Saliency Prediction with Active Semantic Segmentation

Ming Jiang[1]
mjiang@u.nus.edu
Xavier Boix[1,3]
elexbb@nus.edu.sg
Juan Xu[1]
jxu@nus.edu.sg
Gemma Roig[2,3]
gemmar@mit.edu
Luc Van Gool[3]
vangool@vision.ee.ethz.ch
Qi Zhao[1]
eleqiz@nus.edu.sg

[1] Department of Electrical and Computer Engineering
National University of Singapore
Singapore

[2] CBMM, LCSL
Massachusetts Institute of Technology
Istituto Italiano di Tecnologia
Cambridge, MA

[3] Computer Vision Laboratory
ETH Zurich
Switzerland

Abstract

Semantic-level features have been shown to provide a strong cue for predicting eye fixations. They are usually implemented by evaluating object classifiers everywhere in the image. As a result, extracting the semantic-level features may become a computational bottleneck that may limit the applicability of saliency prediction in real-time applications. In this paper, to reduce the computational cost at the semantic level, we introduce a saliency prediction model based on active semantic segmentation, where a set of new features are extracted during the progressive extraction of the semantic labeling. We recorded eye fixations on all the images of the popular MSRC-21 and VOC07 datasets. Experiments in this new dataset demonstrate that the semantic-level features extracted from active semantic segmentation improve the saliency prediction from low- and regional-level features, and it allows controlling the computational overhead of adding semantics to the saliency predictor.

1 Introduction

Saliency models predict the probability distribution of the eye fixations of a human observer over the image pixels. Current state-of-the-art saliency prediction models combine features extracted at pixel level (low-level features), to regional level, and to semantic level. Generally, the higher up in the hierarchy, the larger pixel coverage and the more semantically meaningful are, consistent with the visual pathway in the brain. Combining features at the different levels has been shown to increase the performance of predicting the saliency map, *c.f.* [15, 34].

Dr. Qi Zhao is the corresponding author.

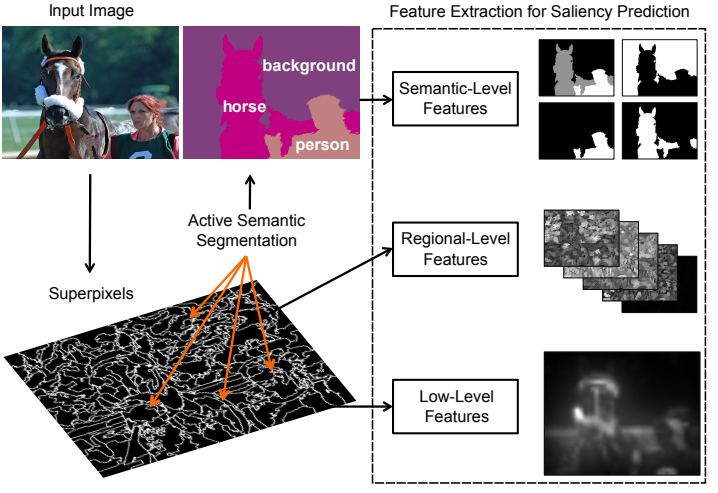


Figure 1: *Overview of the proposed saliency framework.* We introduce new semantic-level features for saliency prediction, based on an active semantic segmentation of the input image. This allows to control the computational overhead of adding semantics to the saliency prediction.

Typically, the features at the semantic level are more computationally expensive to compute than the generic low- and regional-level features. The semantic features are usually implemented by evaluating object classifiers everywhere in the image. As a result, the computational cost of the saliency prediction may become too expensive for applications that demand an efficient prediction of the saliency map. Many of these applications lie in robotics, where real-time is usually a requirement, and computational resources are scarce. Given that saliency prediction in robotics has been extensively used to further speed-up other visual tasks for the robot, *e.g.* [3, 8, 28, 30, 36], it is all the more important that the saliency prediction is also extracted efficiently.

Recently, active semantic segmentation has been introduced to extract a semantic labeling given a budget of time [26]. Active semantic segmentation evaluates object classifiers in a reduced subset of regions in the image, rather than everywhere, and it infers an estimate of the probabilities of the semantic labeling for the whole image. In this paper, we introduce a saliency prediction algorithm that uses active semantic segmentation to efficiently exploit semantic-level features. We introduce new semantic-level features for saliency prediction based on the probabilistic output from active semantic segmentation, which are the probability, uncertainty and the rarity of the semantic class. An overview of our method is illustrated in Figure 1.

