IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 28, 2020, accepted June 29, 2020, date of publication July 2, 2020, date of current version July 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3006700

A Multi-Scale Colour and Keypoint Density-Based Approach for Visual Saliency Detection

ALESSANDRO BRUNO[1], FRANCESCO GUGLIUZZA[2], ROBERTO PIRRONE[3], (Member, IEEE), AND EDOARDO ARDIZZONE[3]

[1]National Centre for Computer Animation (NCCA), Bournemouth University, Poole BH12 5BB, U.K.
[2]Dipartimento di Fisica e Chimica "Emilio Segrè" (DIFC), Università degli Studi di Palermo, 90128 Palermo, Italy
[3]Dipartimento di Ingegneria, Università degli Studi di Palermo, 90128 Palermo, Italy

Corresponding author: Alessandro Bruno (alessandro.bruno@inaf.it)

This work was supported by the OBIND Project, under Grant 086202000366 by Sicily Region, PO FESR 2014-2020, Azione 1.1.5.

ABSTRACT In the first seconds of observation of an image, several visual attention processes are involved in the identification of the visual targets that pop-out from the scene to our eyes. Saliency is the quality that makes certain regions of an image stand out from the visual field and grab our attention. Saliency detection models, inspired by visual cortex mechanisms, employ both colour and luminance features. Furthermore, both locations of pixels and presence of objects influence the Visual Attention processes. In this paper, we propose a new saliency method based on the combination of the distribution of interest points in the image with multiscale analysis, a centre bias module and a machine learning approach. We use perceptually uniform colour spaces to study how colour impacts on the extraction of saliency. To investigate eye-movements and assess the performances of saliency methods over object-based images, we conduct experimental sessions on our dataset ETTO (Eye Tracking Through Objects). Experiments show our approach to be accurate in the detection of saliency concerning state-of-the-art methods and accessible eye-movement datasets. The performances over object-based images are excellent and consistent on generic pictures. Besides, our work reveals interesting findings on some relationships between saliency and perceptually uniform colour spaces.

INDEX TERMS Eye-movements, interest points, saliency map, visual attention.

I. INTRODUCTION

The human visual process starts outside the brain with the projection of the light onto the retina. The retina is a thin layer of neural tissue including the rods and cones, which are responsible for dim light and daylight vision. Thanks to the overall architecture of our visual system, we can transmit and receive up to 10 billion bits of information per second [1]. Therefore, the lack of storage capacity of our brain concerning the huge amount of information going from our eyes towards the cerebral cortex to be processed at much higher levels. However, due to the limits of our brain, we cannot simultaneously perform complex analysis on all the input visual information [1]. For a given scene, the detection of the most critical visual subset occurs as one of the most important tasks of the Human Visual System (HVS). When a person performs any visual task (watching TV, driving a car) the eyes flick rapidly from place to place to inspect the visual scene. While observing a scene, saccadic eye movements allow for the central part of the vision (fovea) to fall upon the region of interest of a picture. Vision is not uniform across our field of view, and acuity decreases with eccentricity as shown in [2]. Scientific researches also show physiological evidence proving that human brain employs visual attention to select regions from images to serialize the perception of objects [3]. Theeuwes [4] discussed evidence regarding the endogenous (goal-oriented) and exogenous (stimulus-driven) control of attention in visual tasks for observers searching for a particular visual feature such as colour, shape, or brightness.

Attention can be described as the allocation of cognitive resources to information, and it can be divided into five types constituting a hierarchical model:

- Focused attention
- Sustained attention
- Selective attention
- Alternating attention
- Divided attention.

We focused exclusively on selective attention as defined in the scientific literature [1], [5] as ‘‘the ability to selectively

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

maintain the behavioural or cognitive resource on specific stimuli while ignoring the distracting or competing stimuli''.

As shown in [6], there is an intimate connection between visual attention and eye movements. That is the reason why investigations on the salient areas of images have become a critical topic in scientific research.

Snowden *et al.* [2] state that visual attention is mainly guided by two factors: bottom-up and top-down factors. Bottom-up factors are stimulus-driven, derived from the regions of interest that pop out from the visual scene to our eyes. Top-Down factors are more relevant to specific rules, assigned visual tasks and behavioural based goals. Regions of interest that catch human attention are highly discriminating to the centre-surround principle. Visual Saliency aims to imitate the behaviour of HVS by predicting the fixation points of the most critical regions of an image from a perceptual point of view. Visual Saliency is a multidisciplinary branch of research; it lies on the progress achieved by different sciences such as Psychology, Neurobiology, Computer Science, Artificial Intelligence, Medicine [7]. In our work, we reveal the most salient subset of an image by setting up the corresponding saliency map. A saliency map is a grayscale map with each pixel falling in the dynamic range [0, 255]. The higher the intensity value, the more salient the location in the visual scene.

As well as visual attention approaches, visual saliency methods can be grouped into three main approaches by considering the visual feature and the visual attention process involved within the extraction of the saliency map: bottom-up, top-down, hybrid.

Visual Saliency bottom-up methods are stimulus-driven, characterized by the so-called ''visual pop-out'' saliency. In these approaches, the exogenous attention is involved with the visual saliency. The centre-surround operation [8] and graph-based activation maps [9] are examples of implemented exogenous attention processes. These methods exploit low-level features of the images such as contrast, texture, colour, intensity to give rise to saliency maps.

Visual saliency top-down methods are based on high-level visual tasks such as text, object, face detection. In the top-down approach [10], a predefined task is given by the object category to be detected.

Hybrid methods are conceived to work on two levels: bottom-up and top-down. The former allows extracting a noisy saliency map, the latter filters out noisy regions in saliency maps created by the bottom-up layer.

We have recently concentrated our efforts on studying the performance of several visual saliency approaches about the task of the object attention. We gathered a collection of eye-fixation point maps for different object images under the name of ETTO (Eye Tracking Through Objects) [11], which is described in greater detail in the eye-tracking datasets section. Visual selective attention includes, among others, location-based and object-based attention [12]). Still, we are interested in studying the different performances of saliency methods for both object-based and generic image datasets. We aim to analyze visual saliency performance regarding how visual attention processes select features that are part of an object. Several experiments conducted over the years showed observers prefer to make an eye movement towards the other end of the same fixated object rather than to an equidistant end of a different object. In fewer words, that is a preference to make eye shifts within the same object rather than between objects [13]. In [14] some researchers dealt with the analysis of those processes which determine whether or not an object in our environment captures our attention.

As revealed by findings and researches over the last decade [15]–[17], the relations between low-level features of the image and visual attention processes need to be further inspected. For this purpose, our investigations addressed the roles played by colour and scale features in detecting the visual saliency of a given image. In this regard, it is well known from the scientific literature that colour plays an essential role in the extraction of saliency maps and more generally on the overall visual attention processes [18]. Several investigations have also been conducted over the last decades on the bottom-up visual attention processes concerning the impact of colour, contrast, texture and multiscale local keypoints on the generation of an early saliency map on the visual cortex V1 [19]. Many past studies have emphasized the importance of borders in colour and luminance surface representations in the early visual cortex V1 [20]. Models inspired by the above aspects aim to mimic V1's simple cell mechanisms by computing centre-surround differences with distinct colour and luminance opponency. Both CIE L*a*b* and CIE L*u*v* colour spaces employ the opponent colour encoding. Euclidean distances are used to provide the spaces with a colour-difference formula for evaluating colour differences in perceptual relevant units [21]. As revealed by Sharma *et al.* in [22], while CIE L*a*b* colour space lacks some sensitivity over monochromatic dark regions close to black, CIE L*u*v* colour space is shown to be more sensitive. We want to extract saliency maps from images following those attentional principles representing the neuronal activities in V1 towards those areas in the brain responsible for the eye movements. In this regard, we set up a saliency extraction method which is based on the multiscale analysis of keypoint density maps over CIE L*u*v* colour space. Then we combine the saliency maps out of each channel into the final saliency map with polynomial regression trained over 100 images randomly picked up from five different eye-tracking datasets.

It is observed that most of the recent saliency detection methods are based on deep learning approaches. They achieve high accuracy levels in detecting saliency maps but need high-performance systems for training and testing steps as well as a considerable number of images and the corresponding eye-fixation data. We want to point out that, other than deep learning methods, we just need only a subset of eye-fixation point data to train our architecture. We do not need any optimization step because the so-called bottom-up attention processes inspire our technique. Our contributions in this work are as follows: a new saliency method based on the multiscale analysis of keypoint density maps over CIE L*u*v* colour space; an extended analysis of the colour impact on the saliency extraction; a comparison study between our method and popular methods in the state-of-the-art on different eye-tracking datasets; a study case showing relationships between saliency and colour on object-based and generic image datasets.

II. RELATED WORKS

In this section, we want to report a list of state-of-the-art methods which are focused on visual saliency and taking into account both perceptual and computational sides. Furthermore, we provide the paper with a subsection focused on eye-tracking datasets we employed to conduct our experimental sessions.

A. RELATED TECHNIQUES

In the first section of this paper we report saliency methods to be divided into three main groups because of the inspirational visual attention principles: bottom-up, top-down, hybrid [23]. Nowadays, a further classification of saliency techniques can be observed in the scientific literature concerning different image processing, computer vision and artificial intelligence approaches. In this regard, most of the approaches in the topic of visual saliency can be divide into traditional approaches-based and deep learning based saliency detection methods.

