2008 IEEE International Conference on
Robotics and Automation
Pasadena, CA, USA, May 19-23, 2008

Visual Saliency Model for Robot Cameras

Nicholas J. Butko, Lingyun Zhang, Garrison W. Cottrell, and Javier R. Movellan

Abstract— **Recent years have seen an explosion of research on the computational modeling of human visual attention in task free conditions, i.e., given an image predict where humans are likely to look. This area of research could potentially provide general purpose mechanisms for robots to orient their cameras. One difficulty is that most current models of visual saliency are computationally very expensive and not suited to real time implementations needed for robotic applications.**

Here we propose a fast approximation to a Bayesian model of visual saliency recently proposed in the literature. The approximation can run in real time on current computers at very little computational cost, leaving plenty of CPU cycles for other tasks. We empirically evaluate the saliency model in the domain of controlling saccades of a camera in social robotics situations. The goal was to orient a camera as quickly as possible toward human faces. We found that this simple general purpose saliency model doubled the success rate of the camera: it captured images of people 70% of the time, when compared to a 35% success rate when the camera was controlled using an open-loop scheme. After 3 saccades (camera movements), the robot was 96% likely to capture at least one person. The results suggest that visual saliency models may provide a useful front end for camera control in robotics applications.

I. Introduction

There has recently been a large amount of scientific research to develop computational models of visual saliency [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. The computational output of these models is a value at each pixel of an image or video sequence (Figure 1) that indicates whether that region is likely to be fixated by humans when the task is to simply look at the image or video. Typically these methods are evaluated on how well they predict the actual specific locations that humans have fixated in eye-tracking experiments where the only instruction is “look” or “watch”. This area of research is of potential interest to social robotics for two reasons: First, a robot that orients its eyes in a manner similar to humans is likely to give an impression of intelligent behavior and facilitate interaction with humans. Second, such models may orient the robot towards regions of the visual scene that are likely to be relevant.

Unfortunately the currently existing models of visual saliency are typically too slow, requiring seconds, if not minutes, to analyze single video frames at very reduced resolution. Here we describe and evaluate a very fast and computationally-lightweight adaptation of a recently published model of visual-salience. The model can comfortably provide saliency maps in about 10 ms per video frame on a modern low-end computer, thus being particularly suitable for robotic applications. We show that the algorithm provides a useful front end for robotics cameras, effectively using foveal information to orient the camera towards likely regions of interest.

II. Previous Models of Visual Saliency

Several Bayesian approaches have been developed recently that provide a computational foundation to the notion of visual saliency. While at first sight these models may appear very different from each other, they can be seen as special cases of the same formalism. In particular many of these approaches implicitly or explicitly define the saliency of a pixel x as a function of the probability that this pixel renders an object of a category of interest, given the available image, i.e.,

$$\begin{aligned} s(x) &\overset{\text{def}}{=} \log p(C_x = 1|f_x) \\ &= \log p(f_x|C_x = 1) + \log p(C_x = 1) \\ &\qquad - \log p(f_x) \end{aligned} \tag{1}$$

where $s(x)$ is the saliency of pixel x and f_x is a feature vector that summarizes the information on image pixels in the neighborhood of x, and C_x is a binary random variable that takes value 1 if pixel x renders an object from the category of interest.

This formulation can be used to compare the choices made by the existing Bayesian approaches. For example, Torralba *et al.* [3] use the $p(C_x = 1)$ term to model class specific location distributions, *i.e.* the density $p(C_x = 1)$ differs for every x depending on the location of x in the image plane, *e.g.* clouds may be more probable a priori toward

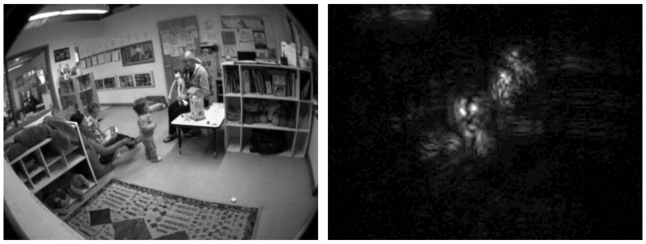

Fig. 1. The purpose of visual saliency algorithms is to quantify the importance of attending to each visual location. Saliency algorithms are often evaluated on how well they predict human eye-fixation data.

