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

Target-directed attention: Sequential decision-making for gaze planning

Julia Vogel and Nando de Freitas
Laboratory of Computational Intelligence
Department of Computer Science
University of British Columbia, Vancouver, BC, Canada
{vogel,nando}@cs.ubc.ca

***Abstract*— It is widely agreed that efficient visual search requires the integration of target-driven top-down information and image-driven bottom-up information. Yet the problem of gaze planning – that is, selecting the next best gaze location given the current observations – remains largely unsolved. We propose a probabilistic system that models the gaze sequence as a finite-horizon Bayesian sequential decision process. Direct policy search is used to reason about the next best gaze locations. The system integrates bottom-up saliency information, top-down target knowledge and additional context information through principled Bayesian priors. This results in proposal gaze locations that depend not only the featural visual saliency, but also on prior knowledge and the spatial likelihood of locating the target. The system has been implemented using state-of-the-art object detectors and evaluated on a real-world dataset by comparing it to gaze sequences proposed by a pure bottom-up saliency-based process and to an object detection approach that analyzes the full image. The target-directed attention system is shown to result in higher object detection precision than both competitors, to attend to more relevant targets than the bottom-up attention system, and to require significantly less computation time than the exhaustive approach.**

I. Introduction

Imagine a robot trying to find a laptop in an office. Most likely, once it enters the room, it stops and starts exhaustively scanning all locations in its field of view. In settings with large amounts of image or video data, the shortcomings of traditional approaches to object detection become apparent: the system analyzes the input image at all pixel locations either by filtering the full image [1] or by using a sliding window classifier [2]. This approach is not only slow, but also prone to producing many false positives at unlikely image locations.

A more natural way of thinking about visual search is a reasoning system that points the agent to those image locations at which the target object is most likely to be. This would reduce processing time since a high-quality object detector only has to be employed at a small number of promising image locations. In addition, such an approach would reduce the false positive rate because unlikely – but visually similar image locations – are never visited. Indications about likely target locations are available at large: low-level attention [3] points to locations that are visually salient based on bottom-up features; target-driven, top-down models of the target (e.g. the interest model of Gould et al. [4] or gist-like features [5] trained on the target object) provide coarse, but fast information about the object likelihood; in addition, information about the context of the target such as spatial location priors [6] is available. This scenario is particularly relevant for robots with a low-resolution, wide field-of-view camera planning image acquisition for a high-resolution, pan-tilt-zoom camera [4], [7].

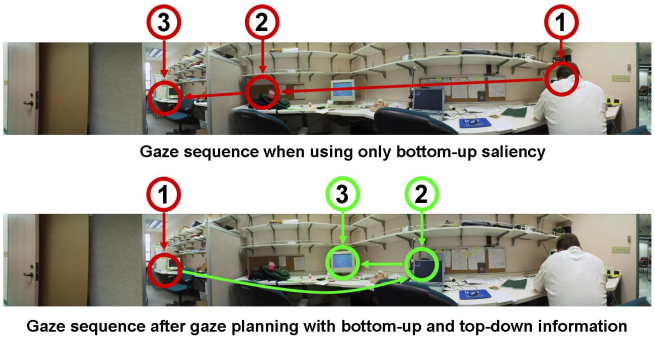

Fig. 1. Typical gaze sequences before and after gaze planning. The numbers specify the gaze order. Green circles indicate detections of the high-quality object detector. Red circles indicate non-detections (both true and false negatives).

We propose a target-directed attention system that reasons about the next best location to attend to. The bottom plot in Figure 1 shows a typical result. The goal is to plan a gaze sequence in the image that attends to and subsequently detects (i.e. recognizes) as many monitors as possible. After gaze planning, the system attends to all three monitors in the image, two of which are detected by the employed high-level detector (more details in Section VI). The proposed system integrates bottom-up saliency, top-down target information, and spatial target context based on the information mentioned in the previous section. Furthermore, gaze sequences are modeled as finite-horizon Bayesian sequential decision process. This permits the system to plan the next best gazes based on its knowledge about the world and its history of previous gazes and observations.

In combination with state-of-the-art object detectors, our approach is tested quantitatively on large, cluttered, panoramic scenes of the Caltech Office DB [8]. We compare our target-directed attention system to an attention system that is based on bottom-up visual saliency only [9] (BU only) and to a system that exhaustively analyzes the full image [1]. Our experiments show that target-directed attention results in higher precision than both competitors. Our system also

attends to significantly more relevant targets than the BU only approach. In addition, target-directed attention requires only 25% of the computation time of the exhaustive system.

II. Related work

Our approach builds on the computational saliency model of Itti and Koch [9]. The central aspect of their model is a two-dimensional saliency map generated by analyzing low-level image features (color, orientation, intensity). Competition among neurons results in a winning location, i.e. the most salient location. Subsequent inhibitions of that location (inhibition of return) makes the system shift to the next most salient location, thus producing an ordered sequence of gaze locations. Our system uses gaze sequences from this bottom-up saliency map as proposal gazes.