1) TRADITIONAL APPROACHES-BASED SALIENCY DETECTION

The traditional approaches-based Saliency Detection methods are meant to be all methods whose computation mechanisms fall within the following models, Bayesian models [24], Cognitive models [25], Decision theoretic models [26], Spectral analysis models [27], Graphical models [28], Information theory models [29], Learning-based models (supervised learning and unsupervised learning) [30], [31]. The grouping of methods as above was first proposed in [32] and categorises saliency extraction techniques into seven groups. Itti *et al.* [8] proposed a bottom-up approach based on multi-scale analysis of the image. First, multi-scale image features are used to create a topographical saliency map. Then, a dynamical neural network selects the attended salient locations. The principle of centre-surround difference is adopted by Koch and Ullman [33] for the parallel extraction of different feature maps. Harel *et al.* [9] proposed a saliency method (well known as GBVS) based on a biologically plausible graph-based model. The leading models of visual saliency may be organized into three stages: extraction, activation, normalization. Wang *et al.* [34] surveyed the corresponding literature on the low-level methods for visual saliency. It is needed to mention the studies conducted by Bruce and Tsotsos [35] on the saliency, visual attention and visual search processes. A practical method for visual saliency detection based on multi-scale and multi-channel mean has been proposed in [36]. The method applies a two-step approach based on the image decomposition and reconstruction with the wavelet transform. Then a bicubic interpolation algorithm is applied to narrow the filtered image in multi-scale. The saliency values are the distances between the narrowed images and the means of their channels. We employed interest points such as SIFT keypoints to extract saliency maps and texture scale [37]–[40]. [41] proposed a low-resolution saliency estimate based on random colour sampling. The technique presented in [42] integrate two saliency maps computed with object proposals and motion-dominated methods, to obtain a spatio-temporal saliency map. A saliency detection method based on a Kalman filter is proposed in [43], inspired by biological phenomena such as the visual surprise and the saccadic eye movement [44]. By adopting the same approach proposed in [8], each features channel is individually represented with a generated saliency map by using the Kalman filter. All of them are then combined in a final map. Kalinin *et al.* [45] approached the problem of localization of the most informative regions in images checking the similarity of those regions with the response of foveal filters. In cognitive top-down approaches like those proposed in [10] and [46], the visual attention process is considered task-dependent. The observer's will and expectations play a critical role in determining why a point is fixed rather than others. Yang and Yang [47] performed saliency detection with a top-down model that jointly learns a Conditional Random Field (CRF) and a visual dictionary. Kanan *et al.* [48] adopted the SUN framework to detect the salient regions of an image by using global features and top-down components.

Generally, hybrid systems for saliency use the combination of bottom-up and top-down stimuli. In many hybrid approaches [49], a top-down layer is used to refine noisy maps extracted by bottom-up layers. A well known state-of-the-art hybrid approach was proposed by Judd *et al.* [30] in addition to a database [50] of eye-tracking data from 15 viewers. Low, middle and high-level features of the eye-tracking data have been used to train a model of saliency. Eye-tracking methodology is widely used for tasks such as Human-Computer Interaction [51], advertising evaluation [52] and different applications [53]–[55]. Generally speaking, saliency approaches are based on several properties, features and notions belonging to psychology, computer vision, neuroscience, biology and medicine. Yu *et al.* [56] used a paradigm based on the Gestalt grouping cues for object-based saliency detection. Chang *et al.* [57] proposed a method based on a graphical model of the relationships between saliency and objectness. Some methods such as in [58] and [59] approach the object detection using image descriptors and the relative orientation of the camera. Pflüger *et al.* [60] managed to build a method based on the optimisation of rectangle features with Adaboost to simulate eye-movements when one looks at visual artworks. Krejtz *et al.* [61] focused their efforts on quantifying the eye movement transitions between different areas of interest using Shannon's entropy and Markov chains. Toet [62] reported a comparative study that evaluates the performances of 13 state-of-the-art saliency models; and a new metric is also proposed and compared with previous models.

2) DEEP NEURAL NETWORK BASED SALIENCY TECHNIQUES

Over the last few years, because of their success in recognition and classification tasks, many researchers approached visual saliency by adopting deep learning techniques [63] and [64]. Deep learning methods allow extracting simple and complex structures in large datasets using the backpropagation algorithm to tune the representation parameters of deep neural networks. For the reasons mentioned above, visual saliency models can be learned using DCNNs (Deep Convolutional Neural Networks). For instance, Li and Yu [65] introduced a neural network architecture, containing a CNN with fully connected layers responsible for the feature extraction at different scales. Das *et al.* [66] tried to answer the question: do humans and deep networks look at the same regions? For this purpose, they conducted a qualitative and quantitative comparison of the maps generated with the state-of-the-art attention-based models and a task-independent saliency baseline. A fully convolutional neural network was adopted to extract the most salient regions in the method proposed by Kruthiventi *et al.* [67]. Zhao *et al.* [68] offered a survey on deep learning methods for object detection using visual saliency approaches. Obeso *et al.* [69] used a saliency-driven approach to predict visual attention in images and use it to train a Deep Convolutional Neural Network. Huang *et al.* [70] proposed a method based on the adaptation of deep neural networks. Apart from the fully convolutional neural networks, generative adversarial networks are used by Pan *et al.* [71] to predict visual saliency. Other than in [67] and [71], Liu and Han [72] adopted a different deep learning paradigm called long-term recurrent convolutional network for the saliency detection. Cornia *et al.* [73] proposed a saliency map method based on deep learning technique achieving excellent results for the accuracy metrics. The saliency prediction incorporates a network focusing on appropriate locations of the image to refine saliency features. Klein and Frintrop [74] employed the centre-surround difference as a means to detect saliency maps from pictures. Kümmerer *et al.* [75] proposed the so-called DeepGaze model to predict where people look in images. Other than other saliency models that use deep learning techniques, the authors did not employ additional fine-tuning, the saliency extraction relies on a transfer learning over the last layers of the network stack. They also [76] gave a detailed description of the high and low-level contributions to fixation points during the visual attention processes, which are necessary to build up a model for the visual saliency. Niu *et al.* [77] explored the connections of visual salience and emotional salience on eye movement behaviour to evaluate their influence on gaze allocation in scene viewing. Liu *et al.* [78] proposed a pixel-wise contextual attention network (PiCANet) which learns to attend to informative context locations for each pixel selectively. PiCANet has also been incorporated with the U-Net architecture to detect salient objects. A saliency detection framework using multi-cue and high-level differences is proposed in [79]. In greater detail, visual information is grouped into multi-cues vectors to discard the non-salient regions and highlight the salient area. Zeng *et al.* [80] proposed a unified framework to train saliency detection models with diverse weak supervision sources to overcome issues characterizing saliency maps achieved with a single weak supervision source. Other very popular methods in [81], [82] and [83] are based on unsupervised approaches. Eye Movement predictions in 360 Degree Images has emerged lately as a new hot topic in the field of visual saliency, Zhu *et al.* [84] employ spherical harmonics to extract features at different frequency bands and orientations to detect the rare components in the frequency and colour domain. Some 2D visual saliency approaches and methods can be integrated to explore and predict head and eye movement for 360 degree images [85].

B. EYE-TRACKING DATASETS

Eye-tracking technology offers a direct measure of visual attention by recording two kinds of eye movement, fixations and saccades [86]. Eye-tracking is widely used in many different tasks such as the analysis of user behaviour in marketing, advertising effectiveness evaluation, neuroscience, human-computer interaction, gaming, medicine, visualization research and other related disciplines [87], [88]. Fixations indicate where the observer looks, while saccades are movements between two fixations. Saccades and fixations together form the scan path whose data are used to show which regions of an image catch the observer attention. Scan paths can also be considered as key to estimations of cognitive processes during the so-called "free viewing" and task-oriented observations [89].

Eye-tracking technology has improved through the years with the introduction of more accurate instruments and more reliable equipment. Eye-tracking methods can be grouped into four generations [86]:

- First generation - Eye-in-head measurement of the eye consisting of techniques such as scleral contact lens/search coil, electro-oculography
- Second generation - Photo and video-oculography
- Third generation - Analog video-based combined pupil/corneal reflection
- Fourth generation - Digital video-based combined pupil/corneal reflection, augmented by computer vision techniques and digital signal processors (DSPs).

The eye-trackers of the fourth generation that have recently appeared on the market make use of digital optics.

Eye-tracking technology improved its performance in usability, accuracy, and speed by equipping trackers with on-chip Digital Signal Processors (DSPs).

Some recent progress in the eye-tracking technology allowed scientific researchers to collect large quantities of eye-gaze data for different purposes.

Bappy *et al.* [90] collected a dataset named EyeCrowd by recording the eye movements of 16 subjects watching images with various levels of crowd. Judd *et al.* [30] created the MIT dataset, composed of 1003 landscape and portrait images and the corresponding fixation point maps, collected during a free viewing session.

Some recent datasets, such as those collected in [91] and [92], are respectively focused on domains like visual saliency in low-resolution images and web pages.

Ramanathan *et al.* [93] focused on the eye-tracking of face, portrait, nude, action, affect-variant groups giving rise to a dataset composed of 99 fixation point maps. The datasets above are collected during free viewing sessions, that is, the subjects were not assigned any specific visual task when watching. On the other hand, task-driven eye-tracking data are found in [94].

A very popular fixation point dataset containing 2000 images from 20 different categories has been proposed by [95]. Some research groups focused their efforts on gathering eye-movement data from cohort of people affected by some cognitive disorders such as ASD (Authism Spectrum Disorder) [96]–[98].

Along with new topics as the prediction of head and eye-movements in 360-degree images and the recent emerging augmented reality (AR) saliency technique, new data have been gathered to assess the effectiveness of saliency approaches [85], [99].

