

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,200

Open access books available

169,000

International authors and editors

185M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Chapter

Saliency Detection from Subitizing Processing

Carola Figueroa-Flores

Abstract

Most of the saliency methods are evaluated for their ability to generate saliency maps, and not for their functionality in a complete vision pipeline, for instance, image classification or salient object subitizing. In this work, we introduce saliency subitizing as the weak supervision. This task is inspired by the ability of people to quickly and accurately identify the number of items within the subitizing range (e.g., 1 to 4 different types of things). This means that the subitizing information will tell us the number of featured objects in a given image. To this end, we propose a saliency subitizing process (SSP) as a first approximation to learn saliency detection, without the need for any unsupervised methods or some random seeds. We conduct extensive experiments on two benchmark datasets (Toronto and SID4VAM). The experimental results show that our method outperforms other weakly supervised methods and even performs comparable to some fully supervised methods as a first approximation.

Keywords: saliency prediction, subitizing, object recognition, deep learning and convolutional neural network

1. Introduction

For humans, object recognition is a nearly instantaneous, precise, and extremely adaptable process. Furthermore, it has the innate ability to learn new classes of objects from a few examples [1, 2]. The human brain reduces the complexity of incoming data by filtering out some of the information and processes only those things that grab our attention. This, combined with our biological predisposition to respond to certain shapes or colors, allows us to recognize at a glance the most important or outstanding regions of an image. This mechanism can be observed by analyzing which parts of the images humans pay more attention to; for example, where they fix their eyes when they are shown an image [3, 4]. The most accurate way to record this behavior is by tracking eye movements, while the subject in question is presented with a set of images to evaluate. Computational estimation of saliency (or salient or salient regions) aims to identify to what extent regions or objects stand out from their surroundings or background to human observers. Saliency maps can be used in a wide range of applications, including object detection, image and video understanding, and eye tracking. On the other hand, it is known that the human visual system can effortlessly identify the number of objects in the range 1 to 4 by having just one glance [5]. Since

then, this phenomenon, coined later by [6] as subitizing, has been studied and tested in various experimental settings [7].

Therefore, inspired by subitizing and the results obtained in [8, 9], the main objective of this project is to incorporate the subitizing of salient objects (SOS), in order to improve our previous results. This means that the subitizing information will tell us the number of outgoing objects in a given image and thus subsequently provide us with the location or appearance information of the outgoing objects explicitly, and everything will be done within a weakly supervised configuration. It should be noted that when the network is trained with the subitizing supervisions, the network will learn to focus on the regions related to the outgoing objects. Therefore, it will design a saliency subitizing process (SSP) architecture that is responsible for extracting attention regions as saliency map. A second module that is in charge of improving the quality of the saliency masks can be defined as the saliency map update process (SUP), which will basically be in charge of refining the activation regions in an end-to-end way. It will then merge the source images and saliency maps to get the masked images as new inputs for the next refinement. Finally, in this work we propose to design and build a convolutional neural network (CNN), which will basically consist of a process that will be in charge of SSP and a function that will help us in the task of SUP. The first SSP will serve as a support to obtain and calculate the number of outstanding objects and thus extract the saliency maps with their respective locations. Instead, SUP will help us update the saliency masks produced by the first module. The general model of our proposal is shown in **Figure 1**.

However, as this work is a first attempt at the final result, it will only consider the development, experimentation, and explanation associated with step 1.

It briefly summarizes below its main contributions:

- It proposes an approach that generates saliency maps from subitizing of saliency process (SSP),

