Boudry, I. Rotar, P. Dardenne, V. Baeten, and J. A. F. Pierna, "Hyperspectral imaging applications in agriculture and agro-food product quality and safety control: a review", Appl. Spectroscopy Reviews, 48(2), 142–159 (2013).
4. G. Lu, and B. Fei, "Medical hyperspectral imaging: a review", Journal of Biomedical Optics, 19(1), 010901 (2014).
5. H. Liang, "Advances in multispectral and hyperspectral imaging for archaeology and art conservation", Appl. Physics A, 106(2), 309–323 (2012).
6. Li, J., & Gao, W. (Eds.). (2014). Visual saliency computation: A machine learning perspective (Vol. 8408). Springer.
7. M. Cornia, L. Baraldi, G. Serra, R. Cucchiara, "Predicting human eye fixations via an LSTM-based saliency attentive model", 2016 https://arxiv.org/pdf/1611.09571.pdf
8. Q. Wang, P. Yan, Y. Yuan, X. Li, "Multi-spectral saliency detection", Pattern Recognition Letters, 34, 34–41 (2013).
9. J. Zhang, W. Geng, L. Zhuo, Q. Tian, Y. Cao, "Multiscale target extraction using a spectral saliency map for a hyperspectral image", Applied Optics, 55, 8089-8100 (2016).
10. C. Koch, S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry", Hum. Neurobiol. 4, 219-227 (1985).
11. R.L. Canosa, "Modelling selective perception of complex natural scenes", International Journal of Artificial Intelligence Tools,14, 233–260 (2005).
12. A.M. Treisman, G. Gelade, "A Feature-Integration Theory of Attention", Cognitive Psychology, 12, 97–136 (1980).
13. B.W. Tatler, R.J. Baddeley, I.D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time", Vision Res. 45, 643-659 (2005).
14. R. Baddeley, B. Tatler, "High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis", Vision Res. 46, 2824–2833 (2006).
15. X. Hou, L. Zhang, "Saliency detection: A spectral residual approach", IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, 1-8 (2007).
16. N.D.B. Bruce, J.K. Tsotsos, "Saliency, attention, visual search: An information theoretic approach", J. Vision, 9, 1-24 (2009).

17. P. Sharma, F.A. Cheikh, J.Y. Hardeberg, “Saliency map for human gaze prediction in images”, in 16th Color Imaging Conf., Portland, OR, USA, 332-337 (2008).
18. T. Judd, K. Ehinger, F. Durand, A. Torralba, “Learning to predict where humans look”, in IEEE Int. Conf. Computer Vision, 2106–2113 (2009).
19. Borji, Ali. "Saliency prediction in the deep learning era: An empirical investigation." arXiv preprint arXiv:1810.03716 (2018).
20. J. Harel, C. Koch, P. Perona, “Graph-based visual saliency”, Advances in Neural Information Processing Systems, 545–552 (2006).
21. N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, T. Dutoit, “Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis”, Signal Processing: Image Communication, 28, 642-658 (2013).
22. J. Zhang, S. Sclaroff, “Saliency detection: A boolean map approach”, In Proceedings of the IEEE international conference on computer vision, 153-160 (2013).
23. S. Fang, J. Li, Y. Tian, T. Huang, X. Chen, “Learning discriminative subspaces on random contrasts for image saliency analysis”, IEEE transactions on neural networks and learning systems, 28, 1095-1108 (2017).
24. S. Le Moan, A. Mansouri, J.Y. Hardeberg, Y. Voisin, “Saliency for spectral image analysis”, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6, 2472–2479 (2013).
25. Wang, L., Gao, C., Jian, J., Tang, L., & Liu, J. Semantic feature based multi-spectral saliency detection. Multimedia Tools and Applications, 77(3), 3387-3403, (2018).
26. Wang, T., Borji, A., Zhang, L., Zhang, P., & Lu, H. (2017, October). A stagewise refinement model for detecting salient objects in images. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4019-4028).
27. http://halmapr.com/news/pixelteq/2016/07/21/new-spectrocam-swir-640-multispectral-wheel-camera-from-pixelteq.
28. Hardeberg, J. Y. (2004). Filter selection for multispectral color image acquisition. Journal of Imaging Science and Technology, 48(2), 105-110.
29. Martin, T., Brubaker, R., Dixon, P., Gagliardi, M. A., & Sudol, T. (2005, May). 640x512 InGaAs focal plane array camera for visible and SWIR imaging. In Infrared Technology and Applications XXXI (Vol. 5783, pp. 12-20). International Society for Optics and Photonics.
30. https://www.tobii.com/.
31. Schanda, János, ed. Colorimetry: understanding the CIE system. John Wiley & Sons, 2007.
32. J. Hernández-Andrés, J. Romero, A. García-Beltrán, J.L. Nieves, “Testing linear models on spectral daylight measurements”, Applied Optics, 37, 971-977 (1998).
33. Y. Du, C.I. Chang, H. Ren, C.C. Chang, J.O. Jensen, F.M. D’Amico, “New hyperspectral discrimination measure for spectral characterization”, Optical engineering, 43, 1777-1178 (2004).
34. Kümmerer, M., Wallis, T. S., & Bethge, M. (2017). Saliency Benchmarking: Separating Models, Maps and Metrics. arXiv preprint arXiv:1704.08615.
35. A. Borji, D.N. Sihite, L. Itti, “Quantitative analysis of human-model agreement in visual saliency modelling: a comparative study”, IEEE Transactions on Image Processing, 22, 55–69 (2012).
36. Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, F. Durand, “What do different evaluation metrics tell us about saliency models?”, 2016.
37. R.J. Peters, A. Iyer, L. Itti, C. Koch, “Components of bottom-up gaze allocation in natural images”, Vision Research, 45, 2397-2416 (2005)
38. M. Kummerer, T. Wallis, M. Bethge, “Information-theoretic model comparison unifies saliency metrics”, Proceedings of the National Academy of Sciences, 112, 16054–16059 (2015).
39. http://saliency.mit.edu/results_mit300.html