Navalpakkam and Itti [10] extend [9] by introducing top-down gains on the bottom-up saliency map. Statistical top-down knowledge about target and background tunes the bottom-up maps with the goal of maximizing the signal-to-noise ratio. There is evidence that the approach is psychophysically plausible [11]. However, it has not been evaluated on larger scale with datasets commonly used in vision or robotics. Walther and Koch [12] propose a computational system that models attention to salient "proto-objects". Proto-objects are described as volatile, bottom-up units of visual information that can be bound into objects if attended to [13]. However, in [12] no integration with top-down target information is implemented. Tsotsos et al. [14] model visual attention on neuron level by selectively tuning the visual processing network using attentional biases based on top-down information. Instead of incorporating target-specific information directly, Oliva et al. [15] combine bottom-up saliency with contextual priors that model the relationship between context features and the target. Walther et al. [16] use bottom-up saliency to learn and identify particular objects in cluttered scenes.

Early work on selective vision using decision theory includes the TEA-1 system by Rimey and Brown [17]. The system exploits the spatial structure of a scene. Based on handcrafted "goodness" functions, it sequentially collects visual evidence. Similarly, Wixson [18] exploits spatial relationships of the target object with other objects to efficiently search for objects in indoor scenes. Paletta and Pinz [19], and Laporte and Arbel [20] explore decision making for active object recognition and pose estimation. The task is to select a viewpoint in 3D so as to decide on the object label (and pose [20]) with as few views as possible. Due to the large computational complexity of these decision systems, only the approach in [19] is non-myopic. Minut and Mahadevan [21] propose a two-layered architecture to simulate selective attention for visual search tasks. The location of the next gaze is primed coarsely using reinforcement learning and subsequently more finely using bottom-up visual saliency. Gould et al. [4] present a method using peripheral-foveal vision for identifying and tracking objects in an easy dynamic environment. Their approach uses a learned attentive interest map and is based on a myopic policy.

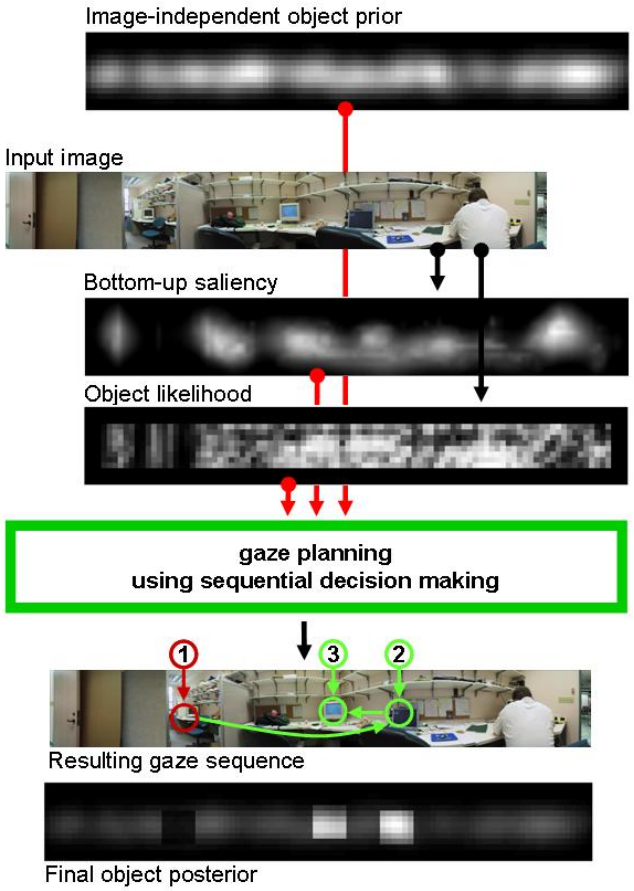

Fig. 2. Inputs and outputs of our selective attention system. The inputs consist of a new image and a prior over object locations, which is learned off-line from a large image database. In addition, we have a likelihood model from which observations (detections) can be simulated and a saliency map [9]. The output of the planner is the sequence of gazes and the posterior distribution. The posterior distribution can be used to focus the computational resources for tasks like detection and recognition where the posterior is high.

III. The sequential gaze planning approach

Our goal is to plan a sequence of gazes so as to minimize the uncertainty in the posterior distribution over the location of objects in the scene. To do so, we adopt a principled Bayesian decision theoretic framework [22]. This approach results in a concentrated posterior distribution over locations, which allows us to conduct subsequent tasks, such as detection and recognition, more efficiently. As a result of this process computational resources only need to be allocated to parts of the image where the posterior is high. Our experiments will show that sequential gaze planning is indeed more computationally efficient than traditional sliding window approaches widely adopted in computer vision.

