Visual saliency on the road: model and database dependent detection

P. Duthon[1,2] J-C. Quinton[2,3] M. Colomb [1]

[1] Cerema, Département Laboratoire de Clermont-Ferrand, 8-10 rue Bernard Palissy, F-63017 Clermont-Ferrand Cedex 2, France.
[2] Clermont University / CNRS (UMR 6602), Pascal Institute, BP 80026, F-63171 Aubière, France.
[3] Université Grenoble Alpes / CNRS (UMR 5224), Laboratoire Jean Kuntzmann, BP 53, F-38041 Grenoble, France.

pierre.duthon@cerema.fr

Résumé

Dans le contexte routier, les objets d'intérêt (saillants ou non) doivent être efficacement détectés quelles que soient les conditions afin d'assurer la sécurité, que ce soit pour des systèmes d'assistance à la conduite ou des véhicules autonomes. Neufs modèles de saillance représentatifs de l'état de l'art sont évalués sur deux bases de données issues du contexte routier (perception humaine et robotique). Bien qu'elle ne soit pas suffisante pour la détection, la saillance visuelle bottom-up fournit des informations pertinentes, d'autant plus en la contrôlant pour ses biais classiques.

Mots clés

Saillance visuelle, Contexte routier, Attention visuelle, Détection de cibles, Analyse de scène.

Abstract

In the road context, objects of interest (salient or not) must be efficiently detected under any condition to ensure safety, for both driver assistance systems and autonomous vehicles. Nine representative state-of-the-art saliency models are evaluated on driving databases (human perception vs. robotics). Although not sufficient for robust detection, bottom-up saliency provides important information, especially when controlling for the classical biases.

Keywords

Saliency, Road context, Visual attention, Target detection, Scene Analysis.

1 Introduction

Vision research studies usually make the distinction between two complementary kinds of processes when dealing with human vision [19]. On the one hand, there are slow, often sequential, and complex high-level processes such as object recognition or visual search. On the other hand, there are low-level mechanisms able to quickly preselect areas of interest within the field of view, on which high-level processes can prioritarily focus. These mechanisms accounting in part for both covert selection of information and overt gaze orientation are simulated by a wide range of models [2]. These models produce saliency maps of visual scenes, where saliency can be basically defined as the tendency for an area to pop out of its context.

Saliency algorithms are often better described as bottom-up, starting from sensory information to more abstract representations, only processing low-level features (intensity, color, orientation, time) with quite simple filters (contrasts, center-surround opposition), yet with a multiscale approach [12]. They thus do not rely on more complex features (from pattern recognition, face recognition, learning), and except for a few [9], do not incorporate top-down processes. At their origin, saliency models were aimed at reproducing human behavior, and results were compared and evaluated in regards to human scanpaths [2]. Scanpaths are sequences of saccadic eye movements and fixations on an image, that usually focus on salient features (as defined above). This is however only true during free-viewing tasks, where the impact of the task on saccades is limited [18]. This has progressively led to the emergence of a new perspective on saliency, defined as a segmentation problem solving method, accompanied by the introduction of new models to be used in computer vision, and thus robotics [3]. Although reference maps (ground truth) for foreground/background segmentation are statistically generated from human decisions, the aim of segmentation algorithms is not to reproduce the exact human behavior and its complex spatiotemporal unfolding.

Now focusing on road context, studies have shown that driving is mainly task driven (involving a lot of top-down processes) [18], and important elements for driving are in-

deed not necessarily salient. Although traffic signs may be designed to be salient, a pedestrian crossing the road may not be. Bottom-up mechanisms nevertheless play a role as a filter, and thus facilitate detection. In this paper, we more specifically evaluate how much bottom-up saliency may contribute to the detection of objects of interest in the road context. In addition to their validity as a predictor of human behavior, saliency algorithms can also be tested as an efficient pre-processing step in autonomous driving systems [1]. Current computational architectures indeed allow the embedded parallel implementation of saliency algorithms on robots, which can then benefit from increased robustness of visual features to scene modification. Several benchmarks of saliency algorithms can be found in the literature [3, 14], using a variety of databases. Some of these databases were specifically created with a human behavior simulation purpose (please refer to [2] for a review). Others are designed to evaluate models with segmentation purpose, yet none of them provides road specific content, in neither of the aforementioned contexts.

