Multimodal Saliency-based Attention for Object-based Scene Analysis

Boris Schauerte, Benjamin Kühn, Kristian Kroschel and Rainer Stiefelhagen

***Abstract*— Multimodal attention is a key requirement for humanoid robots in order to navigate in complex environments and act as social, cognitive human partners. To this end, robots have to incorporate attention mechanisms that focus the processing on the potentially most relevant stimuli while controlling the sensor orientation to improve the perception of these stimuli. In this paper, we present our implementation of audio-visual saliency-based attention that we integrated in a system for knowledge-driven audio-visual scene analysis and object-based world modeling. For this purpose, we introduce a novel isophote-based method for proto-object segmentation of saliency maps, a surprise-based auditory saliency definition, and a parametric 3-D model for multimodal saliency fusion. The applicability of the proposed system is demonstrated in a series of experiments.**

***Index Terms*— audio-visual saliency, auditory surprise, isophote-based visual proto-objects, parametric 3-D saliency model, object-based inhibition of return, multimodal attention, scene exploration, hierarchical object analysis, overt attention, active perception**

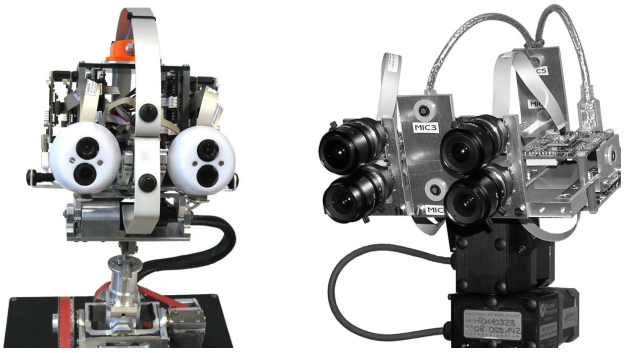

Fig. 1. The Karlsruhe Humanoid Head (left; [1]) which serves as exemplar for the employed experimental sensor platform (right): two wide angle Point Grey Dragonfly2 cameras are mounted above two foveal Dragonfly2 cameras. Beyerdynamics MCE60 omnidirectional microphone pairs are mounted on the front, side, and rear of the setup. The sensor frame is mounted on top of a Directed Perception PTU-D46 pan-tilt unit.

I. INTRODUCTION

Attention is the cognitive process of focusing the processing of sensory information onto salient, i.e. potentially relevant and thus interesting, data. Since robots have limited processing capabilities, attention has attracted an increasing interest in the field of robotics, for example, to enable efficient scene exploration and facilitate real-time processing of the sensory information. Consequently, computational models of attention based on visual and auditory saliency gained increasing interest in theory and applications.

Efficient analysis of complex scenes under the presence of the limited resources of a robotic platform is a key problem of computational perception. In this paper, we present our system for saliency-driven scene analysis which combines audio-visual saliency-based attention with hierarchical, knowledge-driven object analysis and object-based world modeling. We focus the processing onto salient proto-objects, i.e. primitive object hypotheses that are rendered by salient regions, present in the scene. To this end, we steer the sensor setup towards these (proto-)object hypotheses in order to optimize the acoustic and visual perception. The former benefits from the improved sensor alignment with respect to sound sources while the latter takes profit of a complementary pair of cameras that provides a higher resolution and less distorted images of the (proto-)objects (see Fig. 1). The focused proto-object regions are subsequently validated and analyzed using a multimodal hierarchical, knowledge-driven approach. The information about analyzed objects as well as the proto-object hypotheses are managed in an object-based world model. This, in combination with a parametric representation of the proto-object regions that is also used for (multimodal) saliency fusion, enables us to realize object-based inhibition of return.

The remainder of this paper is organized as follows: First, in section II, we provide a brief overview of novel aspects presented in this paper and related work. In section III, we present our system which consists of four components: the processing of visual and acoustic data, the saliency-based fusion and exploration, and the hierarchical analysis of object hypotheses. In section IV, we describe the performed experiments and discuss the achieved results. We conclude with a brief summary and outlook in section V.

II. RELATED WORK

During the last decade computational models of attention have attracted an increasing interest in the field of robotics (see, e.g., [2]–[9]) and diverse other application areas (see, e.g., [10]–[14]). The definition of saliency defines which parts of a signal attract the attention. Unfortunately, only few applicable models for auditory attention exist (e.g., [12], [13]). Most closely related to our work is the model described in [12] which is based on the well-known visual saliency model of Itti et al. [15] and, most notably, has been suc-

This work is supported by the German Research Foundation (DFG) within the Collaborative Research Program SFB 588 “Humanoide Roboter”.

Boris Schauerte, Benjamin Kühn and Rainer Stiefelhagen are with the Institute for Anthropomatics, Karlsruhe Institute of Technology (KIT), Adenauerring 2, 76131 Karlsruhe, Germany. {rainer.stiefelhagen, kuehn, schauerte}@kit.edu

Rainer Stiefelhagen and Kristian Kroschel are with the Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB), Fraunhoferstraße 1, 76131 Karlsruhe, Germany. kristian.kroschel@iosb-extern.fraunhofer.de

