Saliency and Object Detection

Phutphalla Kong
Department of Information and Communication Engineering
Institute of Technology of Cambodia
Phnom Penh, Cambodia
phutphalla@itc.edu.kh

Matei Mancas
Numediart Institute
for Creative Technologies
Faculty of Engineering (FPMs),
University of Mons (UMONS)
Mons, Belgium
matei.mancas@umons.ac.be

Seng Kheang
Department of Information and Communication Engineering
Institute of Technology of Cambodia
Phnom Penh, Cambodia
sengk@itc.edu.kh

Bernard Gosselin
Circuit Theory and Signal Processing Lab
Faculty of Engineering (FPMs),
University of Mons (UMONS)
Mons, Belgium
bernard.gosselin@umons.ac.be

***Abstract*— Visual attention allows the human visual system to effectively deal with the huge flow of visual information acquired by the retina. Since the years 2000, the human visual system began to be modelled in computer vision and it became part of artificial intelligence: while learning focuses on repetitive data which can easily be modeled, computational attention focuses on abnormal, rare and surprising data which can hardly be learnt. Attention is a product of the continuous interaction between bottom-up and top-down information. While the bottom-up information has been extensively investigated through saliency models, top-down influence on visual attention has been less investigated. This paper intends to study the influence of object-based (faces and text) top-down information on bottom-up saliency maps. It proposes a simple yet effective fusion scheme that can be applied on any bottom-up saliency model depending on the object detector effectiveness and the object size. The evaluation results show that it is possible to highly improve classical bottom-up saliency models with the arrival of better object detectors. In the future, such attention models can become as effective as deep-learning based attention models while keeping them more generic and avoiding underestimating bottom-up features.**

Keywords— saliency, top-down attention, bottom-up attention, visual attention, data fusion, eye-tracking, eye fixations.

I. Introduction

Computational visual attention tends to mimic human visual attention and focuses on the more informative and important parts of images. It has been the subject of various studies in a wide range of research fields such as psychology, neuroscience or computer vision. In computer vision, the main approach to the implementation of visual attention includes bottom-up and top-down information[1], however, while bottom-up attention was investigated a lot [4][5][6][7][8][9][11], there were only a few experiments using top-down information in the literature. This is probably because before the arrival of deep learning in attention in 2014, the top-down detectors were not good enough, and afterwards most researchers focused on obtaining an end-to-end deep learning saliency model which naturally integrates top-down information.

There are various kinds of top-down information which can be used in addition with bottom-up saliency [1] such as location-based, contextual-based or object-based models. In this paper, we focused specifically on object-based top-down attention and especially faces and text.

The combination of face detection and low-level saliency provides already results improvements in [2]. The linear combination was weighted to give to faces the same weight that each one of the three bottom-up conspicuity maps (orientation, colour, intensity) which means that the face map global weight was quite low. This helped the authors to deal with false positives from the face detector used at that time which was not optimal. In [10], the author showed that the high-level features such as faces and people can enhance the model performance, but there was no any precise information related to the relative importance of those features. The author also stated that using a bad object detector could clearly decrease the model performance if it produces too many false positives. In [3], the authors dealt with the importance of people and cars for saliency detection. In [16], the authors introduced the idea of the use of object symmetry as top-down attention in images. In [12], target object features from the Pascal VOC object database are learned using a CRF-modulated dictionary. The saliency maps were really focused on the objects with a very high weight.

In computer vision, Deep Neural Networks (DNNs) have changed the saliency paradigm since 2014. The deep features were first used in eDN model [17]. DeepGaze1 model [18] then showed that DNN features trained on object recognition are very useful for saliency detection. This finding seems logical as objects are most of the time regions of interest. Since then a variety of models used fine-tuned mixes of features from several deep learning models. These DNN-based models naturally incorporate top-down information during learning (such as faces and text for example). It seemed that DNNs were the perfect solution to improve the performance of classical bottom-up models.

However, in [19], the authors showed that the importance of bottom-up attention was underestimated by DNN-based models. Indeed, a simple bottom-up model can outperform a state-of-the-art DNN model on images containing less top-down information. This demonstrated that DNNs too much neglect the bottom-up aspect of visual attention and are mostly trained to detect the very attractive top-down objects than to really detect saliency. Moreover, they cannot easily adapt to images which are very different from the ones they were trained on and finally DNNs have the structural issue to provide a result that cannot really be explained in an explicit way.