Figure 2 is an overview of our approach combining bottom-up and top-down information. In particular, the Bayesian decision theoretic framework allows us to integrate, in a statistically coherent way, the following elements:

1) A *prior distribution,* $p(\mathbf{x}_0)$*, over object locations,* $\mathbf{x}_0 \in \mathbb{R}^2$. This prior is either learned from large collections of images or is constructed using expert knowledge

1) Obtain proposal gaze sequence of length $numGazes$ from bottom-up, winner-take-all saliency [9].
2) Enumerate all possible gaze orderings: $permutations(1 : numGazes)$
3) For $k = 1 : NumberOfPossibleGazeOrders$:
   - For $i = 1 : N$:
     - For $t = 1 : T$:
       a) Select the t^{th} gaze location $\mathbf{x}_t^{(i)}$ in the current gaze order.
       b) Generate detection observations as described in Section V-C: $\mathbf{y}_t^{(i)} \sim p(\mathbf{y}_t|\mathbf{x}_t^{(i)})$.
       c) Update the object's location posterior $p(\mathbf{x}_t|\mathbf{y}_{1:t}^{(i)}, \pi)$ using Bayes rule to combine the prior (posterior at gaze $t-1$) and the likelihood.
     - Evaluate the cost function $C_{k,i}^{\pi}$.
   - Evaluate the cost function $C_k^{\pi} = \frac{1}{N}\sum_{i=1}^{N} C_{k,i}^{\pi}$.
4) Choose the gaze sequence with the minimum expected cost.

Fig. 3. Pseudo-code of sequential gaze planning in the open-loop control (OLC) setting. Here, N denotes the number of Monte Carlo samples and T is the planning horizon. In replanning with open-loop feedback control (OLFC), the system uses the present gaze location and the estimated posterior distribution (instead of the prior) as the starting point for the simulations. It is implicit in the pseudo-code that we freeze the random seed generator so as to reduce variance.

(Section V-B).

2) A *model of the sensors (detectors)*. This likelihood model, $p(\mathbf{y}_t|\mathbf{x}_t)$, indicates the probability of positive ($\mathbf{y}_t = 1$) and negative ($\mathbf{y}_t = 0$) detections at a specified location $\mathbf{x}_t$ during the t^{th} gaze (Section V-C).
3) An estimate of the *posterior distribution,* $p(\mathbf{x}_t|\mathbf{y}_{1:t})$*, over object locations given the information gathered in gazes 1 to* t. Since the the domain of $\mathbf{x}$ is two-dimensional, a sensible strategy for computing the posterior is to discretize this domain. If we are planning T steps into the future, this posterior update is repeated T times.
4) A *policy* π. The policy is a sequence of gaze actions. In particular, we consider all permutations (that is, sequences of gazes) of the most promising locations recommend by a saliency map [9]. The learned policy corresponds to the most promising sequence (Section V-A).
5) A *cost function*, C^{π}, that encodes the objective of increasing the information in the posterior distribution as quickly as possible. This cost function is discussed below.

Having defined the problem as one of quickly improving the information in the posterior distribution, the following is a natural cost function:

$$C^{\pi} = \mathbb{E}_{p(\mathbf{x}_{0:T}, \mathbf{y}_{1:T}|\boldsymbol{\pi})}\left[\sum_{t=1}^{T} \lambda^{t-1} \Delta H\right] \quad (1)$$

$$\Delta H = H(p(\mathbf{x}_t|\mathbf{y}_{1:t}, \pi)) - H(p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}, \pi)), \quad (2)$$

where $H(p(\mathbf{x}_t|\mathbf{y}_{1:t}, \pi))$ denotes the entropy of the posterior distribution at time step t and $\lambda \in (0, 1]$ is a discount factor. This cost function is expensive to evaluate and hides an enormous degree of complexity since it is a function of an *intractable filtering distribution* $p(\mathbf{x}_t|\mathbf{y}_{1:t}, \pi)$.

In summary, the policy π is a sequence of possible gaze locations, each of which will entail a specific cost C^{π}. Finding the policy which minimizes the cost is a search problem. We carry out this search using well-accepted methods from sequential decision making, in particular a simulation approach often referred to as direct policy search [23], [24]. This simulation method for reinforcement learning has led to significant achievements in control and robotics [25], [26], [27]. In our setting, given the current policy π, we sample N sequences of observations and corresponding posterior distributions using the PEGASUS methodology [24]. The cost function can then be approximated with the following Monte Carlo estimator:

$$C^{\pi} \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \lambda^{t-1} \Delta H. \quad (3)$$

The algorithm used for this simulation is shown in Figure 3. We obtain Monte Carlo estimates of the cost function for each sequence of gazes and, finally, choose the sequence of lowest cost. The pseudo-code in Figure 3 states that given an unseen image, the system generates first a set of $numGazes$ proposal gaze locations (see Section V-A), which result in $permutations(1 : numGazes)$ possible gaze orderings. The performance of the system $numGazes = min(T, numGazes)$ does not improve for values large than 4. So, using an exhaustive list of permutations is computationally feasible. For larger $numGazes$ or planning horizons T, approximative methods need to be employed [24]. Each of these possible gaze orderings is simulated N times, the cost of each simulation is recorded, and the gaze order with the minimum expected cost is selected. During simulation, the object's location posterior is computed from the object's location prior (Section V-B) and the object likelihood (Section V-C).

