Inhibition of Return in the Bayesian Strategy to Active Visual Search

Kai Welke, Tamim Asfour, Rüdiger Dillmann
Karlsruhe Institute of Technology (KIT)
Institute for Anthropomatics, Humanoids and Intelligence Systems Lab
{welke,asfour,ruediger.dillmann}@kit.edu

Abstract

The inhibition of return mechanism in systems based on visual attention allows to generate a sequence of gaze directions (saccades) from saliency data. In this work we propose such a mechanism that seamlessly integrates with the Bayesian Strategy to attention. Taking into account a memory of attended locations, the requirement for consistency constitutes the drive for the generation of saccades. The general probabilistic model of active saliency is introduced, which is applicable to systems based on saliency maps as well as to landmark based systems. The model is further refined and validated in the context of an active visual search task.

1 Introduction

The application of active camera systems is nowadays common in state of the art humanoid platforms. The popularity of such active systems not only stems from the resulting enhanced visual field but is also attributed to the fact that the gaze is an essential communication channel in human interaction. Robots that have a similar ability of using their gaze as human are more appealing and thus more likely to be accepted in human-centered environments.

The application of an active camera system raises a wide range of problems. In this work we address the problem of directing the gaze of the active system to interesting locations in the context of visual search tasks. The problem of identifying and focusing on interesting locations in the scene is usually referred to as overt visual attention.

The work by Itti et al. [1] has been among the first frameworks for the generation of attention on technical systems. Their work focused on biological plausible processing in the realization of bottom-up attentional mechanisms. In the past years, several new approaches to the generation of attention have been proposed. The most complete view has been provided by the Bayesian Strategy which allows the integration of different aspects of attention, such as bottom-up processing, top-down processing, and scene context ([2], [3]). In the Bayesian Strategy, saliency is generated according to the probability of detecting an object O at a spatial location X given the current visual input J. As shown in [4] the factorization of $P(O = 1, X|J)$ allows to seamlessly integrated the different aspects of attention, each described by one factor. While this strategy allows to generate saliency data, it does not solve the problem of generating gaze sequences from the saliency data in an integrated manner. The question remains: What could be the mechanism that allows to determine a sequence of distinct visual locations which are to be fixated by the active system? In this work we propose and evaluate a novel model that allows to generate sequences of gaze shifts based on the Bayesian Strategy.

The paper is organized as follows. In the next section the general model of active saliency for the generation of gaze sequences is introduced. In Section 3 the application of the proposed model in active visual search is discussed before an experimental validation is provided in Section 4.

2 Active saliency

The proposed model is based on the assumption that transsaccadic memory plays an important role in active visual search [5]. While this assumption is discussed controversially in the context of human visual search, the application of such a memory on a technical system seems to be feasible. The formulation of the Bayesian Strategy is extended in order to include the requirement of a consistent transsaccadic memory of salient locations. For this purpose, the inconsistency of the memory representation corresponding to a spatial location X is described with the random variable I. Thus, the desired posterior in our model becomes $P(O = 1, X, I|J)$, which we will refer to as *active saliency* s_a.

In the following, a distinction is made between peripheral and foveal processing. Object recognition is performed based on foveal processing, while peripheral processing reports changes of the environment that might affect consistency of the memory. The peripheral observation is then defined by the observation of change Z_c. The foveal processing includes the measurement Z_f which updates the believe of object existence and location $bel(O = 1, X)$. Further, the foveal processing results in the validation of memory entities which is represented with the variable V. Reformulating the above model using the introduced observations yields

$$s_a = P(O = 1, X, I|Z_f, Z_c, V).$$

Assuming perfect knowledge of the current gaze direction, OX and I are conditionally independent given all measurements. Further, exploiting the independences from observations and hidden variables as imposed by the above model, the posterior for active saliency under consideration of peripheral perception, foveal perception and memory is given with

$$\begin{aligned} s_a &= P(O = 1, X, I|Z_f, Z_c, V) \\ &= P(O = 1, X|Z_f)P(I|Z_c, V). \end{aligned} \quad (1)$$

The first factor of the above model essentially corresponds to the Bayesian Strategy as proposed in [4]. The second factor allows to integrate the requirement

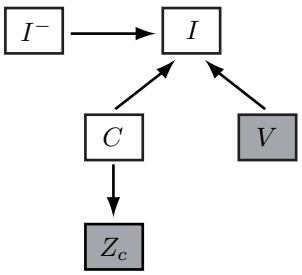

Figure 1. Graphical model of the inconsistency filtering. Inconsistencies I are updated over time under consideration of the measurement of change Z_c and the validation V.

of the consistency of transsaccadic memory and forms the basis for the inhibition of return mechanism. A detailled model for the second factor is defined in the following sections.