[1] Bottom-up information is also known as reflex exogenous reaction, while top-down information is known as reflexive endogenous information.

The authors of [13] provide cues about relative importance of features based on a manual segmentation of the image dataset and is the basis of the work on this paper. In [13], the authors build a saliency model mixing their own bottom-up approach and several higher-level features. While in [13] weights of different features are computed within a specific model and cannot be used in other contexts, the purpose of this paper is to quantify the relative importance of two features proven in [13] to be very influential (faces and text) so that they can be used and integrated very easily to any general bottom-up saliency model, not necessarily in ours. In addition, we also study the size of faces and text assuming that the size is not an independent variable. The question we address here is: how to add in a simple yet effective way an object detector result to any bottom-up saliency model?

The remainder of this paper is organized as follows. In Section II, we describe the high-level features and detectors that are used. In Section III, we deal with the methods and experiments used mix bottom-up and top-down information based on several kinds of detectors. The results and discussion are finally presented in Section IV.

II. High-level Features and Detectors

Low- and high-level information are both important to predict human gaze accurately [19][13]. In [13], the relative importance of different features was used to evaluate the model performance, which was computed by a linear SVM classifier. In terms of importance, it shows that face, text, and gaze direction are the three main features. In addition, colour, orientation, or intensity still have an interesting influence especially when human faces and text were absent [19]. However, in [13], the result cannot be easily applied to any bottom-up attention model, while here we intend to be able to use them in a generic way with any model. In sections II.A and II.B text and faces features are further described.

For the object detectors, we focus on 1) faces and 2) text. Inside those two important features, we try to understand the importance of the size (big faces and text versus small faces and text). The OSIE dataset (Object and Semantic Images and Eye-tracking) [13] provides us with a set of images and the manually segmented masks for text (Fig. 1b) or faces (Fig. 1c) along with the eye tracking fixation map (Fig. 1d). The different detectors that we used are detailed in Section II.C.

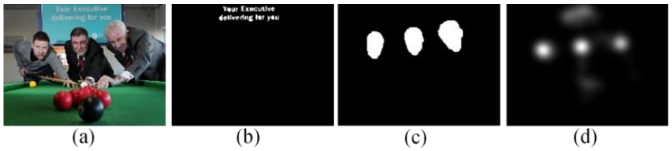
Your Executive delivering for you
Your Executive delivering for you
(a)
(b)
(c)
(d)

Fig. 1. Extracting text and face features and eye-tracking fixation map from OSIE dataset. (a) Input image, (b) Text features, (c) Face features, and (d) Eye-tracking fixation map.

A. Text features

Based on the OSIE dataset, we only select the images containing text (Fig. 2). When checking the corresponding eye-tracking maps, big text seems much more interesting than small text (around twice more interesting). This consideration pushed us to check the difference between big (Fig. 2b) and small (Fig. 2c) text. We used the OSIE masks to separate text regions between big text (more than 29 pixels in height) and small text (less than 29 pixels in height). This threshold depends of course on the image size, but all the images in the database are of the same size here.

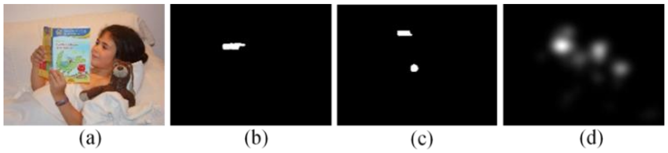

Fig. 2. Defining big and small text features and eye-tracking fixation map from the OSIE dataset. (a) Input image, (b) Big text features, (c) Small text features, and (d) Eye-tracking fixation map.

B. Face features

In the same way, we use the manually segmented masks for the images containing faces (Fig. 3). Even if the difference in terms of eye-tracking maximum is less obvious between big faces and small faces than between big text and small text, we separate big faces (we used a threshold of 76 pixels in height) from small faces (less than 76 pixels in height) the same way as text. Again, the size of the images is always the same here, but for other datasets this threshold should be computed relatively to the mage size to be used on other datasets. Here we only take into account the frontal faces as heads viewed from rear or from the side have less chances to be correctly detected by an automatic face detector.