After the gaze planning system decides on the next best gaze location, the focus of attention is directed to that location and an expensive object detector (see Section V-D) is employed to decide whether the target object is present at that location.

IV. Replanning and open-loop-feedback control

In the previous section, we discussed the algorithm for gaze planning in general. Now, as we progress in the gaze sequence, it is possible to use the newly gathered observations (by running an actual detector; see Section V-D) to update the posterior distribution. This distribution can then be used as the prior for subsequent planning and simulation steps. This process of replanning is known as open-loop feedback control (OLFC) [28]. We can also allow for the planning horizon to recede. That is, as we progress along the sequence of gazes, we keep planning T steps ahead of the current position. This control framework is also known as receding-horizon model-predictive control [29] and is the approach taken in the following experiments.

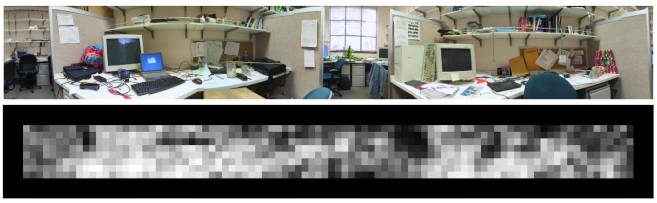

Fig. 4. Image and corresponding object likelihood as obtained with gist features and SMLR classifier

V. Implementation details

In order to simulate the excessive amount of information necessary to be processed by an agent when performing visual search in a new environment, we use 40 randomly selected images from the Caltech Office database [8] for testing and evaluation. The images are 280×1960 pixels and show panoramic (360°) views of office scenes simulating the information available to an agent with an omnidirectional camera. The size of the images makes it unfeasible to analyze them fully in real time. The goal of the experiments is to locate monitors and computer screens in the scenes by minimizing the uncertainty in the posterior distribution of object locations. With this task in mind, we adopt the following implementation of the various elements of the sequential gaze planning approach.

A. Generating proposal gaze locations

The input to the gaze planning system is an ordered set of candidate gaze locations. In our experiments, we use the first $numGazes$ salient locations proposed by the attention system of Itti and Koch [9] (details on in Section II). We use the implementation of bottom-up attention in the Saliency Toolbox of Dirk Walther [12]. After a particular gaze is executed, the corresponding location is inhibited as in the inhibition-of-return in the bottom-up attention system of [9].

B. Estimating the location prior

The location prior is a two-dimensional distribution of likely object locations before any object detection. During gaze planning, the object prior is updated using the simulated observations and the object likelihood and becomes the object posterior. The current object posterior is used as the object prior in the next round of gaze planning. The location prior is estimated through kernel density estimation with a Gaussian kernel from a training set of the Caltech Office database. Figure 5 (top) shows that monitors are most likely to appear in a horizontal band around the center of the image.

C. Likelihood model

The object likelihood encodes the probability $p(\mathbf{y}_t|\mathbf{x}_t)$, the probability that the region around a particular image location $\mathbf{x}_t$ is indicative of the target object. Clearly, this probability could be provided by a full-fledged object detector. But that approach would be too slow for a reasoning system that considers many possible object locations. The main goal at this point of the system is to get a very fast, crude estimate of the object likelihood similar to humans who catch the "gist" of a scene or an image region within milliseconds and in the near absence of attention [30]. Inspired by this behavior, we use the computational implementation of the gist as described in [5] and the sparse multinomial logistic regression classifier (SMLR) of Krishnapuram et al. [31].

Gist feature computation From the training set of the Caltech Office database, we generated 7500 training patches of size 140×140 pixels. This size corresponds to half the height of the image. The training images are gray-scaled and down-sampled by a factor of 2 in order to simulate the coarse spatial information available when quickly scanning a room. For each image, a steerable pyramid transformation is computed (6 orientations and 3 scales), the image is divided into a 4x4 grid, and the average energy of each channel in each grid cell is computed, resulting in 288 features. Subsequently, the dimensionality is reduced by performing PCA and retaining 80 dimensions.

Training the SMLR classifier Of the training set described above, we use 600 positive (i.e. containing monitors of various sizes) and 600 negative patches to train a SMLR classifier with RBF kernel. Inputs are the 80 PCA features. The classification rate on a validation set is 62%.

Computing the object likelihood (Step 3(b) in Figure 3) Around each location $\mathbf{x}$ that the gaze planning system is reasoning about, a window of 140×140 pixels is extracted and downsampled. Gist features as described above are extracted and projected onto the PCA basis. Using these features, the image window is classified with the trained SMLR classifier. The classifier returns a confidence measure between 0 and 1 that can directly be used as object likelihood. The gaze planning system only computes the likelihood at locations relevant for its reasoning. However, for illustrative reasons, Figure 4 shows the likelihood map for a full image. Note the fact that besides locations on and around monitors, the classifier marks general "desk-like clutter" with higher probabilities. This makes sense as monitors are usually located on desks. The computation time per image window is around a tenth of a second.