Our objective in this paper is therefore to test the applicability and limits of purely bottom-up approaches to visual saliency for object detection, while considering two road context applications: the automatic detection of objects for autonomous vehicles (computer vision and mobile robotics) and the simulation of human behavior in driving situations (human vision and psychology). The latter may not only lead to a better understanding of human behavior required to adapt the infrastructure to human specificities, but also to design driving assistance systems relying on joint human-machine visual interactions with the environment. The associated databases not only illustrate the wide range of applications targetted by saliency models, but also demonstrate drastic differences in the visual information provided. Although our study directly relies on road context databases and application constraints, most of our results are generalizable to other domains. Now that we have introduced the applicative context, section 2 sets out the arguments for the selection of representative saliency models, and provides details on the experimental protocol (databases and metrics). Results are then presented in section 3 before reaching the conclusions in the last section.

2 Method

2.1 Model selection

Borji [2] proposed a thorough review of state of art models whose purpose is to model visual attention (human behavior). A taxonomy of saliency models was proposed, adopting the following classes : cognitive models, decision theoretic models, information theoretic models, graphical models, Bayesian models, spectral analysis models, pattern classification models, and miscellaneous. This study is complemented by a benchmark of models whose purpose is to perform image segmentation [3], which we will group in an additional class. As many models are associated to each class, we chose to pick one from each class in order to make our analysis synthetic, yet significant (see Table 1). In addition to testing our selection of models in the road context, we may therefore also be able to put forward their complementarities.

	Reference	Model class	Multiscale	Spatial bias	Bottom-up	Color	Intensity	Orientation
NVT	[13]	Cognitive	✓		✓	✓	✓	✓
VOCUS	[7]	Cognitive	✓		✓	✓	✓	✓
AIM	[5]	Info. Theoretic	∼		✓	✓	✓	
GBVS	[10]	Bayesian	∼	∼	✓	✓	✓	✓
SR	[11]	Spectral			✓		✓	
GAFFE	[17]	Spectral		✓	✓		✓	
MSS	[15]	Miscellaneous	✓		✓		✓	
AWS	[8]	Miscellaneous	✓		✓	✓	✓	✓
CAS	[9]	Segmentation	✓			✓	✓	

Table 1: Summary of selected models.

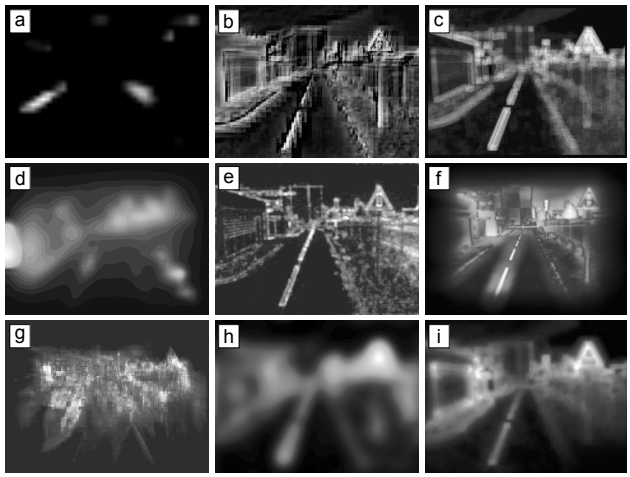

Figure 1: Saliency maps computed on the IPDS picture from Fig.2. a) NVT, b) VOCUS, c) AIM, d) GBVS, e) SR, f) GAFFE, g) MSS, h) AWS, i) CAS.

