DOI 10.34658/9788366741102.11

Prediction of Natural Image Saliency for Synthetic Images

Ewa Rudak, Filip Rynkiewicz,
Marcin Daszuta, Łukasz Sturgulewski, Jagoda Lazarek

Lodz University of Technology
[1] *Institute of Information Technology*
Wólczańska 215, 90-924 Łódź, Poland
marcin.daszuta@p.lodz.pl
[2] *Institute Of Applied Computer Science*
Bohdana Stefanowskiego 18
lukasz.sturgulewski@p.lodz.pl

Abstract. *Numerous saliency models are being developed with the use of neural networks and are capable of combining various features and predicting the saliency values with great results. In fact, it might be difficult to replace the possibilities of artificial intelligence applied to algorithms responsible for predicting saliency. However, the low-level features are still important and should not be removed completely from new saliency models. This work shows that carefully chosen and integrated features, including a deep learning based one, can be used for saliency prediction. The integration is obtained by using Multiple Kernel Learning. This solution is quite effective, as compared to a few other models tested on the same dataset.*

Keywords: *computer games, artificial intelligence, image saliency, Human Visual Attention*

1. Introduction

The first attempts to create computational models of attention were done in 1980 by Anne Treisman and Garry Gelade, the authors of A feature-integration theory of attention, followed by the bottom-up model developed by Itti et al. [1], [2]. Later, the models started being evaluated by comparing them to collected human eye fixation data, yielding the creations of various datasets, such as MIT300 and CAT2000 [3], which contain diverse images, such as natural outdoor images and related human fixation maps. Currently, numerous saliency models are being developed with the use of neural networks and are capable of combining various features and predicting the saliency values with great results. One of them is the state-of-the-art work of Qi Zhao et al. [4] from 2019, which uses scene importance of scene information, affecting the directing of human gaze. A Structure- guided Approach to the Prediction of Natural Image Saliency shows the remarkable improvement in salient region detection that has been made since the development of the first models, which focused mostly on low-level feature extraction such as colour and intensity. Those features are important, yet in most cases they alone do not yield the best results, as can be deduced from the analysis of MIT Saliency Benchmark [3] and various related works. The concept of a saliency map was first proposed by Koch and Ullman [5] [7]. According to them, various elementary feature maps present the conspicuity within a particular feature dimension. Points taken from each of the feature maps, and related to a specific location, can be projected onto a unit of a new map. The combined information obtained from those different feature maps, showing the overall conspicuity of a point or object presented in an image, is then called a saliency map. Koch and Ullman believed that the determination of saliency of a location is primarily based on how it stands out from its surroundings, in terms of simple properties such as depth, colour and orientation. The authors also suggested that in order to combine the feature maps and obtain a saliency map presenting only the most conspicuous regions, neural networks should be used. In their article, they presented the Winner-Take-All network and its two possible implementations. Based on the above-mentioned work, a saliency map could be defined as a two- dimensional representation of the perceptual conspicuousness of corresponding regions in visual space [6]. It can be used for determining which points or objects are the most likely to attract the attention of the human visual system. Fig. 1 is an example of an image together with its saliency map created based on only three features: colour, intensity and orientation. Whether a point is salient is decided based on various saliency models, which will be described in detail in the next chapter of this paper.

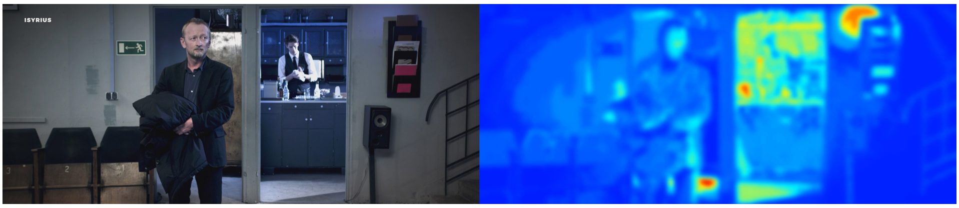


Figure 1. Left: original frame, right: saliency map of frame [7]