Obviously, the better the estimated or computed object likelihood is, the better the gaze planning system will perform. This is a trade-off between available feature information, accuracy, and computation time. We will discuss some ideas regarding the object likelihood in Section VII.

D. Object detector during replanning

Once the gaze planning system has selected a gaze location, a full-fledged object detector is employed in a window of 200x200 pixels around that location. At this stage, any object classifier may be used since the gaze planning system is independent of the gaze execution stage. We use the boosted detector described in [1] with the Matlab implementation provided by A. Torralba [32]. The classifier has been trained for frontal monitors on 400 training images of the LabelMe database [33]. During execution, the detector is run on three scales (scalefactor = 0.7) and all detection scores above a threshold are counted as positives. The detection threshold of $detThres = 50$ has been set using a validation set.

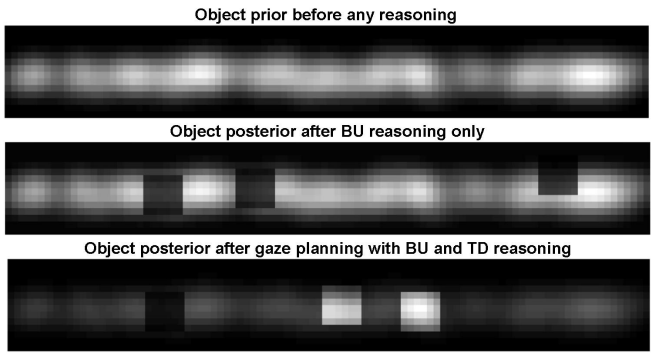

Fig. 5. Object prior and expected object posterior for the two gaze sequences in Figure 1

VI. Results

The goal of the experiments is to detect a maximum number of objects with a minimum number of false positives by minimizing the uncertainty in the posterior distribution over the location of objects in the scene.

In the following, we compare our proposed gaze planning system (GP) to a system that also analyzes only selected locations in the input image, but does not employ any target specific top-down or context information [9]. This bottom-up only system (BU) returns a sequence of generally visually salient locations, but the selection is independent of the task. In addition, we compare gaze planning to object detection by analyzing the full image of 280×1960 pixels (Full images, FI). Here, we use the same object detector with the same parameters (number of scales, scale-factor, detection threshold) as for the localized object detection in Section V-D.

Our hypothesis is that gaze planning is considerably faster than full image analysis and results in less false positives while generating a similar amount of detections proportionally to the number of gazes. This is confirmed by the experiments.

Qualitative results Figure 1 shows typical outputs of the selective attention systems. In both cases, three image locations are analyzed. In the top plot, the system attends to the three most salient image locations based on bottom-up features only and runs the object detector on regions around these locations. All three detections are negative as indicated by the red circles (one miss). In the bottom plot, gaze planning based on top-down and context information in addition to bottom-up information is employed. Here, both the gaze order and the image locations that the system attends to change in a successful way: all three gazes attend to target objects. The object detector correctly detects two of those (green circles) and misses one (red circle).

The example is typical in that gaze planning leads to an increased number of targets being attended to. Depending on the quality of the object detector, most of these targets are also detected.

Figure 5 depicts the difference in the object posterior for the two gaze sequences. The top plot shows the object prior. In the middle, the object posterior after executing the BU gaze sequence can be seen. Three expected non-detections did not change the uncertainty about the object location much. The bottom plot shows the posterior after gaze planning. Here, the entropy of the posterior relative to the prior is clearly lower thus leading to the selection of this particular gaze sequence because the uncertainty in the object location has been reduced most.

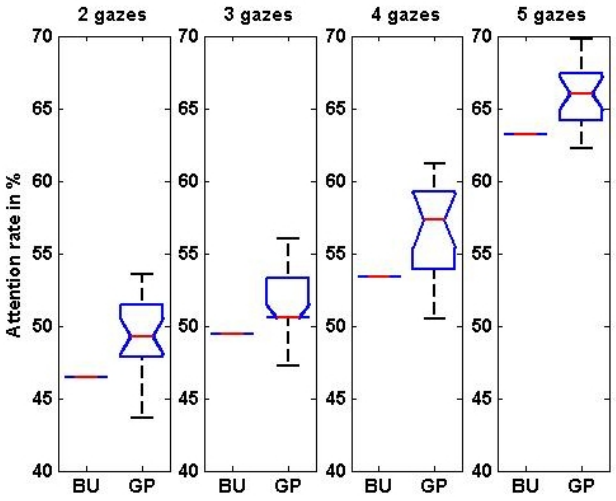

Fig. 6. Attention rate in % when using only bottom-up saliency (BU) or when using gaze planning (GP). Both systems analyze the images at 2, 3, 4, and 5 locations (gazes). Gaze planning uses a planning horizon of $T = 4$.