Authors bios:

Miguel Ángel Martínez is a Post-doc researcher working on multi and hyperspectral high dynamic range image capture and processing in the visible and near-infrared. His research interests are in the field of high dynamic range imaging, spectral imaging, digital image processing and analysis, color and spectral sciences, machine/deep learning, saliency detection, color vision, etc. He is BSc. In Telecommunications Engineering specialized in image and sound, MSc. In Color in Informatics and Media Technology, and PhD. In Physics and Space Sciences.

Sergi Etchebehere is a researcher that worked on detection and classification of salient objects in hyperspectral images. His research interests are spectral image processing and analysis and use of machine learning techniques in computer vision. He currently works at Hewlett Packard.

Eva Valero is Associate Professor at the Department of Optics of the University of Granada. Her recent research interests include hyperspectral imaging, spectral estimation, HDR Imaging, and color and spectral image processing. She is involved in teaching several subjects in the B.S. degrees of Physics, Optics and Optometry, and the CIMET/COSI international master program.

Juan Luis Nieves received M.S. and Ph.D. degrees in physics from the University of Granada, Granada, Spain, in 1991 and 1996, respectively. He is currently Full Professor with the Department of Optics, Science Faculty, at University of Granada, where he conducts research in the Color Imaging Laboratory. His current research interests include computational color vision (color constancy, human visual system processing of spatio-chromatic information), and spectral analysis of color images. Dr. Nieves is the President of the Spanish Color Committee, and representative of this Committee in the International Color Association (AIC), and is currently the coordinator of the Erasmus1 Joint Master Degree "Color in Science and Industry (COSI)."

Table 1: Average and standard deviations over the 9 images for (rows) each model using both original and spectral features, and (columns) each of the metrics.

		AUC	Relative AUC variation	NSS	Relative NSS variation	IG
GBVS	**RGB**	0.792 (0.114)	9.7%	1.176 (0.440)	93.2%	0.526 (0.547)
	Hype	0.868 (0.064)		2.272 (1.348)		
ITTI	**RGB**	0.843 (0.080)	6.7%	1.359 (0.535)	62.4%	0.480 (0.441)
	Hype	0.904 (0.064)		2.356 (1.147)		
BMS	**RGB**	0.631 (0.125)	20.6%	1.093 (0.809)	70.4%	0.551 (1.143)
	Hype	0.761 (0.128)		1.861 (1.237)		
LDS	**RGB**	0.569 (0.144)	7.0%	0.763 (0.604)	70.6%	0.465 (0.230)
	Hype	0.609 (0.087)		1.302 (0.714)		
RARE	**RGB**	0.895 (0.087)	2.1%	2.121 (1.081)	9.4%	-0.053 (0.333)
	Hype	0.914 (0.049)		2.320 (1.007)		

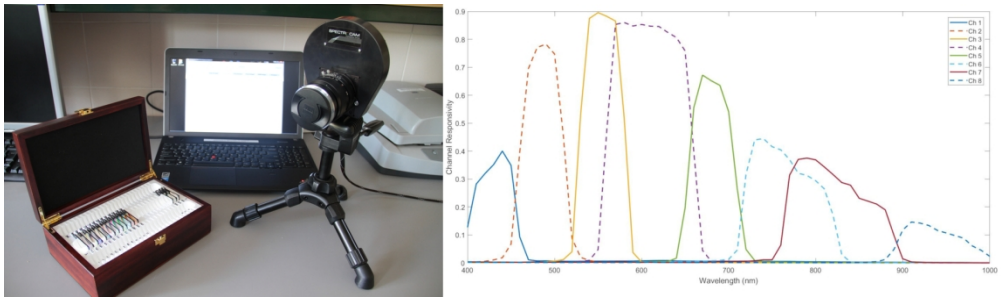

Fig. 1. Left: PixelTeq SpectroCam VIS camera. Right: spectral responsivity of the eight channels used by the Spectrocam VIS camera, computed as the product of the spectral transmittance of each filter by the spectral responsivity of the monochrome sensor.

