Multimodal Saliency-based Attention: A Lazy Robot's Approach

Benjamin Kühn, Boris Schauerte, Kristian Kroschel and Rainer Stiefelhagen

Abstract— **We extend our work on an integrated object-based system for saliency-driven overt attention and knowledge-driven object analysis. We present how we can reduce the amount of necessary head movement during scene analysis while still focusing all salient proto-objects in an order that strongly favors proto-objects with a higher saliency. Furthermore, we integrated motion saliency and as a consequence adaptive predictive gaze control to allow for efficient gazing behavior on the ARMAR-III robot head. To evaluate our approach, we first collected a new data set that incorporates two robotic platforms, three scenarios, and different scene complexities. Second, we introduce measures for the effectiveness of active overt attention mechanisms in terms of saliency cumulation and required head motion. This way, we are able to objectively demonstrate the effectiveness of the proposed multicriterial focus of attention selection.**

Index Terms— **active perception, saliency-based overt attention, and scene exploration**

I. INTRODUCTION

Attention describes the cognitive process responsible for focusing the processing of sensory information onto potentially relevant and thus salient stimuli. Specifically, covert attention refers to the process of focusing the perception on salient stimuli to facilitate real-time processing despite limited computing capacities, while overt attention refers to the act of directing the sense organs towards selected salient stimuli in order to optimize the perception quality. Since both aspects are crucial for autonomous robots in complex, natural scenes, attention has attracted an increasing interest in the field of robotics, mainly for saliency-driven scene exploration and real-time sensor processing (see [1]). However, pure saliency-driven determination of the order in which the salient stimuli are attended often leads to a high amount of ego-motion and erratic motion patterns. Consequently, this leads to high energy costs, wear-and-tear, longer exploration times, and oftentimes artificially looking head motion patterns. Thus, it is necessary to introduce a certain amount robot "laziness" to reduce the energy costs and wear-and-tear, and that may also lead to more natural-like head motion sequences.

We extend our previous work described in [2], which introduced a multimodal attention system for object-based audio-visual scene exploration and analysis. In this contribution, we integrated adaptive predictive gaze control [3] to control a kinematically redundant robot head, i.e. the ARMAR-III humanoid head, generate natural-looking head motion patterns, and take advantage of a new visual saliency algorithm that now also integrates the important influence of motion saliency (see [4]). Most importantly, we present a "lazy" approach of saliency-based scene exploration of newly entered rooms that reduces the amount of necessary (head) ego-motion while it strongly favors to attend the most salient proto-objects as soon as possible. To evaluate our approach, we created a novel data set that consists of 60 recordings (2 sensor setups, 3 scenarios, and $2\times(15+10+5)$ scenes; 36 GigaByte of data). Additionally, we introduce two evaluation measures, i.e. the normalized cummulated saliency and normalized cummulated joint angle distances, as objective measures for specific aspects of active overt attention. This way we are able to show that we can attend to the proto-objects in a scene in an order that still favors proto-objects with a higher saliency while drastically reducing the amount of head servo motion.

II. RELATED WORK

Due to the practical relevance of attention for autonomous robots in complex natural environments, computational attention models have attracted an increasing interest in the field of robotics during the last decade (e. g., [2], [5]–[11]). Accordingly, there exists a wide range of different attention models and the selection of the saliency definition heavily influences which signal components attract the attention.

Since reviewing the existing literature on visual saliency models is beyond the scope of this paper, we have to refer interested readers to recent surveys of computational visual attention models (e.g., [1] and [12]). We use a visual saliency model based on the signature of the DCT transformed quaternion-image [13], [14], which provides state-of-the-art performance in human gaze point prediction and is extremely run-time efficient to compute. As input for the QDCT, we use a quaternion image composed of intensity, color opponents, and motion (see [15]). In coherence theory of visual cognition, proto-objects refer to volatile units of information that can be accessed by selective attention and later validated as actual object [16]. However, it is difficult to determine the image region that approximates the extent of a (proto-)object at the attended location (see [16]). To this end, we analyze the isophote curvature (see [2]) of

This work is a joint collaboration between the Fraunhofer Institute of Optronics, System Technologies and Image Exploitation and Karlsruhe Institute of Technology and was supported by the German Research Foundation within the Collaborative Research Program SFB 588 "Humanoide Roboter".