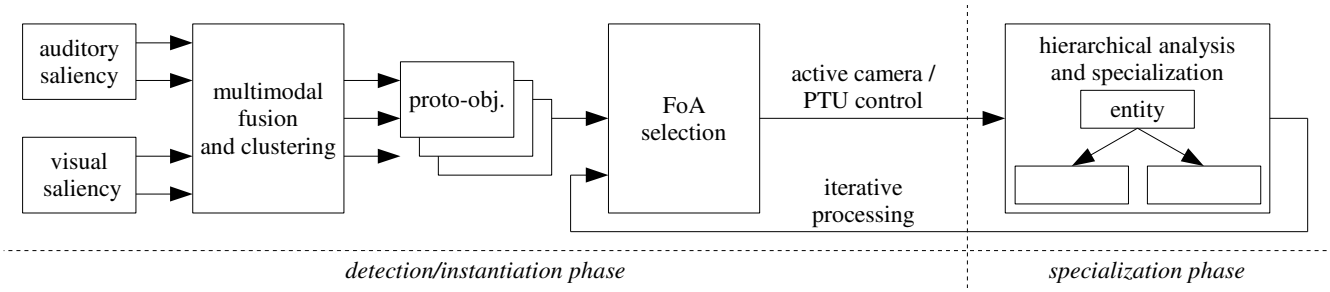

Fig. 2. Schematic overview of the system components. First, the auditory and visually salient signals are determined and encoded in a parametric Gaussian-based 3-D model (Sec. III-A and III-B, resp.). Subsequently, these auditory and visual proto-object hypotheses are clustered and the saliency is fused to enable multimodal focus of attention (FoA) selection and scene exploration (Sec. III-C). To this end, we implemented an object-based inhibition of return mechanism based on an object-based world modeling approach. By regarding each attended and focused proto-object as candidate world model entity, we perform a knowledge-driven, hierarchical refinement and specialization.

cessfully applied for speech processing [13]. However, it is computationally expensive and not perfectly suited to detect salient acoustic events that attract overt attention. For this purpose, we introduce a novel definition of auditory attention based on Bayesian surprise [16]. In contrast, in computer vision a huge amount of saliency models has been proposed in recent years (see [11], [17]). Since reviewing them is beyond the scope of this paper, we recommend reading the survey of computational visual attention in [11]. In this contribution, we apply a visual saliency model that is based on spectral whitening of the image signal (see [3], [18]–[20]), which exploits that the elimination of a signals' magnitude components accentuates narrow spatial events [21].

In coherence theory of visual cognition, proto-objects are volatile units of information that can be accessed by selective attention and subsequently validated as actual objects [22]. Accordingly, a common problem of computational visual attention models is to determine the image region around the selected focus of attention that approximates the extent of a (proto-)object at that location (see [22]). To this end, various conventional segmentation methods are applied (see, e.g., [9], [18], [19], [22]–[25]), e.g. region growing [9] and maximally stable extremal regions [3], [18], and even feedback connections in the saliency computation hierarchy have been introduced [22]. We introduce a novel method that analyzes the isophote curvature (see [26]) of the saliency map to approximate proto-object regions.

Saliency-based overt attention, i.e. the act of directing the sensors towards salient stimuli, and scene exploration for robotic applications has been addressed by several authors in recent years (see, e.g., [3], [5], [7], [9], [27]–[29]). The main difference between (covert) attention mechanisms that operate on still images, i.e. mechanisms that focus the processing of sensory information on salient stimuli, and overt attention realized on robotic platforms is that shifting the focus of attention in the latter leads to (ego-)motion, which can – at least partially – render the previously calculated saliency obsolete (see [9]). It is therefore necessary to use representations that enable storing and updating the information in the presence of ego-motion. To this end, most previous art uses ego-centric models such as an ego-sphere (e.g., [5], [28]). In contrast, we use a Cartesian 3-D reference coordinate system which is attached to a prominent point of the scene (see [30], [31]). Most similar to this model is the 2-D grid representation that was applied in [3]. Most state-of-the-art systems (e.g., [3], [7], [9]) only implement visual attention, which has considerable drawbacks in case of salient events outside the visual field of view (see [10]). Therefore, we also consider acoustic sensor data and present a parametric model suitable for audio-visual saliency fusion, which is most similar to the work presented in [5] and [10]. However, in contrast to [5] and [10], we use a parametric model without spatial quantization such as voxels [10] or ego-centric azimuth-elevation maps [5].

Fusing the information of different sensors and sensor modalities in order to analyze a scene has been addressed throughout the years in several application areas (see, e.g., [30]–[34]). In this contribution, we integrate saliency-driven, iterative scene exploration into the hierarchical, knowledge-driven audio-visual scene analysis approach presented in [30]. This approach uses a combined bottom-up and top-down strategy (see [34] and [33], respectively). Therefore, the multimodal classification and fusion at each level of the knowledge hierarchy is done bottom-up whereas the selection of suitable classification algorithms is done in a top-down fashion. The basis for this exploration and analysis is an object-based world model (see [31]), which follows the approach described in [35]. A notable feature of the chosen object analysis approach is that it facilitates the dynamic adjustment of object-specific tracking parameters, e.g. for mean shift [36], depending on the classification result, e.g. person or object specific parameters.

III. SYSTEM

In the following, we present the components of the proposed system (see Fig. 2). First, we describe how the audio-visual signals are processed (III-A and III-B). Therefore, we explicate how salient signals and corresponding proto-object regions are determined. Then, we explain how the information about salient proto-objects is fused in our 3-D parametric model (III-C). Finally, we describe how the proto-objects in the selected focus of attention are analyzed (III-C).

