DO DEEP-LEARNING SALIENCY MODELS REALLY MODEL SALIENCY?

Phutphalla Kong[1,2], Matei Mancas[2], Nimol Thuon[1], Seng Kheang[1], Bernard Gosselin [2]

[1]Institute of Technology of Cambodia (ITC) – Dept. Information and Communication Engineering
{phutphalla, sengk}@itc.edu.kh, thuonnimol1@gmail.com
[2] University of Mons (UMONS) – Faculty of Engineering (FPMs), Mons, Belgium
{matei.mancas, 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 to predict abnormal, rare and surprising data. Attention is a product of the continuous interaction between bottom-up (mainly feature-based) and top-down (mainly learning-based) information. Deep-learning (DNN) is now well established in visual attention modelling with very effective models. The goal of this paper is to investigate the importance of bottom-up versus top-down attention. First, we enrich with top-down information classical bottom-up models of attention. Then, the results are compared with DNN-based models. Our provocative question is: “do deep-learning saliency models really predict saliency or they simply detect interesting objects?”. We found that if DNN saliency models very accurately detect top-down features, they neglect a lot of bottom-up information which is surprising and rare, thus by definition difficult to learn.

Index Terms— attention, saliency, DNN, bottom-up, top-down, object detection, face detection, text detection

1. INTRODUCTION

Computational visual attention tends to mimic human visual attention and focuses more deeply on the informative and important parts of images. In computer vision, the main approach to the implementation of visual attention includes bottom-up (mainly feature-based information which is known as reflex exogenous reaction) and top-down (learning-based information which refers to reflexive endogenous information). A lot of research was achieved on bottom-up attention models [1],[2],[3],[4],[5],[6],[18] but just only a few on top-down information [11]. It seemed that top-down detectors were not efficient enough to improve results of visual attention models yet.

After an arrival DNN-based models, most researchers have switched their research direction to focus more on obtaining an end-to-end DNN saliency model which naturally integrates top-down information. Since 2014, DNNs have changed the saliency paradigm. The deep features were first used in eDN model [8]. Then, DeepGaze1 model [9] showed that the DNN features trained on object recognition were very useful for saliency detection. This finding seems logical as objects apparently represents the regions of interest in images. Since then, a variety of models used fine-tuned mixes of features from several deep learning models which naturally incorporated top-down information (i.e., faces and texts) during the learning process.

However, in [10], the authors showed that the importance of bottom-up attention was underestimated by DNN-based models. In their experiments, a simple bottom-up model could outperform a state-of-the-art DNN model when the images contained less top-down information. This demonstrated that DNNs too much neglected the bottom-up aspect of visual attention, and they were mostly trained to detect the attractive top-down objects rather than detect saliency itself. Moreover, they could not easily adapt to images in a different context from their training set and they had the structural issue to provide a result that could not really be explained in an explicit way.

In [11], the authors showed that, compared to old detectors which were not accurate enough, current detectors (ie., face detectors), when mixed to bottom-up saliency maps provide significantly better visual attention results. It is therefore possible to integrate the top-down information into classical bottom-up attention models in a hand-crafted way.

This paper aims to investigate the roles of both top-down and bottom-up information in saliency. According to MIT saliency benchmark [14] DNN-based models have populated the top-results showing the importance of top-down attention. However, this paper investigates if those models do not reduce too much the influence of bottom-up information as suggested in [10] turning saliency models into weighted object detectors.

The remainder of this paper is organized as follows. In section 2, we describe a generic top-down framework. In section 3, we will use this framework to check the influence of top-down on bottom-up attention models. Then, a comparison among DNN-based and bottom-up models is described in section 4. Finally, the discussion and conclusion are presented in section 5.

