# SPATIAL BAYESIAN SURPRISE FOR IMAGE SALIENCY AND QUALITY ASSESSMENT

*Ioannis Gkioulekas*[1]*,Georgios Evangelopoulos*[2]*, Petros Maragos*[2]
[1] Harvard SEAS, Cambridge, MA 02138, USA
[2] School of ECE, National Technical University of Athens, Zografou, 15773 Athens, Greece
igkiou@gmail.com [gevag,maragos]@cs.ntua.gr

## ABSTRACT

We propose an alternative interpretation of Bayesian surprise in the spatial domain, to account for saliency arising from contrast in image context. Our saliency formulation is integrated in three different application scenaria, with considerable improvements in performance: 1) visual attention prediction, validated using eye- and mouse-tracking data, 2) region of interest detection, to improve scale selection and localization, 3) image quality assessment to achieve better agreement with subjective human evaluations.

***Index Terms***— Image saliency, Bayesian surprise, visual attention, region detection, image quality assessment.

## 1. INTRODUCTION

The human visual attention system has been for long a subject of research in psychophysics and cognitive sciences, due to its prominent role in biological vision. Significant efforts have also been made in computer vision to construct a computational model of this system, due to the potential for efficient, application-specific and perceptual resource allocation. Attention in this context has been used to achieve critical improvements in applications as diverse as object recognition [1], video summarization and image quality assessment [2–5] among others.

Saliency-based methods have arguably been the most popular in these modeling efforts. Early such methods drew inspiration from biological models of the human visual sensory system [6]. Later approaches have proposed calculating saliency based on a Bayesian framework [7], using the notion of region similarity [8] or natural image statistics [7]. Other models have relied on perceptual coherence theory [1] to make object-based saliency characterizations.

Binding intuition to modeling, several information theoretic measures have been used to quantify saliency, based on ideas stemming from information theory [9–12]. In [9], saliency is measured using the entropy of local feature distributions, under the intuitive notion that high entropy regions exhibit high complexity, and therefore are unpredictable and likely to be fixated by an observer. Self-information, as resulting from a prior local model for each region, is proposed as an alternative measure of unpredictability in [7, 10], where it is argued that high self-information indicates a region unlikely to occur and thus interesting. In [11], saliency is identified by the discriminative power of features with respect to center and surround regions, which in turn is quantified as the mutual information between the features and each region class.

An alternative measure relates to the notion of *Bayesian surprise* [12]. The surprise caused by an observation is defined in a Bayesian sense as the change it brings to an observer's prior beliefs with respect to the phenomenon under consideration. Bayesian surprise is closely related to self-information and mutual information [10, 11], as they all describe different aspects of the contrast of observations to their context, expressed by some prior beliefs. In [12], it was shown that surprise can be used as a measure for saliency induced by changes in the temporal dimension, for example, in video streams. Although surprise is used also as a measure of spatial significance within each frame, saliency only arises due to changes in the temporal neighborhood. Spatial instead of temporal contrast was exploited in [10, 11]. The performance of these models was compared to that of [6], however due to the use of different feature sets (a basis of independent components for the former, a gabor filterbank for the latter), it is unclear to what extent the improvement can be attributed to the information theoretic setting. In this paper, we propose that surprise can also be employed to explain saliency arising from spatial contrast in static images.

In detail, we adopt an information-theoretic approach to study bottom-up spatial saliency. We show how *Bayesian surprise* [12] can be interpreted to explain spatial saliency, and we adapt the model of Itti-Koch [6] for the validation of our hypothesis, using the same features to avoid ambiguities rising from differences in performance. Further, we use surprise to modify the region detector of Kadir-Brady [9] and achieve better localization and scale selection. We also validate our method in the context of image quality assessment, by exploring possible methods for integration with the Structural Similarity Index Metric (*SSIM*) [13].

