SUNDAy: Saliency Using Natural Statistics for Dynamic Analysis of Scenes

Lingyun Zhang (lingyun@cs.ucsd.edu)
Matthew H. Tong (mhtong@cs.ucsd.edu)
Garrison W. Cottrell(gary@cs.ucsd.edu)
Deptartment of Computer Science and Engineering
University of California, San Diego
9500 Gilman Dr., Dept. 0404, La Jolla, CA 92037-0404

Abstract

The notion that novelty attracts attention is core to many accounts of visual saliency. However, a consensus has not been reached on how to best define novelty. Various interpretations of novelty lead to different bottom-up saliency models that have been proposed for static images and more recently for dynamic scenes. In previous work, we assumed that a basic goal of the visual system is to locate targets such as predators and food that are potentially important for survival, and developed a probabilistic model of salience (Zhang, Tong, Marks, Shan, & Cottrell, 2008). The probabilistic description of this goal naturally leads a definition of novelty as *self-information*, an idea that has appeared in other work. However, our notion uses the idea that the statistics used to determine novelty are learned from prior experience, rather than on the current image, leading to an efficient implementation that explains several search asymmetries other models fail to predict. In this paper, we generalize our saliency framework to dynamic scenes and develop a simple, efficient, and online bottom-up saliency algorithm. Our algorithm matches the performance of more complex state of the art algorithms in predicting human fixations during free-viewing of videos.

Introduction

It is of great research interest to understand how the visual system rapidly and efficiently samples the available visual information. One major line of this research stems from the intuition that novel objects or statistical outliers attract attention. Koch and Ullman (1985) introduced the notion of a *saliency map* based around the notion that a region is intrinsically salient if it differs substantially from its surroundings. A number of models stem from a similar intuition that being a local outlier makes a point salient (Itti, Koch, & Niebur, 1998; Gao & Vasconcelos, 2007a; Bruce & Tsotsos, 2006; Torralba, Oliva, Castelhano, & Henderson, 2006). As the small foreground items are often statistically different from the large background, locating statistical outliers in an image can facilitate detecting interesting objects. In addition, as low probability events contain more information (in an information theoretic sense), the definition of saliency as low probability event connects the selective process of visual attention with maximally sampling information.

Since humans live in a dynamic world, video and interactive environments provide a more faithful representation of the task facing the visual system than the static images frequently used in experiments. Studies also show that static measures of saliency do not perform as well as measures that use temporal information in predicting human fixations(Itti, 2005). Thus it is of interest to investigate saliency for dynamic scenes. The notion that statistical outliers attract attention applies equally well to the spatiotemporal domain and again, one sees variants of local outliers. Gaborski, Vaingankar, Chaoji, Teredesai, and Tentler (2004) used mixtures of Gaussians to model what has occurred over a spatiotemporal region of a video; an event is novel and salient if it cannot be accounted for by the model. Gao and Vasconcelos (2007b) extended their static image saliency to dynamic scenes: saliency is measured as KL divergence between the histogram of features in a location and the surround region, with the features implemented as optic flow. Itti and Baldi (2008) related saliency to *Bayesian surprise* which defines saliency as a deviation from what is expected based on a set of internal models of the local visual world.

It is reasonable to assume that one goal of the visual system is to locate targets that are potentially important for survival. In our previous work, we developed a visual saliency model that is based on this simple assumption. From the resulting probabilistic description of this goal, the self-information of the features falls out as bottom-up, task-independent saliency (Zhang et al., 2008). Self-information in this context, learned from natural statistics over development, corresponds with findings that novel items attract attention in visual search (Wolfe, 2001). The reliance of learned natural statistics forms the basis of our model: Saliency Using Natural statistics (SUN). The definition of novelty in SUN, however, is different from that has been used in previous computational saliency models in that statistical outliers are not based only on the current image. In all the models discussed, the statistics were local; for static images, the statistics were gathered solely from the current image, while for video they are gathered over some local spatiotemporal region. In previous work, we showed that feature distributions learned from experience with natural scene images provide a straightforward account for human search asymmetries, a phenomenon that is difficult for models that rely solely on the current image's statistics, as they would find a vertical bar among tilted bars just as salient as a tilted bar among vertical bars. Furthermore, the implementation of SUN performs as well as or better than previous models in predicting human fixations when free viewing images, and is computationally much more efficient(Zhang et al., 2008).