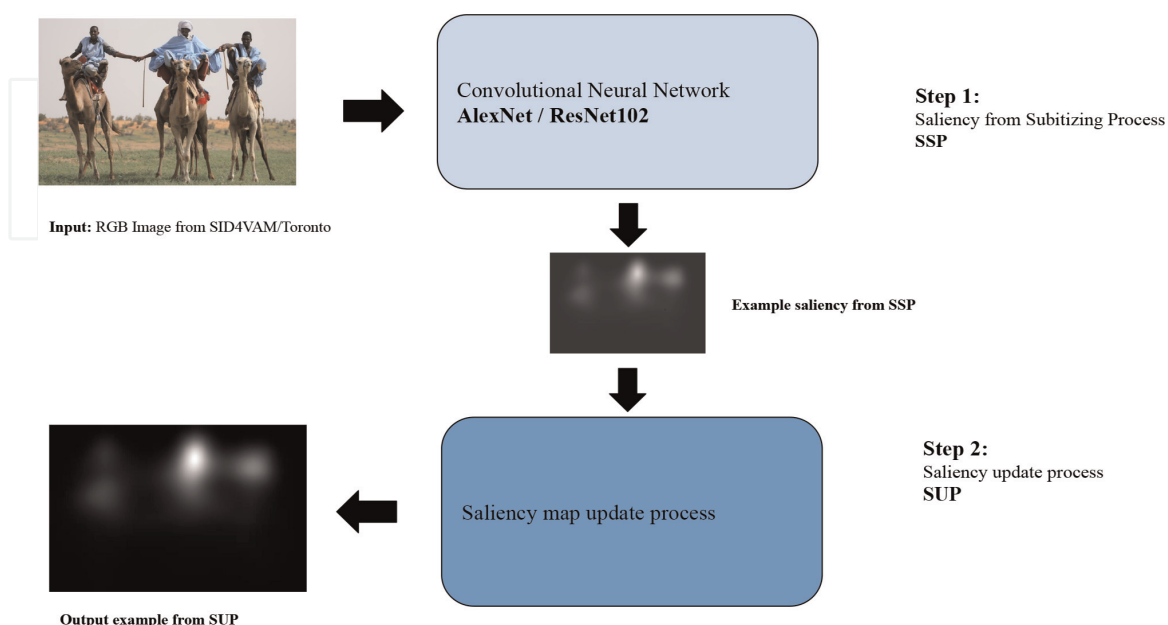


Figure 1. Overview of our proposed method with the saliency subitizing process (SSP) and the saliency updating process (SUP).

- Our saliency does not require any saliency maps for training (like previous works [10, 11]) but instead is trained indirectly in an end-to-end fashion by training the network for image classification with subitizing saliency process (SSP).
- The saliency maps obtained without using any saliency groundtruth data shows competitive results.

The chapter is organized as follows. Section 2 is devoted to review the related work in saliency detection. Section 3 presents our approach. Experimental results are reported in Section 4. Finally, Section 5 contains our conclusions.

2. Related work

Saliency is generally known as local contrast [12], which generally originates from contrasts between objects and their surroundings, such as differences in color, texture, shape. This mechanism measures intrinsically salient stimuli to the vision system that primarily attracts the attention of humans, in the initial stage of visual exposure to an input image [13]. To quickly extract the most relevant information from a scene, the human visual system pays more attention to highlighted regions, as seen in **Figure 1**. Research on computational saliency focuses on the design of algorithms that, like human vision, predict which regions of a scene stand out [14, 15].

Initial efforts to model saliency involved multi-scale representations of color, orientation, and intensity contrast. These were often biologically inspired, such as the well-known works [12, 16]. From that model, a large number of models were based on the manual elaboration of these features to obtain an accurate saliency map [17, 18], either maximizing [19] or learning statistics from natural images [13, 20]. Relevancy research was further driven by the availability of large datasets that enabled the use of machine learning algorithms [21], primarily pre-trained on existing human fixation data.

The question of whether saliency is important for object recognition and tracking has been raised in [22]. More recent methods [23] take advantage of end-to-end convolutional architectures by fine graining on fixation prediction [4, 24, 25]. But the main goal of these works was to estimate a saliency map, not how saliency might contribute to object recognition.

Several works have shown that having the saliency map of an image can be useful for object recognition, for example, [8, 10, 11]. Since the saliency map can help focus attention on the relevant parts of the image to improve recognition, additionally, it can help guide training by focusing backpropagation on relevant image regions. Previous work has shown that saliency modulated image classification (SMIC) is especially efficient for training on data sets with few labeled data [10]. The main drawback of these methods is that they require a trained saliency method. Also, Refs. [8, 9] show that this restriction can be removed and that it can hallucinate the saliency image from the RGB image. By training the network for image classification on the ImageNet dataset [26], it can obtain the saliency branch without using human reference images.

Recently, the progress in the detection of salient objects has grown substantially, mainly benefiting from the development of deep neural networks (CNN). In [27], a CNN based on the use of superpixels for saliency detection was proposed. Instead, Li et al. [28] used multi-scale features extracted from a deep CNN. Zhao et al. [29] proposed a multi-context deep learning framework to detect salient objects with two different CNNs, which were useful for learning local and global information. Yuan

et al. [30] proposed a saliency detection framework, which extracted the correlations between object contours and RGB features of the image. On the other hand, Wang and Shen [31] defined a pyramid-shaped structure to expand the receptive field in visual attention. Hou and Zhang [32] introduced short connections for edge or contour detection. Zhu [33], on the other hand, proposed a visual attention architecture called DenseASPP, to extract information. Chen [34] proposed a spatial attenuation context network, which recursively translated and aggregated the context features in different layers. Tu [35] introduced an edge-guided block to embed boundary information in saliency maps. Zhou [36] proposed a multi-type self-attention network to learn more semantic details from degraded images. However, these methods rely heavily on pixel-based monitoring. Overcoming the scarcity of pixel-based data, it focusses on the saliency detection task.