N. Butko is a Ph.D. Student in the Department of Cognitive Science, University of California San Diego, La Jolla, Ca, 92093-0515. nbutko@cogsci.ucsd.edu

L. Zhang is a Ph.D. graduate from the Department of Computer Science, University of California San Diego, La Jolla, CA, 92093-0404. lingyun@cs.ucsd.edu

G. Cottrell is faculty in the UCSD Department of Computer Science, University of California San Diego, La Jolla, CA, 92093-0404. gary@ucsd.edu

J. Movellan is a Research Scientist in the Institute for Neural Computation, La Jolla, CA, 92093-0523. movellan@mplab.ucsd.edu

978-1-4244-1647-9/08/$25.00 ©2008 IEEE.

the top of the image. It can also take on a different value by switching targets, *e.g.* the distribution $p(C_x = 1)$ when searching for clouds is different from $p(C_x = 1)$ when the category of interest is people. They estimate $p(f_x)$ using a generalized Gaussian fit to the statistics of the specific image being searched.

Bruce & Tsotsos [5] present a model of saliency based on the Shannon information of an event, $-\log p(x)$. They estimate the density $p(f_x)$ using a histogram over a small image region, as opposed to the entire image, as in [3]. Their model implicitly assumes that in general purpose tasks the functions $p(f_x|C_x = 1)$ and $p(C_x = 1)$ are approximately constant with respect to x and they can be ignored for they do not affect the relative saliency of different pixel locations.

Harel *et al.* [6] proposed a model of saliency based on the use of a *dissimilarity* metric. Like [5] the context is free-viewing, and the first two terms become irrelevant in ranking pixels. Like [3] the distribution $p(f_x)$ is estimated based on the histogram of the the current image. However in this case they use a graphical model that weights inter-pixel distance and feature dissimilarity. Probabilities are estimated by sampling, a process that is $O(n^4)$ with n pixels in the image. While this approach matches human free-viewing data well, it is infeasible for calculating salience maps of moderate size in real time.

Zhang *et al.* [7] follow the model in [3], but estimate $p(f_x)$ using frequency counts from a data set of natural images/videos fit to generalized Gaussian distributions. By using features sensitive to local contrast, they are able to replicate saliency effects that in other models require densities to be estimated within each image separately. This makes the model's complexity roughly linear with respect to the number of image pixels, and therefore attractive for real-time implementations, since it does not require recomputing costly frame by frame statistics.

Itti *et al.* [1] proposed a model of visual saliency based on the Feature Integration Theory of human attention [11]. Their model computes many features at each pixel by convolving *e.g.* motion, color, and brightness channels with Difference of Gaussians filters. These are then normalized and half-wave rectified. The different channels are then added together to create a master saliency map. Navalpakkam & Itti [4] define visual saliency in terms of *Signal to Noise Ratio* (SNR). Specifically, the model learns the parameters of a linear combination of low level features that cause the highest expected SNR for discriminating a target from distractors. Itti & Baldi [2] define salience as the KL divergence between the prior distribution that a pixel renders an object of interest and the posterior distribution given the image statistics around that pixel. Specifically, under their model, saliency is proportional to the number of events generated by a Poisson process. A Gamma distribution conjugate prior is maintained over the Poisson distribution's parameters. Spatial saliency detectors estimate the posterior distribution based on map neighbors and temporal saliency detectors estimate the posterior distribution based on subsequent salience of the same pixel. The model is evaluated in terms of its capacity to fit human saccade data in open ended, free-viewing tasks.

Gao & Vasconcelos [9] define saliency as the KL distance between the distribution of a pixel region's filter responses from that of pixels surrounding that region. The distribution of filter responses is estimated as a generalized Gaussian distribution, and a different distribution is fit to each overlapping region of the image.