To carry the analysis of our model, we collected eye tracking data on two popular segmentation datasets, the PASCAL VOC07 [10] and the MSRC-21 [29] datasets. The data will be shared with the community and we hope to facilitate interactions of saliency research with semantic segmentation. Experiments on these datasets, show that saliency prediction accuracy improves significantly when using the features built from active semantic segmentation, and using these features does not compromise the computational efficiency of the algorithm.

2 Related Works

Semantics have been shown to play an important role in improving saliency prediction [9, 20, 34], and recent computational models have incorporated object detectors into saliency models [7, 15, 38, 39]. For example, Cerf *et al.* [7] showed that humans look at faces strongly and rapidly, independently of the task, and adding a face detector into the saliency model consistently boosts prediction performance. Judd *et al.* [15] added pedestrian, and car detectors to the low-level features and face detectors, and combined them with a linear SVM. Zhao *et al.* proposed data driven methods to integrate face detectors and low-level image features with least square regression [38], and boosting [39]. The additional computational cost of using features at multiple levels is remarkable for the semantic-level features, which involve evaluating object classifiers for all regions in the image, and do not scale well with the number of semantic classes that are taken into account. Thus, the systems that are used for efficient applications that rely on saliency prediction, *e.g.* robotic applications [3, 8, 28, 30, 36], usually drop the use of semantic features to maintain a balance between accuracy and efficiency. In this paper, we explore the use of active semantic segmentation, which does not evaluate the object classifiers everywhere in the image, and allows extracting the semantic information given a budget of time.

3 Preliminaries: Active Semantic Segmentation

The active semantic segmentation framework by [26] was introduced to obtain a computationally efficient semantic segmentation pipeline. It is assumed that evaluating object classifiers everywhere in the image is much more computationally expensive than inferring the semantic labeling from a probabilistic model.

3.1 Object Classifiers Evaluated Everywhere

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with nodes in $\mathcal{V}$ and undirected edges in $\mathcal{E}$ be the graph that represents the probability distribution of the semantic labeling, modeled by a Conditional Random Field (CRF) [17]. Let $\mathbf{X} = \{X_i\}_{i \in \mathcal{V}}$ be the set of random variables corresponding to the object class labels of each node. We denote an instance of the random variables as $\mathbf{x} = \{x_i\}$, where x_i takes a value from a set of semantic class labels $\mathcal{L}$. Thus, $\mathbf{x} \in \mathcal{L}^N$, in which N is the cardinality of $\mathcal{V}$ (the number of semantic classes).

The probability density distribution of a labeling modeled with the graph $\mathcal{G}$ is denoted as $P(\mathbf{x}|\theta)$, which can be written as the normalized negative exponential of an energy function $E_\theta(\mathbf{x}) = \theta^T \phi(\mathbf{x})$, in which $\phi(\mathbf{x}) = (\phi_1(x), \ldots, \phi_M(x))^T$ is the vector of potentials of the CRF, and $\theta \in \mathbb{R}^M$ are the parameters of the potentials. We use the common energy function in semantic segmentation with unary potentials and pairwise potentials, expressed as

$$E_\theta(\mathbf{x}) = \theta^T \phi(\mathbf{x}) = \sum_i \psi_i(x_i) + \lambda \sum_{i,j} \psi_{ij}(x_i, x_j). \quad (1)$$

The unary potential, denoted as $\psi_i(x_i)$, is defined as the output of a object classifier that evaluates the semantic class label of node X_i. The pairwise potential, denoted as $\psi_{ij}(x_i, x_j)$ and weighed by λ, is equal to 0 when $x_i = x_j$, and a value that depends on the similarity between regions otherwise. The most probable labeling $\mathbf{x}^\star$ can be obtained by minimizing the energy function $E_\theta(\mathbf{x})$. The calculation of $\mathbf{x}^\star$ is the so called MAP inference, which is commonly done with Loopy Belief Propagation [19] or Graph Cuts [4].

3.2 Object Classifiers Evaluated in a Subset of Regions

Active semantic segmentation is formulated as a probabilistic model with some unknown parameters, which correspond to the object classification scores that have not been evaluated in some regions. Active Semantic Segmentation allows to estimate which regions are best to be selected to evaluate object classifiers. Also, the probabilistic model allows to propagate the semantic information to these regions that have not been observed, and it provides an estimate of the probabilities of the semantic labeling in all the image. These probabilities will be used to design features for saliency prediction, as we show in the next Section.