2.1 Weakly supervised saliency detection

There are many works using weak supervisions for the saliency detection task. For example, Li [37] used the image-level labels to train the classification network and applied coarse activation maps as saliency maps. Wang [38] proposed a weakly supervised two-stage method by designing an inference network to predict foreground regions and global smooth pooling (GSP) to aggregate responses from those predicted objects. On the other hand, Zeng [39] designed a unified network, which is capable of weak monitoring of multiple sources, including image labels, captions, and pseudo-labels. Furthermore, they designed a loss of attention transfer to transmit signals between subnetworks with different supervisions.

Different from the previous methods, it proposes to use subitizing information as weak supervision in the saliency detection task, where it will first study the problem of subitizing of the outgoing object and the relationships between subitizing and saliency detection.

3. Proposed method

This work proposes to design and implement a convolutional neural network, which will consist mainly of saliency subitizing process (SSP). The SSP will help us to count the highlighted objects and at the same time extract the saliency from the maps that will contain the locations (positions) of the objects.

3.1 Subitizing of saliency process (SSP)

It should be noted that the information provided by the subitizing process will indicate the number of outgoing objects in a given image [40]. Therefore, it will not explicitly provide the location or information related to the appearance of the output objects. However, when the network is being trained with subitizing (simulating supervised learning), the network will learn to focus on the regions related to the most salient (or salient) objects. Training images are divided into five categories based on the number of salient objects: 0, 1, 2, 3, and 4 or +. For the same reason, it will design the SSP to extract these regions as if it were a saliency mask. During this process, a classification network will be used for the object subitizing task, in this context ResNet-152 or ResNet50 [41] and AlexNet [42] as “backbone network,” which are pre-trained from the ImageNet dataset [43].

Also, it uses cross-entropy as the classification loss (see Eq. (1)). In order to obtain denser saliency maps, the stride of the last two down-sampling layers is set as 1 in our backbone network, which produces feature maps with 1/8 of the original resolution before the classification layer. In order to enhance the representation power of the proposed network, it also applies two attention modules: channel attention module and spatial attention module, which tell the network “where” and “what” to focus, respectively. Both of them are placed in a sequential way between the ResNet blocks and AlexNet convolutional layers.

$$\mathcal{I} = \sum_{I \in \mathcal{D}} \text{log} p_{c(I)}(y|I) \quad (1)$$

In addition, it applies the technique of the gradient-weighted class activation mapping (Grad-CAM) [44] to extract salient regions as the initial saliency maps, which contains the gradient information flowing into the last convolutional layers during the backward phase. The gradient information represents the importance of each neuron during inference of the network. It assumes that the features produced from the last convolutional layer has a channel size of K . For a given image, let f_k be the activation of unit K , where $k \in [1, K]$. For each class c , the gradients of the score y^c with respect to activation map f_k are averaged to obtain the neuron significant weight a_k^c of class c :

$$a_k^c = \frac{1}{N} \sum_i^m \sum_j^h \frac{\partial y^c}{\partial f_{i,j}^k} \quad (2)$$

where i and j represent the coordinates in the features map $N = m \times h$. With the neuron importance weight a_k^c , we can compute the activation map M^c :

$$M^c = \text{ReLU} \left(\sum_k a_k^c f^k \right) \quad (3)$$

And, finally it adds an activation map with ReLU (rectified linear unit) function layer; this function filters negative gradient values, since only the positive ones contribute to the class decision, while the negative values contribute to other categories. The size of the saliency map is the same as the size of the last convolutional feature maps (1/8 of the original resolution). This process is shown in **Figure 2**.

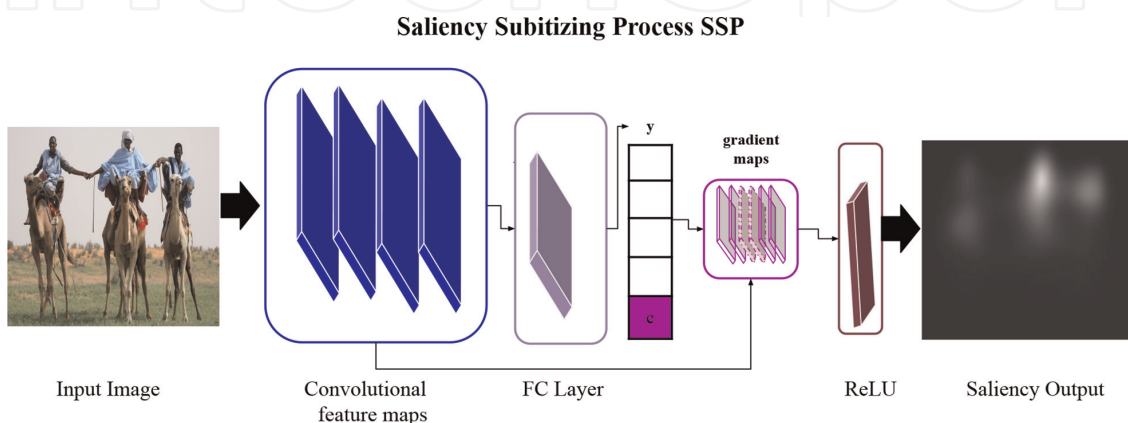


Figure 2.
 The pipeline of the saliency subitizing process (SSP).