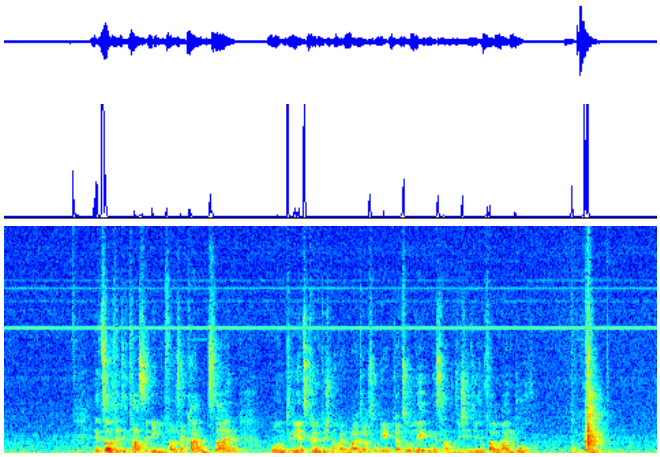

Fig. 3. An example for auditory surprise: an approximately 10 second audio sequence in which a person speaks and places a solid object on a table at the end of the sequence. Top-to-bottom: the measured audio signal (power), the corresponding auditory surprise (the range is clipped at 0.5 for purpose of illustration), and the spectrogram (logarithmic scale).

A. Audio Processing

1) Auditory Surprise: Following the well-established approach of Bayesian Surprise in computer vision [16], we introduce auditory surprise to detect acoustically salient events (see Fig. 3). Therefore, we use the short-time Fourier transform (STFT) to calculate the spectrogram $G(t,\omega) = |F(t,\omega)|^2 = |\mathrm{STFT}(t,\omega)|^2$ of the windowed audio signal $a(t)$, where t and ω denote the discrete time and frequency, respectively. In the Bayesian framework, probabilities correspond to subjective degrees of beliefs in models which are updated according to Bayes rule as data is acquired. At each time step t, we incorporate the new data $G(t,\omega)$ to update the prior probability distribution $P^{\omega}_{\mathrm{prior}} = P(g|G(t-1,\omega),\ldots,G(t-N,\omega))$ of each frequency and obtain the posterior distribution $P^{\omega}_{\mathrm{post}} = P(g|G(t,\omega),G(t-1,\omega),\ldots,G(t-N,\omega))$, where $N \in \{1,\ldots,\infty\}$ allows additional control of the time behavior by limiting the history to $N \neq \infty$ elements if wanted. Using the Gaussian distributions as model, we can calculate the surprise $S_A(t,\omega)$ for each frequency

$$S_{\mathrm{A}}(t,\omega) = D_{\mathrm{KL}}(P^{\omega}_{\mathrm{post}}||P^{\omega}_{\mathrm{prior}}) = \int_{-\infty}^{\infty} P^{\omega}_{\mathrm{post}} \log \frac{P^{\omega}_{\mathrm{post}}}{P^{\omega}_{\mathrm{prior}}} dg \quad (1)$$

$$= \frac{1}{2}[\log \frac{|\Sigma^{\omega}_{\mathrm{prior}}|}{|\Sigma^{\omega}_{\mathrm{post}}|} + \mathrm{Tr}\left[\Sigma^{\omega^{-1}}_{\mathrm{prior}}\Sigma^{\omega}_{\mathrm{post}}\right] - I_{\mathrm{D}} + \quad (2)$$

$$(\mu^{\omega}_{\mathrm{post}} - \mu^{\omega}_{\mathrm{prior}})^T \Sigma^{\omega^{-1}}_{\mathrm{prior}} (\mu^{\omega}_{\mathrm{post}} - \mu^{\omega}_{\mathrm{prior}})] \quad ,$$

where D_{KL} is the Kullback Leibler Divergence and Eqn. 2 is obtained according to the closed form of D_{KL} for Gaussian distributions (see [37]). Accordingly, an observed spectrogram element $G(t,\omega)$ is surprising if the updated distribution $P^{\omega}_{\mathrm{post}}$, which results from incorporating $G(t,\omega)$, significantly differs from the prior distribution $P^{\omega}_{\mathrm{prior}}$. Finally, we calculate the auditory saliency $S_{\mathrm{A}}(t)$ as the mean over all frequencies

$$S_{\mathrm{A}}(t) = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} S_{\mathrm{A}}(t,\omega) \quad . \quad (3)$$

2) Localization: In order to localize acoustic events in the scene, we apply the well-known steered response power (SRP) with phase transform (PHAT) sound source localization [38]. The SRP-PHAT algorithm uses the inter-microphone time difference of arrival (TDOA) of sound signals, which is caused by the different distances the sound has to travel to reach each microphone, to estimate the location of the sound source. To this end, the following inter-microphone signal correlation function is used to determine TDOAs τ of prominent signals at time t

$$R_{ij}(t,\tau) = \int_{-\infty}^{\infty} \psi^{\mathrm{PHAT}}_{ij}(t,\omega) F'_i(t,\omega) F'_j(t,\omega)^* e^{j\omega\tau} d\omega \quad , \quad (4)$$

where F'_i and F'_j are the STFTs of the audio signal at microphone i and j, respectively. The PHAT specific weighting function $\psi^{\mathrm{PHAT}}_{ij}(t,\omega) = |F'_i(t,\omega)F'_j(t,\omega)^*|^{-1}$ can be regarded as a whitening filter and is supposed to decrease the influence of noise and reverberations. Subsequently, we can use the estimated TDOAs to calculate the corresponding positions in the environment.