In this paper, we use spatiotemporal visual features to generalize the static image saliency model to dynamic scenes. We develop an efficient algorithm in which saliency is updated online upon each new frame. The model's performance in predicting human fixations while watching videos is comparable to previous methods, with the advantage of being sub-

stantially simpler.

Saliency is Information

Our definition of bottom-up saliency emerges from a more general goal of the visual attention system: detecting potentially important targets and allocating computational resources to them for further processing. To achieve such a goal, the pre-attentive process must estimate the probability of a target given the visual features at every location in the visual field. We have proposed elsewhere that this probability is visual saliency (Zhang, Tong, & Cottrell, 2007; Zhang et al., 2008).

To make this more explicit, we can calculate the probability that the target is present at a point, z, in the visual field. We use the term "point" loosely here; in this work, it refers to a pixel in an image, but elsewhere it can refer to a single object (e.g. Zhang et al., 2007). This point contains two pieces of information our model makes use of: its location, denoted by $L = l_z$, and the visual features present there, denoted $F = f_z$. If we define C as a binary variable that is 1 when the target is present at the current point and 0 otherwise, the probability of interest is $s_z = p(C = 1|F = f_z, L = l_z)$. Applying Bayes' rule and making the (unwarranted) simplifying assumptions that features and locations are independent and conditionally independent given that $C = 1$, this can be rewritten as:

$$\begin{aligned} s_z &= p(C=1|F=f_z, L=l_z) \\ &= \frac{p(F=f_z, L=l_z|C=1)p(C=1)}{p(F=f_z, L=l_z)} \\ &= \frac{p(F=f_z|C=1)p(L=l_z|C=1)p(C=1)}{p(F=f_z)p(L=l_z)} \\ &= \frac{1}{p(F=f_z)} \cdot p(F=f_z|C=1) \cdot p(C=1|L=l_z) \end{aligned}$$

To compare this probability across locations in an image, it suffices to estimate the log probability (since logarithm is a monotonically increasing function). For this reason, we take the liberty of using the term saliency to refer both to s_z and to $\log s_z$, which is given by:

$$\log s_z = \underbrace{-\log p(F=f_z)}_{\text{Self-info (salience)}} + \underbrace{\underbrace{\log p(F=f_z|C=1)}_{\text{Log Likelihood}} + \underbrace{\log p(C=1|L=l_z)}_{\text{Location prior}}}_{\text{Dependent on target (top-down knowledge)}}$$

The first term on the right side of this equation, $-\log p(F = f_z)$, contains no knowledge of the target and depends only on the visual features observed at the point. In information theory, $-\log p(F = f_z)$ is known as the *self-information* of the random variable F when it takes the value f_z. Self-information increases when the probability of a feature decreases—in other words, rarer features are more informative. While both Torralba et al. (2006) and Bruce and Tsotsos (2006) also define bottom-up saliency as related to self-information, they base their statistics on the current scene (for a more thorough discussion of the differences, see Zhang et al., 2008), while we learn the distributions of the features from previous experience. The remaining terms describe the target appearance and likely locations respectively. Work with SUN's appearance model is described in (Kanan, Tong, Zhang, & Cottrell, in press).

When the organism is not actively searching for a particular target (the *free-viewing* condition), the organism's attention should be directed to any *potential* targets in the visual field, despite the fact that the features and locations associated with the target class are unknown. In this case, the log-likelihood term and location prior are unknown, so we omit these terms from the calculation of saliency. Because the goal of the SUN model is to find potential targets in the surrounding environment, the probabilities should reflect the natural statistics of the environment and the learning history of the organism, rather than just the statistics of the current image. For our bottom-up model, this means that attention will be drawn to novel targets, an idea that has been in the psychology literature for decades at least. For example, Fantz (1964) showed that novel objects attract the attention of infants.

Implementation of Bottom-up Saliency on Dynamic Scenes

In this section, we describe an algorithm that estimates the bottom-up saliency in videos. First we apply a bank of spatiotemporal filters to each video; these filters are designed to be both efficient and in line with the human visual system. The probability distributions of these spatiotemporal features are learned from a set of videos from natural environments. Then for any video, we calculate its features and estimate the bottom-up saliency of each point as $-\log p(F = f_z)$. In the rest of the paper, the features are indexed by pixel coordinates so we drop the index z for notational simplicity.