2. Saliency models

Since 1980, when the feature integration theory was created by Anne Treisman and Garry Gelade [8], a lot has changed and many new saliency models have been created. The aim of developing saliency models is enabling computational prediction of areas that will attract attention based on newly discovered or already known features. All of them require an input image, from which specific feature information is taken and then combined in order to create an output saliency map [9]. At the beginning, saliency models were focused completely on low-level features, which included for instance colour, intensity, orientation and depth. However, those old models failed to find many human eye fixations, which are important and should be taken into account in order to achieve more satisfying results. Recently, saliency benchmarks started presenting saliency models showing enormous improvements due to the usage of neural networks for predicting salient regions in images and videos. This has triggered the emergence of a large number of new saliency models taking advantage of the possibilities of artificial intelligence. Those new models are trained to predict saliency by combining the extraction of features, their integration and the prediction of resulting saliency values. Examples of neural network based models are: DeepGaze, SALICON (which will be described later) and DeepFix [1]. The first of them uses the features obtained from a neural network, which was trained to recognize objects in images. DeepFix is a Convolutional Neural Network learning features in order to predict saliency values for each pixel. Despite the noticeable advance in possibilities of saliency models, there are still things that are missing and that will have to be added to the ones developed in the future. To better understand what still has to be done in order to improve their performance, it is necessary to analyse a few exemplary saliency models, which I decided to base my work on.

2.1. Itti, Koch and Neibur model

This model, despite being quite obsolete and not achieving the best results in saliency benchmarks, is still mentioned in most of the modern related works. It is a bottom-up model, focusing on extraction and combination of low-level features selected from a set of ones that are close to mimicking early visual processing. The authors have based it on the ideas of Koch and Ullman [5] as well as on the feature integration theory [8]. Koch and Ullman proposed certain elementary features, including: colour, intensity, orientation, disparity and direction of movement, represented on various topographical maps. According to them, selective attention is capable of merging information from all these maps into one, called the central representation. In order to obtain the final map, a Winner-Take-All network should be used. The feature integration theory of attention suggests early, automatic and parallel noticing of features present in a stimulus, as well as the separate identification of objects. The first stage was called "preattentive" and occurs before people become conscious of the visible objects. The second stage proposed by Treisman and all. is known as the "focused attention stage", during which all the features noticed before are combined and thus, the whole object can be perceived. For context, the Itti, Koch and Neibur model makes a few assumptions based on the typical instinctive behaviour of human vision when it processes images. For the model to be effective, it must use these assumptions to predict the human visual search strategy. In order to ensure the correct functioning of this solution, it was tested on artificial images. For consistency, these inputs are processed so that they can be spatially scaled, which results in 9 different vertical and horizontal image reduction factors. The use of different scales means that multiscale feature extraction is possible. The R, G and B components are used to determine an intensity image which is necessary for a Gaussian pyramid that uses the scales. The Gaussian pyramid is a hierarchical representation of an image obtained by using Gaussian blur (blurring the image using Gaussian function) and scaling down of subsequent images. While the mentioned RGB components are acceptable for colour representation, there is the potential that the resulting data is not the optimal value for colour processing. Normalization is used to separate the hue from the intensity image but only at suitable intensities. The final step of this process is the creation of 4 colour channels (red, green, blue and yellow) demonstrated by a Gaussian pyramid for each channel.

For visual feature extraction, 3 sets of feature maps are needed. The first set focuses on the difference in intensity between a centre pixel and its surrounding pixels, both bright and dark centre contrasts being computed simultaneously. A second set concerns the colour channels. There are colours that when grouped together are instinctively opposed by our neurons. For example, these include red/green and yellow/blue. The occurrence of such opponency are computed simultaneously to generate the feature maps. The final set is created using orientation. Gabor pyramids determine the contrast in orientation between the centre and surrounding scales. The 3 sets results in 6 intensity maps, 12 colour maps and 24 orientation maps, therefore there are 42 feature maps in total. Each set forms a conspicuity map. A saliency map can be generated by combining these feature maps. Once the conspicuity maps are normalized, they are summed together which creates the saliency map. The maximum salient element is predicted as where the focus of attention will be. The performance of this model proved to be robust when tested by comparing the predictions with results obtained using an eye tracking device.

