Exploiting Visual Saliency Algorithms for Object-Based Attention: a New Color and Scale-Based Approach

Edoardo Ardizzone, Alessandro Bruno, and Francesco Gugliuzza(✉)

Dipartimento dell'Innovazione Industriale e Digitale (DIID),
Università degli Studi di Palermo,
Viale delle Scienze Ed. 6 - 90128 Palermo (PA), Italy
{edoardo.ardizzone,alessandro.bruno15,francesco.gugliuzza}@unipa.it
http://www.diid.unipa.it/cvip

Abstract. Visual Saliency aims to detect the most important regions of an image from a perceptual point of view. More in detail, the goal of Visual Saliency is to build a Saliency Map revealing the salient subset of a given image by analyzing bottom-up and top-down factors of Visual Attention. In this paper we proposed a new method for Saliency detection based on colour and scale analysis, extending our previous work based on SIFT spatial density inspection. We conducted several experiments to study the relationships between saliency methods and the object attention processes and we collected experimental data by tracking the eye movements of thirty viewers in the first three seconds of observation of several images. More precisely, we used a dataset that consists of images with an object in the foreground on an homogeneous background. We are interested in studying the performance of our saliency method with respect to the real fixation maps collected during the experiments. We compared the performances of our method with several state of the art methods with very encouraging results.

Keywords: Visual saliency, object-based attention, SIFT, fixation maps, dataset, eye tracking

1 Introduction

The processing of visual information in human vision system begins in a thin layer made of neural tissue called retina. The architecture of our retina allows us to receive, every second, up to 10 billion bits information, while our cerebral cortex reaches about 10 billion neurons. Due to the lack of storage capability of our brain we cannot simultaneously perform complex analysis on all the input visual information [1]. One of the most important task of the human visual system is to detect the important visual subset, i.e. the salient subset. 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. The movements of the eyes, while observing a scene, are not random: each movement allows the central

part of vision (fovea) to fall upon the region of interest of a picture (this is why vision is not uniform across our field of view and acuity decreases with eccentricity) [2].

There is an intimate connection between visual attention and eye movements; for this reason in the last decades, how, why and when we move our eyes is becoming a major topic in scientific research.

Two main factors drive visual attention: bottom-up factors and top-down factors. Bottom-up factors are derived solely from the visual scene [2]. Regions of interest attracting human attention are sufficiently discriminative with respect to surrounding features. This attentional mechanism is called exogenous [3]. Top-down attention, on the other hand, is driven by cognitive factors such as knowledge, expectations and current goals [4]. Other terms for top-down attention are endogenous [5], voluntary, or centrally cued attention.

The objective of Visual Saliency is to detect the most important regions of an image from a perceptual point of view, i.e. to imitate the behaviour of the human visual system. Visual Saliency studies are included in several research areas, such as Psychology, Neurobiology, Computer Science, Artificial Intelligence, Medicine; our work considers Visual Saliency from a Computer Science point of view. The objective is to build a saliency map revealing the salient subset in an image. It is usually a grayscale map and each pixel falls in the dynamic range [0, 255], where highest intensity values correspond to the most salient pixels of a picture.

In the same way as Visual Attention, Saliency approaches are based on Bottom-up factors and Top-down factors. More in detail, Visual Saliency methods can be grouped in three main approaches: Bottom-up, Top-down, Hybrid.

Bottom-up methods are stimulus-driven. These methods seek for the so-called visual pop-out saliency. Human attention, in these approaches, is considered as a cognitive process concentrating on the most unusual aspects of an environment while ignoring the common aspects. This consideration is implemented by several methods such as center-surround operation [6] and graph based activation maps [7].

Top-down methods for saliency detection are based on high level visual tasks such as Object Detection or Face Detection. In these methods the predefined task is given by the object class to be detected [8].

Hybrid methods are generally structured in two levels: a bottom-up layer gives rise to a noisy saliency map and a top-down layer filters out noisy regions in saliency maps created by the bottom-up layer.