Features

Let r, g and b denote the red, green, and blue components of an input video pixel. The intensity (I), red-green (RG) and blue-yellow (BY) channels are calculated as $I = r + g + b$, $RG = r - g$, $BY = b - \frac{r+g}{2} - \frac{\min(r,g)}{2}$.

The spatiotemporal filters we used are separable linear filters. The feature response function has the form $F = V * g * h$, where V is a channel of the video, g is the component that applies only along the spatial dimensions and h is the component that applies only along the temporal dimension. The filter responses are then used as features.

Difference of Gaussians (DoG) filters are used as the spatial component, g. These linear filters loosely model the response of cells in the lateral geniculate nucleus (LGN) and elsewhere. Mainly we choose these features to keep the implementation as simple as possible in order to verify the power of the underlying model.

The DoG filters are generated using

$$g(x, y; \sigma) = \frac{1}{\sigma^2} \exp\left(-\frac{x^2+y^2}{\sigma^2}\right) - \frac{1}{(1.6\sigma)^2} \exp\left(-\frac{x^2+y^2}{(1.6\sigma)^2}\right).$$

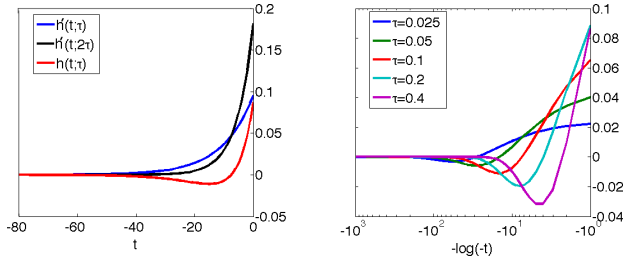

Figure 1: On the left is the temporal filter when $\tau = 0.1$. Plotted are $h'(t;\tau)$ (blue line), $h'(t;2\tau)$ (black line) and $h(t;\tau)$ (red line). The right plot shows the temporal filters for the five time scales used (values of τ of 0.025, 0.05, 0.1, 0.2, and 0.4).

We applied DoG filters to all three channels (I, RG, and BY) using 5 scales ($\sigma = 2,4,8,16$ or 32 pixels), resulting in 15 spatial filters in total.

The temporal filter h takes the form:

$$h(t;\tau) = h'(t;2\tau) - h'(t;\tau)$$

where

$$h'(t;\tau) = \frac{\tau}{1+\tau} \cdot (1+\tau)^t$$

$t \in (-\infty, 0]$ is the frame number relative to the current frame (0 is the current frame, -1 is last frame, etc.) and τ is a temporal scale parameter that determines the shape of the temporal filter. We used 5 temporal scales in our implementation $\tau = 0.025, 0.05, 0.1, 0.2, 0.4$. Figure 1 shows how $h(t;\tau)$ is formed and how it varies with τ. We will refer to $h(t;\tau)$ as a DoE (Difference of Exponentials) due to $h'(t;\tau)$'s similarity with the exponential distribution. We choose DoE as a temporal filter for the following reasons:

- $\lim_{t\to-\infty} h(t;\tau) = 0$. Therefore frames in the distant past do not contribute to the current saliency.
- $\Sigma_{-\infty}^{0} h(t;\tau)dt = 0$. If a part of the scene does not change for a extended period of time, it ceases to be salient.
- $h(t;\tau)$ is largest near $t = 0$ and falls off rapidly. This says that DoE has a strong response to onset and offset of objects.
- The DoE bears some resemblance to the temporal responses of some neurons in LGN of cats(Cai, Deangelis, & Freeman, 1997).
- Using DoE as temporal filters enables very efficient online calculation of the spatiotemporal filter responses (shown below).