2.1 Filtering of inconsistencies

Inconsistencies are filtered over time, where a filtering step involves the integration of the observed change Z_c and occurred validation V. Figure 1 illustrates a graphical model of all involved random variables. In the following, the sample spaces for all variables are defined for N memory entities. For each memory entity a binary random variable is used to encode its state.

- I: Encodes inconsistencies between memory and real world. Inconsistency either takes the value consistent (con) or inconsistent ($\neg con$), thus $I \in \{con, \neg con\}^N$.
- I^-: Previous inconsistencies. $I^- \in \{con, \neg con\}^N$.
- C: Change that occurred in the world as relevant for the memory. The world was either static ($\neg ch$) or changed (ch) with respect to the memory entity, thus $C \in \{ch, \neg ch\}^N$.
- Z: Change of the world as measured by the change sensor. $Z \in \{ch, \neg ch\}^N$.
- V: Encodes whether validation has been performed for the respective entity in order to ascertain its consistency with the real world. $V \in \{val, \neg val\}^N$.

Each random variable encodes the distribution for all memory entities. In order to keep the formalization general, we did not state whether memory entities are stored in a grid based manner as the case in probabilistic saliency map, or in a landmark based manner. For both representations, a common assumption is the independence between entities. Thus, the joint distribution over all variables can be factored to

$$P(I, I^-, C, V, Z) = \prod_{i=1,\dots,N} P(I_i, I_i^-, C_i, V_i, Z_i).$$

According to Fig. 1, the inconsistency for each memory entity i can then be factored using

$$\begin{aligned} &P(I_i, I_i^-, C_i, V_i, Z_i) \qquad (2)\\ &\quad = P(I_i^-)P(I_i|I_i^-, C_i, V_i)P(Z_i|C_i)P(C_i)P(V_i). \end{aligned}$$

The model parameters and the approach for inference of the posterior are introduced in the following for the update of a single entry i.

2.2 Model parameters

The prediction model is defined by the conditional probability of inconsistency I_i given the change C_i, the validation V_i and the inconsistency from the last filtering step I_i^-. This dependency is expressed by the conditional probability table

$$\begin{aligned} &P(I_i = \neg con | I_i^-, C_i, V_i)\\ &\quad = \begin{cases} 0, & \text{if } (I_i^- = con \wedge C_i = \neg ch)\\ 1 - p_v, & \text{if } (V_i = val \wedge I_i^- = \neg con \wedge C_i = \neg ch)\\ 1, & \text{else} \end{cases} \end{aligned}$$

In the above definition, the first statement covers cases where the consistency from the previous step is preserved since no change happened. Validation leads to consistency of the memory if no change happened and the memory was inconsistent, as stated in the second case. The parameter p_v allows to define a confidence for the validation success. In all other cases, the resulting inconsistency equals to one, thus change always overrides validation. This is a necessary statement since the order of the performed validation and measured change is not provided and thus change might have occurred after validation has been performed within the last time interval.

The change sensor is modeled using the forward model $P(Z_i|C_i)$. We define the sensor model for change in a general manner using the sensors sensitivity $w_{c,1}$ and its specificity $w_{c,0}$ in the following way

$$P(Z_i = ch | C_i) = \begin{cases} 1 - w_{c,0} & \text{if } C_i = \neg ch,\\ w_{c,1} & \text{if } C_i = ch \end{cases}.$$

Another parameter required to fully define the probabilistic model is the prior probability of change $P(C_i)$. The prior allows to encode the likelihood of change in the scene p_c. This probability can be used as top-down cue in order to instruct the system to peform more validations in cases, where the scene is changing rapidly. The change prior is defined by

$$P(C_i = ch) = p_c.$$

In summary, the model provides four free parameters which can be used to influence the inference of inconsistencies. The sensitivity $w_{c,1}$ and specificity $w_{c,0}$ of the change sensor as well as the validation probability p_v depend on the system build around the model. These parameters will be defined for the application of active visual search in Section 3. The change prior p_c allows to tune the model according to the volatileness of the current scene.