In our previous work [40] we used an eye-tracker to record the gaze path of 24 observers while viewing each image from a subset of OPED (Object Pose Estimation Database) [100], [101]. It gave rise to a new dataset provided with fixation point maps we made publicly available under the name of ETTO (Eye Tracking Through Objects) [11]. The primary purpose of ETTO is to investigate the relationships between saliency and object visual attention processes. Each image was shown at full resolution for three seconds, separated by one second of viewing a grey screen (we adopted the same experimental approach as in [30]). The database collected in [100], [101] consists of several images with single objects in the foreground and a homogeneous coloured background region. Still, any dataset with a single main object (target) and a limited number of distractors in each image would have been appropriate as well. During the experimental session, viewers sat at a distance of 70 cm from a 22-inch computer screen with a resolution of 1920 x 1080 pixels. ETTO has been used here to assess the effectiveness of saliency methods based on different computational and perceptual approaches concerning the object attention process. Here ETTO is also used to assess the perceptually uniform space colours as CIE L*a*b* and CIE L*u*v*.

III. METHODS

In this section, we report our findings and studies with some methodological novelties based on colour features and new multiscale inspection. Here we would like to highlight that we develop methods for the extraction of saliency with pattern recognition and image processing techniques taking into account biologically inspired principles. The interest points we mentioned in the introduction section are extracted near spikes, borders, edges, contours of objects in the image and go under the name of keypoints. We take inspiration by the bottom-up visual attention processes about the impact of colour, contrast, texture and multiscale local keypoints on the generation of an early saliency map on the visual cortex V1. In the next two subsections we give a more detailed description of the most important steps of our proposed method, that is, the extraction of KDMs (Keypoint Density Maps), the Training of Polynomial Regression and extraction of Saliency Maps.

A. EXTRACTION OF KEYPOINT DENSITY MAPS

The overall architecture of our new proposed saliency extraction is based on the extension of our Keypoint Density Maps (KDMs) [37], [39], [40]. Our method projects KDMs onto a more biologically inspired multiscale architecture. Furthermore, we combine it with a centre bias module and a polynomial regression (see Fig. 1). The saliency map is then extracted using CIE L*a*b* and CIE L*u*v* space colour to assess the impact of colour on visual saliency.

We start this section by describing the full scheme behind KDMs as we already proposed [39]. The Keypoint Density Map $KDM_{c,k}$ is built by counting the number of keypoints inside a sliding window of varying size $k = 1, 2, \ldots, k_{\max}$ in each colour channel c. Given that $m_{c,k}$ is the "mode" that is the most frequent value in $KDM_{c,k}$, we define the mode vector $MV_c = [m_{c,1}, \ldots, m_{c,k_{\max}}]$ for all the KDMs computed for channel c.

In the past, to improve the effectiveness of our algorithm, we added colour information [40] from two colour spaces such as HSV, and CIE L*a*b*. Experimental evidence showed several drawbacks of HSV. We did not notice the improvements we expected by adding HSV colour information. Scientists noticed that HSV is not biologically inspired and does not take into account the colour opponent process [102].

The higher performance achieved through the use of CIE L*a*b* space prompted us to undertake a more in-depth investigation on perceptually uniform colour spaces (CIE L*a*b*, CIE L*u*v*) to assess the effectiveness of their employments in the eye-movement predictions and spot any differences among them in their contribution to the visual saliency detection.

In this work, we compare the performances of SIFT and SURF keypoint detectors when extracting saliency maps with KDMs method. Furthermore, we adopted two perceptually uniform colour spaces (CIE L*a*b* and CIE L*u*v*). We also improved the KDM generation and smoothing by adding a circular mask to the sliding window, and a Gaussian filter in place of the arithmetic mean filter. A multi-scale approach was implemented and will be explained below. As a consequence, in the following equations c can be considered as taking values in either $\{L, u^*, v^*\}$ or $\{L, a^*, b^*\}$.

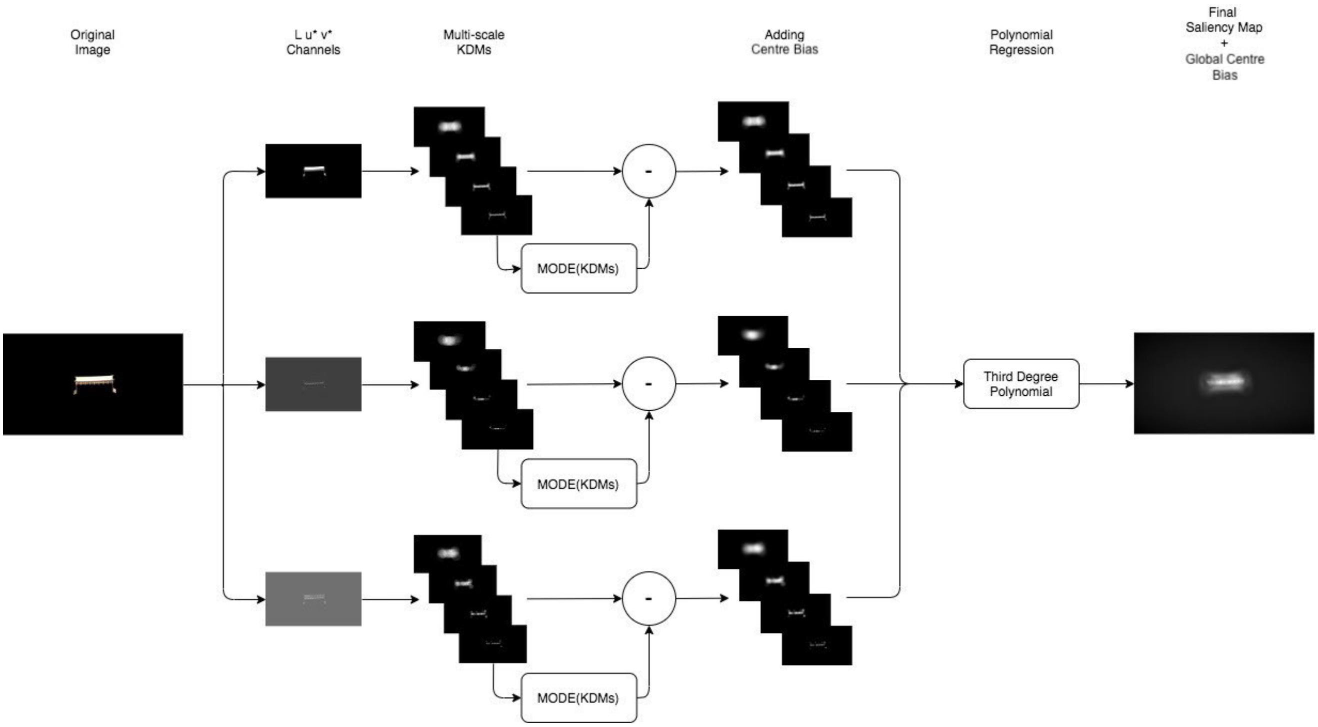

FIGURE 1. **Processing steps of our algorithm for an image in the ETTO data set.**

Local keypoints are extracted on each channel of the colour space, then KDMs are computed as a measurement of the spatial distribution of keypoints over the entire image.

KDM algorithm needs a scale factor to work: this parameter is a function of the size of the image [40].

The spatial distribution of keypoints gives a measure of the behaviour of texture inside the image. Texture allows to read the response of the image concerning several features such as contrast, scale, orientation, edges, object boundaries. According to our method, we extract salient regions by emphasizing rare events in textured areas. One of the basic concepts is that the spatial distribution of keypoints inside an image allows for describing texture variations all over the image. In our previous works, we applied this principle by using SIFT keypoints only on grayscale images (the standard SIFT algorithm [103] is suitable for grayscale images only). Encouraging results prompted us to approach a method including colour information because of its importance in cognitive terms. As mentioned above, in our previous release [40] we tried to use HSV space colour, which did not allow us to achieve the expected result improvements in both effectiveness and accuracy. We focused on the extraction of saliency maps by using KDMs in all the three channels of two perceptually uniform colour spaces: CIE L*a*b* and CIE L*u*v*.

We extended the KDMs, used to detect the texture scale in regular and near regular textures, to the extraction of the saliency maps of an image by applying them to all of the three colour channels. We remark that the spatial distribution of keypoints inside images can be used in the texture variation description. Levels of the image roughness in both fine and coarse-textured regions can be very different. In a fine-textured area, it is expected to find a more significant number of keypoints than in coarse regions. Here we use keypoint density maps to identify the most salient areas. Once the keypoints are extracted in each colour channel c, four KDMs at different scales k are calculated for each channel by selecting a "main" scale factor, as in (1); the other three k are selected by simply subtracting 1, 2, 3 to the main k value. If $k = 1$ is reached before computing all the four maps, no more maps are calculated for that channel.

$$k = 2^{\left\lfloor \log_2 \left(\frac{\min(M,\ N)}{4} \right) \right\rfloor} \tag{1}$$

Here, M and N are the image dimensions.

B. POLYNOMIAL REGRESSION AND SALIENCY MAP

For each KDM, a Saliency Map $SM_{c,k}$ is computed at each scale k for each colour channel c by taking the point-wise difference between the $KDM c, k$ and its mode value $MV_c[k] \equiv m_{c,k}$.

$$SM_{c,k}(x, y) = \left| KDM_{c,k}(x, y) - MV_c[k] \right| \tag{2}$$

Each SM obtained using (2) is multiplied with a centre bias map (Fig. 1); we call $SM^{cb}_{c,k}$ the SM at scale k for each channel c after the application of the centre bias.