Additional criteria for selecting the models include: applicability to static images (not only videos), generation of saliency maps (not only sequences of fixations, thus eliminating graphical models), bottom-up processing (no task driven and top-down mechanisms, thus eliminating decision theoretic models). Although a lot of models are eliminated this way, at least one remains in most of the categories. The most representative models of each class were then prioritized. Within a class, representativity is here defined as 1) being a usually referenced member of a large branch of models within the taxonomy, and/or 2) obtaining good scores in benchmarks compared to their counterparts. Within the cognitive models, VOCUS [7] was selected as driving context oriented variant of the original Neuromorphic Vision Toolkit (NVT) model [13], which was also included as an historical reference. Spectral residual (SR) was kept as the representant of spectral models [11], as well as GAFFE, which is an hybrid model combining filters in the spectral domain and contrast processing in the spatial domain [17]. Finally, AIM was selected as a mem-

	VELER	IPDS
Sample size	72 pictures	53 pictures
Original domain	Human perception	Autonomous vehicles
Source	Photographs (with digital editions)	Video camera (fixed on vehicle frame)
Variability	Context, inserted road users, vehicle lights	Frame sequence in open environment
Image size	1000 x 565 px	640 x 480 px
Targets	Traffic signs, ground marking, vulnerable users	Traffic signs, ground marking

Table 2: Summary of IPDS [16] and VELER [6] databases.

ber of the information theoretic models [5], and GBVS for the Bayesian models [10]. Two miscellaneous models were also selected, because they adopt different points of view on saliency and rely on very different mechanisms. The first one is based on multi-scale symmetry operators (MSS) [15], while the other (AWS) relies on decorrelation (using principal component analysis) and distinctiveness (Hotelling's T2 measure) [8]. Despite our road context which dictates a focus on the rapid detection of areas of interest for further processing, and because segmentation models obtain very high scores on some benchmarks [3], Context-Aware Saliency (CAS) [9] was included for performance comparison as one of the best in its category. All algorithms were run using default parameters, as defined in the source code publicly made available or personally provided by the original authors. For the GBVS model, only the first saliency map after a central fixation was exploited. Saliency map samples are provided in Fig.1. Two models are added as controls: a model of center bias (with generated saliency map following a Gaussian profile function of excentricity), and a uniformly random model (providing a lower bound for the selected metrics and performance).

2.2 Databases

Two road context databases are used to evaluate the models. They originate from completely different areas of the research field, thus allowing us to analyze the influence of the chosen database on the results. To illustrate the differences, one picture of each database is provided on Fig.2. The VELER database was designed to study the detection rate of vulnerable road users according to their position and context [6]. This database contains 72 images pictures of urban driving scenes, where vulnerable road users have been inserted with controlled parameters. The insertion of various road users, as well as the presence or absence of daytime running lights, leads to different versions of urban photographs (originally taken from a human perspective). The second database is a subset of IPDS, a multisensory and publicly available dataset, with data acquired from a mobile robotic transportation platform navigating on Cézeaux campus (Clermont-Ferrand) [16]. The subset contains 53 images from various video sequences captured from a fixed camera on the vehicle frame, images in which relevant targets are visible (traffic signs, ground marking).

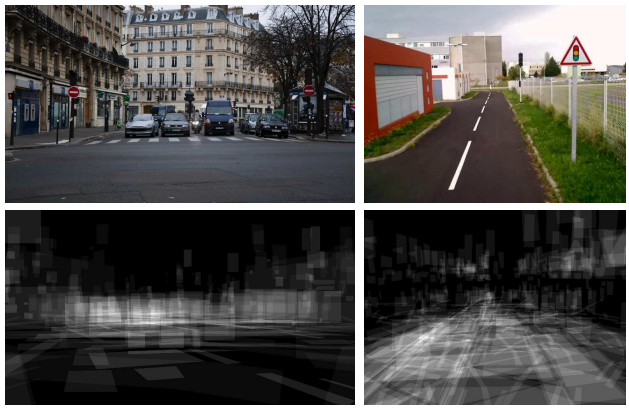

Figure 2: Representative sample picture and Mean Annotation Position (MAP) for VELER [6] (left) and IPDS [16] (right) databases.