Quantitative results In Figures 6, 7, and 8, we compare the performance of the attention system based on bottom-up saliency, the gaze planning system, and the detection on full images quantitatively. All data points with box-whisker plots have been repeated 20 times in order to analyze statistical significance. Both the system based on bottom-up saliency only and the full image analysis are deterministic systems and thus do not generate box-whisker plots.

Figure 6 depicts the attention rate (percentage of relevant objects being attended to) of the selective attention systems. This is the most relevant performance measure for our system because it is a measure that is independent of the subsequently employed high-resolution object detector. It shows that our gaze planning system attends to more relevant objects in the scene than the BU system. The figure also shows that the attention rate increases when the systems attend to more locations in the image. E.g. , the systems attend to five distinct image locations in the right-most plot. In all four cases, gaze planning significantly outperforms the system using bottom-up saliency only.

Figure 7 shows that gaze planning consistently detects more relevant objects than BU saliency. The fact that full image analysis achieves a higher detection rate is not surprising because the attention systems only analyze up to five (!) locations in the image. Note that gaze planning with five gazes in Figure 6 attends to a higher percentage of relevant objects than the full image analysis detects in Figure 7. So gaze planning with a large number of attended locations

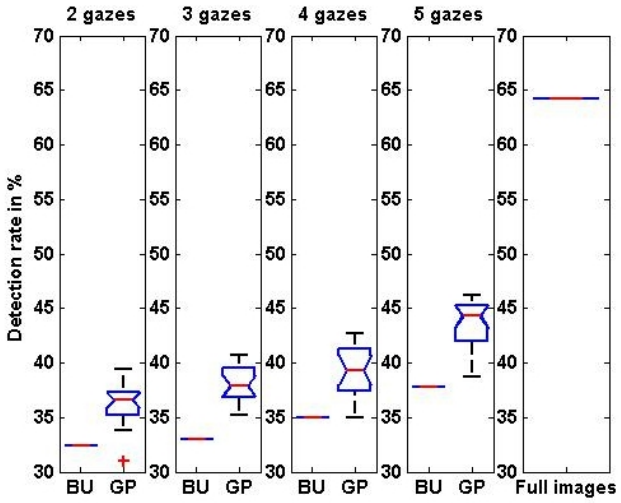

Fig. 7. Detection rate in % when using bottom-up saliency only (BU), when using gaze planning (GP), or when analyzing the full image (Full images). Both selective attention systems analyze the images at 2, 3, 4, and 5 locations (gazes). Gaze planning uses a planning horizon of $T = 4$.

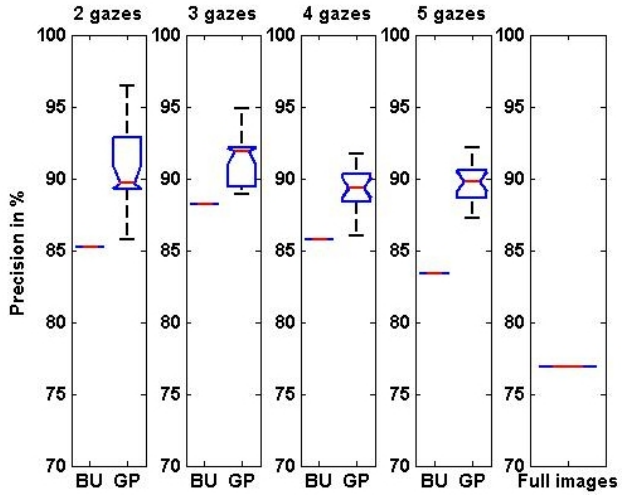

Fig. 8. Precision in % when using bottom-up saliency only (BU), when using gaze planning (GP), or when analyzing the full image (Full images). Both selective attention systems analyze the images at 2, 3, 4, and 5 locations (gazes). Gaze planning uses a planning horizon of $T = 4$.

using a higher-quality object detector is expected to reach the performance of full image analysis in Figure 7.

Another advantage of gaze planning is that it reduces the number of false positives significantly by analyzing only selected image locations. This results in the high precision of our system shown in Figure 8. With a precision of around 90%, gaze planning outperforms both bottom-up attention and analysis of full images. Since the number false positives is so low (< 10 false positives out of 200), small changes in the false positive rate make the precision vary in Figure 8.

Computational complexity Table I shows the difference in computation time of the three approaches. Analyzing the full images takes about four times longer than analyzing the images at only four selected locations whereas gaze planning takes only about 30 seconds longer than BU only.

bottom-up only	183 seconds
gaze planning	214 seconds
full images	826 seconds

TABLE I

COMPARISON OF THE AVERAGE COMPUTATIONAL COMPLEXITY PER IMAGE OF THE THREE APPROACHES. BOTH SELECTIVE ATTENTION SYSTEMS ANALYZE THE IMAGE AT 4 LOCATIONS (GAZES). GAZE PLANNING USES A PLANNING HORIZON OF $T = 2$.