Inspired by the fact that human gaze patterns follow a normal distribution in the natural viewing condition [104] we combined the saliency maps out of the keypoint density maps by using polynomial regression in the 2D subspace. In order, the twelve saliency maps (3 channels × 4 scales) are passed to a third-degree polynomial without mixed terms to perform pixel-by-pixel multi-channel and multi-scale fusion, as shown in (3).

$$SM(x, y) = a_0 + \sum_{c} \sum_{k=1}^{4} \sum_{l=1}^{3} a_{c,k,l} \left[SM_{c,k}^{cb}(x, y) \right]^{l} \tag{3}$$

We also added a global centre bias to tackle the human-gaze centre bias, that is, a simple Gaussian blob centred in the middle, as described by [30].

IV. EXPERIMENTAL RESULTS

In this section, we compare our saliency methods to both traditional and more modern techniques based on deep learning, using our dataset and generic image data sets. We show the performance of our approach with various image types enclosed in generic data sets, as emotional photos, landscapes, hand-drawn sketches, etc. The visual attention data sets we used for this study, all accompanied by eye-tracked fixation locations, are the following:

- ETTO (Eye-Tracking Through Objects - our data set) [40]
- MIT1003 [30]
- MIT CAT2000 (limited to the fixation maps of only 18 users) [95]
- FIGRIM Fixation Dataset [105]
- EyeCrowd [106].

The saliency algorithms used for comparison are the following:

- Our new multi-scale CIE L*u*v* SURF-based method
- Our legacy CIE L*a*b* SIFT-based method [40]
- Itti-Koch-Niebur [8]
- GBVS [9]
- ConvLSTM-based Saliency Attentive Model with a VGG-16 network (SAM-VGG) [73]
- ConvLSTM-based Saliency Attentive Model with a ResNet-50 network (SAM-ResNet) [73]
- Ensembles of Deep Networks (eDN) [31].

We also report the performance of a fixed centred Gaussian distribution as a baseline and choose Normalized scan path Saliency (NSS) as our comparison metric. NSS is well balanced and binarization-independent [1]; we report other metrics [107] to show the improvements and the effectiveness of our method better.

The Itti-Koch-Niebur algorithm we chose for comparison is an enhanced version released with the GBVS Toolbox [108], which also contains the official GBVS code. The reference implementations of ConvLSTM-based and eDN models have been downloaded respectively from [109] and [110]. All the algorithms reported above have been used with their default parameters. As explained in the previous section, the main scale factor k in our method is computed according to equation (1), while the coefficients of the polynomial in equation (3) are learned through the regression.

The whole algorithm has been implemented in MathWorks MATLAB, the polyfitn toolbox [111] has been used to perform the regression and evaluate the polynomial. All the saliency maps have been calculated on an Intel Core i7-4770 computer with four cores (8 threads) and 16 GB of RAM without the use of GPU computing. On average, the execution time required to calculate the saliency map of a 1920 × 1080 RGB image is 15.4071 s. The algorithm runs as a single thread without any GPU assistance, and it could be speeded up taking advantage of faster mathematical libraries, multiprocessing and GPU computation).

Our main research goal was not to create a robust saliency detector. Rather, we were aimed at investigating the impact of perceptually uniform colour spaces such as CIE L*a*b* and CIE L*u*v* on the extraction of saliency maps. So we deliberately stressed our method using both a small training set and reduced computational power. We used all the five ground truth data sets composed of real fixation maps already presented in this section. Our training set is made by 100 images that is 20 randomly selected samples from each data set, while tests have been carried out on the entire data sets. A 10-fold cross-validation was used for training, and the best model has been selected. During training, the multi-scale maps have been pre-weighted with a centre bias map and scaled to the same spatial resolution (512 x 512) before passing them as an input to the polynomial. The only hyper-parameter in our regression model is the degree of the polynomial. We adopted a greedy approach in this respect, trying degrees from 3 to 5. As it was expected, both degree 4 and 5 produced overfitting, and we resorted to a third degree polynomial.

As reported in Tables 1, 2, 3, 4, 5, our method always got excellent NSS results when compared to traditional unsupervised methods on MIT1003, CAT2000 and EyeCrowd datasets, and exceeded Ensembles of Deep Networks' NSS on the same datasets; comparable results as the other saliency methods have been achieved on ETTO dataset.

We measured the performance of our legacy algorithm on the sample dataset in both CIE L*a*b* and CIE L*u*v* spaces, with and without circular masking and Gaussian filtering. Our method on CIE L*u*v* colour space + circular masking + Gaussian filtering (see Fig. 2) reached the best results; therefore subsequent improvements were only tested on this method. More in detail, the CIE L*u*v* + SURF + centre bias algorithm performed equally or better than other variations on the sample dataset because it manages to detect salient features with a fewer number of keypoints (Fig. 2).

The performances of "mixed" methods (CIE L*a*b* + SURF and CIE L*u*v* + SIFT) have not been reported in the tables because they were always lower than those of the method mentioned above.

In the light of the results obtained in our experiments, we recall some concepts from the scientific literature concerning CIE L*a*b* and CIE L*u*v* colour. It is well-known [22] that CIE L*a*b* mainly focuses on differences and common perceptual descriptors of colours.

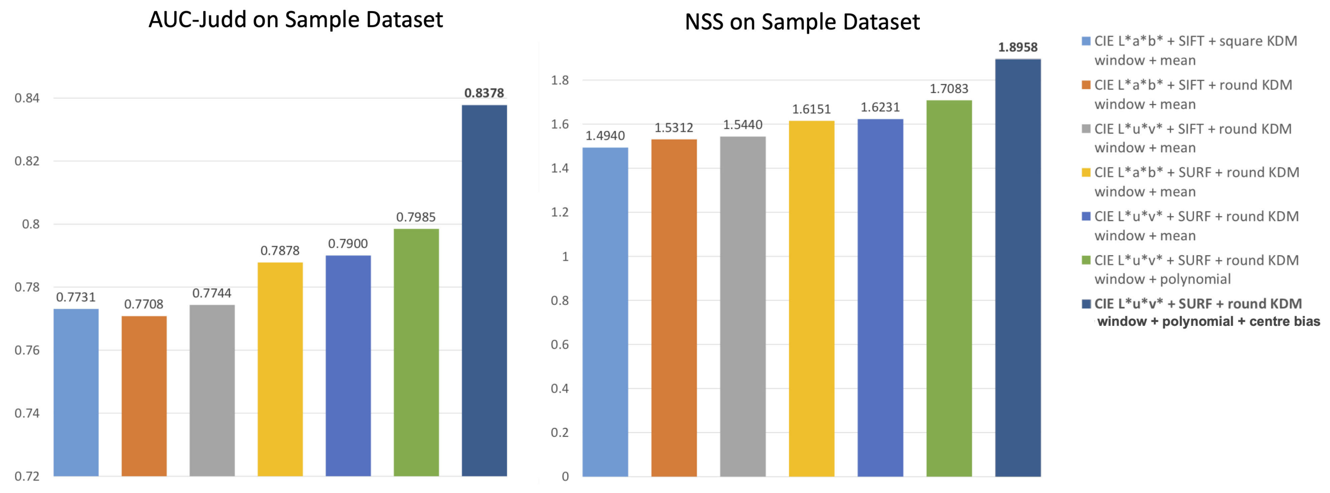

FIGURE 2. Performance graphs of different versions of our saliency model on the sample dataset.

TABLE 1. Results in various metrics on ETTO dataset.

	Saliency method	AUC-Borji	AUC-Judd	CC	KL[1]	NSS	SIM
Traditional	**Our new method**	**0.9363**	**0.9518**	0.6844	0.8904	**4.4399**	0.4993
	Our previous method	0.8253	0.9208	0.6982	2.0256	4.1197	**0.5987**
	Itti-Koch-Niebur	0.9296	0.9257	0.4670	2.0465	3.1424	0.2050
	GBVS	0.9344	0.9511	**0.7158**	**0.8468**	4.7342	0.5062
	Gaussian centre	0.8807	0.9356	0.2080	3.2756	1.3563	0.0726
Deep Learning	eDN	**0.9494**	**0.9556**	0.4171	2.3551	2.8225	0.1480
	SAM-VGG	0.9089	0.9520	0.7061	0.7705	4.8767	0.5685
	SAM-ResNet	0.9164	0.9534	**0.7814**	**0.6412**	**4.9985**	**0.6134**

[1] Lower is better.

TABLE 2. Results in various metrics on MIT1003 dataset.

	Saliency method	AUC-Borji	AUC-Judd	CC	KL[1]	NSS	SIM
Traditional	**Our new method**	0.7953	0.8151	0.4169	1.3739	**1.4071**	0.3519
	Our previous method	0.7304	0.7459	0.2971	1.5973	1.0065	0.3206
	Itti-Koch-Niebur	0.7623	0.7750	0.3307	1.4805	1.1029	0.3226
	GBVS	0.8151	**0.8288**	**0.4175**	**1.2969**	1.3819	**0.3627**
	Gaussian centre	**0.7988**	0.8162	0.3288	1.6990	1.0089	0.2691
Deep Learning	eDN	**0.8471**	0.8579	0.4096	1.5453	1.2969	0.2976
	SAM-VGG	0.8409	0.9197	0.8322	0.7745	3.1733	0.6733
	SAM-ResNet	0.8442	**0.9267**	**0.8689**	**0.7363**	**3.3384**	**0.7126**

[1] Lower is better.