Kienzle *et al.* [10] used a data-driven approach, using human eye movement data on general purpose tasks to learn features that are highly discriminative of regions that are commonly scanned by humans versus regions with low scanning rates.

III. Real-Time Implementation

In this paper, we propose a simplified version of Zhang *et al.*'s model [8] designed to operate in real time at little computational cost. In [8], Zhang extends the model in [7] to temporally dynamic scenes, and characterizes the video statistics around each pixel using a bank of spatio-temporal filters with separable space-time components, i.e., the joint spatio-temporal impulse response of these filters is the product of a spatial and a temporal impulse response. In [8] the spatial impulse responses are Difference of Gaussians (DoG), which model the properties of neurons in the lateral geniculate nucleus (LGN). The surround Gaussian has radius twice the size of the center Gaussian, and each subsequent scale is twice the size of the previous scale. At the smallest scale the radius is 1 pixel and the spatial impulse response at scale i is

$$\begin{aligned} g(i) &= \frac{1}{2\pi(2^{i-1})^2}\exp\left(-\frac{x^2+y^2}{2(2^{i-1})^2}\right) \\ &\quad - \frac{1}{2\pi(2^{i})^2}\exp\left(-\frac{x^2+y^2}{2(2^{i})^2}\right) \end{aligned} \tag{2}$$

The temporal impulse responses are Difference of Exponentials (DoE), which can be implemented recursively in a very efficient manner:

$$h(t;\tau) = \hat{h}(t;2\tau) - \hat{h}(t;\tau) \tag{3}$$

where $\hat{h}(t;\tau) = \frac{\tau}{1+\tau}\cdot(1+\tau)^t$, $t \in (-\infty, 0]$ is the relative frame number to current frame (0 is the current frame, -1 is last frame, etc.) and τ is a temporal scale parameter. The τ of the first scale is a parameter to the model, and it doubles with each successive temporal scale.

The probability distribution of the features $p(f)$ is estimated by collecting filter responses over natural videos, fitting a generalized Gaussian distribution for each individual filter, and combining the distribution across temporal and spatial scales assuming conditional independence.

For the real-time implementation explored in this paper we simplified Zhang's model in the following ways:

1) We used only image intensity channels, not color channels.

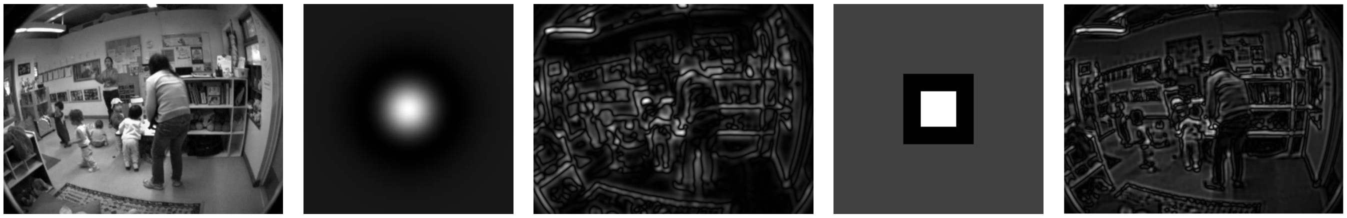

Fig. 2. Difference of Gaussians filter, and the Difference of Boxes approximation. The filters are typical of those used in this paper, with the $r_{center} = 1/2\ r_{surround}$. The filters are respectively applied to the original image (left). Absolute filter responses are shown.

Algorithm 1 Initialize Saliency

```
$NS \Leftarrow 5$   {Parameter: # of Spatial Scales}
$NT \Leftarrow 5$   {Parameter: # of Temporal Scales}
Min$\sigma \Leftarrow 1$   {Parameter: Smallest Box Filter Radius $\in [1, \infty)$}
Min$\tau \Leftarrow 1$   {Parameter: Smallest Time Parameter $\in (0, \infty)$}
$\sigma[1] \Leftarrow$Min$\sigma$
$\tau[1] \Leftarrow$Min$\tau$
for $i = 1$ to $NS$ do
    $\sigma[i+1] \Leftarrow 2\sigma[i]$
end for
for $j = 1$ to $NT$ do
    $\tau[j+1] \Leftarrow 2\tau[j]$
end for
for all $Exp[i, j]$ do
    $Exp[i, j] \Leftarrow \vec{0}$   {$Exp$ has $(NS+1, NT+1)$ vectors the size of the salience map.}
end for
```