BU	$T = 1$	$T = 2$	$T = 3$	$T = 41$
182	205	214	241	248

TABLE II

COMPUTATION TIME IN SECONDS FOR ATTENTION USING ONLY BOTTOM-UP SALIENCY (BU) AND GAZE PLANNING (PLANNING HORIZON $T = 1$, $T = 2$, $T = 3$, AND $T = 4$).

The approaches have been tested on a 3GHz Linux machine with 2GB of memory and averaged over 40 images.

Dependency on the planning horizon In our experiments with gaze planning, we implemented receding-horizon model-predictive control (Section IV): After each planning step, the system runs a detector on the first gaze of the planned gaze sequence, updates its object posterior, and uses it as prior in its next planning step. The planning horizon T specifies the number of steps that the system is planning ahead. Although a larger planning horizon does not increase computation time significantly (Table II), it is still rewarding to analyze the benefit of larger planning horizons.

Figure 9 depicts the dependency of the precision on the planning horizon T. For comparison, the figure also shows the performance of the system based on bottom-up saliency only having no planning horizon and being myopic and the analysis of the full images. At any time during planning, the gaze planning system is selecting from four possible proposal gazes. Under these conditions, the experiments show that an increase of the planning horizon results only in moderate performance improvement. The largest performance increase appears between $T = 1$, i.e. myopic planning, and $T = 2$. So there is a slight advantage of not being myopic. We expect that the improvement due to longer planning horizon increases when the system selects from a larger set of proposal gazes. This will be the case as the complexity of the task increases. We plan to investigate this further in future work.

VII. DISCUSSION AND FUTURE WORK

We presented a framework for gaze planning using sequential decision-making and integrating various kinds of information probabilistically (bottom-up, top-down, context). The system demonstrates visual search on a real-world dataset and is shown to result in a higher precision than both an attention system based on bottom-up information only and a system that exhaustively analyzes images. In addition, our system attends to significantly more relevant

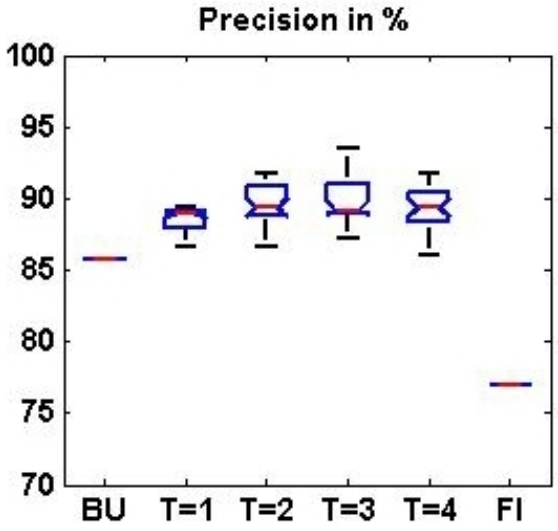

Fig. 9. Precision in % using bottom-up saliency only (BU), gaze planning with planning horizons $T = 1$, $T = 2$, $T = 3$, and $T = 4$, and full image analysis (FI). Both selective attention systems analyze the images at 4 locations (gazes).

objects than the BU only attention system. Also, the proposed gaze planning system is four times faster than the traditional approach of analyzing the image exhaustively. The presented framework is principled in that it is independent of particular object likelihoods or object detectors. Very specific object knowledge (i.e. looking for red mugs or striped animals) would permit to specify a very accurate object likelihood and would thus increase the attention rate. Better high-quality object detectors will improve the detection rate at the attended locations. After the training of the object detectors, gaze planning does not require any additional training and is thus complementary to existing approaches to object detection.

Besides being computationally more efficient and more precise, target-directed attention also relates to the human visual system. Research has shown that human visual attention integrates bottom-up, image-driven and top-down, target-based components [34], [3]. So our gaze planning system might also be interpreted from the perspective of human perception. The system combines low-level vision (bottom-up saliency [9]) with coarse scene layout/gist to guide focused attention as conceptualized in the "triadic architecture" by Rensink [13]). Psychophysical experiments to study these conjectures are within our future research goals.

Other future projects include the learning of relevant low-level input features (see [10]) or the object likelihood during gaze planning, and the implementation of the system on a robot with a pan-tilt-zoom unit. Our Bayesian sequential decision approach also allows naturally for the inclusion of dynamic models so as to treat moving objects. This will be important when considering robotic applications.

REFERENCES

[1] A. Torralba, K. P. Murphy, and W. T. Freeman, "Sharing visual features for multiclass and multiview object detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 854–869, 2007. [Online]. Available: http://people.csail.mit.edu/torralba/iccv2005/boosting/boosting.html

[2] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *Intl. J. Computer Vision*, vol. 38, no. 1, pp. 15–33, January 2000.

[3] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, March 2001.