4. Experiments

This section discusses the advantage of SSP that help us to learn counting of salient objects and extract coarse saliency maps with the precise locations of the target objects.

4.1 Experimental setup

Datasets. The saliency maps have been computed for images from a distinct eye-tracking dataset, corresponding to 120 real scenes (Toronto) [19] and 230 synthetic images with specific feature contrast (SID4VAM) (see **Table 1**) [45]. These images datasets have been computed with our approach, a supervised artificial model that specifically computes high-level features (DeepGazeII, ML-Net (multi-level net), SAM (saliency attentive model), salGAN), and models biological inspiration (IKN (Itti, Koch, and Niebur) [16], AIM [19] (saliency based on information maximization), SDLF (saliency detection by using local features) [20], and GBVS (graph-based visual saliency) [13]).

Networks architectures. It evaluates approach using two network architectures: AlexNet [42] and ResNet-152 [41]. It is modified to meet our requirement. In both cases, the weights were pretrained on ImageNet and then fine-tuned on each of the datasets mentioned above. The networks were trained for 70 epochs with a learning rate of 0.0001 and a weight decay of 0.005. The top classification layer was initialized from scratch using Xavier method [46]. The SSP consists of four convolutional layers for AlexNet and four residual blocks for ResNet-152.

Comparison. This work compares its proposal with other models (see **Tables 2** and **3**—rows 8) from fixation data. For instance, DeepGazeII summed the center baseline, whereas in ML-Net and SAM, the learned priors are used for modulating the result of the network.

4.2 Results

4.2.1 First experiment: Multiple networks

In order to evaluate how accurately the saliency map is able to match the location of human fixations, it used a set of metrics previously defined by [17].

In **Table 4** we show results of area under ROC (AUC), correlation coefficient (CC), normalized scanpath saliency (NSS), Kullback-Leibler divergence (KL), and similarity (SIM) for every network for all datasets.

The area under ROC (AUC) is considered as true positives, the saliency map values coincide with a fixation and false positives, and the saliency map values that have no fixation then compute the area under the curve. Similarly, the NSS computes the

Data Set	Type	# Images	# PP	pxva	Resolution
Toronto	Indoors and outdoors	120	20	32	681x511
SID4VAM	Synthetic pop-out	230	34	40	1280x1024

pxva: pixels per 1 degree of visual angle, PP: participants.

Table 1.
Characteristics of eye-tracking datasets.

Method	AUC	KL ↓	SIM	sAUC	InfoGain
IKN [16]	0.782	1.249	0.366	0.650	-0.024
AIM [19]	0.716	1.612	0.314	0.663	-0.580
SDLF [20]	0.703	1.518	0.304	0.664	-0.398
GBVS [13]	0.803	1.168	0.397	0.632	0.077
DeepGazeII [24]	0.838	1.367	0.325	0.763	-0.200
SAM-ResNet [4]	0.725	2.420	0.516	0.666	-1.555
SalGAN [47]	0.818	1.272	0.435	0.715	0.392
Our Approach (SSP)	0.740	1.409	0.399	0.597	-0.399
GroundTruth (Humans)	0.954	0.000	1.000	0.902	2.425

Table 2.

Comparison of our saliency output with standard benchmark methods over real image **Toronto** dataset for saliency prediction. (Top) Baseline low-level saliency models. (Bottom) State-of-the-art deep saliency models. Best score for each metric is defined as **bold** and TOP-3 scores are italicized.

Method	AUC	KL ↓	SIM	sAUC	InfoGain
IKN [16]	0.678	1.748	0.380	0.608	-0.233
AIM [19]	0.566	14.472	0.224	0.557	-18.181
SDLF [20]	0.607	3.954	0.322	0.596	-3.244
GBVS [13]	0.718	1.363	0.413	0.628	0.331
DeepGazeII [24]	0.610	1.434	0.335	0.571	-0.964
SAM-ResNet [4]	0.673	2.610	0.388	0.600	-1.475
SalGAN [47]	0.662	2.506	0.373	0.593	-1.350
Our Approach (SSP)	0.741	1.658	0.445	0.633	-0.122
GroundTruth (Humans)	0.882	0.000	1.000	0.860	2.802

Table 3.

Comparison of our saliency output with standard benchmark methods over synthetic image **SID4VAM** dataset for saliency prediction. (Top) Baseline low-level saliency models. (Bottom) State-of-the-art deep saliency models. Best score for each metric is defined as bold and TOP-3 scores are italicized.