2.3 Metrics

Humans being very accurate in detecting targets in such databases, we asked human observers to perform segmentation on all pictures, without any time constraint. Since we wanted to make sure an increased saliency value on any part of an object would be considered valid, we relied on hand-defined bounding polygons for moving vehicles (cars, trucks, buses), motorcycles, pedestrians, cyclists and traffic elements. This distinction based on known dynamics was made because motorcycles, pedestrians and cyclists represent the most vulnerable road users, while parked four-wheeled vehicles are not of direct importance. All traffic signs and ground marking were selected since they provide a lot of information for road users.

These (binary) reference maps then need to be matched against the (graded) saliency maps produced by the different models. Many metrics have been introduced and extensively used in the literature, some of which directly applicable to our problem, including: correlation coefficient (CC), normalized scan-path saliency (NSS), area under the ROC curve (AUC), shuffled AUC (sAUC), Kullback-Leibler measure (KL-div) and Earth Mover Distance (EMD) [4]. As the properties they measure might differ and their convergent validity may be limited, we chose to rely on a set of metrics to draw sound conclusions. In order to make our results easily comparable to most of the existing benchmarks, we chose the most common metrics, starting from the CC metric. Albeit simple, this metric gives a good estimate of the accuracy of saliency maps. We also selected the AUC metric, which complementarily takes into account both precision and recall rates. Borji and colleagues have shown that these two metrics are very sensitive to a center bias [4], and we added sAUC as a bias corrected version of the AUC. The reader may note that the NVT model produces quasi binary saliency maps, which are heavily penalized by AUC and sAUC metrics (associated results thus need to be interpreted with caution).

To control for center bias, we computed Mean Annotation

Position (MAP) [3] separately on each database, as the average reference map for all pictures. Thus, if a position is registered in a lot of images, there will be a high intensity at this location on the MAP. Conversely, a uniform MAP reflects the absence of center bias. As demonstrated by the MAPs on Fig.2, the VELER database demonstrates a much strong center bias than IPDS. This difference is easily explained by the fact that the focal point was whether chosen by a human photographer (VELER) or indirectly determined by the pose of the robot (IPDS). In addition to the controlled perspective and framing of the VELER database, reflected by the horizon line on the MAP, inserted elements (pedestrians, cyclists or motorcycles) are placed either near the center or on the sides of the pictures. Although camera height and tilt are also kept constant in the IPDS dataset, pictures are captured at all times on a moving vehicle (on a roughly flat course), and eccentricity is thus accentuated by the variability in target to camera distances or vehicle orientation. When the vehicles moves forward, ground marking will descend from the horizon line to finally disappear at the bottom of the image (when rolling over it). Despite the lack of a clear center bias in IPDS, we thus face a more complex position bias, betraying the regularities in the movement and environment. Our study allows to estimate the impact of these drastic differences in MAPs on the detection performance.

2.4 Procedure

All selected models were applied on every picture of both databases. After evaluating their performance using the introduced metrics, the same was done after applying a correction for center bias. The correlation matrix between the results of all models was also computed to check whether different approaches would encounter difficulties on different images. To further refine our findings, a fine-grained analysis at the image level was done to test which picture characteristics lead to differences in performance.

3 Results

3.1 Models ranking

Models are compared in Fig.3. AIM, GAFFE and AWS models prove to be more effective for both metrics and on both databases. Gaussian, GBVS and GAFFE models take into account the center bias, and thus produce good results when the bias is present (VELER), but are conversely penalized when it is not (IPDS). Spectral residual (SR) is very effective on IPDS only, due to its sensitivity to the narrow and well contrasted dotted lines of the ground marking. Even though they remain effective, VOCUS and MSS display below average results. Itti's NVT model produces almost binary saliency maps, which lead to artificially low AUC and sAUC scores, thus justifying its seemingly bad performance in comparison with others for these metrics.