The eye tracking technique has been typically adopted to examine human visual attention. The location of the eye fixation reflects attention, while the duration of the eye fixation reflects processing difficulty and amount of attention [9]. Specifically, fixation duration varies depending on types of information (e.g. texts or graphics) and types of tasks (e.g. reading or problem solving).

In our work we used an eye tracker to record the gaze path of thirty observers while viewing each image of a dataset [10]. Each image is show at full resolution

for three seconds, separated by one second of viewing a gray screen (we adopted the same experimental approach of Torralba et al. [11]). The database [10] consists of several images with single objects in the foreground and homogeneous background color, but any dataset with a single main object (target) and a limited number of distractors in each image would have been appropriate as well. The viewers, while observing the images, sat at a distance of 70 cm from a 22 inch computer screen of 1920x1080 resolution.

We used the eye tracking data to create a ground truth made of fixation point maps showing where viewers look in the first three seconds of observation.

Our contributions in this work are three: a new available ground truth of fixation maps; a new saliency method, extending our previous study on visual saliency; a correlation study between saliency maps and the object attention process.

2 State of the Art

Models and approaches for visual saliency detection are inspired by human visual system mechanisms. As we reported in the first section of this paper, saliency methods can be divided in three main groups: Bottom-up, Top-down, Hybrid. In [6] the authors proposed a bottom-up approach based on multi-scale analysis of the image. In greater detail, multi-scale image features are used to create a topographical saliency map, then a dynamical neural network selects the attended locations with respect to the saliency values. The principle of center-surround difference is adopted in [12] for the parallel extraction of different feature maps. In [7] Harel et al. propose 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. [13] survey the corresponding literature on the low-level methods for visual saliency.

An effective method [14] for visual saliency detection based on multi-scale and multi-channel mean has been proposed by Sun et al. The image is decomposed and reconstructed by using wavelet transform and 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. SIFT [15] keypoints density maps have been proposed in our previous works to extract saliency maps and texture scale [16] [15] [17].

In Top-down approaches [8] [18], the visual attention process is considered task dependent, and the observer's expectations and wills analyzing the scene are the reason why a point is fixed rather than others. In [19] the authors perform saliency detection with a Top-down model that jointly learns a Conditional Random Field (CRF) and a visual dictionary.

Generally Hybrid systems for saliency use the combination of bottom-up and top-down stimuli. In many hybrid approaches [20] [21], a Top-down layer is used to refine the noisy map extracted from the Bottom-up layer. For example the Top-down component in [20] is face detection. Chen et al. [21] used a combi-

nation of face and text detection and they found the optimal solutions through branch and bound technique. A well known state-of-the-art hybrid approach was proposed by Judd et al. [11] in addition to a database [22] of eye tracking data from 15 viewers. Low, middle and high-level features of this data have been used to train a model of saliency.

Yu et al. in [23] used a paradigm based on Gestalt grouping cues for object-based saliency detection.

In the last years several researchers focused their attention on deep learning approaches for saliency prediction because high-quality visual saliency model can be learned by using deep convolutional neural networks (CNNs). For instance, in [24] the authors introduced a neural network architecture, which has fully connected layers on top of a CNNs responsible for feature extraction at different scales.

The authors of [25] reported a comparative study that evaluates the performances of 13 state-of-the-art saliency models. A new metric is also proposed and compared with previous models. In [26] the authors give some formal definitions on three different type of approaches (Bottom-up, Top-down, Hybrid) and an overview on existing methods. Furthermore, the authors offer a description of publicly available datasets and the performance metrics used.

3 Proposed Methods

3.1 Eye Tracking Data Acquisition

We chose 5 different objects from the Object Pose Estimation Database (OPED) [27] [10], and for each object we selected 19 views having a 130° fixed vertical angle and an horizontal angle ranging from 0° to 180° in 10° increments. The resulting 95 images have been thoroughly filtered to attenuate noise that could shift human attention and padded to fill the 22" screen at 1920x1080 resolution. We showed the images to 24 users (males and females between 21 and 34 years old) placed at approximately 70 cm from the screen.