Dataset	Model	AUC-Judd	AUC-Borji	CC	NSS	KL↓	SIM
	AlexNet	0.7655	0.7298	0.4603	1.3888	1.5155	0.3955
Toronto	ResNet152	0.7911	0.7443	0.5440	1.6391	1.6891	0.4410
	AlexNet	0.6910	0.7366	0.3889	1.4106	1.7152	0.4385
SID4VAM	ResNet152	0.7015	0.7723	0.3910	1.1155	1.9890	0.3996

Table 4.

Benchmark of our method with different networks (top 1 networks are italicized).

average normalized saliency map that coincides with fixations. Other metrics such as CC, KL, and SIM compute the score upon the region distribution statistics of all pixels (KL calculating the divergence and CC/SIM the histogram intersection or similarity of the distribution).

After computing the saliency maps for all datasets (see in **Table 4**) with AlexNet and ResNet152, we observed that metric scores vary considerably depending on dataset or network. AlexNet is shown to provide best results for pop-out patterns (SID4VAM), whereas ResNet152 shows overall higher scores with real images of Toronto dataset.

4.2.2 Second experiment: Qualitative results

These saliency prediction results show that our model has robust metric scores on both real and synthetic images for saliency prediction. Again, we would like to stress that our model is not trained on fixation prediction datasets (**Figure 3**). Its model with subitizing supervision performs best on detecting pop-out effects (from visual attention theories [16]) while performing similarly for real image datasets (**Figure 4**). Some deep saliency models use several mechanisms to leverage (or/and train) performance for improving saliency metric scores, such as smoothing/thresholding (see **Figure 4**, row 4). It is also considered that some of these models are already fine-tuned for synthetic images (e.g., SAM-ResNet [4]). *Our approach* (which has not been trained in these type of data sets) has shown to be robust on these two distinct scenarios/domains.

4.3 Evaluation benchmark of saliency estimation

Here, we compare the saliency estimation that is obtained after only performing Step 1 in **Figure 1** with existing saliency models (see **Table 5**). This saliency estimation is trained without access to any groundtruth saliency data.

Saliency prediction metrics assign a score depending on how well the predicted saliency map is able to match with locations of human fixations (see definitions in Borji et al. [17]). It selected the area under ROC (AUC), Kullback-Leibler divergence (KL), similarity (SIM), shuffled AUC (sAUC) and information gain (IG) metrics considering its consistency of predictions of human fixation maps. It compares scores

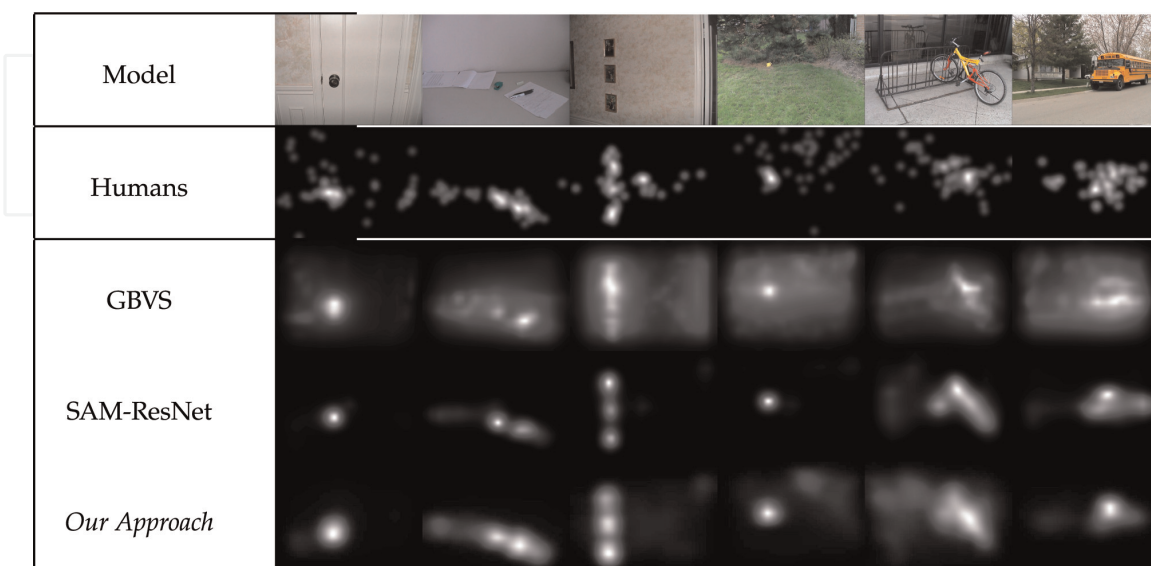


Figure 3. Qualitative results for real images (Toronto dataset). Each image is represented in a different column and each model saliency map in each row. The ground truth density map of human fixations is represented in the second row.