2.2. Boolean Map Saliency (BMS)

In 2013, a new bottom-up saliency model was presented and described by J.Zhang and S.Sclaroff [10]. Unlike many previous ones, it does not focus on extracting simple features like contrast and rarity; instead it takes into account global topological cues described in Gestalt psychological studies. These studies suggest that figures present in an image are more likely to be noticed than elements of background and that focal attention is not necessary for figure-ground assignment [11]. The authors of this work decided to focus on surroundedness cue, which is supposed to have an influence on the figure-ground segmentation. The main idea of surroundedness is that areas are perceived as figures when they can be seen as surrounded by others. In order to measure it, Boolean Map Saliency (called BMS for short) uses image characterization by a set of Boolean maps, inspired by the Boolean Map Theory of visual attention [12]. The computation of an attention map is based on binary image processing techniques, after which a set of randomly selected Boolean maps is used to model the saliency. The result can be further used for salient object detection. The Boolean Map Theory suggests the existence of two visual attention limitations, namely access and selection. The first one is related to the constraints on the available ways to create a spatial representation dividing the visual field into two subsets (known as the Boolean map). These subsets represent regions that are either selected or not. The selection refers to directing of ways in which the map is created, which can be done either through the selection of one feature value per dimension or by using intersection or union to combine the Boolean map with the result of the single feature selection.

The saliency is modelled according to the following equation:

$$\overline{A} = \int A(B)p(B|I)dB \tag{1}$$

Where $\overline{A}$ is the mean attention map, I is the input image, B is a Boolean map and $A(B)$ is an attention map.

The algorithm can be described in the following way:

- An input image I is taken
- A set of Boolean maps is generated
- Attention maps are computed for each of the Boolean maps
- A mean attention map is created
- Post-processing is applied in order to obtain a saliency map

The set of Boolean maps is obtained by randomly thresholding the feature maps of the input image. In order to compute the attention map, 1 is assigned to the union of surrounded regions (ones with closed outer contours) and 0 to the rest of the map, which later has to be normalized. Boolean Map Saliency scored good results in many tests and takes the average of 0.38 seconds per image to process on a computer with only 2 GB of memory, meaning that it can be used on most of the devices. This model is ranked very high in the MIT saliency benchmark as compared to other bottom-up models not using neural networks.

2.3. Saliency in context (SALICON)

The creators of SALICON [13], aware of the semantic gap between the conventional saliency models visual attention prediction and actual human behaviour, focused their work on narrowing it with the use of Deep Neural Networks (DNN). By the time this new model was developed, some other ones were already using object detection in order to improve the results of saliency prediction. Unfortunately, despite being more advanced than the traditional models, based on low-level features, each of those detectors was trained only for a specific category. Thus the need to create a model that would be capable of learning and recognizing various semantics emerged. Deep Neural Networks are known for their ability to learn image representations [14]. They have constrained connections between their numerous layers and are complex, therefore enabling the occurrence of various nonlinearities. Qi Zhao et al. introduced a DNN architecture integrating saliency prediction to a network that had been trained to recognize objects. During the learning process, it was possible to apply different evaluation metrics, such as KLD, CC and NSS. The first of these was later chosen for further training due to it showing the best performance, especially with AUC and sAUC scores, commonly used in saliency benchmarks. The architecture of the work was based on three DNNs used for object recognition: GoogLeNet, AlexNet and VGG-16. As a result, the latter showed the best performance and therefore was used in comparison to other saliency models together with KLD. SALICON achieved high results when compared to other state-of-the-art models and still persists in one of the top positions in the MIT saliency benchmark. A free demo of this DNN-based saliency model is available on the website [15] together with a database of images that can be used. It is also possible to upload any other image and test this solution.