2. GENERIC TOP-DOWN FRAMEWORK

2.1. Top-down information

Classical bottom-up models use image features (ie. luminance, chrominance, texture) to detect locally contrasted or globally rare regions that will be used in the next sections. This section proposes a naïve yet generic top-down information framework that can be added to any bottom-up saliency model. In [11], the authors demonstrated that an object detector could bring remarkable improvement result to saliency maps on condition that such detector is 1) "good enough" (especially with few false positives) and it is 2) "specific enough" (general-purpose object detectors may include objects less likely to attract visual attention [12]).

Here, we use a bunch of existing detectors for face, text, person, animal, and transportation detection. The current DNN-based object detectors have become very good and, based on [11], we hypothesize that the use of this set of good detectors bringing specific top-down information will improve the results of overall saliency maps. All those detectors are used on all the images to keep our method generic and the final results include issues due to false positives or false negatives. To get quantitative results, two different datasets are used. The first one is the Object and Semantic Images and Eye-tracking (OSIE) dataset [15] containing more than 700 images along with the eye-tracking and object segmentation. This dataset is also used for our generic top-down framework parameters tuning. The second dataset is the MIT300 dataset [14] containing 300 images, which is used for comparing our proposed model with various state-of-the-art visual attention models.

In the following sections, we present the different top-down information and the way they are brought together in our generic top-down model.

2.1.1. Face detection

The face detection algorithms available in [13] were used in our study. The first algorithm uses the Histogram of Oriented Gradients (HOG) features combined with a linear classifier (SVM), while the second one uses a Convolutional Neural Network (CNN). The CNN-based face detector outperforms the HOG-based detector on the OSIE dataset especially on the badly exposed faces (Fig.1).

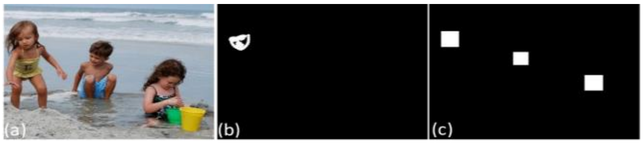

Fig. 1. Comparison results between HOG and CNN-based face detectors. (a) Input image, (b) Result of HOG-based face detector, and (c) Result of CNN-based face detector.

2.1.2. Text detection

Connectionist Text Proposal Network (CTPN) [16] is used as text detector in our framework. The CTPN detects a text line in a sequence of fine-scale text proposals directly in convolutional feature maps. The sequential text proposals are connected by a recurrent neural network, outcoming in an end-to-end trainable model. Moreover, CTPN works reliably on multi-scale and multi-language texts without any additional post-processing step (Fig.2).

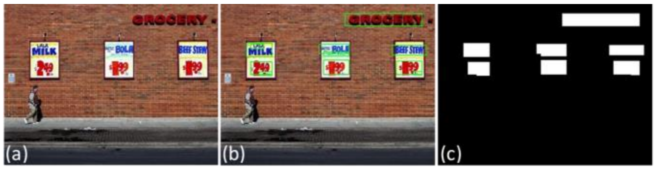

Fig. 2. Result of text detection. (a) Input image, (b) Text detection (green bounding-boxes), and (c) Binary text masks.

2.1.3. Object detection

A state-of-the-art real-time object detection from [17] were used, which could detect over 9000 objects in a reliable way. Many different detection classes are available, but only three categories (i.e., person, animal and transportation) were selected and used in our experiments. As a result, these classifiers provide three different maps with binary masks for persons, animals and transportations (Fig.3).

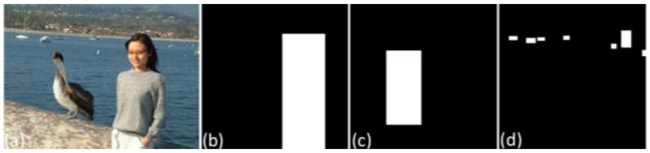

Fig. 3. Result of object detection. (a) Input image, (b) Person detection, (c) Animal detection, and (d) Transportation detection (here the small boats in the back are detected).