## 2. SPATIAL INTERPRETATION OF BAYESIAN SURPRISE

The mathematical formulation of the notion of surprise was introduced in [12], where it was defined as follows: given a prior distribution $P(M)$ over a (discrete) space of models $\mathcal{M}$ describing a phenomenon observed, and the posterior distribution $P(M|D)$ after new data $D$ is obtained for this phenomenon through an observation, then the surprise incurred by $D$ relative to the space $\mathcal{M}$ is given by the Kullback-Leibler divergence $K(\cdot \,\|\, \cdot)$ of the prior from the posterior distribution,

$$\begin{aligned} S(D, \mathcal{M}) &= K(P(M) \,\|\, P(M|D)) \\ &= \sum_{M \in \mathcal{M}} P(M) \log \frac{P(M)}{P(M|D)}. \end{aligned} \tag{1}$$

Bayesian surprise can be used for images to explain the saliency of regions that, compared to the rest of the image, exhibit irregular characteristics. Images demonstrate a consistency of characteristics between spatially neighboring regions, unless a significant change occurs in the content of these regions, such as the sudden appearance of an object of different visual features not present in other regions. Due to this, before actually observing a region, an observer expects it

This work was performed while the first author was at NTUA. Research was supported in part by the EU under the research program Dictasign with grant FP7-ICT-3-231135.


(a) region 1 (b) image (c) region 2

**Fig. 1**: Spatial interpretation of surprise for two regions (yellow circles) and their respective context (rings between red and yellow circles). Intensity distributions are shown for the interior (red) and context (blue) of regions 1 (surprise = 0.14) and 2 (surprise = 2.99).

to most likely be of the same characteristics as the rest of the image. Thus, the *spatial context* of the region allows the observer to create a prior model about its actual content. Such a prior model will be inadequate for regions where events occur, hence the observer will need to compensate for them by creating a posterior model significantly different from the prior. The observation of these regions evokes surprise to the observer, both intuitively and formally as previously defined. Under this intuition, it is appealing to define as salient the regions that are surprising with regards to their context.

The bottom-up, visual cues of each region can be described in terms of image features, and beliefs about them in terms of distributions over all possible feature values. To formulate this notion using the formal definition of Bayesian surprise, the phenomenon observed are the feature values inside the region under examination, and the model space is the set $\mathcal{M}$ of all possible values for that feature. The visual context of the region implies a prior distribution $P(M)$ over the model space before the region is observed. After the region is observed, a posterior distribution $P(M|D)$ is formed, where $D$ is the data acquired from the observation of the region. The surprise incurred by $D$ relative to the space $\mathcal{M}$ is then given by (1). This alternative interpretation is demonstrated in Fig. 1, where the periphery of a circular region serves as its context.

Under this formulation, our proposed saliency measure resembles that of [11], which under some assumptions takes the form of an average of the divergence between the feature distribution in each of the surround and center regions from the distribution in their union. The similarity stems from the fact that we do not explicitly use Bayes' rule to update from the prior to the posterior, instead we directly identify them with the distributions in the surround and center regions respectively. However, the two metrics are conceptually different: mutual information assumes that observations carry information about deciding between two different classes, whereas surprise suggests that observations provide information on how an observer should adjust his prior belief. This distinction would have been clearer if we adopted a parametric model for the feature distributions, which would allow the use of Bayes' rule, but here we preferred to make no assumptions about the form of distributions.

## 3. VISUAL ATTENTION USING SURPRISE

In order to validate our proposed framework, we augment the visual attention model of [6] using surprise. In this model, feature maps are formed for intensity, color and local orientation, and fused into a final saliency map using center-surround differences and a process for the amplification of isolated peak responses. A winner-take-all neural network then produces fixations to the most salient regions.