2.4. Structure guided saliency model

One of the most recent works exploring new ways of achieving improved results in visual attention prediction is A Structure-guided Approach to the Prediction of Natural Image Saliency. It underlines the importance of scene structure, which is said to provide contextual information in directing gaze. Based on studies [16] on the perception of scene layout properties, it has been discovered that human beings rapidly and automatically analyse the structure of a scene when searching for objects. This suggests that from the first moment when a person becomes exposed to a visual stimulus, the anticipation of the important content begins. These layout properties include openness, depth and perspective. The study on the effects of scene structure on directing gaze was done by first building a dataset of images showing common outdoor scenes and recording eye movement data from 15 subjects and by using mouse tracking. The second method involves recording mouse movement data capable of mimicking the way in which human beings see scenes. It was proposed by Qi Zhao et al. [17], based on the method used for website analysis and required specific design of the stimuli in order to make users move their mice when shifting attention. Through observations of the results it was possible to determine which scene level features have an impact on directing gaze by humans. The dataset consisted of 2500 natural images depicting outdoor scenes selected from the SUN2012 database (examples of natural images from this database are presented in Fig. 4.8). All of them were then rescaled to 256 x 256. Afterwards, the eye tracking data was collected for 500 of them and mouse tracking was used on the remaining 2000 in order to reduce the cost of data collection.

This saliency model incorporates two of the low-level features suggested by Itti and Koch, namely colour and intensity. Moreover, it also uses Gabor filters, which are said to be useful for texture representation and discrimination, and the Boolean Map Saliency model, due to the fact that it still can be considered to be very advanced and accurate.

The integration of various features is done using multiple kernel learning (MKL) once all of them have been extracted. The algorithm applied in this case is called simpleMKL. Both MKL and simpleMKL will be described later. The final saliency map (S) is obtained by:

$$S = \{\max(f(x), 0) * g\} \circ L \tag{2}$$

Where g is a Gaussian mask, L is the DTS map and $\circ$ is the Hadamard product operator.

The structure guided model was tested on Toronto and MIT datasets and evaluated using AUC, sAUC, CC and NSS metrics, which will be described later. The experimental results show that the proposed scene structural features improve the performance of saliency prediction in the case of natural outdoor scenes and can be used in combination with other models, including deep learning ones. For instance, the NSS score of Boolean Map Saliency model alone, tested on the MIT dataset, was equal 1.386. However, when this model was combined with the structure guided one, this NSS value changed to 1.689.

2.5. Proposed solution

Based on the analysis of existing solutions, especially the structure guided saliency model, it was possible to determine which models are more and which are less successful in accurately predicting conspicuousness of various regions in an image. As suggested in [4], deep learning saliency models combined with the possibilities of the structure guided approach might lead to even better results. Based on this idea and the feature integration theory [8], I decided to create a model that would benefit from its ability to find salient regions using a combination of low-level features, a deep learning algorithm and the scene structure features. The first of the features that proved worth including was Boolean Map Saliency. Due to the fact that this model scored well in the MIT saliency benchmark, despite not using neural networks, being based on a simple algorithm and a relatively easy implementation, it appeared beneficial for a new model. An example of an image and its saliency map created using the BMS algorithm is presented below (Fig. 1 and Fig. 2).