Algorithm 2 Calculate Saliency $s(x)$

```
Require: $NS, NT, \sigma, \tau, Exp$ initialized in Algorithm 1. $Exp$ is updated in this Algorithm.
$SaliencyMap \Leftarrow \vec{0}$
$Im \Leftarrow$ get downsampled frame from camera
$BoxFilt[1] \Leftarrow$ Filter $Im$ with box-filter, width=$2\sigma[1]+1$
for $i = 1$ to $NS$ do
    $BoxFilt[i+1] \Leftarrow$ Filter $Im$ with box-filter, width=$2\sigma[i+1]+1$
    $DoB[i] \Leftarrow BoxFilt[i] - BoxFilt[i+1]$
    $Exp[i, 1] \Leftarrow \frac{\tau[1]}{1+\tau[1]} DoB[1] + \frac{1}{1+\tau[1]} Exp[i, 1]$
    for $j = 1$ to $NT$ do
        $Exp[i, j+1] \Leftarrow \frac{\tau[j+1]}{1+\tau[j+1]} DoB[i] + \frac{1}{1+\tau[j+1]} Exp[i, j+1]$
        $DoE[i, j] \Leftarrow Exp[i, j+1] - Exp[i, j]$
        $SaliencyMap \Leftarrow SaliencyMap +$ abs$(DoE[i, j])$
    end for
end for
return $SaliencyMap$
```

2) The DoG filters were approximated by difference of box filters DoB (See Figure 2).[1]

3) The filter impulse response distribution was modeled as a Laplacian distribution with unit variance, a special case of the generalized Gaussian distribution.[2]

As in Zhang's original model, we assume an open-ended visual search task, i.e. we don't have prior knowledge about where in an image generally interesting objects will appear, or what they will look like. Under these conditions the location prior $p(C_x = 1)$ and the object appearance model $p(f_x|C_x = 1)$ are approximately constant with respect to x and thus can be ignored.

The approach is pseudocoded in Algorithms 1&2. In Algorithm 2, all arithmetic operations are vector operations.

The computational complexity was roughly linear with respect to n, the number of pixels, as well as NS and NT, the number of spatial scales and temporal scales. Tables I&II show the time needed to compute saliency on a frame varying each of these three complexity dimensions. The computations were performed on a Mac Mini with a 1.87 GHz Intel Core Duo processor. Box filter operations were performed with Apple's vImageBoxConvolve_Planar8 function. Vector algebra operations were performed using the BLAS library. The time was measured in absolute (wall) time, but since the processor was dual core, the process-specific times were nearly identical. In practice our implementation is orders of magnitude faster than those reported in the literature. For example, the popular Saliency model of Itti & Baldi [2] requires ≈ 1 minute for each 30×40 pixel video frame, while the model proposed here takes 11 milliseconds for each 120×160 pixel video frame.

In order to ensure that the simplifications in our approach still maintain the important properties of other visual saliency algorithms, we compared its performance to the model of Itti & Baldi [2]. The task was to predict human eye fixation on videos in a free viewing task; the data were those originally used in [2]. The performance of our algorithm (0.633 AROC) was very similar to that of Itti & Baldi (0.647 AROC). This is also comparable with Zhang's original algorithm, and so very little performance is sacrificed making the three approximations above.

[1] DoB are types of box-filters, a computationally efficient class of filters that have been used with much success recently in visual object classification [12]

[2] In the generalized Gaussian case we have $-\log p(f) = \sum |f_i/\sigma_i|^{\theta_i}$. This becomes $-\log p(f) = \sum |f_i|$ under our Laplacian with $\sigma_i = 1$ approximation.