To integrate surprise in this architecture, we use the feature maps to calculate surprise values for each spatial location and calculate surprise for circular regions by considering their periphery as the region's visual context, following center-surround, early vision strategies. The posterior and prior distributions are then estimated from the normalized histograms of the corresponding feature values in the two regions, respectively, as shown in Fig. 1. The radii of the center and surround regions are chosen adaptively as a fraction of the dimensions of the input image. The surprise maps thus formed for each feature are then used by the rest of the processing scheme.

| Metric | Dataset 1 (40 images, mouse-tracking data) | | | Dataset 2 (120 images, eye-tracking data) | | |
|---|---|---|---|---|---|---|
| | [6] | Surprise | [10] | [6] | Surprise | [10] |
| correlation | 0.1382 | 0.2627 | 0.2410 | 0.1861 | 0.2687 | 0.3116 |
| 1−Hausdorff | 0.4628 | 0.5075 | 0.4380 | 0.4596 | 0.4852 | 0.2543 |
| overlap | 0.1991 | 0.2282 | - | 0.2046 | 0.2397 | - |

**Table 1**: Average values of the three evaluation metrics over images of datasets 1 and 2 for different saliency models.

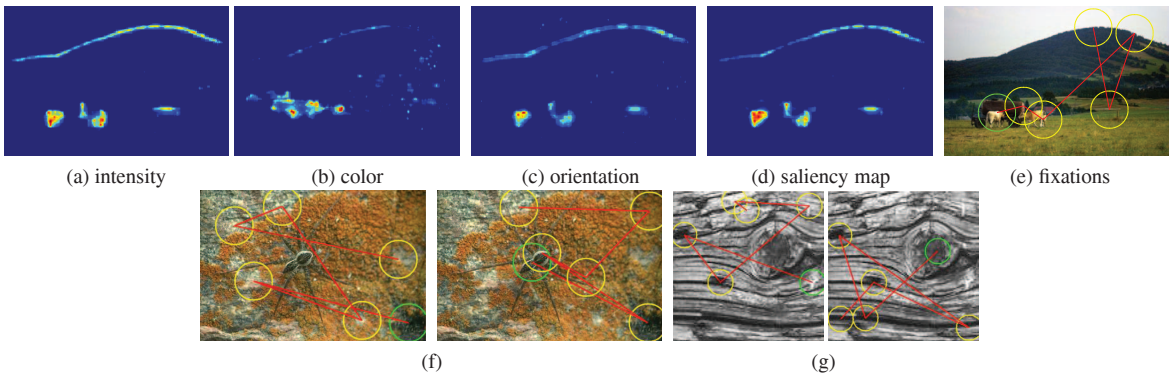
Calculating surprise values for each feature map separately is, in this case, an approximation we make for computational efficiency. This approximation is exact if each feature channel $D_i, i = 1, \ldots n$ is assumed to be independent of the others. Then, if $\mathbf{D}$ is the $n-$ dim feature vector and $\mathbf{M}$ the $n - D$ discrete space of models, it can be easily proven that $S(\mathbf{D}, \mathbf{M}) = \sum_{i=1}^{n} S(D_i, \mathcal{M}_i)$. The alternative would be to use high-dimensional histograms to estimate the joint distribution of values for all features, which would be expensive and produce poor estimates due to the curse of dimensionality. Although the assumption of independence does not generally hold, we can expect the selected features to be significantly non-redundant. Informal experimentation with the two approaches, using only the color channels, has shown that the difference in performance is minimal, whereas the improvement in computational efficiency is substantial.

**Validation:** We test our system with respect to its ability to predict human eye movements, using two different datasets. The first set (denoted as Dataset 1) includes mouse-tracking data collected from 21 to 31 naive observers for each of 40 images using the interface of *ValidAttention*[1] [14]. Saliency maps were created from this data by averaging the responses of all observers and applying gaussian smoothing. The second set (Dataset 2) is the eye fixation data used in [10], which is publicly available[2].