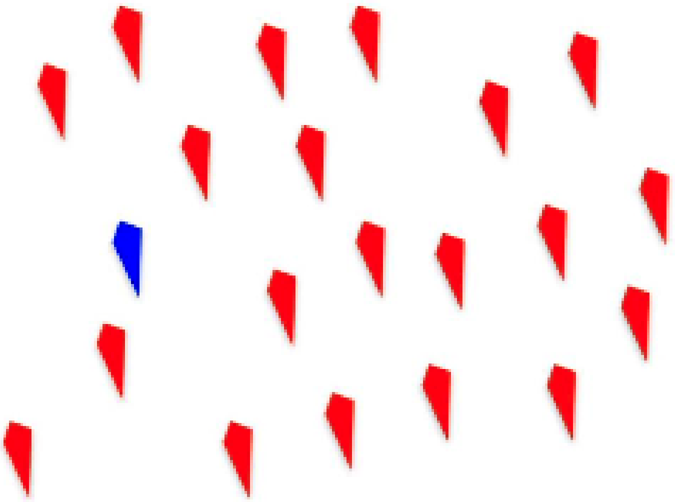


Figure 2. A synthetic image taken from the MIT database

The implementation of this algorithm was written in C++ with the use of OpenCV library and can be compiled in Matlab by using a Mex file. The user

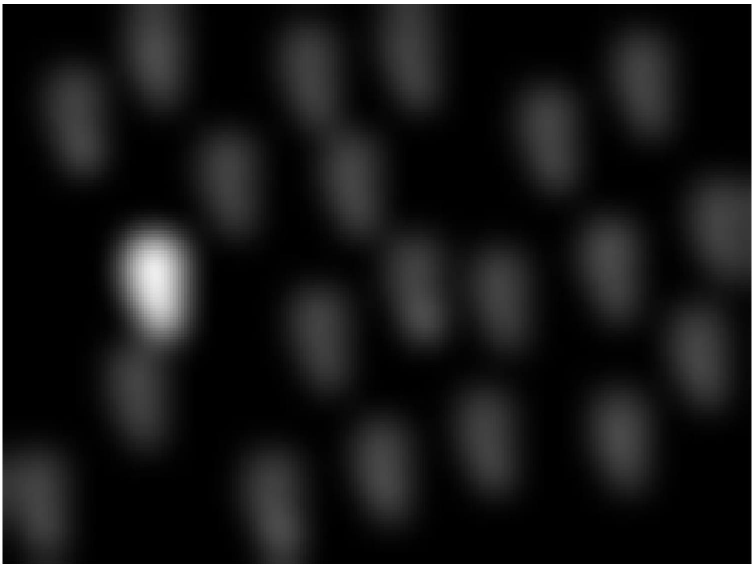


Figure 3. A Boolean Saliency Map, correctly predicting the stronger conspicuousness of an element with a different colour.

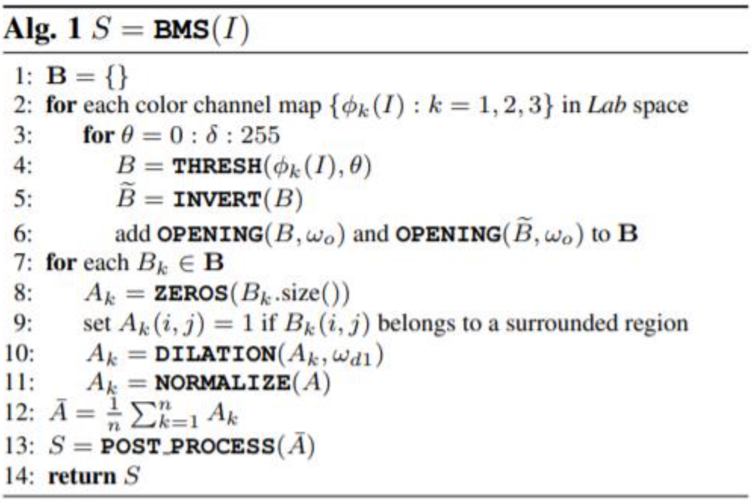
has to specify certain parameters for output generation. The selected values are presented in table 1.

Parameter	**value**
Sampling step size δ	8
Dilation width ω_{d1}	7
Dilation width ω_{d2}	9
Standard deviation for Gaussian blurring σ	9
Colour space	2(Lab)
Whitening	1 (true)
Maximum dimension	400

Table 1. The parameters and values chosen for the BMS feature map

The default parameters, suggested by the authors of the Boolean Map Saliency model were used. Changing values of sampling step size and maximum dimensions would lead to changes in accuracy and speed.