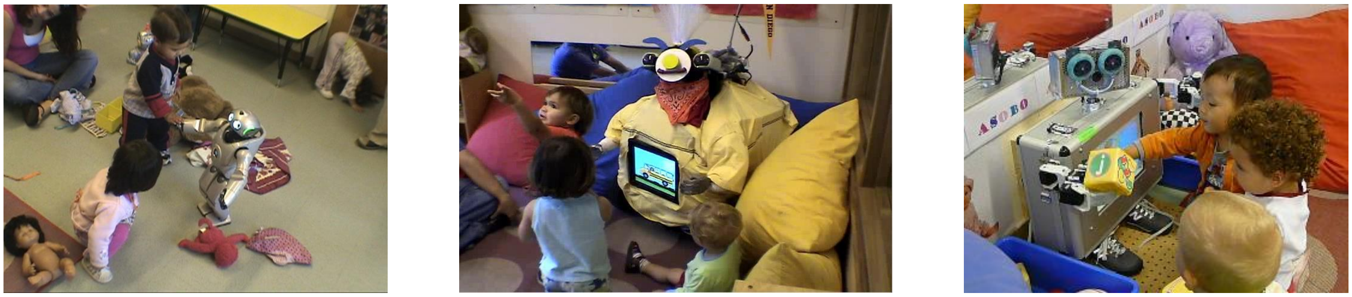

Fig. 3. Three robot members of the RUBI project. **Left:** QRIO is a humanoid robot prototype on loan from Sony corporation. **Center:** RUBI-1, the first prototype developed at UCSD. **Right:** RUBI-3 (Asobo) the third prototype developed at UCSD. It teaches children autonomously for weeks at a time

TABLE I

PROCESSING TIME NEEDED TO COMPUTE SALIENCY MAP AS A FUNCTION OF IMAGE SIZE (5 SPATIAL / 5 TEMPORAL SCALES).

	80 × 60	160 × 120	320 × 240	640 × 480
Time	2.93 ms	10.82 ms	44.96 ms	214.82 ms

TABLE II

PROCESSING TIME NEEDED TO COMPUTE SALIENCY MAP OVER VARIOUS SPATIOTEMPORAL SCALES (160 × 120 PIXELS).

Space\Time	1 Scale	2 Scales	3 Scales	4 Scales	5 Scales
1 Scale	1.32 ms	1.64 ms	1.95 ms	2.26 ms	2.82 ms
2 Scales	2.04 ms	2.71 ms	3.36 ms	3.93 ms	4.62 ms
3 Scales	2.81 ms	3.81 ms	4.72 ms	5.90 ms	7.06 ms
4 Scales	3.35 ms	4.65 ms	5.77 ms	7.58 ms	8.95 ms
5 Scales	3.88 ms	5.32 ms	6.77 ms	9.29 ms	10.82 ms

IV. FIELD STUDY

As part of the RUBI project [13], [14] for the past three years our laboratory has been conducting field studies with social robots immersed at the Early Childhood Education Center at UCSD. The goal of these studies is to explore the possibilities of social robots to assist teachers in early childhood education (Figure 3). One critical aspect of these robots is to be able to find and orient towards humans. While we have already developed powerful algorithms for detecting the presence of humans using video [15], they tend to be computationally expensive and thus best suited for scanning a small foveal region of a scene. As such we were interested in investigating whether a lightweight saliency model could be used on peripheral regions to help orient the fovea towards the most promising regions of the visual scene.

A 2 degree of freedom (pan and tilt) robot camera was constructed using an iSight IEEE1394 640x480 camera with a fisheye lens (160° FOV), 2 Hitech HS-322HD servo motors, and a Phidgets servo control card operated by a Mac Mini (1.87 GHz Intel Core Duo). The robot camera was placed in Room 1 of the UCSD's Early Childhood Education Center (ECEC), where the RUBI project is taking place. The camera was located on a bookshelf above the reach of the children (18–24 months old). The system collected data continuously for 9 hours during one day's operation of ECEC, from 7:30am–4:30pm.