Our modification to the visual attention system of Itti-Koch [6] is evaluated against the original[3] and the system proposed by Bruce-Tsotsos [10]. Performance is measured by examining the similarity between the estimated saliency maps and those from the ground-truth data. As similarity metrics, we use the *sample correlation*, which takes values in the interval $[-1, 1]$, and the 1-complement of the normalized to $[0, 1]$ *grayscale Hausdorff distance*. We also examine the matching of eye fixations between regions produced by the computational and the ground-truth saliency map, measuring the *normalized overlap* area. We only consider the first six fixations of each sequence, as later fixations in the ground-truth would be increasingly influenced by top-down attentive mechanisms, e.g. Fig. 2.

In Table 1, we show the average value of each of the above metrics over the two image datasets, for the three models under consideration. Higher values indicate better performance. Note that, the third metric is not used for the comparison with model [10], as this does not produce fixations. It can be seen that our modification consistently outperforms the original model [6]. This is evident in the presence of strong texture features, e.g. in the two examples of Fig. 2(f),(g), where our surprise-based model correctly identifies the most salient object of the scene, without being distracted to regions of high feature variation. This can be understood if we consider that

[1] http://tcts.fpms.ac.be/~mousetrack
[2] http://www-sop.inria.fr/members/Neil.Bruce
[3] http://www.saliencytoolbox.net



(a) intensity (b) color (c) orientation (d) saliency map (e) fixations

(f) (g)

**Fig. 2**: First row: surprise features, saliency map and fixations for a natural image. Second row: fixations for two textured images by [6] (left) and surprise (right). The fixation in green is the first, and the red lines connect consecutive fixations. Best seen in color.

surprise uses the statistics of feature values and not the local values individually when calculating saliency, and therefore succeeds in characterizing regions as non salient where these statistics do not change. Compared to the model of [10], our method performs better for both metrics in Dataset 1, but only with respect to Hausdorff distance in Dataset 2. Although solid conclusions regarding the performance compared to [10] cannot be drawn, the improvement of performance over the model [6] confirms the hypothesis that surprise interpreted in the spatial domain can be used to quantify saliency.

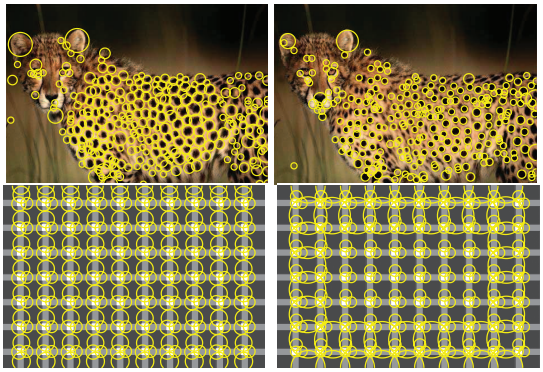
## 4. SURPRISE-BASED REGION DETECTION

Saliency has been previously related to the task of region detection, i.e. the identification of *points of interest*, to be used in subsequent stages of processing, for example for object recognition, scene classification, image matching and retrieval [15]. A well-known approach for the combination of the two paradigms is the *salient region detector* proposed by Kadir-Brady in [9]. There, saliency is equivalent to rarity and for its measurement, a two-fold criterion is used that considers separately the spatial and scale dimensions. For spatial saliency, complexity is considered an indication of rarity, and therefore the entropy $\mathcal{H}_D$ of the local distribution for a feature $D$ is used for its measurement (see also Sec. 2). For saliency across different scales, the rarity of the characteristics exhibited at a certain scale is identified with the statistical dissimilarity between the local distributions of $D$ at different scales. Dissimilarity between distributions at different scales at each point of the image is quantified using the inter-scale saliency measure (for the discrete case)

$$\mathcal{W}_D(s) = \frac{s^2}{2s-1} \sum_{d \in D} |P_D(d|s) - P_D(d|s-1)|, \qquad (2)$$

where $P_D(d|s)$ is the local distribution of feature $D$ at scale $s$. For both measures, intensity is used as feature $D$, and local distributions at different scales are estimated using the normalized histograms in circular regions of varying radius. Then, regions of interest are selected at the local maxima of $\mathcal{H}_D$, and the corresponding scale is selected by maximizing the product $\mathcal{H}_D \cdot \mathcal{W}_D$, thus simultaneously achieving automatic spatial localization and scale selection.