The acquisition procedure for each user was as follows: the Tobii EyeX [28] running at 60 Hz refresh rate was calibrated to the user, then each image was shown for 3 seconds while capturing all user's saccades and fixations, separated by 1 second of neutral gray screen, to keep the results consistent to those of previous works in literature [11].

Fixation Map Generation Data acquired from the Tobii EyeX include three arrays of the same length: an array of positions looked at by the left eye, an array of positions looked at by the right eye, and an array of sampling times. The fixation points are calculated averaging data of both eyes and converting the result to screen coordinates, then full resolution maps are built by adding

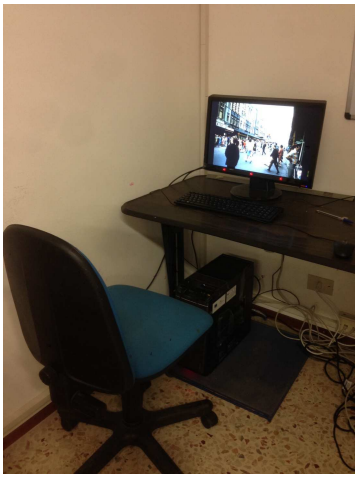

Fig. 1. The setup used for eye fixation data acquisition.

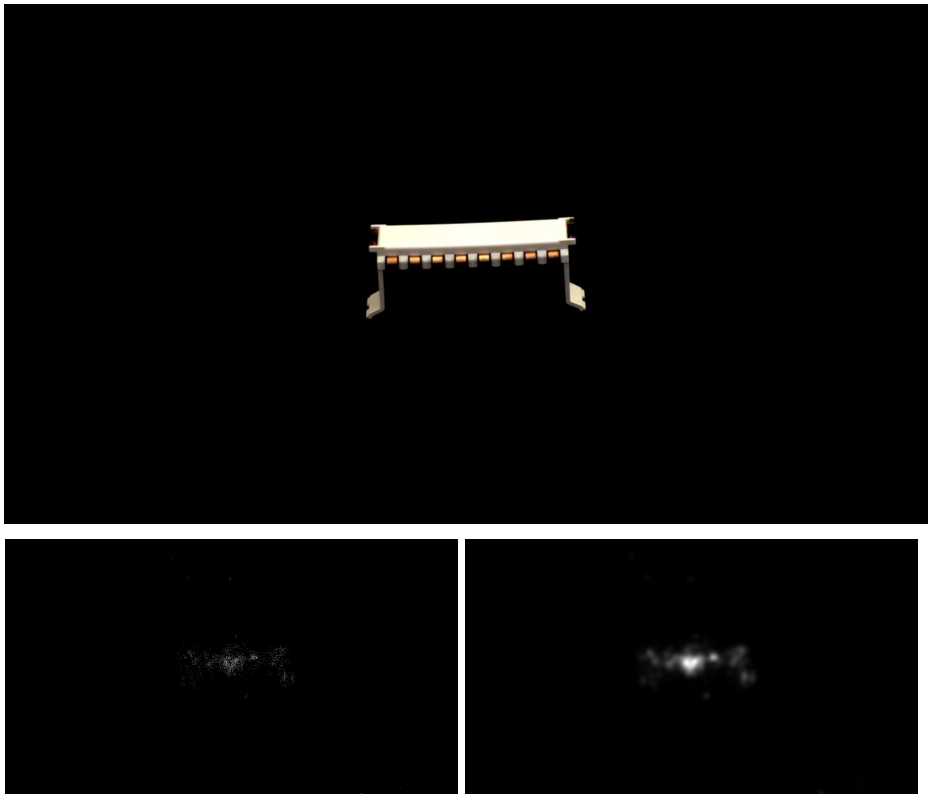

Fig. 2. An image from the OPED and its fixation map before and after Gaussian blurring.

one to a pixel's value each time a user has looked at said pixel. The map are subsequently averaged using a Gaussian convolution kernel and normalized to [0, 1].