TABLE 3. Results in various metrics on CAT2000 dataset.

	Saliency method	AUC-Borji	AUC-Judd	CC	KL[1]	NSS	SIM
Traditional	**Our new method**	0.7800	0.8005	**0.4946**	1.0046	**1.2644**	**0.5073**
	Our previous method	0.7536	0.7660	0.4157	1.1018	1.0658	0.4850
	Itti-Koch-Niebur	0.7578	0.7667	0.4131	0.9715	1.0625	0.4647
	GBVS	**0.7900**	0.8012	0.4864	**0.8615**	1.2458	0.4982
	Gaussian centre	0.8134	**0.8363**	0.4504	1.1785	1.0805	0.4021
Deep Learning	eDN	**0.8426**	**0.8510**	0.4933	**0.9990**	1.2145	0.4477
	SAM-VGG	0.7516	0.8394	0.6280	1.1121	1.7642	0.5764
	SAM-ResNet	0.7619	0.8438	**0.6550**	1.1586	**1.8256**	**0.5910**

[1] Lower is better.

CIE L*a*b* is considered to deal with perceptual colour differences from a numerical perspective. Some follow-ups in colourimetry occurred on CIE L*a*b* shortcomings that mainly affects the colour description in the dark regions of colour space.

In this regard, Sharma and Rodríguez-Pardo [22] had already focused their attention on what they called the dark side of CIE L*a*b*, which means to highlight some of the CIE L*a*b* limitations over the perception of colour. Sharma and Rodríguez-Pardo [22] conducted some

TABLE 4. Results in various metrics on EyeCrowd dataset.

	Saliency method	AUC-Borji	AUC-Judd	CC	KL[1]	NSS	SIM
Traditional	**Our new method**	**0.7737**	**0.7842**	**0.4959**	**0.7852**	**1.1020**	**0.5144**
	Our previous method	0.7251	0.7333	0.3745	0.9294	0.8324	0.4742
	Itti-Koch-Niebur	0.6544	0.6993	0.2786	1.0463	0.5843	0.4442
	GBVS	0.7562	0.7641	0.4444	0.8420	0.9671	0.4940
	Gaussian centre	0.7285	0.7373	0.3770	1.0631	0.7788	0.4242
Deep Learning	eDN	**0.7832**	0.7902	0.4709	**0.9517**	1.0068	0.4552
	SAM-VGG	0.7409	0.8429	0.7064	1.5619	1.8898	0.6397
	SAM-ResNet	0.7495	**0.8466**	**0.7242**	1.4527	**1.9682**	**0.6533**

[1] Lower is better.

TABLE 5. Results in various metrics on FIGRIM dataset.

	Saliency method	AUC-Borji	AUC-Judd	CC	KL[1]	NSS	SIM
Traditional	**Our new method**	0.7794	0.8022	0.4315	1.1692	1.2567	0.4065
	Our previous method	0.7094	0.7270	0.2932	1.5519	0.8601	0.3718
	Itti-Koch-Niebur	0.6700	0.7470	0.2820	1.3278	0.8220	0.3717
	GBVS	0.8076	0.8197	**0.4470**	**1.0519**	**1.2599**	**0.4300**
	Gaussian centre	**0.8297**	**0.8577**	0.4291	1.3717	1.1409	0.3435
Deep Learning	eDN	**0.8621**	**0.8721**	0.5031	1.2348	1.3795	0.3735
	SAM-VGG	0.7739	0.8699	0.6166	**1.1518**	2.0299	0.5373
	SAM-ResNet	0.7835	0.8696	**0.6242**	1.1524	**2.0538**	**0.5397**

[1] Lower is better.

experiments showing that for low lightness values (L*), corresponding to dark colour regions, perceptual differences are well represented by the CIE L*a*b* colour space. Chroma values turned out not to be computed as different even when wider bandwidths are subtracted from the Lightness values exhibiting inconsistent behaviour from a perceptual perspective. It would suggest the colour space might not to be reliable for the representation of the HVS perception of colourful regions in some cases. On the other hand, the CIE L*u*v* colour space seems not to suffer from the problems noticed in [22]. Colour differences in the perceptual domain between CIE L*a*b* and CIE L*u*v* are shown in [22] analyzing the take-off angle. It is defined as the angle comprised between the Lightness axis and the direction of a small monochromatic stimulus. CIE L*a*b* dark side is shown by take-off angles approaching 90-degrees for wavelengths in the region from 400 to 450 nm. Conversely, CIE L*u*v* performs better when close to black, with none of the take-off angles approaching 90-degrees. The reader who may be interested in the mathematical treatment is remanded to the original cited article in [22]. The mentioned shortcoming of CIE L*a*b* might be a reasonable explanation for the better performances of the CIE L*u*v* colour space along the edges and boundaries of single objects on a dark background like those in ETTO.

We ran many trials and experiments within our research using different kinds of images. Some of the pictures include centred objects surrounded by dark regions (ETTO dataset). We highlight that our technique is a bottom-up visual saliency approach. It runs better on object images because it tends to reproduce the bottom-up attentional mechanisms. The mentioned mechanism postulated in [3], is designed to respond to areas of high contrast and will frequently select image regions that correspond to objects. Good evidence of this can be found in the experimental results shown in Fig. 3 and Fig. 4. The sample images for the best and worst saliency maps per each dataset are shown. The best results of our method are achieved over images with a single object in the foreground or when a single visual instance sticks out from the background. On the other side, all images with multiple visual instances (see Fig. 4) or more than one object in the foreground do better suit supervised methods. However, it can be observed from tables 1, 2, 3, 4, 5 and in Fig. 2 that our method's overall performances are consistent all over several datasets and different principle-based approaches.

V. DISCUSSION

In this paper, we reported our findings and research on visual saliency, introducing some methodological novelties concerning our previous works. We also studied the impact of perceptually uniform colour spaces such as CIE L*a*b* and CIE L*u*v* in the extraction of saliency maps. We tackled visual saliency using some biologically inspired principles like the difference centre-surround, the multiscale analysis and the human-gaze centre bias. Inspired by human gaze patterns following a normal distribution in the natural viewing condition, we also used polynomial regression in the 2D subspace to combine the saliency maps obtained at different scales. As it is observed in tables 1, 2, 3, 4 and 5 we compared the performance of our proposed method against several state-of-the-art methods, which are based on different approaches and principles. Since it is also our interest to study the different performances of saliency methods over several eye-tracking datasets, we collected several experiments from the most popular datasets such as MIT1003, CAT2000, EyeCrowd, FIGRIM, as well as our dataset, ETTO. It is

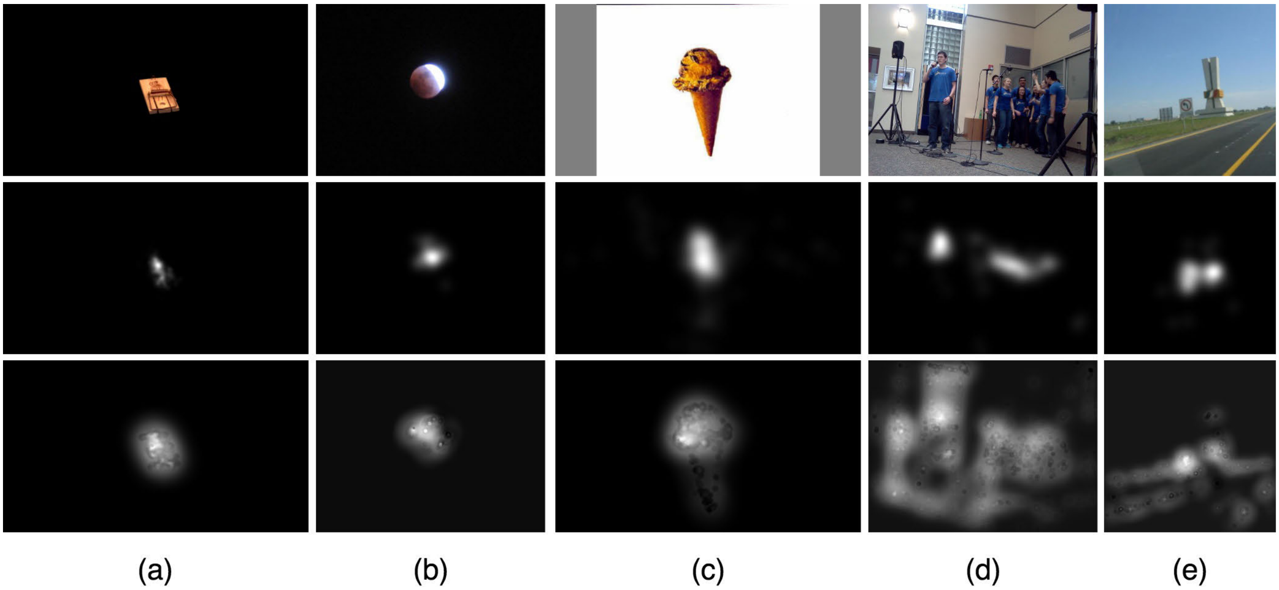

FIGURE 3. Best results per data set. First row: original images; second row: fixation maps; third row: our saliency maps. (a) ETTO: MouseTrapBlack__Curve_030_Rot_010_fhd.png, (b) MIT1003: i2281902585.jpeg, (c) CAT2000: Object\017.jpg, (d) EyeCrowd: 498.jpg, (e) FIGRIM: sun_bcunorujjxrvnqkx.jpg.