We propose the use of surprise instead of entropy as a measure of spatial saliency, thus identifying saliency with contrast to context instead of complexity. In addition to the arguments of Sec. 2, which we validated experimentally, in the particular setting of the salient region detector, the use of surprise is also motivated by its relation

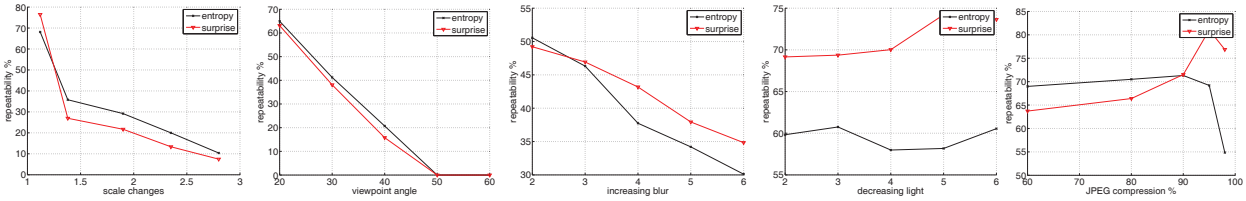


(a) entropy (b) surprise

**Fig. 4**: Regions detected in a natural and a synthetic image.

to the inter-scale saliency measure $\mathcal{W}_D$. By comparing (2) with (1), we see that the scale selection criterion is, in fact, a measure of the surprise incurred by the observation of a region at scale $s$, given a prior from the context at scale $s-1$, with the $\mathcal{L}_1$ distance between two distributions being used as a measure of dissimilarity instead of their Kullback-Leibler divergence. Therefore, the use of surprise as a saliency measure allows for the consistent interpretation of both the spatial localization and scale selection operations of the algorithm within the same theoretical treatment of saliency. Calculation of a surprise-based measure $\mathcal{S}_D$ can be done as in Sec. 3, using only intensity as feature $D$.

In order to gain some insight into the effect of this change, we show in Fig. 4 the regions detected using both the entropy and the surprise-based detector for two characteristic images. Spatial surprise becomes higher as the contrast of selected regions to their context increases. Therefore, at a distinct region in the image, spatial location is selected at its center and scale is adjusted as tight as possible around it, in order for its "distinctiveness" to be maximized. Entropy instead maximizes complexity, and therefore selects scale and center location so that image parts of varying characteristics are included in the extracted region. These observations account for the better scale selection on the blob-like features of the natural image example when using surprise, and the selection of centers instead of corners of distinct regions in the synthetic image example. We note that, although surprise achieves better scale selection and localization, the regions selected by the two detectors are otherwise similar.




(a) (b) (c) (d) (e)

**Fig. 3**: Repeatability of interest points under changes of (a) scale and rotation, (b) viewpoint, (c) blur, (d) lighting, and (e) compression.

**Evaluation:** To compare the performance of the two detectors with respect to the stability of the detected regions under various transformations, we use the experimental framework of [15] [4]. Theoretically, the modified algorithm has the same properties as the original one, i.e. invariance under planar rotation and global intensity shifts, whereas it demonstrates robustness to scale and viewpoint changes and global intensity scalings [9]. In practice, errors due to discretization and density estimation errors are expected to effect the modified algorithm more severely, due to the more extensive calculations for surprise.

The obtained *repeatability* scores are shown in Fig. 3. The entropy-based detector performs better in the cases of combination of scale and rotation and viewpoint change. This is in agreement with our observations on the increased severity of errors when using surprise. On the other hand, our detector performs favorably for changes in lighting conditions and blur, whereas none of the two detectors is clearly superior in the case of JPEG compression. This improvement can be explained if we consider that blur removes local complexity, which is detected using entropy, but maintains the local contrast detected by surprise. Overall, however, the performance of the two detectors is similar.