3.2 Proposed Saliency Map Generation Method

We aimed to improve the method we (Ardizzone et al.) developed in 2011 [15] by adding chroma information to the saliency map generation algorithm based on SIFT [29] Density Maps (SDMs). A SDM is built by counting the number of detected SIFT keypoints inside a sliding window of size k x k centered on each pixel of the image. To obtain a valid saliency map, the SDM is further processed by taking the absolute difference of each pixel with the most frequent value (mode) of the map, rescaling the values to [0, 1] and blurring the result with an average filter which has a window size that is half of that used to build the map (k).

Color-based saliency has been implemented in two ways by harnessing the power of HSV and CIE L*a*b* color spaces. We early found that the optimal SDM window size equation we used in [15]:

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

is unsuitable in object attention because it is calculated on entire image size, while the object only takes a small central portion of it, causing excessive loss of detail in the generated saliency maps. We overcame the problem by first taking the mean of the dimensions of the object bounding boxes in all images used during the data acquisition phase, then applying (1) to the calculated values.

HSV Color Space Saliency In HSV an image is expressed using cylindrical coordinates, where hue is an angular dimension that goes from 0° to 360° and then back to 0°, while saturation and value are linear dimensions. 8-bit RGB images can be easily converted to HSV by projecting the RGB cube on a chromaticity plane in such a way that an hexagon is formed:

$$\begin{aligned}
C_{max} &= \max(R, G, B) \\
C_{min} &= \min(R, G, B) \\
\Delta &= C_{max} - C_{min} \\
H &= \begin{cases} 0 \text{ if } C_{max} = 0 \\ \left(60 \times \frac{G-B}{\Delta} + 360\right) \mod 360 & \text{if } R = C_{max} \\ 60 \times \frac{B-R}{\Delta} + 120 & \text{if } G = C_{max} \\ 60 \times \frac{R-G}{\Delta} + 240 & \text{if } B = C_{max} \end{cases} \\
S &= \frac{\Delta}{C_{max}} \\
V &= \frac{C_{max}}{255}
\end{aligned} \tag{2}$$

We convert hue and saturation from polar coordinates to cartesian coordinates[1], in order to eliminate the discontinuity around zero in hue values:

$$\begin{aligned} X &= S \circ \cos(H) \\ Y &= S \circ \sin(H) \end{aligned} \tag{3}$$

then we rescale the X, Y, V channels to the [0, 1] range for convenience of processing and separately calculate statistically processed SDMs. The three maps are combined into the final saliency map:

$$SM_{HSV} = \frac{1}{3}(SM_H + SM_S + SM_V) \; . \tag{4}$$

Fig. 3. An image from the OPED and its SIFT saliency map calculated in HSV space.

CIE L*a*b* Color Space Saliency HSV space still shows some shortcomings, namely hue and saturation channels are dominated by noise when brightness is low; furthermore, it is not biologically inspired, and does not model the HVS color opponent process [30]. Therefore, we decided to implement SDM calculation also in the CIE L*a*b* space, which is perceptually uniform and designed with color opponency in mind [31].

The processing steps are essentially the same as the previous method: RGB → L*a*b* conversion (D65 illuminant used as reference), channel range rescaling, SDM calculation, statistical processing and fusion. Coordinate transformation has been omitted because this color space does not have mathematical discontinuities.

4 Experimental Results

We generated saliency maps using various methods, as our legacy work [15], Itti-Koch-Niebur [6], GBVS [7], Judd [11], our two new color-based methods and a

[1] Not to be confused with CIE XYZ.