We introduce an indicator vector δ, in which each entry δ_i is set to 1 if the potential of node i has been computed, and 0 otherwise. We use θ_δ to denote the computed potentials, while the unknown potentials are modeled with a probability distribution $f(\theta_i)$. This probability distribution can be learned from training data or fixed by hand. In our case, we define $f(\theta_i)$ as a Gaussian distribution with mean and standard deviation fixed to constant values learned via cross-validation, as in [26]. From each sample drawn from $f(\theta_i)$, we can run MAP inference and obtain a labeling $\mathbf{x}^\star$. This produces a distribution of MAP labelings due to the unknown parameters of the energy function. Let $P(\mathbf{X}^\star = \mathbf{x}|\theta_\delta)$ be this probability distribution of the MAP labeling.

The selection of object classifiers to evaluate is done with a score based on features extracted from $P(\mathbf{X}^\star = \mathbf{x}|\theta_\delta)$. The score ranks the unary potentials that are candidates to be evaluated with object classifiers, and the potentials that rank with a higher score are evaluated. We use the expected entropy score as in [26]. We refer to [26] for further details on Active MAP Inference.

4 Saliency Estimation with Feature Integration

Based on the output of the active semantic segmentation (*i.e.* a set of semantic labelings of the full image that define the probability distribution of the semantics given few observations on the image), we now introduce the saliency prediction model to integrate levels of features including a new set of semantic-level features. We use superpixels instead of pixels as the base representation for saliency prediction, as the visual attention is attracted by an extended region that represents an object or a part of an object, and not by each single pixel [16].

We use a Support Vector Regression (SVR) to learn feature integration for saliency prediction, as suggested by [1]. In order to train the model, ground-truth fixation maps of the images are generated from their eye-tracking data. In the fixation map of an image, all fixated pixels were represented as white and the rest as black. The map is then convolved with an one-degree-sized Gaussian kernel representing the idealized fovea radius [34]. The generated fixation maps are averaged within each superpixel. In the following subsections, we introduce the features that are integrated with the SVR.

4.1 Low- and Regional-Level Features

We use features for the low and regional-levels that are already available in the literature, because they were shown to improve the saliency prediction accuracy [34].

Low-Level Features. We incorporate the state-of-the-art GBVS [12] method to calculate a combined low-level saliency feature. It also acts as a baseline of our saliency model. The GBVS saliency map is generated based on three biologically-plausible saliency channels,

namely color, intensity, and orientation [14]. In the proposed superpixel-based framework, the pixel-wise feature values are averaged within each superpixel.

Regional-Level Features. The Gestalt principles suggest many important factors in the grouping of basic visual elements and the segregation of foreground and background [22]. Locally coherent regions or proto-objects, are more likely to attract attention than others depending on the size and shape. Recent computational models have attempted to predict attention based on such regional-level features [27, 33, 35]. In this work, we use five regional-level features (*i.e*. *size*, *solidity*, *convexity*, *complexity*, *eccentricity*) proposed by [34], which are independent of the semantics, and effective in predicting saliency.

4.2 Features from Active Semantic Segmentation

We now introduce the features that we extract from the set of samples of semantic labelings. Given a budget of time, active semantic segmentation evaluates object classifiers in a subset of regions, selected using the expected entropy score (see Section 3). We compute the following features from a set of 25 semantic labelings sampled from $P(\mathbf{X}^\star = \mathbf{x}|\theta_\delta)$. We use $\{\mathbf{s}^k\}$ to denote this set of MAP labelings drawn from $P(\mathbf{X}^\star = \mathbf{x}|\theta_\delta)$, where k indexes the set of samples.