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

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

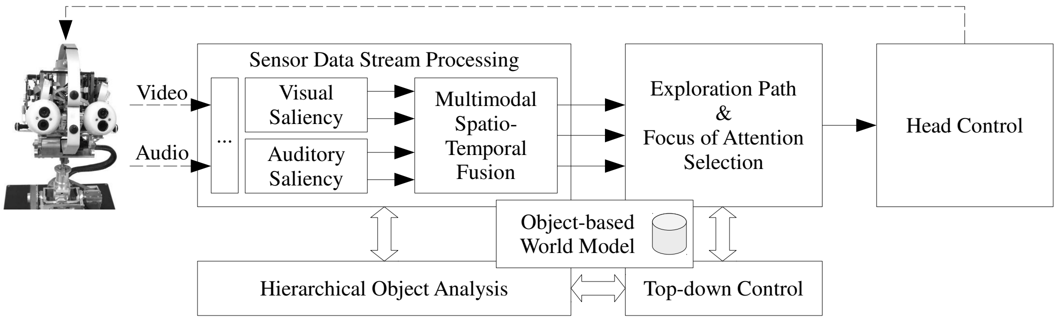
Video
Audio
Sensor Data Stream Processing
...
Visual Saliency
Auditory Saliency
Multimodal Spatio-Temporal Fusion
Exploration Path & Focus of Attention Selection
Head Control
Object-based World Model
Hierarchical Object Analysis
Top-down Control

Fig. 1: Schematic system overview: The visually and auditory salient signal components are determined (Sec. III-B.2 and III-B.3) and represented in a 3-D Gaussian model (Sec. III-B.1). Afterwards, these salient auditory and visual proto-object hypotheses are clustered and fused (Sec. III-B.4) in order to enable multimodal focus of attention selection and scene exploration (Sec. III-C).

the saliency map in order to determine salient regions and roughly approximate the extent of the proto-object region [2].

In contrast to visual saliency, few models to determine salient auditory signals have been proposed that are suitable for robotics applications (e.g., [2], [17], [18]). To this end, we introduced auditory surprise in [2], which is based on the accumulated Bayesian surprise [19] of all frequencies. This model is fast to compute and – according to our experience – it reliably and robustly detects auditory salient events.

Saliency-based overt attention, i.e. directing the robot sensors towards salient stimuli, and saliency-based scene exploration has been addressed by several authors (e. g. [2], [6], [7], [11], [20]). However, most state-of-the-art systems only consider visual attention (e. g. [6], [11]), which makes it impossible to react on salient events outside the visual field of view (see [21]). In contrast, we consider auditory data and use a parametric model suitable for audio-visual saliency fusion [2], which is most closely related to the approaches presented in [7] and [21]. When realizing overt attention it is important to consider that each shift of the overt focus of attention leads to ego-motion, which partially renders the previously calculated information obsolete (see [11]). Thus, it is necessary to enable storing and updating the saliency as well as object information in the presence of ego-motion. To this end, we use a Cartesian 3-D reference coordinate system which is attached to a prominent point of the scene (see [2]; see [22], [23]). This is most similar to the 2-D grid representation that was applied in [6] and differs from most related work that typically uses ego-centric representations such as, most importantly, an ego-sphere (e. g. [7], [20]).

In many publications on overt attention (see, e.g., [2], [5], [7], [24]), the order in which the objects in the scene are attended is only based on the saliency. To this end, in each focus of attention selection step, the location with the highest saliency gains the focus of attention and an inhibition of return mechanism ensures that salient regions not visited twice. However, in many situations, this leads to extensive head motion and thus on real robotic platforms a high energy consumption, wear-and-tear, and a longer time to attend all objects due motor/servo speed limitations. Furthermore, the sometimes erratic head motions oftentimes does not resemble human behavior. Consequently, it is necessary to take other aspects into consideration when deciding which location to attend next. For example, in [6], [8], [25] additional top-down parameters introduced (e.g., task specific or frontier-based exploration parameters). Most related to our work, in [26] and [27] rating functions for object search are used, such as, e.g., a motion cost function for sensor alignment.

Information acquisition and fusion for scene analysis using different sensor setups and modalities has been addressed throughout the years for various application areas (e. g. [21], [23], [28]–[31]; see [32]). In this paper, we extend our previously proposed concept of multimodal, saliency-based, iterative scene exploration and analysis [2] using a multi-objective exploration path approach that determines the object analysis order during the exploration process. To this end, similar to human behavior (see [33]), we also consider saliency information that is currently not available in the field of view of the robot, but in the short-time memory of our world model. In order to analyze attended regions, we use a hierarchical, knowledge-driven analysis strategy (see [28]; [22], [23]), which combines a bottom-up and top-down strategy (see [32] and [30], respectively) and is used in combination with an object-based world model that follows the approach described in [34]. An integrated tracking of world model entities makes it possible to detect changes in their saliency as well as to distinguish novel proto-objects from already attended objects, which both can influence the object analysis order as well as duration and is inspired by reported human behavior (see [33]).