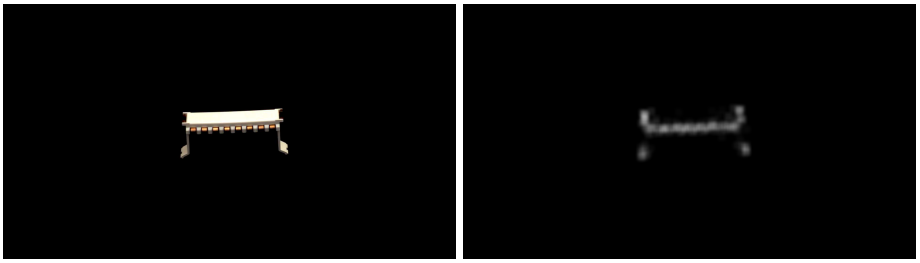

Fig. 4. An image from the OPED and its SIFT saliency map calculated in L*a*b* space.

fixed centered Gaussian distribution as a baseline [11]. We ran tests on our 95 image dataset and its related fixation point and fixation map database, on an Intel Core i7-4770 computer with 4 cores (8 threads) and 16 GB of RAM. For the calculation of GBVS and Itti-Koch-Niebur saliency maps the GBVS Toolbox [32] has been used, as it includes an enhanced implementation of Itti's algorithm; Judd saliency maps were instead generated running Judd's code [22] with its original trained parameters. We binarized saliency and fixation maps at various percentiles [11] [15] (between 0.95 and 0.5) and evaluated the performance of our method in terms of F-measure values:

$$P = \frac{n(M_D \cap M_R)}{n(M_D)}; R = \frac{n(M_D \cap M_R)}{n(M_R)}$$
$$F_1 = 2\frac{P \times R}{P + R} \tag{5}$$

where M_D is the binary version of the detected saliency map, while M_R is the binary version of the reference fixation map. We also calculated Normalized Scanpath Saliency (NSS) values, which is a well balanced, binarization-independent metric [33].

From Fig. 5, we note an opposite trend with respect to natural image saliency model performances reported in other works: in object attention, as saliency threshold reduces the F-measure tends to reduce as well, instead of increasing. The performances of both our models, instead, increase slightly with threshold until they reach a plateau at 90% saliency levels. Our CIE L*a*b*-based method always gets best results in both metrics, while the HSV-based method underperforms at high saliency levels with respect to GBVS and our previous work.

The execution time required for calculating a HSV or L*a*b* saliency map is about 12 seconds for a 1920x1080 image.

5 Conclusion and Future Works

In this paper we presented a new scale and color-based visual saliency method to generate accurate saliency maps when only one object is present in the stim-