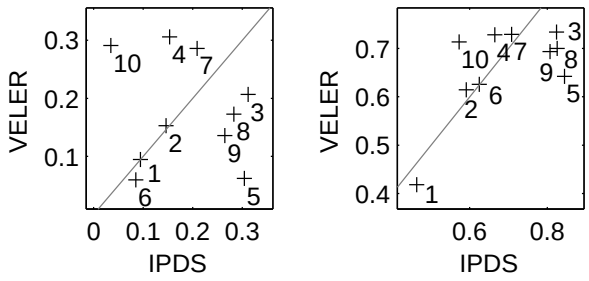

Figure 3: Comparison of results for both databases (gray line represents equal performance) using CC (left) and sAUC (right) metrics, with models: NVT (1), VOCUS (2), AIM (3), GBVS (4), SR (5), MSS (6), GAFFE (7), AWS (8), CAS (9), Gaussian (10).

3.2 Center bias correction

The influence of center bias correction is estimated by scaling the saliency maps pixel-wise by the Gaussian map. Since some algorithms already include this type of correction (GAFFE with post-processing, GBVS intrinsically), applying this correction a second time increases the weight of information near the center of the picture, thus allowing to further check any loss or gain in performance.
From Table 3, the Gaussian correction has a positive impact for all models on the VELER database. Reassuringly, the models that received prior correction are also those which benefit from the smallest improvement. On the contrary, applying a center bias correction on the IPDS pictures degrades the results for all models, which could be predicted from the bad results of the Gaussian model alone on this database. Although the specific shape of the Gaussian profile was arbitrarily chosen and could be optimized, these results mainly reflect the existing differences between the databases, with VELER demonstrating a more focused MAP than IPDS. The MAPs show that the spatial location of targets on the image is very important, and a simple center bias correction is not sufficient. Such position bias should rather be considered as top-down, with target-location associations acquired through learning. Saliency maps could then either be modulated by the expectancy of the target or centered through by a focus mechanism (similar to the overt/covert attentional systems in humans).

3.3 Correlation between models results

In addition to check whether some targets are difficult to detect for all models or a subset, computing the correlation between model results on all pictures for each database also allows understanding the specificities of the models. In this paper, we only describe the more interesting results found for the VELER database. The Fig.4 shows the correlation coefficient for each couple of models on VELER.
The AIM model can be thought as the most representative model, with a high average correlation with the other models. The strong correlation between Gaussian, GAFFE and

	IPDS		VELER	
	CC	sAUC	CC	sAUC
NVT	2.0	-0.0	8.5	0.6
VOCUS	-21.9	-2.5	53.3	3.7
AIM	-17.6	-5.2	45.1	1.8
GBVS	-7.6	-1.9	7.3	0.7
SR	-23.9	-9.4	204.1	6.3
MSS	-18.8	-3.1	138.7	6.1
GAFFE	-11.3	-2.2	11.3	0.6
AWS	-23.9	-7.3	57.1	2.9
CAS	-23.1	-7.2	77.6	2.3
Gaussian	-0.6	0.0	1.8	-0.0

Table 3: Change (in percents) in mean scores obtained through Gaussian (center bias) correction on both databases.

GBVS again reflects the integration of a center bias correction. GAFFE directly includes a final step with Gaussian correction, while GBVS includes a normalization step on a graph of pixel nodes, where the arcs are weighted by pixel-to-pixel distances. Pixels on the edge of the image are disadvantaged, as they are on average more distant to the others. Even though NVT is again isolated because of its quasi-binary nature, it is maximally correlated with the model derived from it (VOCUS).