With the exception of the last property, these properties are all shared with the DoG. Because all filters are linear:

$$\begin{aligned} F(x,y,t;\sigma,\tau) &= V(x,y,t) * g(x,y;\sigma) * h(t;\tau) \\ &= V(x,y,t) * g(x,y;\sigma) * (h'(t;2\tau) - h'(t;\tau)) \\ &= F'(x,y,t;\sigma,2\tau) - F'(x,y,t;\sigma,\tau) \end{aligned}$$

where $F'(x,y,t;\sigma,\tau) = V(x,y,t) * g(x,y;\sigma) * h'(t;\tau)$. This can be calculated efficiently, as:

$$\begin{aligned} F'(x,y,0;\sigma,\tau) &= \frac{F'(x,y,-1;\sigma,\tau)}{1+\tau} \\ &\quad + \frac{\tau}{1+\tau} \cdot V(x,y,0) * g(x,y;\sigma) \end{aligned}$$

To estimate the response to spatiotemporal feature $g(x,y;\sigma) * h'(t;\tau)$ at the current frame, $F'(x,y,0;\sigma,\tau)$, we simply require the spatiotemporal filter response at the previous frame and the spatial filter response at the current frame. Besides the advantage in calculation speed, this also removes the need for memory of earlier frames, a property not enjoyed by previously used spatiotemporal filters. The final response can then be easily calculated by $F(x,y,t;\sigma,\tau) = F'(x,y,t;\sigma,2\tau) - F'(x,y,t;\sigma,\tau)$.

Learning the distribution

As described above, there are 15 features on the spatial dimension: 5 from each channel. On the temporal dimension there are 5 scales and they are combined with each spatial feature. Thus there are in total 75 feature responses. By computing these feature responses on natural videos (about 2 hours of animal/plant documentary videos), we obtained an estimate of the probability distribution over the observed values of each of 75 features.

We used Song's algorithm (Song, 2006) to fit a generalized Gaussian distribution to the estimated distribution for each feature:

$$p(r;\varsigma,\theta) = \frac{\theta}{2\varsigma\Gamma(\frac{1}{\theta})} \exp\left(-\left|\frac{r}{\varsigma}\right|^{\theta}\right).$$

In this equation, θ is the shape parameter, ς is the scale parameter and r is the filter response. This resulted in one shape parameter, $\theta_{i,j}$, and one scale parameter, $\varsigma_{i,j}$, for each of the 75 filters: $i = 1,2,...,15$ is the index for spatial filters, and $j = 1,2,...,5$ is the index for temporal scales. The generalized Gaussians provide an excellent fit to the data.

Taking the logarithm, we obtain the log probability over the possible values of each feature:

$$\log p(F_{i,j}) = -\left|\frac{f_{i,j}}{\varsigma_{i,j}}\right|^{\theta_{i,j}} + const. \qquad (1)$$

These feature responses are not independent, but we proceed as if they are for simplicity. Saliency can then be calculated with a simple formula:

$$\log s = -\log p(F = f) = \sum_{j=1}^{5}\sum_{i=1}^{15} \left|\frac{f_{i,j}}{\varsigma_{i,j}}\right|^{\theta_{i,j}} + const.$$

This equation shows how easily bottom-up saliency is to calculate in SUN; raw filter responses are scaled and shaped by the learned parameters and combined through summation. It's worth repeating that aside from feature selection, all parameters of the model are completely determined by natural statistics.

Table 1: Summary of initial results.

Method	KL	ROC area
Chance	0	0.5
Bayesian Surprise	0.133	0.647
SUN	0.100	0.626
SUN (w/ 8 pixel border)	0.181	0.660
Centered Gaussian	0.441	0.764

Results

We evaluate our saliency algorithm on the human fixation data from (Itti, 2005). Eye movements were recorded from 8 subjects viewing 50 videos from indoor and outdoor scenes, television broadcasts, and artificial environments totaling over 25 minutes of video at 640×480 (at 60.27 Hz, a viewing distance of 80 cm, and with a field of view of $28\,\text{deg} \times 21\,\text{deg}$). Data was collected using an ISCAN RK-464 tracking the right eye. Two hundred eye movement traces were used (four subjects for each video clip). See (Itti, 2005) for more details.

Itti and Baldi (2008) reports results of their saliency measure (Bayesian surprise) on this data set. Under this theory, organisms form models of their environment, and assign probability distributions over the possible models. Upon the arrival of new data, the distribution over possible models is updated with Bayes rule, and the KL divergence between the prior distribution and posterior distribution is measured. The more the new data forces the distribution to change, the larger the divergence. These KL scores of different distributions over models combine to produce a saliency score.