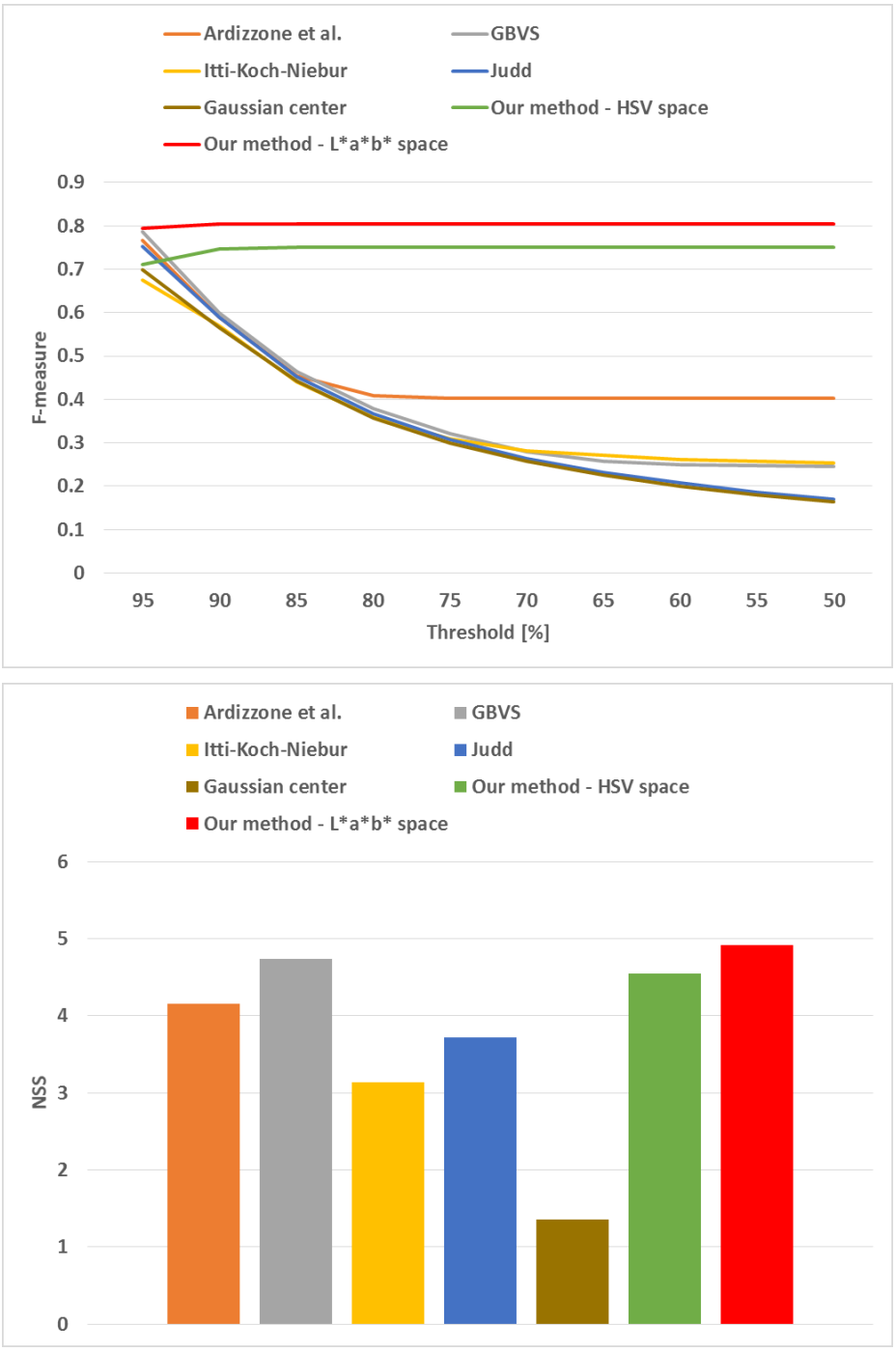

Fig. 5. Performance graphs of various saliency models in terms of F-measure vs. threshold (top) and NSS values (bottom).

ulus scene and we proposed a reference dataset to evaluate algorithm results. Our method, although taking into account only bottom-up features and being unsupervised, performs better than various reference algorithms, some of which exploit top-down features and trained neural networks (Judd et al.). We expect that this method would also give good results on natural images depicting a main object and a limited number of distractors, especially if dissimilar in size to the main object itself.

We believe that the effectiveness of our method comes from its ability to adapt to the effective object size, therefore correctly keeping track of the saliency of small object features.

In our future works, we plan to implement these improvements in our saliency algorithms for natural images and crowded scenes. The extension of this method to crowded scenes is not trivial and will probably require the addition of extra segmentation and object detection steps, to identify the size of all relevant objects in the image and compute the optimal SDM window size. We are also investigating the feasibility of multi-scale approaches using different window sizes on different parts of the image.

References

1. Li, J., Gao, W.: Visual saliency computation: A machine learning perspective. Volume 8408. Springer (2014)
2. Snowden, R., Snowden, R.J., Thompson, P., Troscianko, T.: Basic vision: an introduction to visual perception. Oxford University Press (2012)
3. Egeth, H.E., Yantis, S.: Visual attention: Control, representation, and time course. Annual review of psychology **48**(1) (1997) 269–297
4. Bressler, S.L., Tang, W., Sylvester, C.M., Shulman, G.L., Corbetta, M.: Top-down control of human visual cortex by frontal and parietal cortex in anticipatory visual spatial attention. Journal of Neuroscience **28**(40) (2008) 10056–10061
5. Posner, M.I.: Orienting of attention. Quarterly journal of experimental psychology **32**(1) (1980) 3–25
6. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11) (1998) 1254–1259
7. Harel, J., Koch, C., Perona, P., et al.: Graph-based visual saliency. In: NIPS. Volume 1. (2006) 5
8. Luo, J.: Subject content-based intelligent cropping of digital photos. In: Multimedia and Expo, 2007 IEEE International Conference on, IEEE (2007) 2218–2221
9. Tsai, M.J., Hou, H.T., Lai, M.L., Liu, W.Y., Yang, F.Y.: Visual attention for solving multiple-choice science problem: An eye-tracking analysis. Computers & Education **58**(1) (2012) 375–385
10. `http://www.cvl.isy.liu.se/research/objrec/posedb/`
11. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: Computer Vision, 2009 IEEE 12th International Conference on, IEEE (2009) 2106–2113
12. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. In: Matters of intelligence. Springer (1987) 115–141