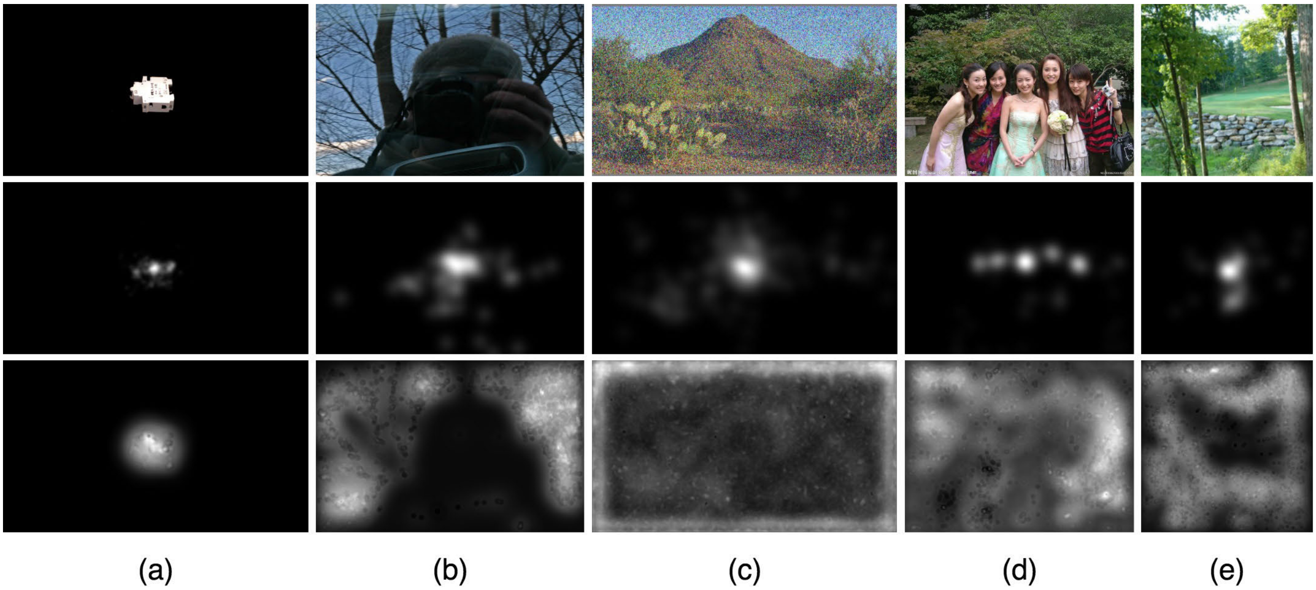

FIGURE 4. Worst results per data set. First row: original images; second row: fixation maps; third row: our saliency maps. (a) ETTO: AutoFuseBlack__Curve_030_Rot_000_fhd.png, (b) MIT1003: i2278136983.jpeg, (c) CAT2000: Noisy\127.jpg, (d) EyeCrowd: 236.jpg, (e) FIGRIM: sun_aokgqlnxzonxwzlk.jpg.

worth to mention that ETTO is a collection of fixation points on images where single objects are in the foreground. The above-mentioned step brought us to evaluate differences among many state-of-the-art methods comparing saliency maps and fixation points. Besides, the experiments with several saliency methods on ETTO images allowed us to highlight the differences among different approaches concerning the object attention process. To show how accurate is the detection of the most important regions from a perceptual viewpoint, we used a method based on pattern recognition and image processing techniques. We compared the results of our approach applied to the datasets mentioned above and against the most popular methods belonging to different classes (deep learning, bottom-up, supervised, unsupervised). Some other considerations can be drawn by looking at the experimental results. Our method overcomes our previous work on all datasets we used; it reaches the best performances over ETTO, which contains object images with a dark background. We guess the reason behind the best performances over ETTO is highly related to the impact of CIE L*u*v* space colour on the saliency extraction. It seems to be more robust, especially in cases of dark coloured regions (like ETTO images). Our experiments confirm some shortcomings of the CIE L*a*b* colour space in those image regions with low lightness values (L*), which correspond to dark background regions. Those regions are observed not to be well represented in the saliency maps in terms of perceptual differences. A further investigation of the temporal range of the eye-movements and their correlations to the different saliency approaches would allow for a better understanding of some necessary details. It mostly concerns the impact of some visual features from the images in the first 100 milliseconds of observation. The refinement of our experiments and comparisons along the temporal range might be useful to map the visual features of the images to the processes of visual attention both the exogenous and endogenous ones. Moving on to the artificial intelligence perspective of our work, we want to mention that only two deep learning methods (SAM-VGG and SAM-ResNet) outperform our method over the more generic image datasets (see tables 2, 3, 4, 5). All other methods' performances such as Itti-Koch-Niebur, eDN, GBVS and our previous method's release are lower than or equal to our current proposed method. Other than deep learning methods, we do not need thousands of eye-fixation point data to train our architecture and any optimization step. As a matter of facts, our technique is mainly inspired by the so-called bottom-up attention processes aiming to reproduce the attentional principles representing the neuronal activities in V1 towards those areas in the brain responsible for the eye movements. Our method is comparable to the Deep Learning ones on images with single objects in the foreground. We guess this aspect might deserve some more in-depth investigations to establish further connections between our approach and the object visual attention-related processes. Last, we believe it is still worth to investigate cues for visual saliency detection improvements in spite of the excellent accuracy achieved by many state-of-the-art methods. We want to focus on techniques which provide a well-balanced trade-off between overall good detection accuracy and acceptable hardware requirements. We have been working on the re-implementation of our novel method in Python, taking advantage of GPU computation frameworks like CUDA (supported by OpenCV). We expect the code rewrite to speed-up the algorithm's execution. That would make it feasible for on-the-fly intra-frame video saliency extraction when used in conjunction with state-of-the-art inter-frame saliency extraction techniques.

ACKNOWLEDGMENT

The authors thank Basilio Fabio Ceraolo for the assistance given on data acquisition, gathering and processing of the ETTO dataset during eye-tracking sessions. A special thanks to Caterina Ciaramitaro for her support during the final step of manuscript editing.

REFERENCES

[1] J. Li and W. Gao, *Visual Saliency Computation: A Machine Learning Perspective*, vol. 8408. Berlin, Germany: Springer, 2014.

[2] R. Snowden, R. J. Snowden, P. Thompson, and T. Troscianko, *Basic Vision: An Introduction to Visual Perception*. London, U.K.: Oxford Univ. Press, 2012. [Online]. Available: https://global.oup.com/ukhe/product/basic-vision-9780199572021

[3] D. Walther, U. Rutishauser, C. Koch, and P. Perona, "Selective visual attention enables learning and recognition of multiple objects in cluttered scenes," *Comput. Vis. Image Understand.*, vol. 100, nos. 1–2, pp. 41–63, Oct. 2005.

[4] J. Theeuwes, "Endogenous and exogenous control of visual selection," *Perception*, vol. 23, no. 4, pp. 429–440, Apr. 1994.

[5] M. M. Sohlberg and C. A. Mateer, *Introduction to Cognitive Rehabilitation: Theory and practice*. New York, NY, USA: Guilford Press, 1989.

[6] N. V. K. Medathati, H. Neumann, G. S. Masson, and P. Kornprobst, "Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision," *Comput. Vis. Image Understand.*, vol. 150, no. 5, pp. 1–30, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314216300339

[7] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," *ACM Trans. Appl. Perception*, vol. 7, no. 1, p. 6, 2010.

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[9] J. Harel, "Graph-based visual saliency," in *Proc. NIPS*, vol. 19, 2006, p. 5.

[10] J. Luo, "Subject content-based intelligent cropping of digital photos," in *Proc. IEEE Multimedia Expo Int. Conf.*, Jul. 2007, pp. 2218–2221.

[11] E. Ardizzone, A. Bruno, and F. Gugliuzza. (2017). *Eye-Tracking Through Objects (ETTO) Dataset*. [Online]. Available: https://github.com/fgugliuzza/saliency/tree/master/etto

[12] M. Matsukura and S. P. Vecera, "Interference between object-based attention and object-based memory," *Psychonomic Bull. Rev.*, vol. 16, no. 3, pp. 529–536, Jun. 2009.

[13] J. Theeuwes, S. Mathát, and A. Kingstone, "Object-based eye movements: The eyes prefer to stay within the same object," *Attention, Perception, Psychophys.*, vol. 72, no. 3, pp. 597–601, Apr. 2010.

[14] C. Stothart, D. J. Simons, W. R. Boot, and T. J. Wright, "What to where: The right attention set for the wrong location," *Perception*, vol. 2019, Oct. 2019, Art. no. 0301006619854302.

[15] G. K. and S. M. L., "The effect of color combination on visual attention and usability of multiple line graphs," *J. Commun. Inf. Sci.*, vol. 1, no. 1, pp. 11–21, Apr. 2011.

[16] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[17] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.

[18] H. P. Frey, C. Honey, and P. Konig, "What's color got to do with it? The influence of color on visual attention in different categories," *J. Vis.*, vol. 8, no. 14, p. 6, Oct. 2008.

[19] D. Berga and X. Otazu, "A neurodynamic model of saliency prediction in v1," 2018, *arXiv:1811.06308*. [Online]. Available: http://arxiv.org/abs/1811.06308

[20] H. S. Friedman, H. Zhou, and R. von der Heydt, "The coding of uniform colour figures in monkey visual cortex," *J. Physiol.*, vol. 548, no. 2, pp. 593–613, Apr. 2003.

[21] X. R. Fdez-Vidal, R. Rodriguez-Sanchez, J. A. Garcia, and J. Fdez-Valdivia, "Integral opponent-colors features for computing visual target distinctness," *Pattern Recognit.*, vol. 33, no. 7, pp. 1179–1198, Jul. 2000.