2.1.4. Context-based top-down information

Besides the three detection models mentioned above, a centered Gaussian function was also added into the image because it plays an important role for natural images.

The image context (Gist) is immediately detected by the viewer [27]. In [26], the author showed the difference between natural image context (where the eye gaze focuses in the center), a website context (where it is attracted more towards the top-left corner and an advertising context (where the behavior is in between the previous two). The OSIE and MIT300 datasets contain mainly natural images, so a centered Gaussian function is the best choice.

2.2. Mixing top-down & bottom-up information

A straightforward combination of the bottom-up model, the centered Gaussian function and the different detectors for face, text, person, animal, and transportation were implemented. The binary masks which are the outputs of all the detectors are smoothed to better fit into the saliency map (SM). The combination begins with the centered Gaussian (Eq.1). Then, the object detectors are added (Eq. 2, Eq. 3, Eq. 4) and mixed together in Eq5. Finally, the faces and text, which have more impact are added only at the end (Eq. 6).

$$\begin{aligned}
&\text{CSM} && = (a*SM*CG^{b}) + (1\text{-}a)*SM && (1)\\
&\text{CTSM} && = (Tra*\text{CSM}) + \text{CSM} && (2)\\
&\text{CASM} && = (Ani*\text{CSM}) + \text{CSM} && (3)\\
&\text{CPSM} && = (Per*\text{CSM}) + \text{CSM} && (4)\\
&\text{COSM} && = (\text{CTSM} + \text{CASM} + \text{CPSM})/3 && (5)\\
&\text{FAPTTX} && = (\text{COSM} + F + w*T)/3 && (6)
\end{aligned}$$

where *a*, *b* are two parameters (found to be *a*=0.75 and *b*=4), SM is the bottom-up saliency map, CG is the centered Gaussian image, *Tra*, *Ani*, *Per*, *F*, *T* are the smoothed masks of transportation, animal, person, face, and text detection, respectively. *w* is a weight set to 0.6.

The *a*, *b* and *w* parameters were found as to optimize the results on the OSIE dataset. At the end, the final saliency map is optimized by blurring as stated in [7].

3. TOP-DOWN VS. BOTTOM-UP INFLUENCE

To investigate on the role of different top-down information, RARE [18] is used as bottom-up model to generate saliency maps (SM). The same saliency metrics as in the MIT300 dataset evaluation [19] were used. These metrics consist of the Correlation Coefficient (CC), Kullback-Leibler Divergence (KLD), Normalized Scanpath Saliency (NSS), Similarity (SIM), and Judd Area Under the ROC curve (AUCJ). The smallest values represent the best results in KLD metric. For the other metrics, higher values are best.

Table 1 shows, on the OSIE dataset, the metric values between the eye-tracking fixation maps and the model output. This output is 1) bottom-up saliency maps (SM) alone on the first line, 2) SM with face detection (F) on the second line, 3) SM with text detection (TX) on the fourth line, 4) SM with animal detection (Ani) on the sixth line, 5) SM with person detection (Per) on the eights line, 6) SM with transportation detection (Tra) on the tenth line and 7) SM with centered Gaussian (CG) on the twelfth line. Results are computed on subsets of images: 279 images with faces, 425 images with text, 138 images with animals, 484 images with persons, 98 images with transportation and 700 images with centered Gaussian. In Table 1, the results in terms of CC metric shows that the faces influence is definitely higher than SM with a 0.15 (0.5631 – 0.4179 ~ 0.15) improvement measured on the 279 images in each of which having at least one face. Besides, it is quite interesting to see the result given by text detection TX (with 0.5478 - 0.4637 ~ 0.08 difference). The centered gaussian and the animals' detection comes after with a 0.04 difference. Strangely, person detection is less useful than animal detection with only 0.01 of difference on the CC metric. This is probably because the eye gaze will only focus on small parts of the body while the face has already been taken by the face detector. Finally, the transportation detector has a negative effect on the result with 0.02 difference. This shows that the objects like cars, buses, bikes, and so on are not really attended or only on very specific parts of these objects. The results for the other metrics are similar to the CC metric.