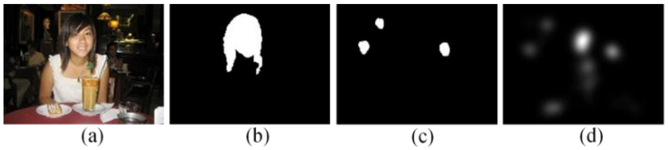

Fig. 3. Defining big and small face features and eye-tracking fixation map from OSIE dataset. Big and small face features are defined by us as the text features. (a) Input image, (b) Big face features, (c) Small face features, and (d) Eye-tracking fixation map.

C. Object detectors

We first define a perfect detector which is simply the human-based masks already segmented in [13].

For faces, we use a state-of-the-art face detector based on the DLIB library [22]. We used the classical Histograms of Oriented Gradients (HOG) feature followed by a SVM classifier which has a good face detection rate [21]. On this detection, we added the face template approach based on a cascade of classifiers from [15] which exhibited good results for frontal faces, with few false positives.

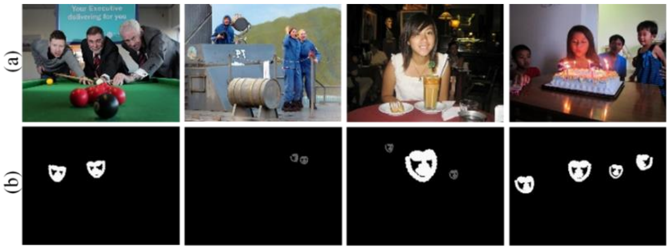

Fig. 4. Applying face detection method from [15]. (a) Input images, (b) Result of face detection. The results contain either big faces (brighter), either small faces, and either both big and small faces.

For text detection, we used an older approach which is integrated into the OpenCV library [14]. This detector used Extremal Regions (ERs) which are robust to several image transformations. A second step is used in the algorithm: OCR helps to improve overall results. However, in this paper we did not use any OCR results. For real-life images, this detector results are poor with both misdetections and false detections.

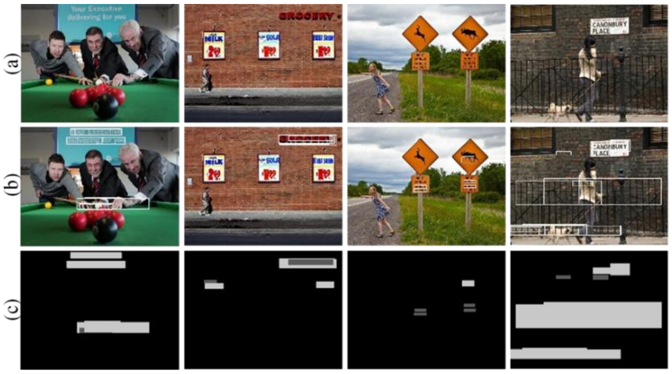

Fig. 5. Applying text detection method from [14]. (a) Input images, (b) Result of text detection inside white bounding boxes, and (c) Converting text detection areas into white to indicate text features. The results contain both big (brighter) and small texts.

Thus, we can compare the results between a perfect detector (for text and face features), a good detector (face detection), and a poor detector (text detection).

III. EXPERIMENT

The experiment intends to provide us with a clear view on how top-down information effects bottom-up information by adding text and face detection. It can be divided into two questions: 1) how to extract weights for different kind of top-down information and 2) how to mix the top-down information to bottom-up in a simple way.

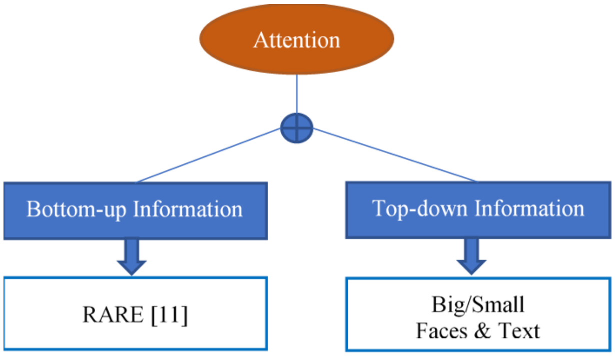

Fig. 6. Components of bottom-up and top-down attention used in our experiment.