Images were processed in real-time. They were received from the camera at 640×480 resolution at approximately 15 FPS (i.e. every 66 msec). For the purpose of computing saliency, they were downsampled to a 160×120 pixel resolution. A saliency map was then computed in six-times-faster-than-real-time for all the pixels ($\approx$ 11 msec, see Table II), using a bank of 5 spatial filters and 5 temporal filters. The DoB spatial filters had odd center widths $\{3, 5, 9, 17, 33\}$ so that they would be defined about a central pixel. The above diameters correspond to radii about the center of $\{1, 2, 4, 8, 16\}$ respectively. The corresponding surround widths were $\{5, 9, 17, 33, 65\}$. The τ temporal parameters were $\{1, 2, 4, 8, 16\}$.

a) Experimental Camera – Saliency Track: At the start of each experiment, the camera was moved to a central location.

Starting 30 frames after any camera movement, on each successive frame, if the maximum saliency pixel exceeded threshold and its location was more than 10 degrees in either the pan or tilt direction from the current fixation point, the servos would reposition the camera so that the maximum saliency pixel in the saliency map was now at approximately the center of the image plane.

15 frames after a movement was initiated (to allow for

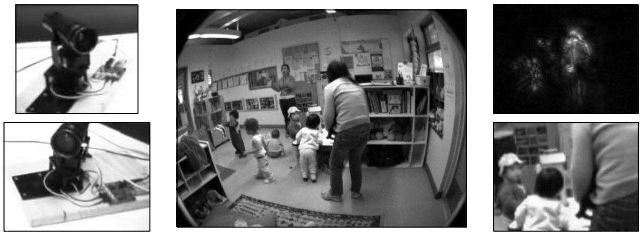

Fig. 4. Experimental Setup: A simple robotic camera (left) collected very wide angle – 160° – images at 640×480 resolution (center) and downscaled them to 160×120 resolution for the purpose of computing a saliency map (top right). The camera then rotated – pan/tilt – so that the maximum saliency pixel was now in the center of gaze. After movement, a 160×120 snapshot of the center of gaze at full resolution was saved as a foveal representation (bottom right). This fovea was coded offline for the presence of people.