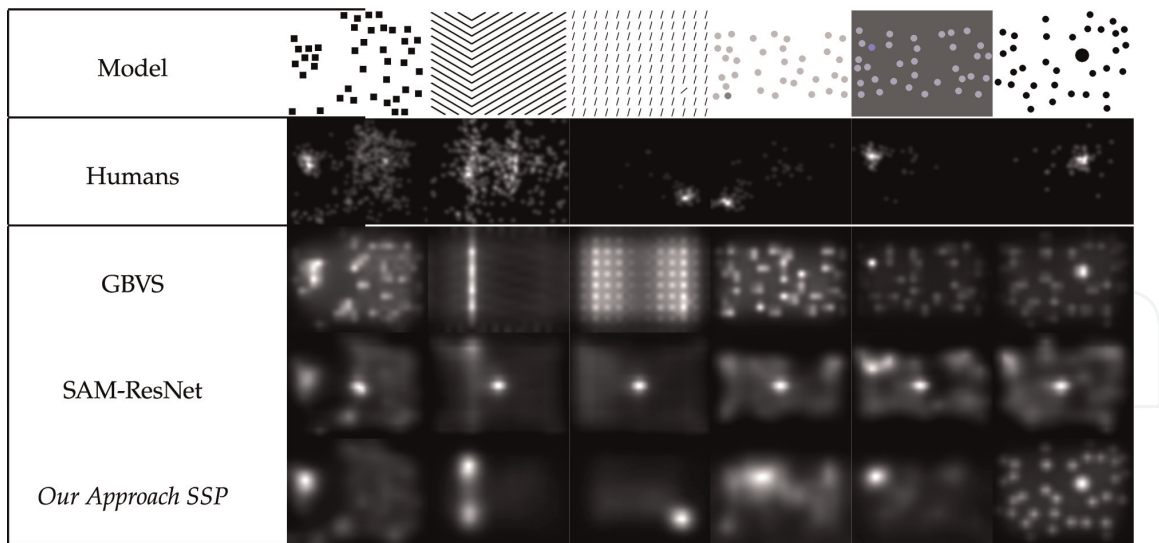


Figure 4. Qualitative results for synthetic images (*SID4VAM* dataset). Each image is represented in a different column and each model saliency map in each row. The ground truth density map of human fixations is represented in the second row.

with classical saliency models, both with handcrafted low-level features (i.e., IKN [16], AIM [19], SDLF [20], and GBVS [13]) and with state-of-the-art deep saliency models (i.e., DeepGazeII [24], SAM-ResNet [4], and SalGAN [47]) mainly pretrained on human fixations. The results are surprising; our method, which has not been trained on any saliency data, obtains competitive results. For the case of *Toronto* (Table 2), the best model is GBVS, followed by our model, which scores in the top 3 of KL and SAM-ResNet and scores slightly higher in InfoGain metric. For the case of *SID4VAM* (Table 3), our approach gets the best scores for most metrics compared with other deep saliency models, being mainly among the top 2 acquiring similar scores to GBVS in most metrics (outperforming it in AUC measures).

These saliency prediction results show that our model has robust metric scores on both real and synthetic images for saliency prediction. Again, we would like to stress that our model is not trained on fixation prediction datasets and our model with subitizing supervision (SUP) performs best on detecting pop-out effects (from visual attention theories [16]), while performing similarly for real image datasets (Figure 4). Some deep saliency models use several mechanisms to leverage (or/and train) performance for improving saliency metric scores, such as smoothing/thresholding (see Figure 4, rows 5). It also considers that some of these models are already fine-tuned for synthetic images (e.g., SAM-ResNet [4]). *Our approach* (which has not been trained in these types of datasets) has shown to be robust on these two distinct scenarios/domains.

5. Conclusions

In this chapter, we proposed a method for the saliency estimation with weak subitizing supervision. We designed a model with the saliency subitizing process (SSP), which generates the initial saliency map using subitizing information. Without any seeds from unsupervised methods, this method outperforms other weakly supervised methods and even performs comparable to some fully supervised methods.

#	Name	Year	Features/Architecture	Mechanism
1	IKN	1998	DoG (color+intensity)	—
2	AIM	2005	ICA (infomax)	max-like
3	GBVS	2006	Markov chains	graph prob.
4	SDLF	2006	Steerable pyramid	local+global prob.
5	ML-Net	2016	VGG-16	Backprop.(finetuning)
6	DeepGazeII	2016	VGG-19	Backprop.(finetuning)
7	SAM	2018	VGG-16/ResNet-50 + LSTM	Backprop.(finetuning)
8	SalGAN	2017	VGG-16 Autoencoder	Finetuning+GAN Loss
#	Name	Learning	Training Data (#img)	Bias/Priors
1	IKN	—	—	—
2	AIM	Unsupervised	Corel (3600)	—
3	GBVS	Unsupervised	Einhauser (108)	graph norm.
4	SDLF	Unsupervised	Oliva (8100)	scene priors
5	ML-Net	SALICON (10 k), MIT (1003)	learned priors	—
6	DeepGazeII	Supervised	SALICON (10 k), MIT (1003)	center bias
7	SAM	Supervised	SALICON (10 k) & others	Gaussian priors
8	SalGAN	Supervised	SALICON (10 k), MIT (1003)	—

DoG: difference of Gaussians, ICA: independent component analysis, C-S: center-surround, max-like: max-likelihood probability, BCE: binary cross-entropy, GAN: generative adversarial network.