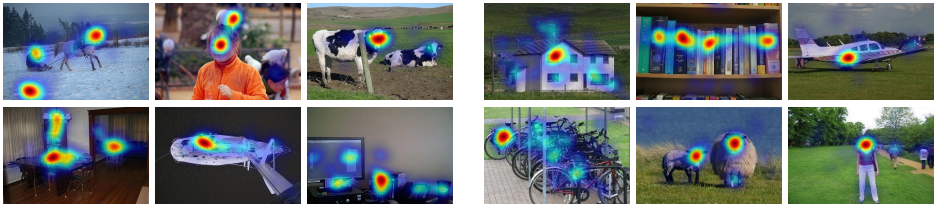
Label Probability. It has been shown that certain object categories attract attention more strongly and rapidly than others [7, 34]. To define a list of object categories of interest, we propose to leverage on current datasets and borrow the object categories in the PASCAL VOC07 [10] and the MSRC-21 [29] datasets that are commonly used in the community. Particularly, there are 20 classes in the VOC07 dataset including person, chair and dog, and 21 classes in the MSRC-21 dataset including face, body, tree, water. The label probability is computed as the normalized distribution of class labels from the samples of the active semantic segmentation, $\{\mathbf{s}^k\}$. Thus, for each superpixel we extract one feature per semantic class, which for class c and superpixel indexed by i is denoted as $p_i(c)$, and it is

$$p_i(c) = \frac{1}{|\{\mathbf{s}^k\}|} \sum_k \mathbf{I}[s_i^k = c], \tag{2}$$

where $|\{\mathbf{s}^k\}|$ is the number of samples in the set, and $\mathbf{I}[]$ is the indicator function. Note that active semantic segmentation allows to compute this probability in all the image even though we only evaluated object classifiers in few locations in the image.

Semantic Uncertainty. Several studies point out that the uncertainty about the information at low- and regional-levels plays a crucial role in understanding human gaze allocation [24, 25]. We explore the uncertainty about the semantic labeling. We evaluate the Shannon entropy of the label probability in a superpixel, which is $\sum_c p_i(c)\log_2(p_i(c))$.

Semantic Rarity. Global rarity captures the likelihood that certain image features represent a distinct object compared to the background in natural scenes. While recent models started to encode global rarity of low-level features over an entire visual scene [2, 5, 37], rarity at the semantic-level has not been explored yet. We first compute a final semantic labeling from the set of samples of the active segmentation. This is done by assigning to each superpixel the class label with higher probability, *i.e*. the class label c that maximizes $p_i(c)$. Let M be the number of predicted semantic classes in the final semantic labeling, and let A_l be the area of each semantic class, where $l = 1,\ldots,M$. The rarity of each class is defined as $-\log(S_l/\sum_{l=1}^{M} S_l)$. This generates one feature per superpixel, which is the rarity corresponding to the superpixel class in the final semantic labeling.



(a) (b)

Figure 2: *Sample images from (a) VOC07 and (b) MSRC-21 datasets*. The eye-fixation density maps from all subjects are overlaid on the images in heat maps.

Object Center. One scenario where discrepancy often happens between a low-level based computational saliency model and human behavior is that the model tends to highlight object boundaries that are normally with high pixel-level contrast. In comparison, humans tend to look at the center regions of an object. The final labeling from the active semantic segmentation allows estimating the object centers. Thus, for each semantic class, we estimate the object center as the centroid of the corresponding image region. Then, to generate a feature that models the object center preference, we evaluate the distance from the superpixel to its center object. We use a Gaussian function to evaluate the spatial distance, of size approximately 2 degrees of visual field, which is about 15 pixels in MSRC-21, and 35 pixels in VOC07. In VOC07, the object center is not computed on the background.

5 Eye-Tracking on VOC07 and MSRC-21 Datasets

We introduce the dataset with semantic labelings and eye-fixations ground-truth. We conducted eye-tracking experiments on two popular semantic segmentation datasets, with semantic segmentation ground-truth already available, namely the PASCAL VOC07 [10] and the MSRC-21 [29]. The eye-tracking data is publicly available at `http://www.ece.nus.edu.sg/stfpage/eleqiz/bmvc15.html`. We show in Figure 2 sample images from VOC07 and MSRC-21 datasets with eye fixations overlaid. The motivation of the data collection procedure is to validate our proposed method, as well as to facilitate research across saliency and semantic segmentation.

All participants viewed the full set of images freely at a 57 cm distance in front of a 22-inch LCD monitor. Their eye movements were recorded by an Eyelink 1000 (SR Research, Osgoode, Canada) eye-tracking device, at a sample rate of 1000 Hz. The screen resolution was 800×600, and the images were uniformly scaled for the full-screen presentation. In the experiment, each image was presented for 3 seconds, followed by a drift correction that required subjects to fixate in the center and press the space key to continue.