Saliency maps were sampled at the target location of a saccade at the time the saccade was initiated. By histogramming the number of actual fixations for each value of salience, a distribution of saliency was formed for human fixations. This could be compared with the distribution of fixations over saliency for random "fixations" chosen uniformly over the image. By looking at the KL divergence ("distance") between the two distributions, we get the KL score we report (Itti & Baldi, 2008). The farther from the distribution of salience over random locations, the better.

We also use the ROC area evaluation method, where each saliency map is treated as a binary classifier that, for a given threshold, classifies points with salience above the threshold as fixated and those below the threshold as not fixated. By varying the threshold and comparing the performance at each threshold to the human fixations (which are treated as ground truth), an ROC curve is obtained. The area under the curve reflects how well the saliency map predicts the human fixations (Gao & Vasconcelos, 2007a; Bruce & Tsotsos, 2006).

The results are shown in Table 1. "Chance" indicates baseline performance. The KL score for Bayesian Surprise is smaller than that reported in (Itti & Baldi, 2008) because they use an extra step of taking the maximum in a local window on the saliency map. We found this step systematically increases all the reported measurements but does not change our qualitative conclusions. We observed border artifacts in our saliency map because of filter convolutions extending beyond the image. We therefore set the border of our map to zero remove the invalid portion of the convolution and approximate the borders present in (Itti & Baldi, 2008). Surprisingly, this drastically improved the evaluation scores. This was most apparent in the KL measurement; modifying the border had large effects on the random-saccading distribution of salience, but little effect on the distribution of salience for human saccades. Hence, depending on how edges were handled, we could report performance that was better or worse than Itti and Baldi (2008).

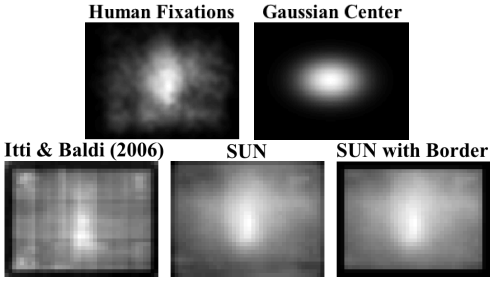

Figure 2: Center bias and border effects. Top left: Overall average of human fixations on the Itti (2005) dataset. Top right: A 2D gaussian fit to the fixation data from (Bruce & Tsotsos, 2006). Bottom row: The average saliency map over all frames of the Itti (2005) dataset for three models: Bayesian Surprise, SUN, and SUN with an 8 pixel zeroed-out border.

This appears to be a function of a phenomenon in such data sets known as *center bias* (Parkhurst & Niebur, 2003; Tatler, Baddeley, & Gilchrist, 2005; Zhang et al., 2008). Hence we decided to look at how well a "saliency measure" based on a simple Gaussian fit to the distribution of human saccades from another data set (on the static images in (Bruce & Tsotsos, 2006)) would perform on this data set (see Figure 2). This simple technique drastically outperformed our results and the the surprise model using these metrics (Table 1). The reason for this is clear by visual inspection of the data in Figure 2. The human data is highly center-biased, and so adding a larger border increases performance. The width of the border added to SUN in the right hand image was approximately equal to the darkened borders of Itti and Baldi (2008), and led to our model outperforming theirs. This finding for fixations while viewing video is consistent with earlier studies involving static images, showing that a simple model that predicts that saliency falls off with distance from the center of the screen outperforms other models (Le Meur, Le Callet, & Barba, 2007; Zhang et al., 2008). It is hard to tell whether the difference between two algorithms is due to the model or simply differences in treating the filter responses on the border. Clearly, a better method of evaluation is needed.

The fundamental problem is that sampling image locations uniformly is not at all indicative of how human saccades tend to be distributed. (Parkhurst & Niebur, 2003; Tatler et al., 2005) have suggested that the random fixations should be

Table 2: Summary of results with shuffled metric.