The function getAttentionMap() is responsible for the computation of attention maps, which is achieved by assigning 1 to surrounded regions and 0 to the remaining parts of the map. The algorithm is shown in Fig. 4



```
Alg. 1 $S = \mathbf{BMS}(I)$
$\mathbf{B} = \{\}$
for each color channel map $\{\phi_k(I) : k = 1, 2, 3\}$ in Lab space
    for $\theta = 0 : \delta : 255$
        $B = \mathtt{THRESH}(\phi_k(I), \theta)$
        $\widetilde{B} = \mathtt{INVERT}(B)$
        add $\mathtt{OPENING}(B, \omega_o)$ and $\mathtt{OPENING}(\widetilde{B}, \omega_o)$ to $\mathbf{B}$
for each $B_k \in \mathbf{B}$
    $A_k = \mathtt{ZEROS}(B_k.\mathrm{size}())$
    set $A_k(i, j) = 1$ if $B_k(i, j)$ belongs to a surrounded region
    $A_k = \mathtt{DILATION}(A_k, \omega_{d1})$
    $A_k = \mathtt{NORMALIZE}(A)$
$\bar{A} = \frac{1}{n} \sum_{k=1}^{n} A_k$
$S = \mathtt{POST\_PROCESS}(\bar{A})$
return $S$
```

Figure 4. The BMS algorithm

Another of the chosen low-level features was intensity. As one of the features used in the Itti-Koch model, it complements the BMS or colour map and is relatively easy to compute. It has been selected due to its ability to detect certain salient areas, which were not highlighted in the case of the remaining features. In order to obtain the intensity map, a function compute_ittifeature() is called. It uses the main function from the Graph-Based Visual Saliency implementation [18] with the useIttiKochInsteadOfGBVS parameter set to 1, resulting in computation of three Itti-Koch features: colour, intensity and orientation. Selected ones are then resized and saved in a form of images. The structure guided model does not yet appear in the MIT saliency benchmark, however, authors have provided information about the scores obtained by using three evaluation metrics: shuffled AUC, NSS and CC after testing the model on two datasets: MIT and Toronto. Below the evaluation results for the MIT dataset are shown (Tab. 2) , as presented in [4].

Model	NSS	sAUC	CC
Judd	1.396	0.597	0.578
BMS	1.386	0.687	0.537
Structure guided	1.742	0.725	0.671

Table 2. Evaluation results for selected algorithms

This state-of-the-art model extends the possibilities of previously implemented ones by adding structural features, which are very useful and positively affect the results in case of outdoor scenes. There are, however, cases in which this model fails to correctly classify a scene or does not successfully predict the location of a convex part or the vanishing point. It also would not improve the results for indoor scenes or ones presenting a single object. Various existing models incorporate face (both human and animal), object and text detection due to the fact that it has been proved that the human visual system searches for them and they do attract visual attention [19]. Thus, I have decided to include an additional feature capable of detecting conspicuous areas by using a deep learning algorithm. SALICON, as already mentioned, has an architecture based on Deep Neural Network, able to learn high-level features for saliency prediction. It is able to correctly detect conspicuous regions such as faces, text and animals and is rated as one of the best in the MIT saliency benchmark. In order to decide whether this model would serve as a valuable feature, it had to be tested on various images and the results compared to those of other chosen features. For this purpose, several images from available databases were used on the SALICON demo [15], proving the usefulness of the model and its correctness, especially in the case of face detection. An example of a correct face detection is presented in Fig. 5.