[22] G. Sharma and C. E. Rodríguez-Pardo, "The dark side of cielab," *Proc. SPIE Int. Soc. Opt. Photon.*, vol. 8292, Feb. 2012, Art. no. 82920D.

[23] K. Duncan and S. Sarkar, "Saliency in images and video: A brief survey," *IET Comput. Vis.*, vol. 6, no. 6, pp. 514–523, Nov. 2012.

[24] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, p. 32, Dec. 2008.

[25] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, no. 9, pp. 1395–1407, Nov. 2006.

[26] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," in *Proc. 17th Int. Conf. Neural Inf. Process. Syst.*, Cambridge, MA, USA: MIT Press, 2004, pp. 481–488. [Online]. Available: http://dl.acm.org/citation.cfm?id=2976040.2976101

[27] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 996–1010, Apr. 2013.

[28] A. A. Salah, E. Alpaydin, and L. Akarun, "A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 420–425, Mar. 2002.

[29] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao, "Simulating human saccadic scanpaths on natural images," in *Proc. CVPR*, Jun. 2011, pp. 441–448.

[30] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2106–2113.

[31] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2798–2805.

[32] A. Mahdi, M. Su, M. Schlesinger, and J. Qin, "A comparison study of saliency models for fixation prediction on infants and adults," *IEEE Trans. Cognit. Develop. Syst.*, vol. 10, no. 3, pp. 485–498, Sep. 2018.

[33] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," in *Matters of Intelligence*. Berlin, Germany: Springer, 1987, pp. 115–141.

[34] L. Wang, S.-L. Dong, H.-S. Li, and X.-B. Zhu, "A brief survey of low-level saliency detection," in *Proc. Int. Conf. Inf. Syst. Artif. Intell. (ISAI)*, Jun. 2016, pp. 590–593.

[35] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *J. Vis.*, vol. 9, no. 3, p. 5, Mar. 2009.

[36] L. Sun, Y. Tang, and H. Zhang, "Visual saliency detection based on multi-scale and multi-channel mean," *Multimedia Tools Appl.*, vol. 75, no. 1, pp. 667–684, Jan. 2016.

[37] E. Ardizzone, A. Bruno, and G. Mazzola, "Visual saliency by keypoints distribution analysis," in *Proc. Int. Conf. Image Anal. Process.* Berlin, Germany: Springer, 2011, pp. 691–699.

[38] E. Ardizzone, A. Bruno, and G. Mazzola, "Saliency based image cropping," in *Proc. Int. Conf. Image Anal. Process.* Berlin, Germany: Springer, 2013, pp. 773–782.

[39] E. Ardizzone, A. Bruno, and G. Mazzola, "Scale detection via keypoint density maps in regular or near-regular textures," *Pattern Recognit. Lett.*, vol. 34, no. 16, pp. 2071–2078, Dec. 2013.

[40] E. Ardizzone, A. Bruno, and F. Gugliuzza, "Exploiting visual saliency algorithms for object-based attention: A new color and scale-based approach," in *Proc. Int. Conf. Image Anal. Process.*, 2017, pp. 191–201.

[41] M. M. I. Lie, G. B. Borba, H. Vieira Neto, and H. R. Gamba, "Joint upsampling of random color distance maps for fast salient region detection," *Pattern Recognit. Lett.*, vol. 114, pp. 22–30, Oct. 2018.

[42] Z. Tu, Z. Guo, W. Xie, M. Yan, R. C. Veltkamp, B. Li, and J. Yuan, "Fusing disparate object signatures for salient object detection in video," *Pattern Recognit.*, vol. 72, pp. 285–299, Dec. 2017.

[43] S. Roy and P. Mitra, "Visual saliency detection: A Kalman filter based approach," 2016, *arXiv:1604.04825*. [Online]. Available: http://arxiv.org/abs/1604.04825

[44] N. Bigdely-Shamlo, A. Vankov, R. R. Ramirez, and S. Makeig, "Brain activity-based image classification from rapid serial visual presentation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 16, no. 5, pp. 432–441, Oct. 2008.

[45] K. B. Kalinin, M. V. Petrushan, and A. I. Samarin, "Relation of descriptor completeness and distinctiveness to localization of the most informative regions in images," *Perception*, vol. 44, nos. 8–9, pp. 1029–1039, Aug. 2015.

[46] V. Sundstedt, A. Chalmers, K. Cater, and K. Debattista, "Top-Down Visual Attention for Efficient Rendering of Task Related Scenes.," in *Proc. VMV*, 2004, pp. 209–216.

[47] J. Yang and M.-H. Yang, "Top-down visual saliency via joint CRF and dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 576–588, Mar. 2017.

[48] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell, "SUN: Top-down saliency using natural statistics," *Vis. Cognition*, vol. 17, nos. 6–7, pp. 979–1003, Aug. 2009.

[49] L.-Q. Chen, X. Xie, X. Fan, W.-Y. Ma, H.-J. Zhang, and H.-Q. Zhou, "A visual attention model for adapting images on small displays," *Multimedia Syst.*, vol. 9, no. 4, pp. 353–364, Oct. 2003.

[50] T. Judd, K. Ehinger, F. Durand, and A. Torralba. (2009). *Learning to Predict Where Humans Look*. [Online]. Available: http://people.csail.mit.edu/tjudd/WherePeopleLook/index.html

[51] V. Gentile, S. Sorce, A. Malizia, D. Pirrello, and A. Gentile, "Touchless interfaces for public displays: Can we deliver interface designers from introducing artificial push button gestures?" in *Proc. Int. Work. Conf. Adv. Vis. Interface*, New York, NY, USA, 2016, pp. 40–43, doi: 10.1145/2909132.2909282.

[52] G. Hervet, K. Guérard, S. Tremblay, and M. S. Chtourou, "Is banner blindness genuine? Eye tracking Internet text advertising," *Appl. Cognit. Psychol.*, vol. 25, no. 5, pp. 708–716, Sep. 2011.

[53] A. T. Duchowski, "A breadth-first survey of eye-tracking applications," *Behav. Res. Methods, Instrum., Comput.*, vol. 34, no. 4, pp. 455–470, Nov. 2002.

[54] S. A. Oliveira, S. S. Alves, J. P. Gomes, and A. R. R. Neto, "A bi-directional evaluation-based approach for image retargeting quality assessment," *Comput. Vis. Image Understand.*, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314217302035

[55] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. M. Snoek, "VideoLSTM convolves, attends and flows for action recognition," *Comput. Vis. Image Understand.*, vol. 166, pp. 41–50, Jan. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314217301741

[56] J.-G. Yu, G.-S. Xia, C. Gao, and A. Samal, "A computational model for object-based visual saliency: Spreading attention along gestalt cues," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 273–286, Feb. 2016.

[57] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 914–921.

[58] L. Czáni and M. Rashad, "The use of IMUs for video object retrieval in lightweight devices," *J. Vis. Commun. Image Represent.*, vol. 48, pp. 30–42, Oct. 2017.

[59] L. Czáni and M. Rashad, "Lightweight active object retrieval with weak classifiers," *Sensors*, vol. 18, no. 3, p. 801, Mar. 2018.

[60] H. Pfläger, B. Höferlin, M. Raschke, and T. Ertl, "Simulating fixations when looking at visual arts," *ACM Trans. Appl. Perception*, vol. 12, no. 3, pp. 1–20, Jul. 2015.

[61] K. Krejtz, A. Duchowski, T. Szmidt, I. Krejtz, F. González Perilli, A. Pires, A. Vilaro, and N. Villalobos, "Gaze transition entropy," *ACM Trans. Appl. Perception*, vol. 13, no. 1, pp. 1–20, Dec. 2015.

[62] A. Toet, "Computational versus psychophysical bottom-up image saliency: A comparative evaluation study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2131–2146, Nov. 2011.

[63] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[64] B. Sullivan, "Charniak, E. An introduction to deep learning," *Perception*, vol. 48, no. 8, pp. 759–761, Aug. 2019, doi: 10.1177/0301006619857273.

[65] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5012–5024, Nov. 2016.

[66] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" *Comput. Vis. Image Understand.*, vol. 163, pp. 90–100, Oct. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314217301649

[67] S. S. S. Kruthiventi, K. Ayush, and R. V. Babu, "DeepFix: A fully convolutional neural network for predicting human eye fixations," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4446–4456, Sep. 2017.

[68] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[69] A. M. Obeso, J. Benois-Pineau, M. S. G. Vázquez, and A. A. R. Acosta, "Saliency-based selection of visual content for deep convolutional neural networks," *Multimedia Tools Appl.*, vol. 78, no. 8, pp. 9553–9576, Apr. 2019.

[70] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 262–270.

[71] J. Pan, C. Canton Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i-Nieto, "SalGAN: Visual saliency prediction with generative adversarial networks," 2017, *arXiv:1701.01081*. [Online]. Available: http://arxiv.org/abs/1701.01081

[72] N. Liu and J. Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3264–3274, Jul. 2018.

[73] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an LSTM-based saliency attentive model," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5142–5154, Oct. 2018.

[74] D. A. Klein and S. Frintrop, "Center-surround divergence of feature statistics for salient object detection," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2214–2219.

[75] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "DeepGaze II: Reading fixations from deep features trained on object recognition," 2016, *arXiv:1610.01563*. [Online]. Available: http://arxiv.org/abs/1610.01563

[76] M. Kummerer, T. S. A. Wallis, L. A. Gatys, and M. Bethge, "Understanding Low- and high-level contributions to fixation prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4789–4798.