Please note that the STFT F' of the sound source localization uses a lower temporal-resolution than the STFT F of the salient event detection. This is due to the fact that we require real-time performance and, on the one hand, want to detect short-timed salient events while, on the other hand, require sufficiently large temporal windows for robust correlations. Therefore, the window length of the localization is a multiple of the salient event detections' window length. Accordingly, we have to aggregate the saliency of all detection windows that are located within the localization window. We use the maximum as aggregation function, because we want to react on short-timed salient events, instead of suppressing them.

3) Parametrization: Since the sound source localization tends to noisy detections, we perform spatio-temporal clustering to remove outliers and improve the accuracy of the localization. Accordingly, we can use the mean of each cluster o as location estimate μ_o and calculate the corresponding co-variance matrix Σ_o. Consequently, each detected acoustically salient (proto-)object hypothesis o is described by its saliency s_o, the estimated location μ_o, and the co-variance matrix Σ_o, which encodes the spatial uncertainty.

B. Video Processing

1) Visual Saliency: To calculate the bottom-up visual saliency, we use a phase-based approach that was inspired by [19] and [20]. However, since the influence of the spectral residual is negligible in many situations when compared to pure spectral whitening [20], we apply spectral whitening to the image $I(t)$ to obtain the saliency map

$$S_V(t) = g * \mathrm{FT}^{-1}\left\{e^{\mathrm{i}\Phi(\mathrm{FT}\{I(t)\})}\right\} \quad (5)$$

with the Fourier transform FT and an optional smoothing filter g. Please note that the image can either be given in a quaternion-based multi-channel representation [20] or as single-channel, e.g. gray-level, image [19].

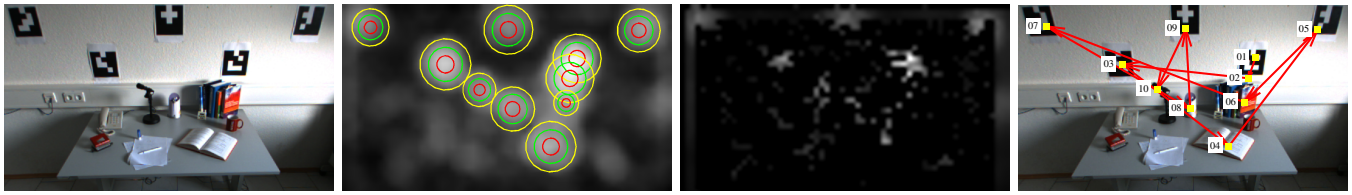

Fig. 4. Left-to-right: an exemplary scene image, the saliency map, the accumulator, and the first 10 salient proto-object regions that are selected by the location-based inhibition of return (see Sec. III-B). The estimated Gaussian weight descriptors are depicted as overlay on the saliency map (illustrated as circles with center μ_o and radii $r \in \{\sigma_o, 2\sigma_o, 3\sigma_o\}$ in {red, green, yellow}, respectively. Please note that the value range of the saliency map and the accumulator is attenuated for purpose of illustration.

2) Proto-Object Regions: In order to estimate the proto-object regions, we analyze the isophote curvature (see [26]) of the saliency map (see Fig. 4). Here, isophotes represent (closed) curves of constant saliency within the saliency map. Assuming a roughly (semi-)circular structure of salient peaks, we can then determine the center of each salient peak as well as the corresponding pixels, i.e. defined as pixels whose gradients point towards the peak center. This way, we are able to efficiently extract salient regions, even in the presence of highly varying spatial extent and value range, partially overlapping peaks and noise. To this end, we analyze the local isophote curvature κ of the saliency map $S_V(x,y;t)$

$$\kappa = -\frac{S_{cc}}{S_g} = -\frac{S_y^2 S_{xx} - 2S_x S_y S_{xy} + S_x^2 S_{yy}}{(S_x^2 + S_y^2)^{3/2}} \quad , \qquad (6)$$

where S_{cc} is the second derivative in the direction perpendicular to the gradient and S_g is the derivative in gradient direction. Accordingly, S_x, S_y and S_{xx}, S_{xy}, S_{yy} are the first and second derivatives in x and y direction, respectively. Exploiting that the local curvature is reciprocal to the (hypothetical) radius r of the circle that generated the saliency isoline of each pixel, i.e. $r(x,y) = 1/\kappa(x,y)$, we can estimate roughly the location of each peak's center. Therefore, respecting the isophote orientation and direction, we calculate the displacement vectors $D = (D_x, D_y)$ with

$$D_x = \frac{S_x\left(S_x^2 + S_y^2\right)}{S_{cc}} \quad \text{and} \quad D_y = \frac{S_y\left(S_x^2 + S_y^2\right)}{S_{cc}} \qquad (7)$$

and the resulting hypothetical peak centers $C = (C_x, C_y)$ with

$$C_x = P_x - D_x \quad \text{and} \quad C_y = P_y - D_y \quad , \qquad (8)$$

where the matrices P_x and P_y represent the pixel abscissae and ordinates, i.e. the pixel (x,y) coordinates, respectively.

Thus, we can calculate a saliency accumulator map A_S in which each pixel votes for its corresponding center. The most salient regions, i.e. corresponding to the extents of the proto-objects in the image (see, e.g., [19]), in the saliency map can then be determined by selecting the pixels of the accumulator cells with the highest voting score. By choosing different weighting schemes for the voting, we are able to implement divers methods for assessing the saliency of each region. In our implementation, we use the saliency as weight and normalize each accumulator cell by division by the number of pixels that voted for the pixel. However, due to noise and quantization effects, we additionally select pixels that voted for accumulator cells within a certain radius. Unfortunately, the initially selected pixels of our (proto-)object regions are contaminated with outliers caused by (background) noise. Therefore, we apply convex peeling (see [39]) to remove scattered outliers and eliminate regions whose percentage of detected outliers is too high.