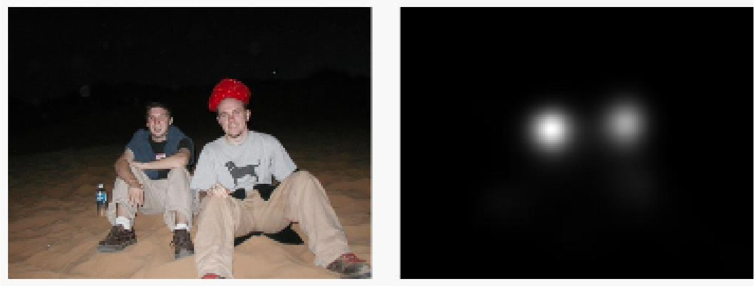


Figure 5. An example of correctly detected faces in an image by SALICON

The chosen features should not require a lot of time to compute, therefore making the feature integration process easier and faster.

2.6. Results

The proposed model was used to obtain various saliency maps of outdoor and indoor scenes from the selected dataset. The results have shown the capability of this solution, proving that it can achieve good results without the need to combine a large amount of different features, which positively affects the overall performance of the model. Tab. 3 presents the comparison of time needed to train the model in order to obtain final saliency maps for the presented solution and all features used in the structure guided model. Both results were obtained on a computer with 16 GB of memory and an Intel Core i5 2,4 GHz 8th generation processor.

Model	**Estimated training time**
Proposed solution	1923.16 seconds
Structure guided	u

Table 3. Comparison of estimated training time for feature integration.

This shows that the suggested model requires a lot less time for the feature integration process than the structure guided one (less than 30% of the time required by the more complex model). The new model only uses 5 features, as compared to one that takes into account 11. In order to evaluate the results, it was necessary to carefully analyse them and to notice differences between both models, that could provide hints on which features exactly are either more or less beneficial for the overall score. First thing that should be noticed is the feature negatively affects certain saliency maps by making them less distinctive. It was not used in the model suggested by me, therefore all the resulting maps have better defined salient regions. Most of the results that show the correctly found most salient regions in images presenting human beings or animals were positively affected by the deep learning feature. However, in many cases other, less conspicuous regions that attract human visual attention were not shown on saliency maps.

In case of scenes without elements such as people, animals and text, the proposed model was able to find the most salient regions, which is the result of including SALICON feature. However, it failed to detect the most salient element in a synthetic, which was correctly marked in the structure guided model. This suggests that the colour feature should be incorporated into my model in the future to resolve similar cases, as BMS alone might not serve as a strong enough feature in the training feature.

Due to the presence of structure features, the proposed model was still capable of detecting salient regions due to the presence of convex parts and vanishing points. The failures in these cases were caused by failures in detecting these elements by the structure feature itself. The analysis of results was necessary in order to find out which feature has a positive influence on the final saliency maps. It has shown that the colour feature should be included in the future and the training might be more biased towards the structure and deep learning features in order to improve the final result. The visualizations of Normalized Scanpath Saliency of obtained maps also had to be analysed. More yellow regions indicate higher similarity between the fixation and saliency maps. The NSS visualizations were created for all the saliency maps for both models.

Additionally, the NSS score was calculated for each of the obtained saliency maps due to the fact that it is a lot easier to compare the numerical values received by various models. The results are presented in Tab 4.

Saliency model	NSS score
My model	0.6745
Structure guided	0.5404
SALICON (alone)	0.6415

Table 4. Calculated NSS scores for different saliency models

Based on the presented results it is possible to draw a conclusion that the saliency model described in this paper is able to correctly predict salient regions in images, using only 5 features and therefore shortening the necessary training time.