Table 1. Results using RARE model (OSIE dataset) on the number of images (on a total of 700) where at least an object is detected. The result with bold-fonts represents the best result in comparison.

Maps	Metrics				
(images)	CC	KLD	NSS	SIM	AUCJ
SM (279)	0,4179	1,1548	1.4118	0.4115	0.8291
F (279)	**0.5631**	**0.939**	**1.8914**	**0.5165**	**0.8525**
SM (425)	0,4637	1,0492	1,4626	0,439	0,8311
TX (425)	**0,5478**	**0,9011**	**1,787**	**0,4995**	**0,8544**
SM (138)	0,4754	1,1183	1,7178	0.4202	0,8516
Ani (138)	**0,5111**	**1,0425**	**1,8565**	**0.4716**	**0,8629**
SM (484)	0,4587	1,0971	1,57	0,4262	0,8412
Per (484)	**0,4699**	**1,0594**	**1,6185**	**0,4626**	**0,8433**
SM (98)	**0,5152**	**0,9998**	**1,8336**	0,4471	**0,8636**
Tra (98)	0,4902	1,0135	1,7608	**0,4748**	0,8579
SM (all)	0.4683	1.0597	1.5364	0.4364	0.8365
CG (all)	**0.5001**	**0.9738**	**1.6231**	**0.4679**	**0.8472**

To check how general our framework is, we tested it on four different bottom-up saliency models such as AIM [21], AWS [22], GBVS [23], and RARE [18]. Table 2 shows, for each model, the results of the bottom-up saliency map alone (SM) and the SM added with our framework (FAPTTX). RARE model has the best results but they are very close to AWS. GBVS and AIM are less good. One can see that the best improvement is achieved for AIM which for example gains about 0.22 in CC metric and seems to be the one capturing the less top-down attention. AWS and RARE both improve at 0.16 (CC metric). Finally, GBVS only improves about 0.13 (CC metric) probably because it already has the centered gaussian included in the bottom-up model.

Table 2. Correlation result using several models (OSIE dataset)

Model		Metrics				
		CC	KLD	NSS	SIM	AUCJ
AIM	SM	0.3251	1.5241	1.0717	0.3454	0.7733
	FAPTTX	0.5392	1.1186	1.7311	0.407	0.8496
AWS	SM	0.4583	1.1171	1.4855	0.4268	0.8219
	FAPTTX	0.6161	0.8313	2.029	0.4995	0.8708
GBVS	SM	0.438	1.088	1.3496	0.425	0.8159
	FAPTTX	0.5608	0.9379	1.8104	0.4828	0.8488
RARE	SM	0.4683	1.0597	1.5364	0.4364	0.8365
	FAPTTX	**0.6235**	**0.8162**	**2.0868**	**0.5192**	**0.8719**

4. DNN-BASED VS. BOTTOM-UP MODELS

The new DNN-based attention models occupy the first places on benchmarks such as the one of MIT300. In this section, we check what happens when we come up with a bottom-up model augmented with our top-down framework.

4.1. Qualitative comparison

To first make a qualitative comparison, for the DNN-based model we use SAM-ResNet [24] and Salicon [25] which are two state-of-the-art DNN-based models. We notice that they provide better results than our proposed approach especially if the scenes contains humans (Fig.4). However, they provide poorer results than RARE model if the scene is complex with unknown objects (Fig.5).

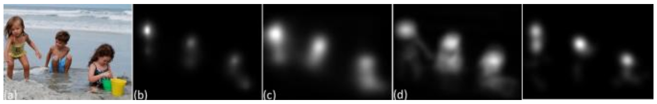

Fig. 4. Result where DNN-based models are better than bottom-up models. (a) Input image, (b) Result of SAM-ResNet, (c) Result of Salicon, (d) Result of our model and (e) Eye tracking map.