PASCAL VOC07 [10]. This dataset contains 422 images equally divided in training and validation sets, and 210 test images. It contains 20 different labeled object classes plus background. We recruited 14 university students (5 females and 9 males, aged between 19 and 28) to participate in the eye-tracking experiment.

MSRC-21 [29]. The dataset includes a total of 591 images, split in training and testing sets. It has fully labeled images with 21 different classes in which the background is divided into several classes such as road, water and sky. Eye-tracking data were collected from 14 university students (8 females and 6 males, aged between 20 and 32) viewing all images.

6 Experiments

In this section, we report experiments to show the effectiveness of the proposed model. After introducing the implementation details, we evaluate the performance of our method.

6.1 Implementation Details

The implementation details for active semantic segmentation and saliency prediction are reported below.

Superpixel Segmentation. The images are first over-segmented using the SEEDS superpixel algorithm [32]. The VOC07 images are over-segmented with about 600 superpixels, and the ones of MSRC-21 with about 300, as reported in [26].

Unary Potentials. In MSRC-21 dataset we use the features and classifiers reported in [18]. In VOC07 we use the ground-truth labels for unary potentials. When using the true semantic label, we take the most occurring ground-truth label for each superpixel and assign it to the superpixel. The unary potentials to be instantiated are selected as in [26], until a percentage of the superpixels is observed (this percentage is indicated in each experiment). We implemented the score based on the expected entropy reward.

Pairwise Potentials. The pairwise potential is the common modulated Potts model with color difference. The parameter to modulate is the negative exponential of the difference between the mean of the RGB color of the superpixels that are spatially connected in the image. We learn the slope of the negative exponential as we did with the aforementioned Gaussian distribution of the unknown potentials.

Inference. We use α-expansion graph cuts [4] to compute the MAP labeling in a complete energy function. We compute 25 samples of semantic labels from the active semantic segmentation, to compute the features for saliency prediction.

SVR for Saliency Prediction. We use all superpixels in the training images as samples to learn the linear SVR, which allows the regressor to automatically ignore the samples with small regression errors. Superpixels are normalized to have zero mean and unit standard deviation in the feature space. We integrate the features with a L2-regularized SVR [11], fixing the C parameter of the SVR to 1. We have also tested lasso-type algorithms for the same purpose, but no advantages were found in the targeted tasks.

6.2 Performance Evaluation

In the following, we analyze the contributions of features at each level, by comparing our model with the different combinations of feature sets. Then, we compare our model to state-of-the-art methods. We use three complementary evaluation measures (*i.e.* AUC [31], NSS [23], and CC [21]) that are commonly used to evaluate saliency models.

Comparison to Baselines. We design baselines incrementally adding feature sets from the low-level GBVS model to the semantic-level, and learn an SVR model with each set of features. We evaluate them by varying the percentage of evaluated object classifiers over the superpixels, to show how saliency prediction performance changes with different computational costs. From Figure 3, we can observe that for both datasets, the use of semantic information improves the performance of saliency prediction, suggesting the importance of semantic content in predicting gazes. As expected, there is an increase in the performance with more observed superpixels, and full observability of the object classifiers in all superpixels achieves the best results. The results show that the proposed semantic features are able

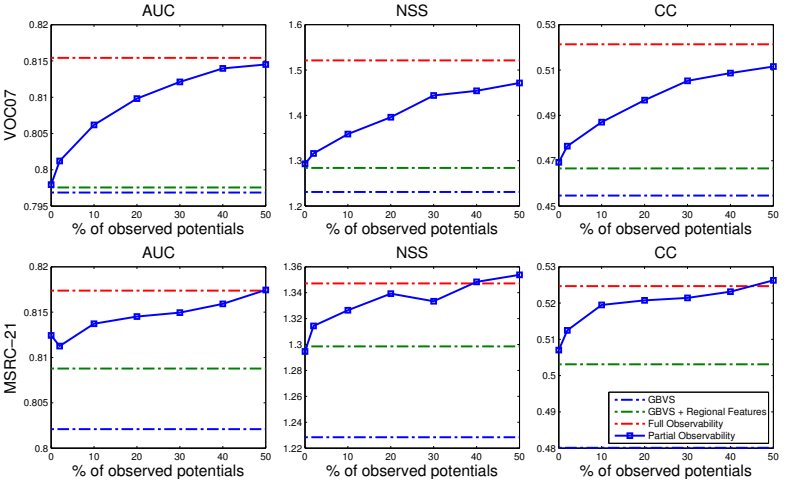


Figure 3: *Quantitative comparison of the baselines*. Performance evaluation of saliency prediction in the VOC07 and MSRC-21 datasets, with various percentages of observations.

to capture useful semantic information for saliency prediction, since with a low percentage of observed superpixels, the saliency performance is close to the full observability model.