3. Conclusions

Most of the state-of-the-art saliency models use neural networks in order to improve their performance. In fact, it might be difficult to replace the possibilities of artificial intelligence applied to algorithms responsible for predicting saliency. However, the low-level features are still important and should not be removed completely from new saliency models. The presented solution does not require a lot of computing time for feature integration process, which is certainly one of its advantages. The limitation of features to only a few ones has also positively affected the easiness of result analysis it was a lot simpler to determine which feature could possibly be problematic or beneficial for the whole model. This work shows that carefully chosen and integrated features, including a deep learning based one, can be used for saliency prediction. The integration is obtained by using Multiple Kernel Learning. This solution is quite effective, as compared to a few other models tested on the same dataset. However, improvements can be added in the future, including more biased learning process to focus on more important features (pointing to more salient regions) and the low-level colour feature based on Itti-Koch model. The failure cases occurred when the separate features failed, thus any improvements in the algorithms used to obtain the feature saliency maps should also increase the probability of successful salient region prediction in suggested model. Nevertheless, even without any additional changes it can detect faces, text, scene structural features and objects, which are generally known to be conspicuous.

References

[1] Bylinskii, Z., Recasens, A., Borji, A., Oliva, A., Torralba, A., and Durand, F. Where should saliency models look next? volume 9909, pages 809–824. 2016. ISBN 978-3-319-46453-4. doi:10.1007/978-3-319-46454-1_49.

[2] Itti, L., Koch, C., and Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20:1254 – 1259, 1998. doi:10.1109/34.730558.

[3] Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., and Torralba, A. Mit saliency benchmark.

[4] Liang, H., Jiang, M., Liang, R., and Zhao, Q. A structure-guided approach to the prediction of natural image saliency. *Neurocomputing*, 378:441–454, 2020. doi:10.1016/j.neucom.2019.09.085.

[5] Koch, C. and Ullman, S. *Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry*, pages 115–141. Springer Netherlands, Dordrecht, 1987. ISBN 978-94-009-3833-5. doi:10.1007/978-94-009-3833-5_5. URL https://doi.org/10.1007/978-94-009-3833-5_5.

[6] Veale, R., Hafed, Z., and Yoshida, M. How is visual salience computed in the brain? insights from behaviour, neurobiology and modeling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372:20160113, 2017. doi:10.1098/rstb.2016.0113.

[7] Rogalska, A. and Napieralski, P. The visual attention saliency map for movie retrospection. *Open Physics*, 16(1):188–192, 2018. doi:doi:10.1515/phys-2018-0027. URL https://doi.org/10.1515/phys-2018-0027.

[8] Treisman, A. M. and Gelade, G. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980. ISSN 0010-0285. doi:https://doi.org/10.1016/0010-0285(80)90005-5. URL https://www.sciencedirect.com/science/article/pii/0010028580900055.

[9] Koehler, K., Guo, F., Zhang, S., and Eckstein, M. What do saliency models predict? *Journal of vision*, 14, 2014. doi:10.1167/14.3.14.

[10] Zhang, J. and Sclaroff, S. Saliency detection: A boolean map approach. pages 153–160. 2013. doi:10.1109/ICCV.2013.26.

[11] Palmer, S. *Vision Science: From Photons to Phenomenology*, volume 1. 1999.

[12] Huang, L. and Pashler, H. A boolean map theory of visual attention. *Psychological review*, 114 3:599–631, 2007.

[13] Huang, X., Shen, C., Boix, X., and Zhao, Q. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. pages 262–270. 2015. doi:10.1109/ICCV.2015.38.

[14] Krizhevsky, A., Sutskever, I., and Hinton, G. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 2012. doi:10.1145/3065386.

[15] Jiang, M., Huang, S., Duan, J., and Zhao, Q. Salicon: Saliency in context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1072–1080. 2015. doi:10.1109/CVPR.2015.7298710.

[16] Ross, M. G. and Oliva, A. Estimating perception of scene layout properties from global image features. *Journal of vision*, 10 1:2.1–25, 2010.

[17] Jiang, M., Huang, S., Duan, J., and Zhao, Q. Salicon: Saliency in context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1072–1080. 2015. doi:10.1109/CVPR.2015.7298710.

[18] http://www.vision.caltech.edu/ harel/share/gbvs.php.

[19] Bylinskii, Z. and Tilke Judd, A. T. F. D., Aude Oliva. What do different evaluation metrics tell us about saliency models? *arXiv*, 2017.