[4] S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Messner, G. Bradski, P. Baumstarck, S. Chung, and A. Y. Ng, "Peripheral-foveal vision for real-time object recognition and tracking in video," in *International Joint Conference for Artificial Intelligence IJCAI'07*, Hyderabad, India, January 2007.

[5] A. Torralba, K. Murphy, W. Freeman, and M. Rubin, "Context-based vision system for place and object recognition," in *International Conference on Computer Vision ICCV'03*, October 2003.

[6] A. Torralba and P. Sinha, "Statistical context priming for object detection," in *International Conference on Computer Vision ICCV'01*, July 2001.

[7] D. Kragic and M. Björkman, "Strategies for object manipulation using foveal and peripheral vision," in *International Conference on Computer Vision Systems ICVS'06*, 2006.

[8] M. Fink and P. Perona, "The Full Images for Natural Knowledge, Caltech Office DB," California Institute of Technology, Tech. Rep. CSTR:2003.008, 2003.

[9] 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, May 2000.

[10] V. Navalpakkam and L. Itti, "Ann integrated model of top-down and bottom-up attention for optimizing detection speed," in *Conference on Computer Vision and Pattern Recognition CVPR'06*, New York, USA, June 2006.

[11] ——, "Search goal tunes visual features optimally," *Neuron*, vol. 53, no. 4, pp. 605–617, Feb 2007.

[12] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, pp. 1395–1407, 2006.

[13] R. Rensink, "The dynamic representation of scenes," *Visual Cognition*, vol. 7, no. 1/2/3, pp. 17–42, 2000.

[14] J. Tsotsos, S. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, pp. 507–545, 1995.

[15] A. Oliva, A. Torralba, M. Casthelano, and J. Henderson, "Top-down control of visual attention in object detection," in *Intl. Conf. on Image Processing ICIP*, 2003, pp. 253–256.

[16] D. Walther, U. Rutishauser, C. Koch, and P. Perona, "Selective visual attention enables learning and recognition of multiple objects in cluttered scenes," *Computer Vision and Image Understanding*, vol. 100, pp. 41–63, 2005.

[17] R. Rimey and C. Brown, "Control of selective perception using bayes nets and decision theory," *International Journal of Computer Vision*, vol. 12, no. 2/3, pp. 172–207, 1994.

[18] L. Wixson and D. Ballard, "Using intermediate objects to improve the efficiency of visual search," *Intl. J. Computer Vision*, vol. 12, no. 2-3, pp. 209–230, 1994.

[19] L. Paletta and A. Pinz, "Active object recognition by view integration and reinforcement learning," *Robotics and Autonomous Systems*, vol. 31, no. 1-2, pp. 1–18, 2000.

[20] C. Laporte and T. Arbel, "Efficient discriminant viewpoint selection for active bayesian recognition," *International Journal of Computer Vision*, vol. 68, no. 3, pp. 267–287, July 2006.

[21] S. Minut and S. Mahadevan, "A reinforcement learning model of selective visual attention," in *Proc. 5th Intl. Conf. on Autonomous Agents*, 2001, pp. 457–464. [Online]. Available: citeseer.ist.psu.edu/article/minut01reinforcement.html

[22] D. Lindley, *Making Decisions*, 2nd ed. Wiley, 1980.

[23] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3, pp. 229–256, 1992.

[24] A. Y. Ng and M. I. Jordan, "PEGASUS: A policy search method for large MDPs and POMDPs," in *Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 2000.

[25] N. Kohl and P. Stone, "Policy gradient reinforcement learning for fast quadrupedal locomotion," in *IEEE International Conference on Robotics and Automation*, 2004.
[26] A. Ng, A. Coates, M. Diel, V. Ganapathi, J. Schulte, B. Tse, E. Berger, and E. Liang, "Inverted autonomous helicopter flight via reinforcement learning," in *International Symposium on Experimental Robotics*, 2004.
[27] J. Peters and S. Schaal, "Policy gradient methods for robotics," in *IEEE Intl. Conf. on Intelligent Robotics Systems*, 2006.
[28] D. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific, 2007.
[29] J. Maciejowski, *Predictive Control with Constraints*. Prentice Hall, 2002.
[30] L. Fei-Fei, R. Van Rullen, C. Koch, and P. Perona, "Rapid natural scene categorization in the near absence of attention," *Proceedings of the National Academy of Sciences of the USA*, vol. 99, no. 14, pp. 9596–9601, July 2002.
[31] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink, "Learning sparse bayesian classifiers: multi-class formulation, fast algorithms, and generalization bounds," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 957–968, June 2005.
[32] A. Torralba, iCCV 2005 short course. A simple object detector with boosting, http://people.csail.mit.edu/ torralba/iccv2005/boosting/boosting.html.
[33] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "Labelme: a database and web-based tool for image annotation." MIT AI Lab, Tech. Rep. AIM-2005-025, 2005.
[34] J. Wolfe, "Guided search 2.0: A revised model of visual search." *Psychonomic Bulletin & Review*, vol. 1, no. 2, pp. 202–238, 1994.