Method	KL	ROC area
Chance/Gaussian	0	0.5
Bayesian Surprise	0.034	0.581
Dynamic Saliency	0.041	0.582

drawn with the location distribution of human fixations. We therefore modified the KL measurement to account for this. Instead of forming the baseline (comparison) distribution of salience by counting how often each salience value occurs at random locations sampled from the image, we form it instead by counting how frequently salience values occur at human fixations in an image. We use the locations of human fixations from a *different frame of the video* and measure the salience values at those locations. Then we compare the distribution of salience values at the locations humans fixated in each frame to the salience values in that same frame, but using fixations from a different frame. I.e., the comparison distribution is created by shuffling the frames of the saliency maps over each movie, giving them the human *spatial* distribution but not the *temporal* distribution. Put another way, rather than determining whether subjects looked at the most salient location in each frame, we instead measure whether they look at a fixated point *when* it is most salient. This has the desired effect of causing the simple static measure of fitting a Gaussian to the human distribution to have a score of zero; since this static version of saliency does not change, shuffling has no effect on the distribution of salience value counts. We modified the ROC metric similarly. A related method was proposed in (Tatler et al., 2005) and used in (Zhang et al., 2008) for static images. However, the temporal component makes the metric more stringent than when shuffling a set of independent images - for video, this necessitates accurate prediction of the timing of fixations. As discussed in (Carmi & Itti, 2006) this metric *underestimates* the model performance since the center of the screen for pictures and video genuinely tends to be the most salient part of the scene due to cameraman (or director) bias. However, this is still a useful measure for *relative* model comparisons, serving as a lower bound assessment of models' prediction ability.

Nevertheless, our method continues to do better than chance, and slightly better numerically than Itti and Baldi's surprise model (Itti & Baldi, 2008) on this data set, as shown in Table 2. Scores of both models may appear low, but the strictness of our evaluation metric needs to be remembered; we're evaluating whether the model predicts *when* a fixation will be made to a location, not simply *where* as in Table 1. This demonstrates that saliency models outperform the baseline Gaussian after compensating for the center bias.

Figure 3 shows the saliency maps on some frames of different videos. This provides some insight into what SUN finds salient. For instance, motion is one of the most salient features, as shown clearly in the first row; the person that is moving while talking is far more salient than the face of the person who is calmly listening.

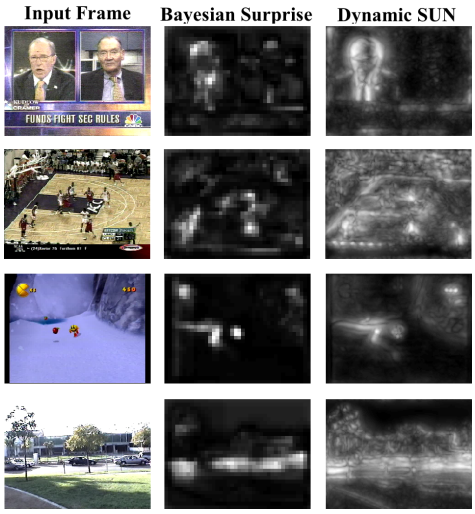

Figure 3: The saliency maps for several frames of video from (Itti, 2005).

Despite the similarity in performance, our model is significantly simpler. We use 15 spatial filters and learn 75 distributions offline. Bayesian Surprise, in contrast, uses 72 spatial filters and must maintain 432,000 distributions that must be updated with each frame. This difference in complexity has consequences on runtime; on a Pentium 4 3.8 GHz dual core PC with 1 GB RAM, SUN runs through a video of about 500 frames in minutes while Bayesian Surprise requires hours. A version of SUN designed to run in faster than real time with only modest decreases in performance is described by Butko, Zhang, Cottrell, and Movellan (2008) and was shown to have applications in a social robotics environment[1].

Discussion

In this paper we generalized our principled probabilistic measure of saliency (Zhang et al., 2008) to video. In our formulation, bottom up saliency emerges as the self-information of visual features when estimating the probability of finding a target. We designed a feature space that can be calculated very efficiently, which leads to a simple, fast algorithm.

Our findings also agree with (Parkhurst & Niebur, 2003; Tatler et al., 2005) in pointing out some disadvantages of using some of the previously proposed evaluation metrics. Data collected in a lab often show a strong center bias that confounds proper evaluation of the results. By shuffling the frames but maintaining the patterns of fixations, we effectively remove the effects of this bias. However, as (Carmi & Itti, 2006) points out, there is also a central bias introduced by having humans center the camera on interesting parts of the scene - the center is inherently more likely to be salient.