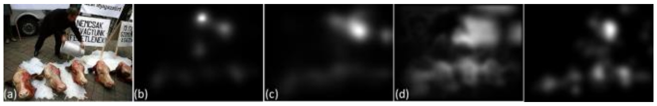

Fig. 5. Result where DNN-base models is less good than bottom-up models. (a) Input image, (b) Result of SAM-ResNet, (c) Result of Salicon, (d) Result of our model and (e) Eye tracking map.

4.2 Quantitative comparison

4.2.1. OSIE dataset

While the DNN-based models are overall better that the proposed model, we looked more in details on the images. Our experiment shows that on the OSIE dataset RARE bottom-up model alone is better than SAM-ResNet for 5.7% of the images and RARE augmented with our generic framework is better than SAM-ResNet on 14.3% of the images. If we only take the images where our model has a CC metric higher than 0.05 compared to SAM-ResNet (which will boost the model of about 10 places on a benchmark such as MIT300), our approach is still much better than SAM-ResNet on 9% of the images. We used here the CC metric as it is one which is not favorable to our approach (see Table 5 showing that KLD is the best metric four our model). It appears that DNN-based models might be inferior in learning bottom-up data while they are better in detecting objects usually salient (top-down information).

4.2.2. MIT300 dataset

According to MIT300 benchmark [20], our model has the best results compared to all bottom-up models (Table 4). It is still surpassed by some DNN-based models (Table 5), but a lot of those models are now less good than ours. This shows that a bottom-up model, simply augmented with some top-down information can be better than all the other bottom-up models and even better than number of other DNN-based models depending on the metrics. For example, for the KLD metric by this date, our model is better (lower KLD value) than about 18 DNN-based models while less good than only 7 DNN-based models.

Table 3. Comparing result between SAM-ResNet and ours

Model	Metrics				
	CC	KLD	NSS	SIM	AUCJ
SAM alone	**0.7713**	**1.3726**	**3.1023**	**0.651**	**0.9026**
SAM+FAPTTX	0.7546	1.6081	2.8205	0.6226	0.8943
Ours (FAPTTX)	0.6235	0.8162	2.0868	0.5192	0.8719

Table 4. Comparing result between bottom-up models and ours

Model	Metrics				
	CC	KLD	NSS	SIM	AUCJ
Ours	**0.6166**	**0.7179**	**1.6762**	**0.5472**	**0.8388**
BMS	0.55	0.81	1.41	0.51	0.83
OS	0.54	0.84	1.41	0.51	0.82
GBVS	0.48	0.87	1.24	0.48	0.81

Table 5. Comparing result between DNN-based models and ours

Model	Metrics				
	CC	KLD	NSS	SIM	AUCJ
DSCLRCN	0.8	0.95	2.35	0.68	0.87
SALICON	0.74	0.54	2.12	0.6	0.87
SAM-Rest	0.78	1.27	2.34	0.68	0.87
Ours	**0.6166**	**0.7179**	**1.6762**	**0.5472**	**0.8388**
SalNet	0.58	0.81	1.51	0.52	0.83
eDN	0.45	1.14	1.14	0.41	0.82
GoogLeNet	0.49	0.99	1.26	0.45	0.81
JuntingNet	0.54	0.96	1.43	0.46	0.80

5. DISCUSSION AND CONCLUSION

The purpose of this paper is to understand the difference in visual attention computation between classical bottom-up saliency models and DNN-based saliency models and the relative importance of bottom-up and top-down information.

Our results show that the influence of the main objects in images is the following: 1) face detection is the most important, 2) text detection (with about half of the importance of face detection), 3) animal detection (about half less important than text detection). The influence of person and transportation detection is marginal or even negative, because the viewer gaze probably focus on small parts of persons or cars but not everywhere on their bounding boxes. Concerning bottom-up information, we show that for almost 6% of the images in the OSIE dataset, a bottom-up model alone can better predict the gaze than DNN-based models. This means that the bottom-up information still remains important and should not be neglected in visual attention, especially in the complex and crowded images where it is hard to identify faces.