To further refine our results, we compared pairs of models at the picture level, and here chose to contrast VOCUS (representative of the NVT family of models) with AIM (which highly performed and was qualified as the most representative), which are only moderately correlated. Fig.5 shows the CC scores of VOCUS vs. AIM for each image of the VELER database. While AIM performs on average better than VOCUS, and especially on pictures 53b, 06 and 42, VOCUS yet clearly outperforms AIM on pictures 11b, 21 and 19 (see images on Fig.6). AIM wins for high contrast images (e.g. saturated sky, very dark road) while VOCUS better detects the targets on less contrasted and better exposed images. On these latter images, buildings in the background provide a gradual transition from the sky to the road, and large size targets are also often present (e.g. bus on image 19). Such targets are then easily detected by VOCUS thanks to its multi-scale architecture, thus boosting its CC score. Reciprocally, VOCUS also detects a contrasted sky as the most prominent element, and although the sky may be salient, it is not important from a driving perspective (for safety or navigation in urban environments). To put it briefly, the dynamic range, the structure of the environment, as well as the size of the targets play an important role in the evaluation of saliency models.

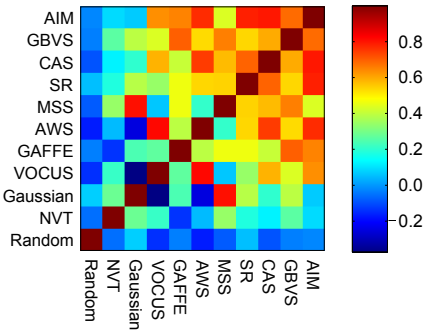

Figure 4: Correlation coefficients between the models results for CC metric on VELER.

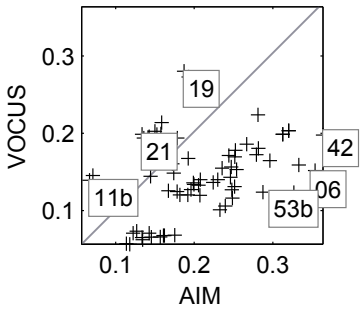

Figure 5: VOCUS vs AIM results for CC metric on VELER.

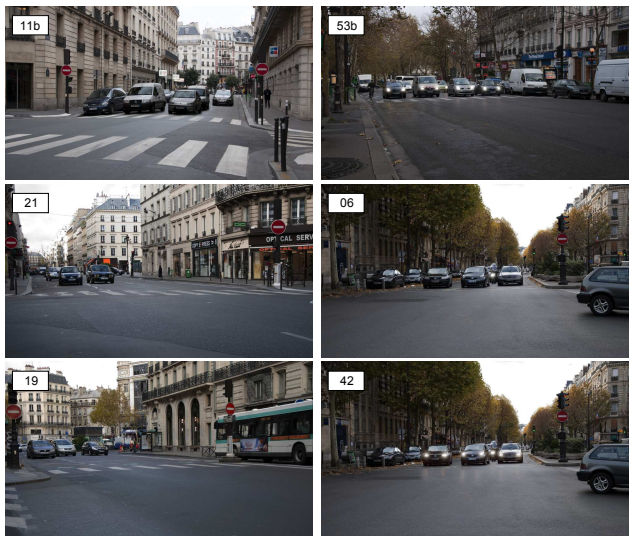

Figure 6: Images with CC differences between models, to the advantage of VOCUS (11b, 21 and 19 on the left) or AIM (53b, 06 and 42 on the right).

4 Discussion

Detection capabilities of a set of representative saliency models have been evaluated using complementary metrics and driving oriented databases, in which important elements (for safety or navigation in our context) may not always be the most salient. AIM, AWS and GAFFE models demonstrated the best results on both databases, followed by CAS and GBVS models. Other models show database dependent performance, with SR for instance achieving top performance on IPDS database while scoring very low on VELER. The take-home message is that the task (e.g. targets) and context (e.g. urban or not, point of view) are key elements that can drastically alter performance, and that saliency models should thus be evaluated and selected in regards to the target application.

In addition, we brought out characteristics of the image

databases which largely impact detection performance, including: 1) Position bias (not limited to center bias) which can be estimated using Mean Annotation Position maps (MAPs); 2) Dynamic range and colors, which depend of acquisition devices, their calibration and settings (thus also depending on shooting conditions), and which could be corrected through the use of dedicated devices (e.g. luminance-meters).