[77] Y. Niu, R. M. Todd, M. J. Kyan, and A. K. Anderson, "Visual and emotional salience influence eye movements," *TAP*, vol. 9, no. 3, pp. 1–13, 2012.

[78] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3089–3098.

[79] J. Wu, H. Yu, J. Sun, W. Qu, and Z. Cui, "Efficient visual saliency detection via multi-cues," *IEEE Access*, vol. 7, pp. 14728–14735, 2019.

[80] Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, M. Qian, and Y. Yu, "Multi-source weak supervision for saliency detection," 2019, *arXiv:1904.00566*. [Online]. Available: http://arxiv.org/abs/1904.00566

[81] S. Frintrop, T. Werner, and G. M. Garcia, "Traditional saliency reloaded: A good old model in new shape," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 82–90.

[82] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

[83] J. Zhang and S. Sclaroff, "Saliency detection: A Boolean map approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 153–160.

[84] Y. Zhu, G. Zhai, X. Min, and J. Zhou, "The prediction of saliency map for head and eye movements in 360 degree images," *IEEE Trans. Multimedia*, early access, Dec. 12, 2019, doi: 10.1109/TMM.2019.2957986.

[85] Y. Zhu, G. Zhai, and X. Min, "The prediction of head and eye movement for 360 degree images," *Signal Process., Image Commun.*, vol. 69, pp. 15–25, Nov. 2018, doi: 10.1016/j.image.2018.05.010.

[86] A. T. Duchowski, "Eye tracking techniques," in *Eye Tracking Methodology*. Berlin, Germany: Springer, 2017, pp. 49–57.

[87] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl, "State-of-the-art of visualization for eye tracking data," in *Proc. EuroVis*, 2014, pp. 1–10.

[88] J. Hasic, A. Chalmers, and E. Sikudova, "Perceptually guided high-fidelity rendering exploiting movement bias in visual attention," *TAP*, vol. 8, no. 1, pp. 1–6, 2010.

[89] A. Samarin, T. Koltunova, V. Osinov, D. Shaposhnikov, and L. Podladchikova, "Scanpaths of complex image viewing: Insights from experimental and modeling studies," *Perception*, vol. 44, nos. 8–9, pp. 1064–1076, Aug. 2015.

[90] J. H. Bappy, S. Paul, and A. K. Roy-Chowdhury, "Online adaptation for joint scene and object classification," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 227–243.

[91] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2014, pp. 740–755.

[92] C. Shen and Q. Zhao, "Webpage saliency," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 33–46.

[93] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua, "An eye fixation database for saliency detection in images," in *Proc. ECCV*, 2010, pp. 30–43, 2010.

[94] K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva, "Modelling search for people in 900 scenes: A combined source model of eye guidance," *Vis. Cognition*, vol. 17, nos. 6–7, pp. 945–978, Aug. 2009.

[95] A. Borji and L. Itti, "CAT2000: A large scale fixation dataset for boosting saliency research," 2015, *arXiv:1505.03581*. [Online]. Available: http://arxiv.org/abs/1505.03581

[96] H. Duan, G. Zhai, X. Min, Y. Fang, Z. Che, X. Yang, C. Zhi, H. Yang, and N. Liu, "Learning to predict where the children with asd look," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 704–708, doi: 10.1109/ICIP.2018.8451338.

[97] H. Duan, G. Zhai, X. Min, Z. Che, Y. Fang, X. Yang, J. Gutiérrez, and P. L. Callet, "A dataset of eye movements for the children with autism spectrum disorder," in *Proc. 10th ACM Multimedia Syst. Conf.*, Amherst, MA, USA, M. Zink, L. Toni, and A. C. Begen, Eds., Jun. 2019, pp. 255–260, doi: 10.1145/3304109.3325818.

[98] H. Duan, X. Min, Y. Fang, L. Fan, X. Yang, and G. Zhai, "Visual attention analysis and prediction on human faces for children with autism spectrum disorder," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 3, pp. 1–23, Jan. 2020.

[99] Y. Zhu, D. Zhu, Y. Yang, H. Duan, Q. Zhou, X. Min, J. Zhou, G. Zhai, and X. Yang, "A saliency dataset of head and eye movements for augmented reality," 2019, *arXiv:1912.05971*. [Online]. Available: http://arxiv.org/abs/1912.05971

[100] F. Viksten, P.-E. Forssén, B. Johansson, and A. Moe. (2009). *Object Pose Estimation Database*. [Online]. Available: http://www.cvl.isy.liu.se/research/objrec/posedb/

[101] F. Viksten, P.-E. Forssen, B. Johansson, and A. Moe, "Comparison of local image descriptors for full 6 degree-of-freedom pose estimation," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 2779–2786.

[102] S. Engel, X. Zhang, and B. Wandell, "Colour tuning in human visual cortex measured with functional magnetic resonance imaging," *Nature*, vol. 388, no. 6637, pp. 68–71, Jul. 1997.

[103] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[104] Y. Sugano and A. Bulling, "Self-calibrating head-mounted eye trackers using egocentric visual saliency," in *Proc. 28th Annu. ACM Symp. User Interface Softw. Technol.*, 2015, pp. 363–372.

[105] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva, "Intrinsic and extrinsic effects on image memorability," *Vis. Res.*, vol. 116, pp. 165–178, Nov. 2015.

[106] M. Jiang, J. Xu, and Q. Zhao, "Saliency in crowd," in *Proc. ECCV*, 2014, pp. 17–32.

[107] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, 2018.

[108] J. Harel. (2012). *A Saliency Implementation in MATLAB*. [Online]. Available: http://www.vision.caltech.edu/ harel/share/gbvs.php

[109] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. (2017). *Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model*. [Online]. Available: https://github.com/marcellacornia/sam

[110] E. Vig, M. Dorr, and D. Cox. (2014). *eDN-Saliency*. [Online]. Available: https://github.com/coxlab/edn-cvpr2014

[111] J. D'Errico. (2012). *Polyfitn*. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/34765-polyfitn

ALESSANDRO BRUNO received the master's degree in computer science and the Ph.D. degree in computer vision from the University of Palermo, in 2008 and 2012, respectively. During his Ph.D. Scholarship at the University of Palermo (IT), he focused on texture and local keypoint analysis. After his Ph.D., he worked at the Computer Vision and Image Processing Lab (CVIP) headed by Prof. Ardizzone from the University of Palermo, from 2012 to 2017. From 2012 to 2019, he taught basics of computer science at the School of Medicine, University of Palermo, as a Contract Lecturer. In 2019, he was a Research Visitor of the Imaging Group headed by Prof. J.-P. Muller at MSSL from the University College London (UCL). He has dealt with biomedical imaging, image forensics, visual attention, visual saliency, and remote sensing. As a Postdoctoral Research Fellow at INAF, he worked on detecting cloud masks from remote sensing imagery and gamma-hadron separation in cosmic ray analysis using deep learning architectures. He is currently working as a Research Associate at the National Centre for Computer Animation (NCCA), Bournemouth University. His main research interests include visual attention, human–computer interaction, and biomedical imaging. He is a member of CVPL (already GIRPR) and the Italian Association for Research in pattern recognition, computer vision, and machine learning. In 2018, he won the Fellowship at the Italian National Institute for Astrophysics (INAF) and a position as a postdoctoral research fellow, in 2019.

FRANCESCO GUGLIUZZA received the B.Sc. and M.Sc. degrees in computer science from the University of Palermo, in 2012 and 2015, respectively. In 2015, he started a three-year Ph.D. Scholarship at the University of Palermo, where he majored in computer vision and image processing techniques. He dealt with topics, such as remote sensing, visual saliency, and biomedical imaging. On the 15th of February 2019, he defended his Ph.D. thesis titled Methods and Techniques for Multi-Source Data Analysis and Fusion. Since 2019, he has been also a Red Hat Certified Administrator in the Red Hat OpenStack and OpenShift Administration. He is currently working as a Postdoctoral Researcher with the University of Palermo. He is a member of CVPL (already GIRPR) and the Italian Association for Research in pattern recognition, computer vision, and machine learning.

ROBERTO PIRRONE (Member, IEEE) received the degree in electronic engineering from the University of Palermo, in 1991, and the Italian Research Doctorate degree in computer engineering, in 1995. He was Assistant Professor of computer engineering, from 1999 to 2004. He has been an Associate Professor of computer engineering with the Department of Engineering (DI), University of Palermo, since 2004. He is currently the Head of the Computer-Human Interaction Laboratory (CHI-Lab) at DI. His main research interests include artificial intelligence and machine learning with applications in medical imaging, bioinformatics, and natural language processing. He teaches big data to computer engineering graduate students, web and mobile programming to undergraduate students in computer engineering, and foundations of computer science to undergraduate students in communication science. He organized several scientific events in the field of artificial intelligence. He is a coauthor of more than 130 indexed articles. He received several grants from the EU and the Italian University Ministry.

EDOARDO ARDIZZONE is currently a Full Professor of computer systems with the Department of Engineering (DI), University of Palermo, Italy. He teaches image processing at the graduate course of computer engineering of the University of Palermo. He is the author or coauthor of more than 180 scientific articles. He has been responsible of research units in Palermo involved in many research projects in his interest domains. His current research interests include image processing and analysis, medical imaging, image restoration, and content-based image and video retrieval. He is a member of CVPL (already GIRPR and IAPR-IC) and the Association of Italian Researchers in the area of pattern recognition and image analysis.