Finally, we show that mixing a bottom-up model with our naïve top-down information framework leads on the MIT300 saliency benchmark to the best results among all bottom-up models and overtakes number of DNN-based models especially on KLD which measures the probability distribution resemblance with eye-tracking. Considering the fact that DNN-based models results cannot be explained and that they seem to neglect bottom-up information, future work is to see how a DNN model can be mixed with bottom-up models to consider both the good top-down detection of DNN-based models and the necessary bottom-up attention from classical models.

6. ACKNOLEDGEMENTS

Supported by ARES-CCD (program AI 2014-2019) under the funding of Belgian university cooperation.

7. REFERENCES

[1] 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.

[2] R. Rosenholtz, "*A Simple Saliency Model Predicts a Number of Motion Popout Phenomena*," in Vision Research 39, 19:3157–3163, 1999.

[3] 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.

[4] 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.

[5] T. Avraham and M. Lindenbaum, "*Esaliency: Meaningful Attention using Stochastic Image Modeling*," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 99(1), 2009.

[6] 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.

[7] N. Riche, "*Study of Parameters Affecting Visual Saliency Assessment*," in: Mancas M., Ferrera V., Riche N., Taylor J. (eds) From Human Attention to Computational Attention. Springer Series in Cognitive and Neural Systems, vol 10. Springer, New York, NY, 2016

[8] 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.

[9] 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.

[10] 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.

[11] K. Phutphalla, M. Matei, K. Seng, and G. Bernard, "*Generic and Effective Visual Attention*," in Proceeding on ICPRAI, May 2018.

[12] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "*Fusing Generic Objectness and Visual Saliency for Salient Object Detection*," in ICCV, 2011.

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

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

[15] 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.

[16] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "*Detecting Text in Natural Image with Connectionist Text Proposal Network*," in ECCV, 2016.

[17] J. Redmon and A. Farhadi, "*Yolo9000: Better, faster, stronger*," arXiv preprint arXiv:1612.08242, 2016.

[18] Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., and Dutoit, T. (2013), "*Rare2012: A Multi-scale Rarity-based Saliency Detection with Its Comparative Statistical Analysis*," Signal Processing: Image Communication, 28(6), 642-658.

[19] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, F. Durand, "*What do different evaluation metrics tell us about saliency models?*" arXiv:1604.03605, 2016.

[20] T. Judd, F. Durand, and A. Torralba, "*A Benchmark of Computational Models of Saliency to Predict Human Fixations*," MIT technical report, 2012

[21] N. Bruce and J. Tsotsos, "*Attention based on Information Maximization*," Journal of Vision, 7(9):950–950, 2007.

[22] A. Garcia-Diaz, V. Leboran, X. R. Fdez-Vidal, and X. M. Pardo, "*On the Relationship between Optical Variability, Visual Saliency, and Eye Fixations: A Computational Approach*," Journal of Vision, 12(6), 2012.

[23] J. Harel, C. Koch, and P. Perona, "*Graph-based Visual Saliency*," in NIPS, 2006.

[24] M. Cornia, L. Baraldi, G. Serra, R. Cucchiara, "*Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model*," arXiv preprint, arXiv:1611.09571v3, 2017.

[25] X. Huang, C. Shen, X. Boix, and Q. Zhao, "*Salicon: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks*," In Proceedings of the IEEE International Conference on Computer Vision, pages 262–270, 2015.

[26] M. Mancas, "*Relative influence of bottom-up and top-down attention*," Attention in Cognitive Systems, Vol. 5395 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2009.

[27] A. Oliva, A. Torralba, "*Building the gist of a scene: the role of global image features in recognition*," Progress in Brain Research, Elsevier, Volume 155, Part B, 2006.