Table 5.
Description of saliency models.

Finally, as this work is a first approximation, future work would be to verify how its saliency map would improve if the SUP update module were added.

Acknowledgements

We thank the support from FOVI21001 “Fomento a la Vinculación Internacional para Instituciones de Investigación Regionales (ANID, Chile),” Agencia Nacional de Investigación y Desarrollo and ALBA Research Group (Algorithms and Database) 2130591 GI/VC, “Ayudantes para el Fortalecimiento de Investigación FACE 2022,” and “Proyecto de Reinserción” DIUBB 2230508 IF/RS of the University of Bío.

Additional information

This chapter is a continuation of my PhD thesis, previous works related to the subject of saliency, and the use of the subitizing technique, because it had already been tested by other works, and in this way, the saliency estimation of my previous works was improved.


IntechOpen

Author details

Carola Figueroa-Flores
Department of Computer Science and Information Technology, Universidad del Bío Bío, Chile

*Address all correspondence to: cfigueroa@ubiobio.cl

IntechOpen

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Sun X, Yao H, Ji R, Liu XM. Toward statistical modeling of saccadic eye-movement and visual saliency. *IEEE Transactions on Image Processing*. 2014; **23**(11):4649-4662
- [2] Vincent BT, Tatler BW. Systematic tendencies in scene viewing. *Journal of Eye Movement Research*. 2008:1-18. eyemovement.org. DOI: 10.16910/jemr.2.2.5
- [3] Achanta R, Hemami S, Estrada F, Susstrunk S. Frequency-tuned salient region detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009. pp. 1597–1604
- [4] Cornia M, Baraldi L, Serra G, Cucchiara R. Predicting human eye fixations via an LSTM-based saliency attentive model. *IEEE Transactions on Image Processing*. 2018b;**27**(10):5142-5154
- [5] Stanley J. The power of numerical discrimination. *Nature*. 1871;**3**:367-367. DOI: 10.1038/003367b0
- [6] Kaufman EL, Lord Miles W, Whelan RT, Volkman J. The discrimination of visual number. *The American Journal of Psychology*. 1949;**62**:498-525
- [7] Whalen J, Gallistel CR, Gelman R. Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*. 1999;**10**:130-137. DOI: 10.1111/1467-9280.00120
- [8] Flores CF, Raducanu BC, Berga D, van de Weijer J. Hallucinating saliency maps for fine-grained image classification for limited data domains. *VISIGRAPP (4: VISAPP)*. 2021. pp. 163-171
- [9] Figueroa-Flores C, Berga D, van de Weijer J, Raducanu B. Saliency for free: Saliency prediction as a side-effect of object recognition. *Pattern Recognition Letters*. 2021:1-7. DOI: 10.1016/j.patrec.2021.05.015
- [10] Figueroa-Flores C, Gonzalez-Garcia A, van de Weijer J, Raducanu B. Saliency for fine-grained object recognition in domains with scarce training data. *Pattern Recognition*. 2019;**94**:62-73
- [11] Murabito F, Spampinato C, Palazzo S, Giordano D, Pogorelov K, Riegler M. Top-down saliency detection driven by visual classification. *Computer Vision and Image Understanding*. 2018:67-76. *Understanding*
- [12] Itti L, Koch C. Computational modeling of visual attention. *Nature Reviews. Neuroscience*. 2001;**2**:194-203. DOI: 10.1038/35058500
- [13] Harel J, Koch C, Perona P. Graph-based visual saliency. In: *Advances in Neural Information Processing Systems 19 (NIPS 2006)*. No. 19. Cambridge, MA: MIT Press; 2007. pp. 545-552. Available from: <https://resolver.caltech.edu/CaltechAUTHORS:20160315-111145907>. ISBN: 0-262-19568-2
- [14] Li Y, Hou X, Koch C, Rehg JM, Yuille AL. The secrets of salient object segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2014. pp. 280–287
- [15] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, et al. Show, attend and tell: Neural image caption generation with visual attention. In: *International Conference on Machine Learning*. 2015. pp. 2048–2057