To do so, we choose the top-down information as described in the previous section (Fig. 6, bottom-right), while the bottom-up saliency comes from the RARE model [11] (Fig. 6, bottom-left). This approach was purely bottom-up (no additional centred gaussian or learning-based information), and it considered both local information and global information through a rarity approach. By considering the MIT saliency benchmark [23], this model is bellow most of the DNN-based models.

A. *Top-down features weight*

While in [13], the features weight is computed by the means of a classifier, we choose here to use the experimental data that we have in the OSIE dataset to extract a meaningful individual weight for each feature of interest. For that purpose, we decided to measure the average maximum eye gaze attractivity on all the OSIE images for big and small text and face masks. As it can be seen in the schema in Fig. 7, the eye-tracking map is multiplied by the binary mask which will provide the eye-tracking intensity on the object of interest. Then the maximum of these values is averaged over all the images in the dataset providing a weight for the given feature. For weight of big text, after 75 images we are stabilized between 0.75 and 0.78, and for weight of small text, after 75 images we are stabilized between 0.31 and 0.34. For weight of big face and small face, after 75 image we stabilized between 0.81 and 0.84 and between 0.64 and 0.67, respectively. As a result, between 75 images and 100 images are enough to get stable weights which do not depend a lot on the images we add.

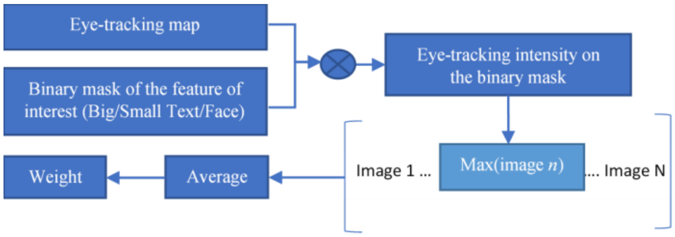
Eye-tracking map
Binary mask of the feature of interest (Big/Small Text/Face)
Eye-tracking intensity on the binary mask
Image 1 ...
Max(image n)
... Image N
Weight
Average

Fig. 7. Object binary mask is used along with the eye-tracking map to extract the maximum eye-tracking value for the object. This value, averaged on all the images will provide a weight for a given object.

B. *Top-down and bottom-up fusion*

Once the weight was computed for one of the objects among small text (ST), big text (BT), small faces (SF), and big faces (BF), the question is how to make a fusion between this information and the bottom-up saliency map.

First, as described in Fig. 8, for each feature, we split the small and big masks and then smooth them in order to obtain and image close to the fuzzy bottom-up saliency map.

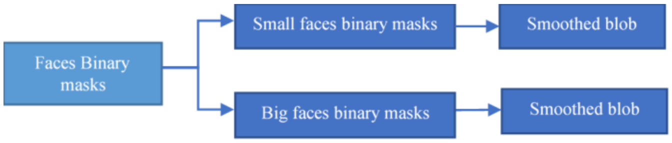

Fig. 8. For face and text, the binary masks are split between big and small masks and then low-pass filtered to provide a smoothed result before being fused with the bottom-up saliency map.

We made linear combinations between bottom-up information (saliency maps) and top-down information (text and face detection). We generated results of saliency maps using RARE [11] and 1) text detection using both [14] and the masks as described in section II.A and 2) face detection using both [15] and the masks as described in section II.B.

To make a simple fusion between bottom-up saliency maps (SM) and top-down information (faces alone, text alone or both), we used linear combinations which are easy to implement. The weights were either the same for text (big and small) and faces (big and small), either different by using the results that we obtained in section III.A which are given the following formulas:

$$(SM + ATF) \,/\, 2 \tag{1}$$

$$(SM + BTF) \,/\, 2 \tag{2}$$

$$(SM + STF) \,/\, 2 \tag{3}$$

$$SM + (wSTF \,/\, wBTF) * STF + BTF \tag{4}$$

where *SM* is bottom-up saliency maps computed from RARE [11], *ATF* is either all text (big and small), either all face (big and small), either all text and face depending on the experiment, *BTF* is either big text, either big faces, either all big text and big faces depending on the experiment, *STF* is either small text, either small faces, either all small text and small faces depending on the experiment. *wSTF* and *wBTF* use the weights found in section III.A for either small text or small faces (*wSTF*) and for either big text or big faces (*wBTF*).