13. Wang, L., Dong, S.L., Li, H.S., Zhu, X.B.: A Brief Survey of Low-Level Saliency Detection. In: Information System and Artificial Intelligence (ISAI), 2016 International Conference on, IEEE (2016) 590–593
14. Sun, L., Tang, Y., Zhang, H.: Visual saliency detection based on multi-scale and multi-channel mean. Multimedia Tools and Applications **75**(1) (2016) 667–684
15. Ardizzone, E., Bruno, A., Mazzola, G.: Visual saliency by keypoints distribution analysis. In: International Conference on Image Analysis and Processing, Springer (2011) 691–699
16. Ardizzone, E., Bruno, A., Mazzola, G.: Scale detection via keypoint density maps in regular or near-regular textures. Pattern Recognition Letters **34**(16) (2013) 2071–2078
17. Ardizzone, E., Bruno, A., Mazzola, G.: Saliency based image cropping. In: International Conference on Image Analysis and Processing, Springer (2013) 773–782
18. Sundstedt, V., Chalmers, A., Cater, K., Debattista, K.: Top-Down Visual Attention for Efficient Rendering of Task Related Scenes. In: VMV. (2004) 209–216
19. Yang, J., Yang, M.H.: Top-down visual saliency via joint CRF and dictionary learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(3) (2017) 576–588
20. Tsotsos, J.K., Rothenstein, A.: Computational models of visual attention. Scholarpedia **6**(1) (2011) 6201
21. Chen, L.Q., Xie, X., Fan, X., Ma, W.Y., Zhang, H.J., Zhou, H.Q.: A visual attention model for adapting images on small displays. Multimedia systems **9**(4) (2003) 353–364
22. `http://people.csail.mit.edu/tjudd/WherePeopleLook/index.html`
23. Yu, J.G., Xia, G.S., Gao, C., Samal, A.: A computational model for object-based visual saliency: Spreading attention along Gestalt cues. IEEE Transactions on Multimedia **18**(2) (2016) 273–286
24. Li, G., Yu, Y.: Visual Saliency Detection Based on Multiscale Deep CNN Features. IEEE Transactions on Image Processing **25**(11) (2016) 5012–5024
25. Toet, A.: Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(11) (2011) 2131–2146
26. Duncan, K., Sarkar, S.: Saliency in images and video: a brief survey. IET Computer Vision **6**(6) (2012) 514–523
27. Viksten, F., Forssén, P.E., Johansson, B., Moe, A.: Comparison of local image descriptors for full 6 degree-of-freedom pose estimation. In: Robotics and Automation, 2009. ICRA'09. IEEE International Conference on, IEEE (2009) 2779–2786
28. Gibaldi, A., Vanegas, M., Bex, P.J., Maiello, G.: Evaluation of the Tobii EyeX Eye tracking controller and Matlab toolkit for research. Behavior Research Methods (2016) 1–24
29. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
30. Engel, S., Zhang, X., Wandell, B.: Colour tuning in human visual cortex measured with functional magnetic resonance imaging. Nature **388**(6637) (1997) 68–71
31. Sharma, G.: Color fundamentals for digital imaging. In: Digital Color Imaging Handbook. CRC Press (2002)
32. `http://www.vision.caltech.edu/~harel/share/gbvs.php`
33. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? arXiv preprint arXiv:1604.03605 (2016)