To extract all salient proto-object regions that attract the attention, we apply a location-based inhibition of return (see [40]) mechanism on the saliency map (see, e.g., [5], [8], [15], [18]). To this end, we use the accumulator to select the most salient proto-object region and inhibit all pixels within the estimated outline by setting their saliency to zero. This process is repeated until no further prominent salient peaks are present in the map.

3) Parametrization: For each extracted salient proto-object region $o \in O(S_V(t))$, we derive a parametric description by fitting a Gaussian weight function f_o. We assume that the Gaussian weight function encodes two distinct aspects of information: the saliency s_o as well as the (uncertain) spatial location and extent of the object μ_o and Σ_o, respectively. Consequently, we decompose the Gaussian weight function:

$$f_o^G(x) = \frac{s_o}{\sqrt{(2\pi)^D \det(\Sigma_o)}} \exp\left(-\frac{1}{2}(x-\mu_o)^T \Sigma_o^{-1}(x-\mu_o)\right) \qquad (9)$$

with $D = 2$. Exploiting the stereo setup, we can estimate the depth and project the 2-D model into 3-D. This way, we obtain a 3-D model for each visually salient proto-object region that follows the representation of the detected acoustically salient events, see Sec. III-A.3. However, we have to make assumptions about the shape, because the spatial extent of the object in direction of the optical axis can not be observed. Thus, we simplify the model and assume a spherical model in 3-D and, accordingly, a circular outline in 2-D, i.e. $\Sigma_o = I_D \sigma_o$ with the unit matrix I_D.

C. Exploration and Analysis

1) Saliency Fusion: After the detection and parametrization of salient auditory and visual signals, we have a set of auditory and visual proto-object hypotheses represented in a Gaussian notation at each point in time t. To reduce the influence of noise as well as to enable multimodal saliency fusion, we perform a cross-modal spatio-temporal mean shift clustering [36] of the auditory and visual Gaussian representatives. Accordingly, we obtain a set of multimodal

clusters $C(t)$, each of which can be interpreted as a (saliency-weighted) Gaussian mixture model. Consequently, we estimate the center of each cluster $c \in C(t)$ according to

$$\mu_c = \mathbb{E}[c] = \sum_{c_i \in c} w_{c_i} \mu_{c_i} \quad \text{with} \quad w_{c_i} = \frac{s_{c_i}}{\sum_{c_i \in c} s_{c_i}} \tag{10}$$

and $(\mu_{c_i}, \sigma_{c_i}, s_{c_i}) = c_i \in c$ (see Sec. III-A.3 and III-B.3). Subsequently, we can estimate the auditory s_c^{A} and visual s_c^{V} saliency of each cluster

$$s_c^{\mathrm{A}} = \sum_{c_i \in c} 1_{\mathrm{A}}(c_i) w_{c_i}^{\mathrm{A}} f_{c_i}^{\mathrm{G}}(\mu_c) \text{ and } s_c^{\mathrm{V}} = \sum_{c_i \in c} 1_{\mathrm{V}}(c_i) w_{c_i}^{\mathrm{V}} f_{c_i}^{\mathrm{G}}(\mu_c) \tag{11}$$

with

$$w_{c_i}^{\mathrm{A}} = \frac{1_{\mathrm{A}}(c_i) s_{c_i}}{\sum_{c_i \in c} 1_{\mathrm{A}}(c_i) s_{c_i}} \text{ and } w_{c_i}^{\mathrm{V}} = \frac{1_{\mathrm{V}}(c_i) s_{c_i}}{\sum_{c_i \in c} 1_{\mathrm{V}}(c_i) s_{c_i}} \quad , \tag{12}$$

where 1_{A} and 1_{V} are indicator functions that encode whether an element is an auditory or visual proto-object hypothesis, respectively. Finally, we combine the audio and visual saliency to obtain the cross-modal saliency s_c. We follow the results reported in [41] and use a linear combination for cross-modal integration, which has been shown to be a good model for human overt attention and is optimal according to information-theoretical criteria [41]. However, other biologically plausible combination schemes (see [41]) can be realized easily. Since the cross-modal saliency depends on the range of the saliency in each modality, we have to normalize the range of the auditory and visual saliency to obtain value ranges that are suitable for multimodal combination. Hence, we empirically determined suitable parameters for truncation and normalization of the value ranges.

2) Object-based Inhibition of Return: In order to iteratively attend and analyze the objects present in the scene, we use the detected salient (multimodal) proto-objects to realize an object-based inhibition of return mechanism. Therefore, at each decision cycle, the most salient proto-object cluster that is not related with an already attended and analyzed proto-object gains the overt focus of attention. To this end, we use a Euclidean distance metric to relate the proto-object clusters with previously inspected object entities that are managed in an object-based world model (see [31]). However, there is an exception to this bottom-up selection rule: higher-level functions can queue specific objects to regain the overt focus of attention, which has a higher priority and thus allows an integrated top-down control of overt attention.