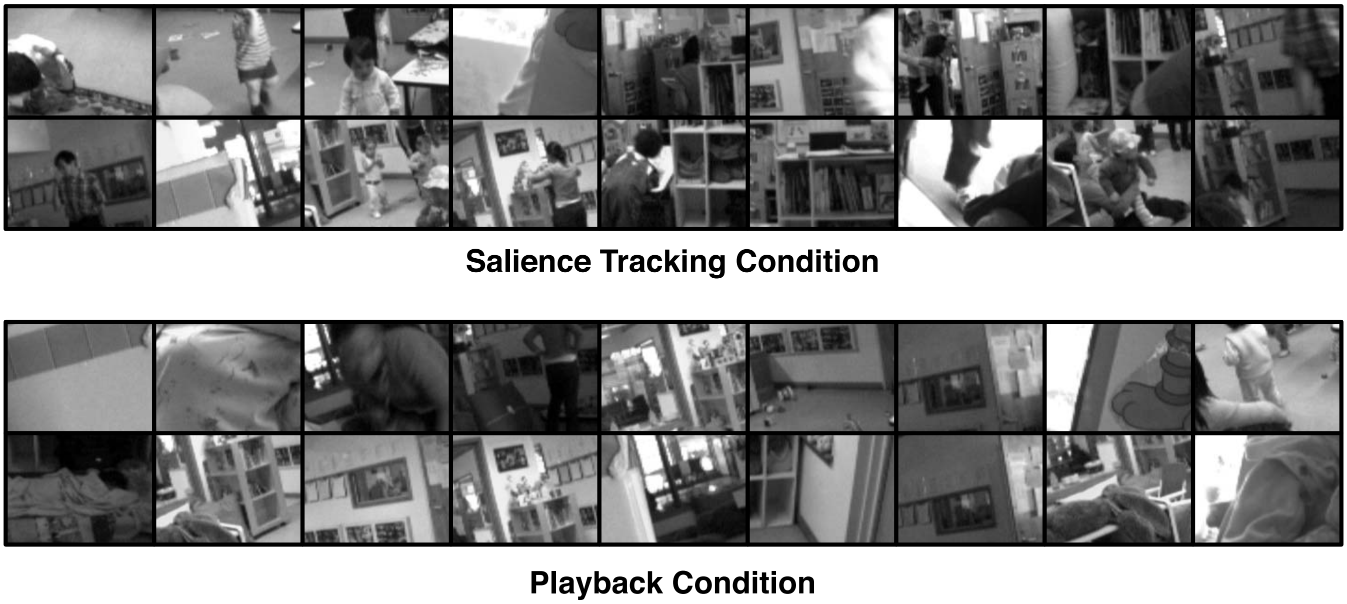

Fig. 5. Center of attention (fovea) in saliency tracking condition and playback condition. In each case, 18 images were chosen randomly from the whole set, and so the sample is representative. Many more people are attended in the saliency condition than the playback condition.

the movement's completion), an image of the camera's view was saved. Additionally, a foveal view containing the center 160×120 pixels of the high resolution 640×480 image was saved, simulating the foveal region over which high level but computationally expensive perceptual primitives could operate (*e.g.*, person detection, expression recognition).

b) Control Camera – Playback: An additional camera control condition was implemented. In this condition the camera played back in open-loop the exact same movements as in the previous salience-directed movement condition. This served as a control with the same motion statistics as the salience condition, but the movements were not caused directly by current events in the world. In addition to preserving the motion statistics, the playback framework served to tie together in the two conditions the implicit prior on the "location of the class of generally interesting objects," or $p(C_x = 1)$ in Equation 1. Thus the only difference between the two conditions was that one was caused by features that were unlikely in natural statistics, *i.e.* ones for which $-\log p(F_x)$ was high.

Each condition ran sequentially for 3 minutes at a time. A pair of conditions salience and playback would take about 6 minutes. There was an additional 3 minute break between cycles. In all, 64 cycles were completed and 4964 images were collected.

V. Analysis of results

After the experiment a subset of the foveal center-images was chosen randomly and uniformly from all 4964 collected images. The subset of images was coded by 4 coders. Two of the coders were authors of this paper and two were naïve third parties. The coders were instructed to label the number of people they could see in each 160×120 foveal image. The coding was done in a double-blind fashion: the images were ordered randomly across labels and time collected. All coders, including the authors, were given no extra information to indicate which images came from which condition. All coders labeled 1050 images (510 saliency condition, 540 playback condition) in the same order.

The average Pearson correlation between the four coders across the 1050 labels was 0.8723. We marked a foveal snapshot as "containing a person" if two or more coders agreed that there was a person in the snapshot.

A. Results

It should be noted that the control condition in our experiment was designed to be particularly difficult, much harder than random search. For example, in the control condition, the camera oriented toward regions of space that had been salient in the experimental condition. These regions tended to have people in the experimental condition and thus were still likely to have people at control time. In spite of this, the experimental camera (Saliency Tracking) performed much better than the control camera (Playback). In the Salience Tracking condition, 68.04% of foveal images contained people. In the Playback condition, only 34.81% of foveal images contained people. Thus by orienting toward salient events in the image plane, the camera attended to people twice as often as just looking in the places where people are likely to appear. A random sample of images from both conditions is shown in Figure 5.

Note that with a detection rate of 68% per saccade, after 3 saccades, we are 96.8% likely[3] to have seen at least one

[3] Assuming people are always present. This figure is an underestimate and the true rate will be higher given presence of people because this average performance figure includes even times when there are no people to be seen, such as nap time or when children are playing outside. $96.8\% = 1 - (1 - .68)^3$

person. A post processing algorithm operating over these saccades would review $(3 * 160 \times 120)$ pixels, representing more than an 81% reduction in search time.

Most importantly, the salience algorithm is fast and efficient. Salience was calculated in less than 11 ms for each 67 ms frame grab, leaving over 83% of CPU cycles to be dedicated to other tasks important to the function of the robot, including sophisticated visual post-processing.