We can extract similar conclusions for both datasets. The difference in the results between VOC07 and MSRC-21 is that in VOC07, for low percentage of observed superpixels, the performance is not as relatively high as in MSRC-21. The same can also be observed from the regional features. These may be caused by occlusions, object scale, and intra class variability that are much stronger in VOC07, and hence, the propagation of the semantic information under partial observability is not as effective as in MSRC-21. Another reason is that the background classes in VOC07 are not labeled, and hence, our model can not capture any semantic information from objects in the background. This may be important for saliency prediction since some objects in the background may receive eye-fixations.

Comparison to state-of-the-art. We compare the proposed model with several state-of-the-art saliency algorithms: the classic Itti-Koch model [14] (denoted as Itti), the baseline GBVS model [12], AIM [6], SUN [37], and the Image Signature (IS) [13]. Many saliency models implicitly or explicitly blur the output saliency maps to increase their performance, as current saliency evaluation metrics are sensitive to the blurring. To make the comparison fair, we use the standard deviation of the Gaussian blurring in viewer's degree of visual field as a parameter, and explicitly cross-validate this parameter on the validation set to optimize the comparative performance of each algorithm. We restrict this parameter between 0 and 2 to prevent the over-blurred maps from becoming a model of the global center bias.

Figure 4 illustrates the model performance with different blurring parameters, and the optimal scores of all models are compared in Table 1. The results shows that our model with active semantic segmentation is competitive with state-of-the-art, under various percentages of observed superpixels. We can see that the performance of the baseline GBVS is higher than most low-level models, this is because GBVS intrinsically models the center-bias. Note that with the semantic-level features we proposed, our model outperforms the low-level feature integrated in the model, *i.e*. GBVS. This result agree with previous findings about that the semantic features are useful for predicting human visual attention [7, 15, 34, 38, 39].

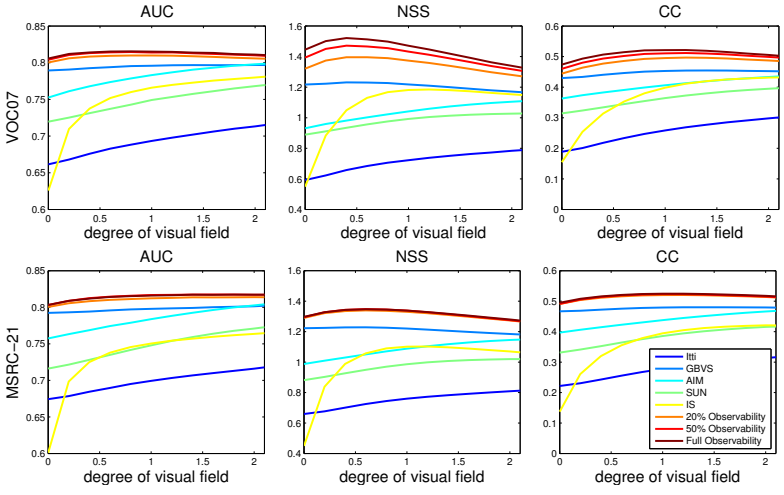


Figure 4: *Comparison to state-of-the-art.* Saliency prediction performances under different Gaussian blurring conditions. Note that the methods have different sensitivities to the blurring. We compare our model to: Itti-Koch model [14] (denoted as Itti), the baseline GBVS model [12], AIM [6], SUN [37], and the Image Signature (IS) [13].

Table 1: *Summary of the results.* Performance of the different models under optimized blurring conditions.

Model Name	**VOC07**			**MSRC-21**		
	AUC	NSS	CC	AUC	NSS	CC
Itti [14]	0.715	0.788	0.301	0.718	0.812	0.316
GBVS [12]	0.797	1.231	0.455	0.802	1.229	0.480
AIM [6]	0.799	1.108	0.435	0.804	1.148	0.468
SUN [37]	0.770	1.028	0.397	0.773	1.020	0.417
IS [13]	0.781	1.185	0.432	0.765	1.102	0.421
20% Observability	0.810	1.396	0.497	0.814	1.340	0.520
50% Observability	**0.815**	1.472	0.512	**0.817**	**1.348**	0.523
Full Observability	0.815	**1.521**	**0.521**	0.817	1.347	**0.525**

Finally, note that only observing 20% of the regions in the image, the saliency prediction accuracy is similar as when evaluating classifiers everywhere in the image. Thus, 20% is a good compromise in terms of efficiency and accuracy, since it achieves much higher accuracy than without using semantic-level features, and similar levels of accuracy with a 5× speed up as with 100% observability.