- [16] 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*. 1998;**20**(11):1254-1259
- [17] Borji A, Itti L. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2013;**35**(1): 185-207. DOI: 10.1109/tpami.2012.89
- [18] Bylinskii Z, DeGennaro EM, Rajalingham R, Ruda H, Zhang J, Tsotsos JK. Towards the quantitative evaluation of visual attention models. *Vision Research*. 2015;**116**:258-268. DOI: 10.1016/j.visres.2015.04.007
- [19] Bruce NDB, Tsotsos JK. Saliency based on information maximization. In: *Proceedings of the 18th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press. 2005. pp. 155–162
- [20] Torralba A, Oliva A, Castelhano MS, Henderson JM. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*. 2006;**113**(4):766-786. DOI: 10.1037/0033-295x.113.4.766
- [21] Borji A, Sihite DN, Itti L. What/where to look next? Modeling top-down visual attention in complex interactive environments. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. 2014;**44**(5):523-538
- [22] Han S, Vasconcelos N. Biologically plausible saliency mechanisms improve feedforward object recognition. *Vision Research*. 2010;**50**:2295-2307
- [23] Borji A. Saliency prediction in the deep learning era: Successes and limitations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2019;**669**-700. DOI: 10.1109/TPAMI.2019.2935715
- [24] Kümmerer M, Wallis TSA, Bethge M. DeepGaze II: Reading fixations from deep features trained on object recognition. *ArXiv Preprint ArXiv: 1610.01563*. 2016
- [25] Pan J, Canton C, McGuinness K, O'Connor NE, Torres J, Sayrol E, et al. SalGAN: Visual saliency prediction with generative adversarial networks. In *arXiv*. 2017
- [26] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*. 2015;**115**(3):211-252
- [27] Qin Y, Lu H, Xu Y, Wang H. Saliency detection via cellular automata. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE Computer Society. 2015. pp. 110–119
- [28] Li C, Yuan Y, Cai W, Xia Y, Dagan Feng D. Robust saliency detection via regularized random walks ranking. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015. pp. 2710-2717
- [29] Zhao J, Sun S, Liu X, Sun J, Yang A. A novel biologically inspired visual saliency model. *Cognitive Computation*. 2014;**6**(4):841-848. DOI: 10.1007/s12559-014-9266-z
- [30] Yuan Y, Li C, Kim J, Cai W, Feng DDF. Reversion correction and regularized random walk ranking for saliency detection. *IEEE Transactions on Image Processing*. 2017;**1**:1-8. DOI: 10.1109/TIP.2017.2762422
- [31] Wang W, Shen J. Deep visual attention prediction. *IEEE Transactions on Image Processing*. 2018;**27**(5):2368-2378

- [32] Hou X, Zhang L. Saliency detection: A spectral residual approach. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. 2007. pp. 1–8
- [33] Zhu W, Liang S, Wei Y, Sun J. Saliency optimization from robust background detection. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014. pp. 2814–2821
- [34] Chen C, Tang H, Lyu Z, Liang H, Shang J, Serem M. Saliency modeling via outlier detection. *Journal of Electronic Imaging*. 2014;23(5):53023
- [35] Tu Z, Ma Y, Li C, Tang J, Luo B. Edge-guided non-local fully convolutional network for salient object detection. 2019. Retrieved from: <http://arxiv.org/abs/1908.02460>
- [36] Zhou Z, Wang Z, Lu H, Wang S, Sun M. Multi-type self-attention guided degraded saliency detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07 SE-AAAI Technical Track: Vision). 2020. pp. 13082–13089. DOI: 10.1609/aaai.v34i07.7010
- [37] Li G, Yu Y. Deep contrast learning for salient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. 2016. pp. 478–487
- [38] Wang L, Lu H, Wang Y, Feng M, Wang D, Yin B, et al. Learning to detect salient objects with image-level supervision. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. pp. 3796–3805. DOI: 10.1109/CVPR.2017.404
- [39] Zeng Y, Zhuge Y, Lu H, Zhang L, Qian M, Yu Y. Multi-source weak supervision for saliency detection. 2019. Retrieved from: <http://arxiv.org/abs/1904.00566>
- [40] He S, Jiao J, Zhang X, Han G, Lau RWH. Delving into salient object subitizing and detection. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017. pp. 1059–1067
- [41] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2016. pp. 770–778
- [42] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*. 2012. pp. 1097–1105
- [43] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2009. pp. 248–255
- [44] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *ICCV*. 2017. pp. 618–626
- [45] Berga D, Fernández-Vidal XR, Otazu X, Pardo XM. SID4VAM: A benchmark dataset with synthetic images for visual attention modeling. *ICCV*. 2019: 8788–8797
- [46] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research*. 2010. Vol. 9. pp. 249–256. Available from: <http://proceedings.mlr.press/v9/glorot10a.html>
- [47] Pan J, Cristian C, Kevin K, O’Connor NE, Torres J, Sayrol E, et al. SalGAN: Visual saliency prediction with generative adversarial networks. 2017