To also test the impact of the detector accuracy, we divided our experiment into three different parts: perfect detector, good detector, and bad detector. For good and bad detector, we use the state-of-the-art detector [15] and [14], respectively while for the perfect detectors we used the masks from [13] as shown in Fig. 8 for both face and text.

IV. Results and Discussion

A. Weights for face and text

To get ideal weights for big and small text and face, we used the method described in Fig. 7. As a result, we obtained a weight of big text (wBT) = 0.7871, of small text (wST) = 0.3221, of big face (wBF) = 0.8159, and of small face (wSF) = 0.6457. We can see that the difference between big text and small text is more important than big face and small face. Big face is a little more important than big text, but the difference is not very significant.

B. Perfect detector

For the perfect detector, we did three experiments. In the first one, we combined the bottom-up saliency map (SM) with text alone, then with face alone, and finally with both text and face.

We used several metrics to evaluate the bottom-up attention model object-based top-down attention by making correlation between some different results (from fusion algorithms) and the eye-tracking Fixation Maps (FM). For those metrics, we used Correlation Coefficient (CC), Kullback-Leibler Divergence (KLDiv), Normalized Scanpath Saliency (NSS), Similarity, and Area Under the ROC curve from Judd (AUC_J). Those metrics provide some complementarity and are well described in [20]. For CC, NSS, Similarity, and AUC_J, the higher value is the best, for KLDiv, the lower value is the best.

The results are summarized given the three different experiments in Tables I (SM with text alone), II (SM with face alone), and III (SM with both text and face). The first line corresponds to the comparison between the bottom-up saliency map (SM) and the eye tracking fixations map (FM). The second line corresponds with the comparison of all features (all text (AT) in Table I, all faces (AF) in Table II and all text and face (ATF) in Table III). The third line is a comparison between FM and big text (BT) in Table I, FM and big face (BF) in Table II, and between FM and big text and face (BTF) in Table III. The fourth line in Tables I and II represent the comparison between FM and small text (ST) and FM and small faces (SF), respectively. The final line is the comparison of the weighted fusion for big and small text (wBST) in Table 1, for big and small face (wBSF) in Table II and all big and small text and face (wTF) in Table III.

A first global remark is that all metrics are very coherent, and they provide almost the same relative rank for all measures.

TABLE I. Result between Bottom-up and Text Detection

Correlation	**Metris**				
	CC	***KLDiv***	***NSS***	***Similarity***	***AUC_J***
SM, FM	0.4683	1.0591	1.5364	0.4364	0.8365
AT, FM	0.5042	1.0140	1.7013	0.4514	0.8452
BT, FM	0.5058	1.0151	1.7008	0.4504	0.8444
ST, FM	0.4666	1.0587	1.5420	0.4378	0.8372
wBST, FM	**0.5061**	**1.0127**	**1.7081**	**0.4517**	**0.8454**

In Table I, it indicates that adding information about small text brings nothing to the result because result of small text is a little less good than the bottom-up saliency map alone in some metrics. On the other hand, adding big text provides an important and improvement result in the CC metric compared to result when adding both small and big text (AT). For all metrics, the use of the weights provides the best results of all.

TABLE II. Result between Bottom-up and Face Detection

Correlation	**Metrics**				
	CC	***KLDiv***	***NSS***	***Similarity***	***AUC_J***
SM, FM	0.4683	1.0591	1.5364	0.4364	0.8365
AF, FM	0.5352	0.9790	1.8334	0.4581	0.8494
BF, FM	0.5277	0.9876	1.7890	0.4553	0.8478
SF, FM	0.4762	1.0514	1.5851	0.4396	0.8379
wBSF, FM	**0.5354**	**0.9786**	**1.8348**	**0.4582**	**0.8494**

In Table II, it indicates that adding information about small face brings this time a small improvement to the bottom-up saliency map alone and is never negative. However, adding big face provides an important result improvement. Moreover, adding both big and small face also brings improvement. The best case, for all metrics, the use of the weights provides the better results of all.