2.3 Inference of inconsistencies

The inference of the inconsistency of memory entities is formulated based on the model defined in Section 2.1 and its parameters discussed in Section 2.2. Using the prior believe of inconsistency of a memory entity $bel(I_i^-)$, the observation of validation $V_i = v_i$ and of

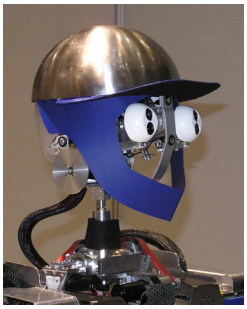

Figure 2. The Karlsruhe Humanoid Head is equipped with a 3 DoF active camera system and offers one perspective and one foveal camera pair.

measured change $Z_i = z_i$ the posterior believe $bel(I_i)$ can be calculated by performing the marginalization

$$P(I_i|Z_i = z_i, V_i = v_i) \tag{3}$$
$$= \frac{\sum_{I_i^-,C_i} P(I_i, I_i^-, C_i, V_i = v_i, Z_i = z_i)}{\sum_{I_i,I_i^-,C_i} P(I_i, I_i^-, C_i, V_i = v_i, Z_i = z_i)}.$$

The posterior believe is calculated using the factorization from (2).

3 Application in active visual search

The model proposed in the previous section is now put into the context of active visual search. The active visual search task is performed using the Karlsruhe Humanoid Head [6] as depicted in Fig. 2. The head provides an active camera system combined with foveal and peripheral camera pairs that allow to actively search for objects in the scene.

The goal of the active visual search task consists in establishing and retaining a consistent memory of target object instances by performing object detection in the foveal cameras. Resulting from the narrow field of view of the foveal cameras, the execution of saccades is necessary in order to enhance the field of sight. The sequence of saccades is generated using the proposed active saliency.

3.1 Saliency calculation and representation

The saliency for the active visual search task is generated according to the Bayesian Strategy. Similar to our previous work in [7] saliency is generated based on top-down knowledge of the object appearance alone. As starting point for the object search, candidates of target objects are extracted in the peripheral views. The extracted candidates are stored in a landmark based map where each candidate is represented with a normally distributed location uncertainty X_i and the probability of existence O_i. This memory representation is updated with each foveal observation of the candidates using the Bayes filter

$$P(O = 1, X|Z_f) = \eta P(Z_f|O = 1, X)P(O = 1, X). \tag{4}$$

The above posterior corresponds to the top-down factor of the Bayesian Strategy.

3.2 Extension to active saliency

In order to execute saccades which ensure a consistent memory based on the above saliency representation, the active saliency is now put in the context of the

Figure 3. Setup used for the active visual search task. The goal consisted in building and retaining a consistent memory of the two cereal boxes based on the active saliency.

active visual search task. While the top-down saliency formulated in (4) corresponds to the first factor of the factorization in (1), in the following the second factor is further defined. For each landmark in the memory of object candidates the inconsistency is updated according to (3) separately.

The detection of change Z_c is performed based on the peripheral observations. Each object candidate in memory is monitored over time by building correspondences between observed candidates and stored candidates according to [8]. If the observation differs from the memory representation, change has been detected. The change sensor in the model is assumed to provide perfect specificity $w_{c,0} = 1$. A sensitivity of $w_{c,1} = 0.9$ is used to express the possibility of unobserved change.

The parameter for validation certainty p_v expresses the ability to validate the memory content based on the foveal views. In order to model decreasing validation performance toward the borders of the foveal images, the following validation certainty is used for the left and right foveal image. Let the spatial 2D location of the object within left and right image be defined by $\vec{u}_l$ and $\vec{u}_r$. The validation certainty for each image is then determined using

$$p_{v,l|r} = e^{-\frac{(\vec{u}_{l|r} - \vec{c}_{l|r})^T \Sigma_v (\vec{u}_{l|r} - \vec{c}_{l|r})}{2}}, \tag{5}$$

where $\vec{c}_l$ and $\vec{c}_r$ are the centers of left and right images and Σ_v is a diagonal matrix of variances. The validation certainty p_v is defined by the product of the validation certainty of both cameras

$$p_v = p_{v,l} p_{v,r}. \tag{6}$$

4 Experimental Validation

In the following, the feasibility of the active saliency model is validated. The goal of the validation consists in illustrating the ability to generate gaze sequenced which assure memory consistency based on the top-down saliency data. The evaluation of the top-down generation of saliency is beyond the scope of this work.