155x45mm (300 x 300 DPI)

Fig. 2. Original scene (left), fixation map (center) and segmented ground-truth image (right) for one of the scenes.

Fig. 3. RGB scene and corresponding feature images fed as input to the VAM tested.

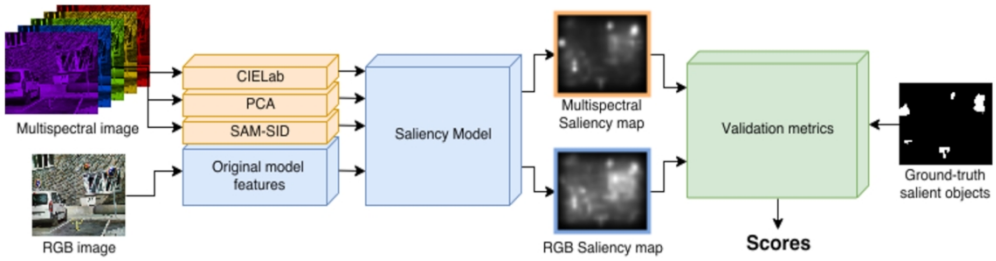

Fig. 4. Illustration of the Work-flow of our experiment. The procedure is repeated for each of the models tested.

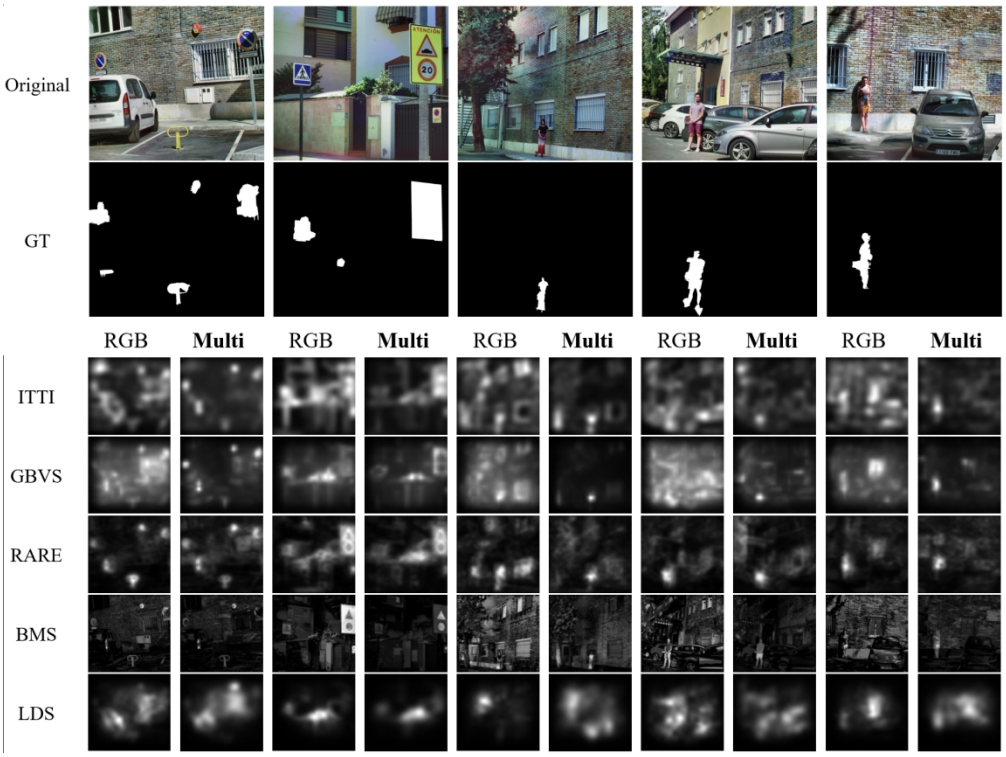

Fig. 5. An example of the saliency maps for each model using both original and spectral features of different images; ground truth (GT) is also shown for comparison

Fig. 6: An example of one of the images (original RGB), with its segmentation ground truth, and its feature images corresponding to principal components 2 and 3 (PCA 2 and 3), and SAM-SID.