3) Hierarchical, Knowledge-driven Proto-object Analysis: After the sensors have been aligned with respect to the proto-object in the current overt focus of attention, the foveal cameras (see Fig. 1) are used to inspect the object. Therefore, we extend the multimodal knowledge-driven scene analysis and object-based world modeling system presented in [30] and [31], to comply with the iterative, saliency-driven focus of attention and exploration mechanism. Most importantly, we replaced the detection and instantiation phase by regarding proto-objects as candidates for world model entities. The attended proto-object region is instantiated as entity and subsequently hierarchically specialized and refined in a knowledge-driven model (see [30], [31]). The analysis of each proto-object is finished, if no further refinement is possible, which marks the end of the decision cycle and initiates the next shift of attention. Within this framework, every entity is tracked which is an important feature of object-based inhibition of return.

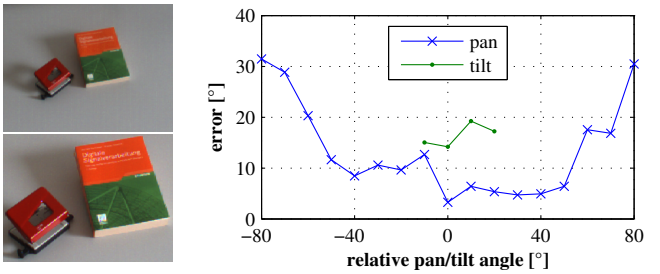

Fig. 5. Left-to-right: exemplary image of an object in the coarse (top-left) and fine (bottom-left) view, respectively. Mean sound source localization error (in °) depending on the pan-tilt-orientation of the sensor setup (right).

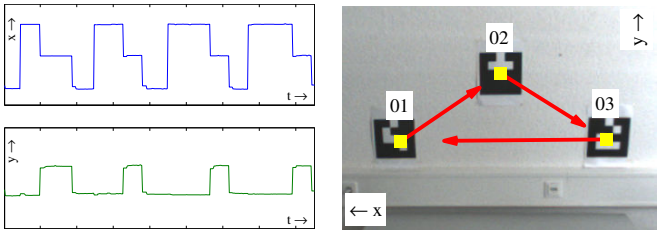

Fig. 6. A short temporal section of the attended x-y position (left) in the cyclic focus of attention shift experiment (see [5]). The positions correspond to the calibration marker locations (right) that lie on the same x-z plane.

IV. EXPERIMENTAL EVALUATION

A. Setup

The sensor setup that was used for the evaluation of the presented system is shown in Fig. 1. The wide angle and foveal cameras have a focal length of 6 mm and 3.5 mm, respectively. The stereo baseline separation between each camera pair is 90 mm. The camera sensors provide a resolution of 640×480 px at a frame rate of 30 Hz. In the evaluation only the front and side omnidirectional microphones are used (see Fig. 1). The distance between the side microphones is approximately 190 mm and the vertical distance between the front microphones is approximately 55 mm. The pan-tilt unit is mounted on a tripod such that the cameras are roughly on eye height of an averagely tall human in order to reflect a humanoid view of the scene.

B. Procedure and Measures

First of all, to demonstrate that overt attention is beneficial and justifies the required resources, we assess the impact of active sensor alignment on the perception quality (Sec. IV-C.1). While the improvement of the image data quality of objects in the focused foveal view compared to the coarse view is easily visible (see Fig. 5), the impact on the acoustic perception depends on several factors, most importantly the sensor setup. Consequently, as reference we evaluate the acoustic localization error with respect to the

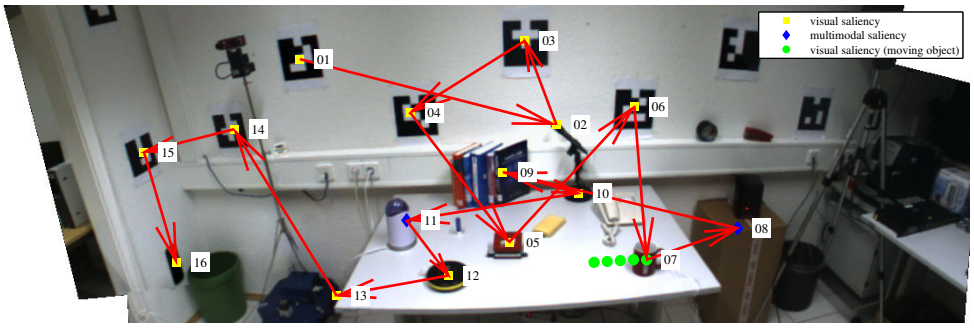

Fig. 7. An example of multimodal scene exploration: The focus of attention is shifted according to the numbers in the stitched image of the scene (only the first 15 shifts are shown). The yellow squares mark objects that attracted the focus solely due to their visual saliency whereas the blue squares (at 08 and 11) mark audio-visually caused shifts of attention. Furthermore, the green dotted lines (at 07) roughly indicate the trajectory of the moved object.

pan-tilt orientation of our sensor setup relative to sound sources, e.g. household devices and speaking persons. For this purpose, the sound sources were placed at fixed locations and the localization was performed with pan-tilt orientations of $\{-80°,\ldots,80°\}\times\{-30°,\ldots,0°\}$ in $10°$ steps (see Fig. 5). We only consider the angular error, because in our experience the distance error is too dependent on the algorithm parameters, implementation, and sampling rate.