Qualitative assessment. The saliency maps generated by these models are demonstrated in Figure 5. We show the results obtained with 20% of observed object classifiers. As can be seen, with the features at the semantic-level that we introduced, our model approximately estimates the locations of salient objects, and thereby predicts the saliency better than the other models.

Computational Cost. The computational cost of the low- and regional- level features is negligible compared to the cost of extracting the semantic information. Thus, the final computational cost of saliency prediction is bounded to the computational cost of the active

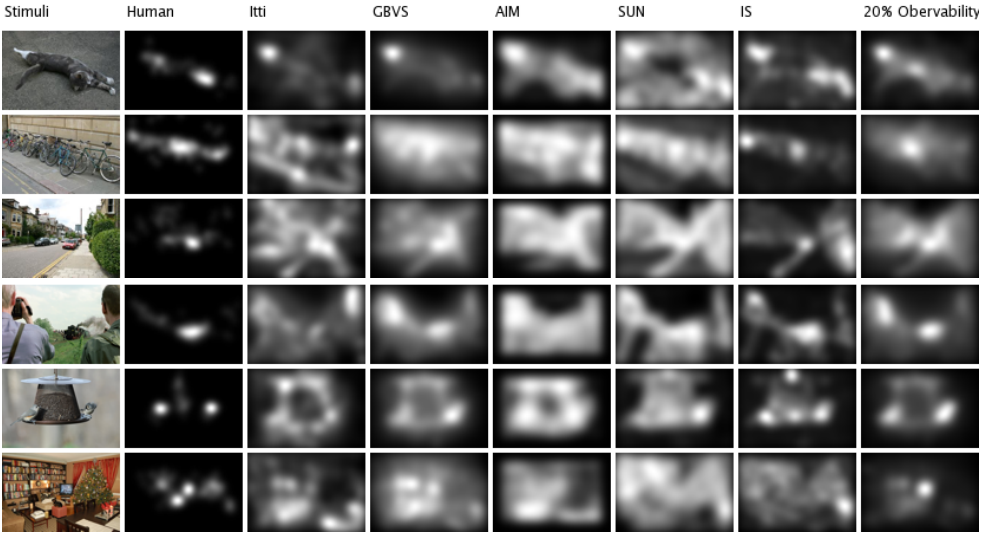


Figure 5: *Qualitative results.* Comparison with the state-of-the-art models and human fixations. Output saliency maps are blurred with optimal Gaussian kernels.

semantic segmentation. Roughly, the computational cost is proportional to the amount of evaluated object classifiers, since they are the computational bottleneck. Thus, with 20% of observed object classifiers, the speed up is of about 5×. We run our experiments in an Intel CPU 2.8GHz i7 with 8 cores. For the MSRC-21 dataset, the computational cost without active semantic segmentation is about 0.3 fps. When evaluating 20% we surpass the 1 fps boundary, namely we achieve 1.3 fps, and with 5%, it is 3.3 fps.

7 Conclusions

We introduced an efficient saliency prediction model using active semantic segmentation. The proposed semantic features can be extracted efficiently given a budget of time. We evaluated them in a new dataset of eye-fixations on two popular datasets for semantic segmentation (MSRC-21 and VOC07). Results demonstrated the effectiveness of the semantic features for saliency prediction under several computational time constraints.

Acknowledgements

The research was supported by the Defense Innovative Research Programme (No. 9014100596), the Ministry of Education Academic Research Fund Tier 1 (No. R-263-000-A49-112), and the ERC Advanced Grant VarCity.

References

[1] L. Bi, O. Tsimhoni, and Y. Liu. Using the support vector regression approach to model human performance. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 41(3):410–417, May 2011.

[2] A. Borji and L. Itti. Exploiting local and global patch rarities for saliency detection. In *Proc. Computer Vision and Pattern Recognition*, 2012.