Overall, our results show comparable performance with Itti and Baldi's surprise model (Itti & Baldi, 2008) in predicting human fixations despite the relative simplicity of the SUN

[1]FastSaliency code is available for download at `http://mplab.ucsd.edu/~nick/NMPT/`

model. The efficiency of the SUN model is due to two main factors: First, we give our model *experience* with other videos that allow us to precompute what is novel, rather than what is "currently unexpected." Second, the particular form of our temporal component, the Difference of Exponentials, allows for a linear, nearly memoryless updating of the salience map. Both of these lead to a model that could plausibly be computed by neurons. Furthermore, the search asymmetries discussed by Zhang et al. (2008) provide additional motivation for using prior statistics. We do not deny that there are effects of more recent history, and in the end, the right answer might be some combination of precomputed statistics and more temporally local statistics.

In our future work, we intend to investigate such issues such as the effects of a foveated retina (currently both Bayesian Surprise and our model are applied to the entire image at the same resolution), and generalizing our notion of salience to one of utility, as in (Nelson & Cottrell, 2007).

Acknowledgements

The authors would like to thank Laurent Itti & Pierre Baldi and Neil Bruce & John Tsotsos for sharing their human fixation data and algorithms. We would also like to thank Dashan Gao, Dan Hill, Honghao Shan, Piotr Dollar, Nick Butko, Javier Movellan, and Tim Marks for helpful discussions. This work was supported by NIMH grant MH57075 to GWC and NSF grant #SBE-0542013 to the Temporal Dynamics of Learning Center, Garrison W. Cottrell, PI.

References

Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in Neural Information Processing Systems 18* (p. 155-162). Cambridge, MA: MIT Press.

Butko, N. J., Zhang, L., Cottrell, G. W., & Movellan, J. R. (2008). Visual saliency model for robot cameras. In *Proceedings of the 2008 IEEE international conference on robotics and automation (icra)* (pp. 2398–2403). Pasadena, CA, USA.

Cai, D., Deangelis, G., & Freeman, R. (1997). Spatiotemporal receptive field organization in the lateral geniculate nucleus of cats and kittens. *Journal of Neurophysiology*, *78*(2), 1045–1061.

Carmi, R., & Itti, L. (2006). The role of memory in guiding attention during natural vision. *J. Vision*, *6*(9), 898–914.

Fantz, R. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, *146*(3644), 668.

Gaborski, R., Vaingankar, V., Chaoji, V., Teredesai, A., & Tentler, A. (2004). Detection of inconsistent regions in video streams. In *Proceedings of SPIE* (pp. 202–210). Bellingham, WA: SPIE.

Gao, D., & Vasconcelos, N. (2007a). Bottom-up saliency is a discriminant process. In *IEEE International Conference on Computer Vision.* Rio de Janeiro, Brazil: IEEE Computer Society.

Gao, D., & Vasconcelos, N. (2007b). The discriminant center-surround hypothesis for bottom-up saliency. In *Advances in Neural Information Processing Systems 20.* Cambridge, MA: MIT Press.

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, *12*(6), 1093-1123.

Itti, L., & Baldi, P. F. (2008). Bayesian surprise attracts human attention. *Vision Research*.

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

Kanan, C., Tong, M. H., Zhang, L., & Cottrell, G. W. (in press). Sun: Top-down saliency using natural statistics. *Visual Cognition*. (DOI: 10.1080/13506280902771138)

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, *4*(4), 219–27.

Le Meur, O., Le Callet, P., & Barba, D. (2007). Predicting visual fixations on video based on low-level visual features. *Vision Research*, *47*(19), 2483–2498.

Nelson, J., & Cottrell, G. (2007). A probabilistic model of eye movements in concept formation. *Neurocomputing*, *70*(13-15), 2256-2272.

Parkhurst, D., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision*, *16*(2), 125-154.

Song, K. (2006). A globally convergent and consistent method for estimating the shape parameter of a generalized gaussian distribution. *IEEE Transactions on Information Theory*, *52*(2), 510-527.

Tatler, B., Baddeley, R., & Gilchrist, I. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research*, *45*(5), 643–59.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, *113*(4), 766–786.

Wolfe, J. (2001). Asymmetries in visual search: An introduction. *Perception & Psychophysics*, *63*(3), 381–389.

Zhang, L., Tong, M. H., & Cottrell, G. W. (2007). Information attracts attention: a probabilistic account of the cross-race adavantage in visual search. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (p. 749-754). Nashville, Tennessee: Cognitive Science Society.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, *8*(7), 1-20.