We perform a couple of experiments to evaluate the behavior of the proposed system, because a quantitative, comparative method to evaluate the performance of an overt attention system does not exist (see [9], [17]). In order to obtain a reliable impression of the performance of our system, we repeated every experiment multiple times with varying environmental conditions such as, e.g., lighting, number of objects, distracting clutter, and timing of events. Inspired by the evaluation procedures in [5] and [9], we investigate and discuss the performance of saliency-driven visual (Sec. IV-C.2 and IV-C.3) as well as multimodal scene exploration (Sec. IV-C.4 and IV-C.5).

C. Results and Discussion

1) Audio-Visual Perception: As can be seen in the error curve depicted in Fig. 5, the angular localization error is minimal if the head faces the target object directly. This can be explained by the hardware setup in which the microphones are nearly arranged on a meridional plane. Interestingly, the curve shows a non-monotonic error progression, which is mainly caused by the hardware that interferes with the acoustic characteristic and perception, e.g. the cameras heavily influence the frontal microphones (see Fig. 1). Additionally, in Fig. 5 we show an example of the coarse and fine, i.e. foveal, view of a focused object to illustrate the improved visual perception, i.e. increased level of detail.

2) Visual Exploration I – FoA Shift: In style of the experiment described in [5, Sec. V–B], we mounted three salient calibration marks on the walls of the office environment and removed other distracting stimuli (see Fig. 6). In this experiment, we benefit from an object-specific lifetime that can be assigned to analyzed objects in our world model. Each object-specific lifetime is continuously evaluated and updated by, e.g., taking the visibility into account. Thus, if an object has expired and is perceived as salient, it can regain the focus of attention. Driven by the implemented inhibition of return mechanism, the three salient marks are explored by shifting the overt attention from one mark to the next most salient mark that is not inhibited. As expected, the achieved behavior corresponds to the cyclic behavior described in [5]. Each attended mark is focused by controlling the pan-tilt-servos and the resulting trajectory is illustrated in Fig. 6.

3) Visual Exploration II – Object-based IoR: Once an object has been analyzed, it is being tracked and inhibited – as long as the object has not been marked for re-focusing by higher-level processes – from gaining the overt focus of attention. In order to test the object-based inhibition of return mechanism, we perform experiments with moving objects in the scene. For this purpose, we place movable objects in the scene, start the exploration, and move objects after they have been analyzed. As expected, smoothly moving objects do not attract the focus, although they are moved to locations that have not been salient before. Naturally, this behavior even remains when motion is integrated as an additional saliency cue. Interestingly, objects that abruptly change their expected motion pattern attract the focus of attention again, because the tracking fails. Although this could be seen as a technical deficit, this behavior is desired for an attention-based system and biologically motivated (see [14]).

4) Multimodal Exploration I – FoA Shift: Following the experimental procedure in [5, Sec. V–C], we examine the behavior in scenes with acoustic stimuli. Therefore, we extend the scenario of the previous experiment (Sec. IV-C.3) and add a single visible sound source, e.g. a blender or a talking person. Our system explores the environment based on visual saliency until the acoustic stimulus begins and the sound source directly gains the focus of attention.

5) Multimodal Exploration II – Scene: Finally, we unite the previously isolated experiments and assess the performance on more complex scenes with several objects, object motion, and auditory stimuli (please see Fig. 7 for an exemplary scene). The system is capable of handling these situations according to our expectations. Most importantly, objects that are auditory and visually salient tend to attract the saliency even if they are not the most salient point in each modality. Furthermore, salient sound sources outside

the visual field of view compete with visually salient stimuli and both are able to attract the overt focus of attention due to the normalized value ranges (see Sec. III-C.1).

V. CONCLUSION AND FUTURE WORK

We presented and evaluated a multimodal attention system for object-based audio-visual scene exploration and analysis. Our model is based on bottom-up attention and the use of proto-object regions as descriptors of salient (proto-)object hypotheses in the scene. For this purpose, we introduced a novel isophote-based saliency map segmentation as well as a surprise-based definition of acoustically salient events. By combining the proto-object regions with information about already attended objects in a parametric 3-D model, we are able to realize audio-visual saliency fusion and seamlessly enable object-based inhibition of return. The implemented inhibition of return mechanism allows to iteratively explore and audio-visually analyze objects in the scene, which allows to improve the perception through active sensor alignment and limit the amount of required processing resources at each point in time. However, several aspects remain as future work. Most importantly, we plan to further investigate integrated computational mechanisms for location- and object-based inhibition of return.

REFERENCES

[1] T. Asfour, K. Welke, P. Azad, *et al.*, "The Karlsruhe Humanoid Head," in *Humanoids*, 2008.

[2] N. Butko, L. Zhang, *et al.*, "Visual saliency model for robot cameras," in *Proc. Int. Conf. Robot. Autom.*, 2008.

[3] D. Meger, P.-E. Forssén, K. Lai, *et al.*, "Curious George: An attentive semantic robot," in *IROS Workshop: From sensors to human spatial concepts*, 2007.

[4] F. Orabona, G. Metta, and G. Sandini, "Object-based visual attention: a model for a behaving robot," in *CVPR Workshop: Attention and Performance in Computational Vision*, 2005.

[5] J. Ruesch, M. Lopes, A. Bernardino, *et al.*, "Multimodal saliency-based bottom-up attention: A framework for the humanoid robot iCub," in *Proc. Int. Conf. Robot. Autom.*, 2008.

[6] D. Figueira, M. Lopes, R. Ventura, and J. Ruesch, "From pixels to objects: Enabling a spatial model for humanoid social robots," in *Proc. Int. Conf. Robot. Autom.*, 2009.