The limitations of this study also put forward limitations in the common use of saliency models, and should thus lead to future work on these aspects: 1) Most models apply only and are evaluated on static images, while the spatiotemporal structure in video sequences may for instance turn moving pedestrians into salient targets; 2) Most saliency algorithms can easily detect part of a traffic sign, but cannot properly segment and identify it without resorting to complementary methods (this of course applies to the detection of objects in any domains); 3) Color balance, contrast or levels impact saliency, but are not usually taken into account, while they could be controlled by simple preprocessing (exposure correction or white balance being for instance somehow performed by the human eye).

More importantly in the driving context, pure bottom-up saliency cannot alone be used for the reliable detection of important elements, as they are not necessarily salient (e.g. pedestrian). Yet our study shows that saliency algorithms can already highlight a lot of potential targets, and thus bring in information that can be further processed and filtered by top-down processes. These latter processes (which usually require iterations to reach a decision, e.g. identity of an object) can continuously modulate bottom-up saliency processes (which naturally apply for parallel implementation). Such modulation could be the estimation of a priori locations of targets (instead of the simple center bias correction we used here), which may in turn integrate information from other sensors (movement speed, GPS, acceleration, orientation of the camera).

5 Acknowledgments

We would like to thank the authors of the different models for their answers to our requests and for sharing their source code. Acknowlegment also goes to Viola Cavallo (IFSTTAR) who has graciously made available the VELER database for this study. This research also received support from the French program "investissement d'avenir" managed by the National Research Agency (ANR), from the European Union (Auvergne European Regional Development Funds) and from the "Région Auvergne" in the framework of the IMobS3 LabEx (ANR-10-LABX-16-01).

References

[1] M. Birem, J.-c. Quinton, F. Berry, and Y. Mezouar. SAIL-MAP : Loop-Closure Detection Using Saliency-Based Features. *IEEE/RSJ International Conference on Intelligent Robots and Systems 2014 (IROS)*, 2014.

[2] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE PAMI*, 35(1):185–207, 2013.

[3] A. Borji, D. Sihite, and L. Itti. Salient object detection: A benchmark. *Computer Vision – ECCV 2012*, pages 414–429, 2012.

[4] A. Borji, D. N. Sihite, and L. Itti. Analysis of scores, datasets, and models in visual saliency prediction Supplementary Material. 2013.

[5] N. Bruce and J. Tsotsos. Saliency based on information maximization. *Advances in neural information processing systems*, 2006.

[6] V. Cavallo and M. Pinto. Are car daytime running lights detrimental to motorcycle conspicuity? *Accident; analysis and prevention*, 49(2012):78–85, Nov. 2012.

[7] S. Frintrop. *VOCUS: A visual attention system for object detection and goal-directed search*, volume 320. 2006.

[8] A. Garcia-Diaz and X. Fdez-Vidal. Decorrelation and distinctiveness provide with human-like saliency. *Advanced concepts for intelligent vision systems*, pages 343–354, 2009.

[9] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE PAMI*, 34(10):1915–26, Oct. 2012.

[10] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *Advances in neural information processing systems*, 2007.

[11] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. *CVPR*, (800), 2007.

[12] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12):1489–506, Jan. 2000.

[13] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI*, 20(11):1254–1259, 1998.

[14] T. Judd, F. d. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. 2012.

[15] G. Kootstra, A. Nederveen, and B. D. Boer. Paying attention to symmetry. *BMVC*, 2008.

[16] H. Korrapati and J. Courbon. 'The Institut Pascal Data Sets': un jeu de données en extérieur, multicapteurs et datées avec réalité terrain, données d'étalonnage et outils logiciels. *Orasis*, 2013.

[17] U. Rajashekar, I. van der Linde, a. C. Bovik, and L. K. Cormack. GAFFE: a gaze-attentive fixation finding engine. *IEEE transactions on image processing*, 17(4):564–73, Apr. 2008.

[18] H. Shinoda, M. M. Hayhoe, and A. Shrivastava. What controls attention in natural environments? *Vision research*, 41(25-26):3535–45, Jan. 2001.

[19] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision research*, 45(5):643–59, Mar. 2005.