In Table III, we did not compute the result for small text and face since it was always smaller than big text and face. We just kept here the best combinations: ATF for all text and face, BTF for only big face and text, and the weighted text and face (wTF). While ATF is always a little better than BTF, the weighted version is even better.

TABLE III. RESULT BETWEEN BOTTOM-UP, TEXT, AND FACE DETECTION

Correlation	**Metrics**				
	CC	***KLDiv***	***NSS***	***Similarity***	***AUC_J***
SM, FM	0.4683	1.0591	1.5364	0.4364	0.8364
ATF, FM	0.5687	0.9284	1.9787	0.4719	0.8595
BTF, FM	0.5632	0.9395	1.9382	0.4684	0.8567
wTF, FM	**0.5691**	**0.9273**	**1.9824**	**0.4723**	**0.8597**

C. *Imperfect detectors*

Here we decide to use two state-of-the-art detectors which imply some misdetections or false detections (especially for the text detector). We don't combine result for both text and face detection since the text-related results are very bad (see Table IV). However, the results from face detection are good because facial landmarks approach [15] can detect the frontal face well although it misses some faces in a scene.

Table IV shows that the misdetections and even more the false detections of the text detector seriously decrease the results compared to the bottom-up saliency map alone. This text detector is not good enough to be used to add top-down information. For all metrics, results of the saliency map alone are better than all text, big text, and weighted.

Table V shows that the face detector, which has a better quality, can provide good improvement of the bottom-up saliency map. However, the difference between a simple average fusion (AF) and the weighted version (wBSF) is not significative. For some metrics such as CC and KLDiv it is even better to just make the global average instead of using the face weights.

TABLE IV. RESULT BETWEEN BOTTOM-UP AND TEXT DETECTION (BAD TEXT DETECTOR)

Correlation	**Metrics**				
	CC	***KLDiv***	***NSS***	***Similarity***	***AUC_J***
SM, FM	**0.4683**	**1.0591**	**1.5364**	**0.4364**	**0.8365**
AT, FM	0.4092	1.1523	1.3416	0.4193	0.8132
BT, FM	0.4071	1.1555	1.3306	0.4184	0.8130
ST, FM	0.4620	1.0641	1.5226	0.4364	0.8347
wBST, FM	0.4101	1.1513	1.3445	0.4196	0.8134

TABLE V. RESULT BETWEEN BOTTOM-UP AND FACE DETECTION (GOOD FACES DETECTOR)

Correlation	**Metrics**				
	CC	***KLDiv***	***NSS***	***Similarity***	***AUC_J***
SM, FM	0.4683	1.0591	1.5364	0.4364	0.8364
AF, FM	**0.5264**	**0.9968**	**1.8122**	**0.4523**	**0.8471**
BF, FM	0.5180	1.0047	1.7508	0.4504	0.8455
SF, FM	0.4763	1.0518	1.6003	0.4381	0.8381
wBSF, FM	0.5258	0.9975	1.8126	0.4520	0.8471

D. *Discussion and conclusion*

In this paper, we show how to simply add top-down information to any bottom-up saliency models in a generic way. Our work focused on both text and face features.

We tested several object detectors (bad, good, and perfect), and we demonstrated that if the detector is not good enough, it is better not to use it at all and only use the bottom-up information (Table IV). If the detector is good, a simple average can be almost as good as a more complex weighted average (Table V). When the detector becomes very good, then the weighted average really makes sense (Tables I to III). This is even more the case when several top-down features are mixed to bottom-up and some might be more important than others.

The size of the top-down object is very important. This was more the case with text where the difference in terms of eye fixations between big text (titles) and small text (description) is very important. Indeed, people provide attention to text because of its cognitive content. While for titles, the cognitive load needed is very small, for blocks of smaller text, they will less attract attention, especially if the beginning of the text (the most attended) has no important information. There is still a difference between big face and small face, but this difference is smaller. If we just consider big text and big face, the weights values of them are almost the same. It is an interesting result as previous results do not consider the difference between big and small text or big and small face [13]. The result was polluted by small text which really decrease a lot the overall text importance.