An additional benefit is derived from saliency's resilience to distorted images: it works well on the entire image plane of a very wide angle camera. However, object identification algorithms are often brittle to the warping caused at the edges of the wide angle lens. By using saliency on a very wide field of view, we can identify from large regions of the real world areas of interest and then point the center of the lens toward them. Objects in the central region are undistorted, and may be discovered easily by our machine perception algorithms.

Although we did not investigate it systematically, the salience algorithm also appears to be robust to lighting conditions. For example, during nap time, the lights of the classroom were turned off, but the robot continued to orient toward teachers walking around the room.

VI. Conclusions

We presented a fast visual saliency algorithm that approximates very well current models of early human visual attention. From a Bayesian point of view the algorithm is designed to find regions of an image plane most likely to be useful in unconstrained conditions, i.e., situations where there is a very large number of potential tasks of interest. The proposed approach matches human eye fixation data almost as well as current state of the art models of early visual attention, yet it is orders of magnitude faster. It can operate in real time in a low end modern computer, leaving plenty of CPU for other operations. This makes the approach ideal for robotic applications.

We presented empirical results from a field study using a robotic camera in daily life conditions. To our knowledge this is the first example of a practical use of current models of early human visual attention to a real time robotics task. The results suggested that models of visual saliency may provide a promising approach for efficient camera orientation in social robotics applications.

Ackowledgments

The authors would like to acknowledge Matthew H. Tong for reproducing the results in [2], and the teachers, parents, and children in Room 1 of the Early Childhood Education Center for their invaluable continued support and involvement in the RUBI project.

This work was supported by UC Discovery Grant 10202, by NSF Science of Learning Center grant #SBE-0542013 to GWC, by NIMH grant R01 MH57075 to GWC, and by NSF IGERT training grant #DGE-0333451 (V. de Sa, PI).

References

[1] 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*, vol. 20, no. 11, pp. 1254–1259, Nov 1998.

[2] L. Itti and P. Baldi, “A principled approach to detecting surprising events in video,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2005.

[3] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, “Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search,” *Psychological Review*, vol. 113, no. 4, pp. 766–786, 2006.

[4] V. Navalpakkam and L. Itti, “An integrated model of top-down and bottom-up attention for optimal object detection,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, New York, NY, Jun 2006, pp. 2049–2056.

[5] N. D. B. Bruce and J. K. Tsotsos, “Saliency based on information maximization,” in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006, pp. 155–162.

[6] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007.

[7] L. Zhang, M. H. Tong, and G. W. Cottrell, “Information attracts attention: A probabilistic account of the cross-race advantage in visual search,” in *Proceedings of the 29th Annual Cognitive Science Conference*, 2007.

[8] L. Zhang, M. H. Tong, N. J. Butko, J. R. Movellan, and G. W. Cottrell, “A bayesian framework for dynamic visual saliency,” (In Preparation), 2007.

[9] D. Gao and N. Vasconcelos, “Bottom up saliency is a discriminant process,” in *IEEE International Conference on Computer Vision*, 2007.

[10] W. Kienzle, F. A. Wichmann, B. Schölkopf, and M. Franz, “A nonparametric approach to bottom-up visual saliency,” in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007.

[11] A. M. Treisman and G. Gelade, “A feature-integration theory of attention,” *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, Jan 1980.

[12] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 2001. [Online]. Available: citeseer.ist.psu.edu/article/viola01rapid.html

[13] J. R. Movellan, F. Tanaka, B. Fortenberry, and K. Aisaka, “The RUBI project: Origins, principles and first steps,” in *Proceedings of the International Conference on Development and Learning (ICDL05)*, Osaka, Japan, 2005.

[14] F. Movellan, J. R.and Tanaka, T. C., R. P., and E. M., “The rubi project: A progress report,” in *Proceedings of the 2nd ACM/IEEE international conference on human-robot interaction*, 2007.

[15] I. Fasel, B. Fortenberry, and J. R. Movellan, “A generative framework for real-time object detection and classification,” *Computer Vision and Image Understanding*, vol. 98, pp. 182–210, 2005.