A simple search task as illustrated in Fig. 3 was chosen as example that allows to illustrate the properties of the inhibition of return mechanism. For this

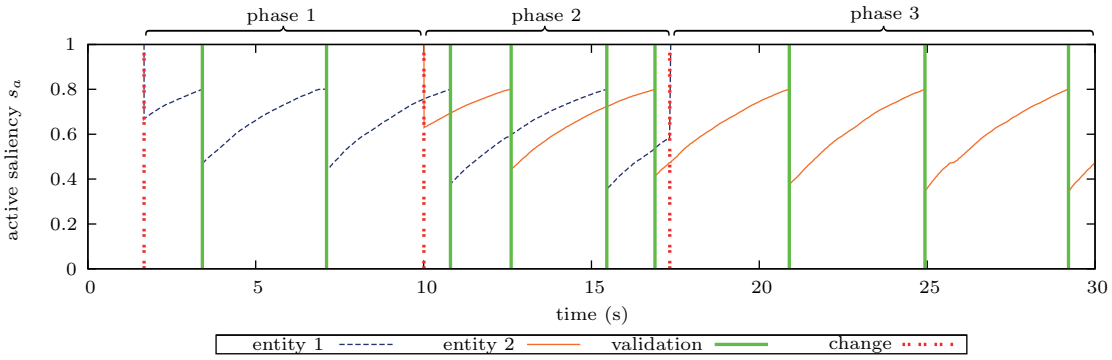

Figure 4. During the active visual search experiment, two object instances were visible to the system. Depending on the current measurement of change and performed validation, the active saliency is inferred. Once the active saliency exceeds the threshold $s_{max} = 0.8$ a saccade is executed.

purpose, two instances of a cereal box were successively brought into the view of the peripheral cameras. Using the proposed model, the active saliency s_a for the stored object instances was continuously updated. A saccade toward an object instance was executed once the active saliency exceeded a given threshold $s_{max} = 0.8$. The active saliency was recorded for two object instances used during the experiment. The time course of active saliency as monitored during the experiment is illustrated in Fig. 4. Time steps where either change was detected or validation was performed are marked with horizontal lines. As can be seen in the illustrated time course, measured change results in an increase of the active saliency and an immediate validation of the corresponding object instance. Validation only is performed, once the active saliency reaches the threshold s_{max}. In the following the different phases of the experiment and their counterparts in Fig. 4 are explained.

The experiment consist of three phases: In the first phase, only one object is visible to the peripheral camera system. For the second phase, an additional instance is positioned in the scene. Finally, the first instance is removed from the scene. The different phases of the experiment are reflected in Fig. 4 with the observation of change. In the first two phases objects are brought into the scene which results in the observation of change and the execution of a saccade in order to detect the new instance based on the foveal views. In the third phase the object is removed from the scene which results in the observation of change and the removal of the object instance from memory after the execution of a saccade and foveal recognition. In between the observed changes the stored object instances are validated frequently, where the frequency is defined by the prior of change p_c, the sensitivity of the change sensor, and the success of the last performed validation. Each time the active saliency exceeds the threshold s_{max} validation is performed.

5 Conclusion

The concept of active saliency takes into account the consistency of memory accumulated during saccadic eye movements. The resulting inhibition of return mechanism is closely coupled to the requirement of a consistent memory. The proposed probabilistic model seamlessly integrates with the Bayesian Strategy to visual attention and provides semantically descriptive parameters. The behavior in terms of gaze sequences as generated by the proposed model has been demonstrated in the context of an active visual search task. While a landmark based representation was used as underlying memory, the proposed approach is formulated in a general way and can also be applied to grid based representations such as probabilistic saliency maps.

Acknowledgement

The research leading to these results has received funding form the European Union Sixths and Seventh Framework Programme FP7/2007-2013 under grant agreement no.270273 and from the German Research Foundation (DFG: Deutsche Forschungsgemeinschaft) under the SFB 588.

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, 1998.

[2] A. Torralba, "Modeling global scene factors in attention." *J Opt Soc Am A Opt Image Sci Vis*, vol. 20, no. 7, pp. 1407–1418, 2003.

[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] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, 2008.

[5] P. D. Graef and K. Verfaillie, "Transsaccadic memory for visual object detail." *Prog Brain Res*, vol. 140, pp. 181–196, 2002.

[6] T. Asfour, K. Welke, P. Azad, A. Ude, and R. Dillmann, "The Karlsruhe Humanoid Head," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, 2008, pp. 447–453.

[7] K. Welke, T. Asfour, and R. Dillmann, "Active multi-view object search on a humanoid head," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2009, pp. 417–423.

[8] ——, "Bayesian visual feature integration with saccadic eye movements," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, 2009, pp. 256–262.