An important result improvement can be obtained by using classical bottom-up attention models to which we can add easily the higher level detected objects. Resulting models will approach novel DNN-based attention approaches while they keep generality. They are also well responding to bottom-up features which are underestimated by DNNs [19]. In addition to that, classical models can have a behaviour which can be explained while DNNs provide results without letting any chance to the programmers to explain why exactly their model works well or not. Being able to explain the reaction of an algorithm might be critical especially for security applications. That is why, for our future work, we will go deeper in the object-based top-down features which can be extracted and the optimal mix with bottom-up saliency maps.

ACKNOWLEDGMENT

This research study is supported by ARES-CCD (program AI 2014-2019) under the funding of Belgian university cooperation.

REFERENCES

[1] M. Matei, "From Human Attention to Computational Attention," pp. 105-121, Springer New York, 2016.

[2] M. Cerf, J. Harel, W. Einhauser, and C. Koch, "Predicting human gaze using low-level saliency combinded with face detection," in J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, NIPS. MIT Press, 2007.

[3] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in IEEE 12th International Conference on Computer Vision, 2009 (pp. 2106–2113). Washington, DC: IEEE Computer Society.

[4] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," in Vision Research, 40:1489–1506, 2000.

[5] R. Rosenholtz, "A simple saliency model predicts a number of motion popout phenomena," in Vision Research 39, 19:3157–3163, 1999.

[6] X. Hou and L. Zhang, “Saliency detection: A spectral residual approach,” in Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, 0:1–8, 2007.

[7] N. D. B. Bruce and J. K. Tsotsos, “Saliency, attention, and visual search: An information theoretic approach,” in Journal of Vision, 9(3):1–24, 3 2009.

[8] T. Avraham and M. Lindenbaum, “Esaliency: Meaningful attention using stochastic image modeling,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, 99(1), 2009.

[9] W. Kienzle, F. A.Wichmann, B. Schölkopf, and M. O. Franz, “A nonparametric approach to bottom-up visual saliency,” in B. Schölkopf, J. C. Platt, and T. Hoffman, editors, NIPS, pages 689–696. MIT Press, 2006.

[10] A. Borji, “Boosting bottom-up and top-down visual features for saliency estimation,” in Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), p.438-445, June 16-21, 2012.

[11] Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., & Dutoit, T. (2013), “Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis,” Signal Processing: Image Communication, 28(6), 642-658.

[12] Y. Jimei, “Top-down visual saliency via joint CRF and dictionary learning,” in Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), p.2296-2303, June 16-21, 2012.

[13] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, “Predicting human gaze beyond pixels,” in Journal of Vision, 14(1):28, pp. 1-20, 2014.

[14] L. Neumann, and J. Matas, “Real-time scene text localization and recognition,” in Computer Vision and Pattern Recognition (CVPR) 2012 IEEE Conference on, vol. 2, pp. 3538-3545, 2012.

[15] V. Kazemi, J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867-1874, 2014.

[16] G. Kootstra, A. Nederveen and B. de Boer, “Paying Attention to Symmetry,” in M. Everingham and C. Needham, editors, Proceedings of the British Machine Conference, pages 111.1-111.10. BMVA Press, September 2008.

[17] E. Vig, M. Dorr, and D. Cox, “Large-scale optimization of hierarchical features for saliency prediction in natural images,” in Computer Vision and Pattern Recognition, 2014. CVPR’14. IEEE Conference on. IEEE, 2014.

[18] M. Kümmerer, L. Theis, and M. Bethge, “Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet,” in International Conference on Learning Representations - Workshop Track (ICLR), 2015.

[19] M. Kümmerer, T. S. Wallis, L. A. Gatys, and M. Bethge, “Understanding low- and high-level contributions to fixation prediction,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct 2017.

[20] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, T. Dutoit, “Saliency and human fixations: State-of-the-art and study of comparison metrics,” in International Conference on Computer Vision (ICCV), pp. 1153-1160, 2013.

[21] Dalal, N., & Triggs, B., “Histograms of oriented gradients for human detection,” in Computer Vision and Pattern Recognition, IEEE Computer Society Conference on Vol. 1, pp. 886-893, June 2005.

[22] Dlib C++ Library. (2017, June 10). Retrieved from http://dlib.net/.

[23] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. MIT Saliency Benchmark.