[3] A. Borji, M.N. Ahmadabadi, B.N. Araabi, and M. Hamidi. Online learning of task-driven object-based visual attention control. *Image and Vision Computing*, 28:1130–1145, 2010.

[4] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004.

[5] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *Advances in Neural Information Processing Systems*, 2006.

[6] N.D.B. Bruce and J.K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *J. of Vision*, 9(3), 2009.

[7] M. Cerf, E.P. Frady, and C. Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *J. of Vision*, 9(12), 2009.

[8] A. Dankers, N. Barnes, and A. Zelinsky. A reactive vision system: Active-dynamic saliency. In *Proc. of the International Conference on Computer Vision Systems*, 2007.

[9] W. Einhäuser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *J. of Vision*, 8(14), 2008.

[10] M. Everingham, L. Van Gool, C. KI Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *Int. Journal of Computer Vision*, 88(2):303–338, 2010.

[11] R.-F. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *J. of Machine Learning Research*, 9:1871–1874, 2008.

[12] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, 2007.

[13] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):194, 2012.

[14] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (11):1254–1259, 1998.

[15] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Proc. IEEE Int. Conf. on Computer Vision*, 2009.

[16] G. Kanizsa. *Organization in vision: Essays on Gestalt perception*. Praeger New York, 1979.

[17] S. L. Lauritzen. *Graphical models*. Oxford statistical science series. London: Oxford University Press, 1996.

[18] A. Lucchi, Y. Li, X. Boix, K. Smith, and P. Fua. Are spatial and global constraints really necessary for segmentation? In *Proc. IEEE Int. Conf. on Computer Vision*, 2011.

[19] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proc. Annual Conference on Uncertainty in Artificial Intelligence*, 1999.

[20] A. Nuthmann and J.M. Henderson. Object-based attentional selection in scene viewing. *J. of Vision*, 10(8), 2010.

[21] N. Ouerhani, R. von Wartburg, H. Hugli, and R. Muri. Empirical validation of the saliency-based model of visual attention. *Electronic letters on computer vision and image analysis*, 3(1):13–24, 2004.

[22] S.E. Palmer. *Vision science: Photons to phenomenology*, volume 1. MIT press Cambridge, MA, 1999.

[23] R.J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18):2397–2416, 2005.

[24] K. Rayner. Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology*, 62(8):1457–1506, 2009.

[25] L. W. Renninger, P. Verghese, and J. Coughlan. Where to look next? eye movements reduce local uncertainty. *J. of Vision*, 7(3):6, 2007.

[26] G. Roig, R. Boix, X.and de Nijs, S. Ramos, K. Kühnlenz, and L. Van Gool. Active map inference in crfs for efficient semantic segmentation. In *Proc. IEEE Int. Conf. on Computer Vision*, 2013.

[27] A. F. Russell, S. Mihalacs, R. Von der Heydt, E. Niebur, and R. Etienne-Cummings. A model of proto-object based saliency. *Vision Research*, 94:1–15, 2014.

[28] C. Scheier and S. Egner. Visual attention in a mobile robot. In *Proc. Int'l Symp. Industrial Electronics*, 1997.

[29] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *Int. Journal of Computer Vision*, 2009.

[30] C. Siagian and L. Itti. Biologically inspired mobile robot vision localization. *IEEE Transactions on Robotics*, 25(4):861–873, 2009.

[31] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45(5):643–659, 2005.

[32] M. Van den Bergh, X. Boix, G. Roig, B. de Capitani, and L. Van Gool. SEEDS: Superpixels extracted via energy-driven sampling. In *Proc. European Conf. on Computer Vision*, 2012.

[33] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006.

[34] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao. Predicting human gaze beyond pixels. *J. of Vision*, 14(1):1–20, January 2014.

[35] J. Yu, J. Zhao, J. Tian, and Y. Tan. Maximal entropy random walk for region-based visual saliency. *IEEE Transactions on Cybernetics*, 44(9):1661–1672, Sept 2014.

[36] Y. Yu, G. K I Mann, and R. G. Gosine. An object-based visual attention model for robotic applications. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 40(5):1398–1412, 2010.

[37] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, and G.W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *J. of Vision*, 8(7), 2008.

[38] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *J. of Vision*, 11(3), 2011.

[39] Q. Zhao and C. Koch. Learning visual saliency by combining feature maps in a non-linear manner using adaboost. *J. of Vision*, 12(6), 2012.