## 5. APPLICATION TO IMAGE QUALITY ASSESSMENT

Saliency has been previously used to improve the performance of image quality metrics [2–5]. It is therefore reasonable to assume that saliency models can be compared in such a context, with better models enabling a metric to achieve closer agreement with subjective human evaluations of quality. Due to space limitations, for our experiments we only use the SSIM metric proposed in [13], although other quality metrics can be considered as done in [3].

To integrate saliency into the fidelity metric, we weight the contribution of each image region to the overall SSIM value based on its perceptual importance. We consider two alternative approaches for weighting, one where all regions are weighted according to their saliency [3] and one where only fixated regions are assigned significantly large weights [5]. For the former, we weight each image patch using either the average or the maximum saliency value over its support. Following the same validation procedure and on the same database as in [5], we report in Table 2 correlation and RMSE values after non-linear regression between DMOS scores and scores produced by the various algorithms. Attention using surprise performs significantly better in improving SSIM compared to the model of [6]. Moreover, our results support the use of fixations as in [5], over that of saliency maps in [3] for visual importance pooling in SSIM. In the latter case, using the maximum instead of average saliency value as a patch weight works better. This is in agreement with the observations made in [1], where it is argued that whether an object will be fixated is determined by the maximum saliency value over the image region it covers.

[4] http://www.robots.ox.ac.uk/~vgg/research/affine/detectors.html

| Metric | SSIM | SSIM + [6] | | | SSIM + Surprise | | |
|---|---|---|---|---|---|---|---|
| | | avrg. | max | fixations | avrg. | max | fixations |
| correlation | 0.9388 | 0.9379 | 0.9378 | 0.9470 | 0.9479 | 0.9479 | 0.9497 |
| rmse | 7.9612 | 8.0202 | 8.0262 | 7.4265 | 7.3638 | 7.3616 | 7.2384 |

**Table 2**: Saliency-based quality metrics on LIVE database [5].

## 6. CONCLUSION

We have presented a new interpretation of Bayesian surprise in the spatial domain, to explain the characterization of image regions that differ from their context as salient. Surprise spatial saliency was employed to improve the computational attention system of Itti-Koch [6] in predicting human eye fixations. Our framework has also been validated in applications such as image region detection and quality assessment. Additional image results can be found at http://cvsp.cs.ntua.gr/research/surprise.

## 7. REFERENCES

[1] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[2] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric," in *ICIP*, 2007.

[3] Q. Ma and L. Zhang, "Saliency-based image quality assessment criterion," *LNCS*, vol. 5226, no. 1124-1133, pp. 1, 2008.

[4] E.C. Larson, C. Vu, and D.M. Chandler, "Can visual fixation patterns improve image fidelity assessment?," in *ICIP*, 2008.

[5] A.K. Moorthy and A.C. Bovik, "Visual importance pooling for image quality assessment," *IEEE JSTSP*, vol. 3, no. 2, pp. 193–201, 2009.

[6] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision research*, vol. 40, no. 10-12, pp. 1489–1506, 2000.

[7] 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*, vol. 8, no. 7, pp. 32, 2008.

[8] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. of Vision*, vol. 9, no. 12, pp. 1–27, 2009.

[9] T. Kadir and J.M. Brady, "Scale, saliency and scene description," *IJCV*, vol. 45, no. 2, 2001.

[10] N. Bruce and J.K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, 2009.

[11] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE PAMI*, vol. 31, no. 6, pp. 989–1005, 2009.

[12] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision Research*, vol. 49, no. 10, pp. 1295–1306, 2009.

[13] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE IP*, vol. 13, no. 4, pp. 600–612, 2004.

[14] M. Mancas, *Computational Attention: Towards attentive computers*, Ph.D. thesis, Faculty of Engineering, Mons, 2007.

[15] K. Mikolajczyk *et. al.*, "A comparison of affine region detectors," *IJCV*, vol. 65, no. 1, pp. 43–72, 2005.