[7] T. Xu, N. Chenkov, K. Kühnlenz, and M. Buss, "Autonomous switching of top-down and bottom-up attention selection for vision guided mobile robots," in *Proc. Int. Conf. Intell. Robots Syst.*, 2009.

[8] B. Schauerte and G. A. Fink, "Focusing computational visual attention in multi-modal human-robot interaction," in *Proc. Int. Conf. Multimodal Interfaces*, 2010.

[9] M. Begum, F. Karray, G. K. I. Mann, and R. G. Gosine, "A probabilistic model of overt visual attention for cognitive robots," *IEEE Trans. Syst., Man, Cybern. B*, vol. 40, pp. 1305–1318, 2010.

[10] B. Schauerte, J. Richarz, T. Plötz, *et al.*, "Multi-modal and multi-camera attention in smart environments," in *Proc. Int. Conf. Multimodal Interfaces*, 2009.

[11] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundation: A survey," *ACM Trans. Applied Perception*, vol. 7, 2010.

[12] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: an auditory saliency map," *Current Biology*, vol. 15, pp. 1943–1947, 2005.

[13] O. Kalinli and S. Narayanan, "Prominence detection using auditory attention cues and task-dependent high level information," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 17, pp. 1009–1024, 2009.

[14] M. Heracles, U. Körner, T. Michalke, *et al.*, "A dynamic attention system that reorients to unexpected motion in real-world traffic environments," in *Proc. Int. Conf. Intell. Robots Syst.*, 2009.

[15] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, pp. 1254–1259, 1998.

[16] L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems*, 2006.

[17] F. Shic and B. Scassellati, "A behavioral analysis of computational models of visual attention," *Int. J. Comp. Vis.*, vol. 73, pp. 159–177, 2007.

[18] B. Schauerte, J. Richarz, and G. A. Fink, "Saliency-based identification and recognition of pointed-at objects," in *Proc. Int. Conf. Intell. Robots Syst.*, 2010.

[19] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2007.

[20] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2008.

[21] A. Oppenheim and J. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, pp. 529–541, 1981.

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

[23] Y. Yu, G. K. I. Mann, and R. G. Gosine, "An object-based visual attention model for robotic applications," *IEEE Trans. Syst., Man, Cybern. B*, vol. 40, pp. 1398–1412, 2010.

[24] J. Schmuedderich, H. Brandl, B. Bolder, *et al.*, "Organizing multimodal perception for autonomous learning and interactive systems," in *Humanoids*, 2008.

[25] F. Orabona, G. Metta, and G. Sandini, "Attention in cognitive systems. theories and systems from an interdisciplinary viewpoint," L. Paletta and E. Rome, Eds., 2008, ch. A Proto-object Based Visual Attention Model, pp. 198–215.

[26] J. Lichtenauer, E. Hendriks, and M. Reinders, "Isophote properties as features for object detection," in *Proc. Int. Conf. Comp. Vis. Pat. Rec.*, 2005.

[27] S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal, "Overt visual attention for a humanoid robot," in *Proc. Int. Conf. Intell. Robots Syst.*, 2001.

[28] K. A. Fleming, R. A. Peters II, and R. E. Bodenheimer, "Image mapping and visual attention on a sensory ego-sphere," in *Proc. Int. Conf. Intell. Robots Syst.*, 2006.

[29] A. Dankers, N. Barnes, and A. Zelinsky, "A reactive vision system: Active-dynamic saliency," in *Proc. Int. Conf. Vis. Syst.*, 2007.

[30] T. Machmer, A. Swerdlow, B. Kühn, and K. Kroschel, "Hierarchical, knowledge-oriented opto-acoustic scene analysis for humanoid robots and man-machine interaction," in *Proc. Int. Conf. Robot. Autom.*, 2010.

[31] B. Kühn, A. Belkin, A. Swerdlow, *et al.*, "Knowledge-driven opto-acoustic scene analysis based on an object-oriented world modelling approach for humanoid robots," in *Proc. 41st Int. Symp. Robotics and 6th German Conf. Robotics (ISR/ROBOTIK)*, 2010.

[32] I. Essa, "Ubiquitous sensing for smart and aware environments," *IEEE Personal Communications*, vol. 7, pp. 47–49, 2000.

[33] J. Holsopple and S. Yang, "Designing a data fusion system using a top-down approach," in *Proc. Int. Conf. Military Comm.*, 2009.

[34] D. Hall and J. Linas, *Handbook of Multisensor Data Fusion: Theory and Practice*. CRC Press, 2008.

[35] M. Baum, I. Gheta, A. Belkin, *et al.*, "Data association in a world model for autonomous systems," in *Proc. Int. Conf. Multisensor Fusion and Integration for Intelligent Systems*, 2010.

[36] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 603–619, 2002.

[37] J. Hershey and P. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *Int. Conf. Acoustics, Speech and Signal Processing*, 2007.

[38] T. Machmer, J. Moragues, A. Swerdlow, *et al.*, "Robust impulsive sound source localization by means of an energy detector for temporal alignment and pre-classification," in *Proc. Europ. Sig. Proc. Conf.*, 2009.

[39] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, pp. 85–126, 2004.

[40] W.-L. Chou and S.-L. Yeh, "Location- and object-based inhibition of return are affected by different kinds of working memory," *The Quarterly Journal of Experimental Psychology*, vol. 61, pp. 1761–1768, 2008.

[41] S. Onat, K. Libertus, and P. König, "Integrating audiovisual information for the control of overt attention," *Journal of Vision*, vol. 7, 2007.