III. SYSTEM

A. System Architecture

1) System Organization and World Model: Our system architecture (see Fig. 1) has a modular structure, which enables a flexible, exchangeable, and extendable organization of the data flow and processing (see [22]). In order to organize the data, we use an object-based world model that consistently manages a-priori knowledge in a static sub-model and dynamically acquired information in a dynamic sub-model (see [23]). The a-priori knowledge contains geometric information about the environment, ontologies, and previously trained classifier models of objects, object attributes, and object classes. The dynamic model manages the perceived information such as, e.g., detected (proto-)objects and persons in the environment that are provided by the perception modules at run-time.

2) Object Analysis: The target of the active head control, i.e. overt attention, is to optimize the perception of the current (proto-)object of interest. In our system, the visual perception benefits from the higher level of detail provided by a second stereo camera pair (foveal) with a narrower field of view than the primary stereo camera pair that has a wide viewing angle (peripheral). Furthermore, the auditory perception benefits from the improved acoustic properties of the aligned microphone array (see [2]). The improved perception can subsequently be exploited by the applied object analysis algorithms. Therefore, we initialize an entity with the information of the current proto-object in the world model and start the hierarchical, knowledge-driven analysis described in [22].

3) Top-down Control: A top-down mechansim is responsible to keep the balance between exploration and analysis. This is necessary, because the complete analysis of an object can require a considerable amount of processing ressources and, most importantly, time, which is especially critical in dynamic scenes. Thus, the time available to analyze each object is dynamically limited depending on run-time constraints, the level of awareness, the type of each entity (e.g., person, object), and the number of proto-objects on the exploration path. If the number of proto-objects is very high, several objects can be marked – supported by the integrated tracking of proto-objects – for a further analysis at a later time. Furthermore, descriptions of unknown objects can be acquired and added to the world model, e.g., via text input or multimodal interaction (see [10]).

4) Head Control: The gaze direction of the robot head is characterized by a 2-dimensional vector of pointing angles, which can be expressed as function of the head kinematics. For the ARMAR-III robot head, a gaze control concept is implemented that uses differential kinematics, which allows the visual tracking of moving targets with unknown and arbitrary trajectory. Additionally, a natural human-like behavior is obtained by specifically applying all available degrees of freedom (DoF) of the robot head (see [3]).

B. Saliency Representation, Estimation and Fusion

1) Proto-Objects: Visually salient regions and auditory salient events in the perceivable space of the sensors lead to a set of auditory and/or visual proto-object hypotheses

$$\{h_1, \dots, h_N\} = \mathcal{H} = \mathcal{H}_A \cup \mathcal{H}_V , \tag{1}$$

where each hypothesis h_i is a 3-tuple consisting of saliency s_{h_i}, spatial mean $\boldsymbol{\mu}_{h_i}$ as well as spatial variance $\boldsymbol{\Sigma}_{h_i}$

$$h_i = (s_{h_i}, \boldsymbol{\mu}_{h_i}, \boldsymbol{\Sigma}_{h_i}) \in \mathcal{H} \quad . \tag{2}$$

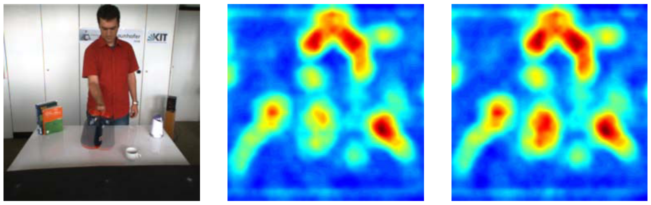

Fig. 2: Left-to-right: original scene, saliency without motion component, and saliency with motion component.

Thus, every proto-object hypothesis h_i is represented in a consistent Gaussian notation

$$f^{\mathrm{G}}_{h_i}(\boldsymbol{x}) = \frac{s_{h_i}}{\sqrt{(2\pi)^3 |\boldsymbol{\Sigma}_{h_i}|}} e^{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{h_i})^T \boldsymbol{\Sigma}^{-1}_{h_i} (\boldsymbol{x} - \boldsymbol{\mu}_{h_i})} . \tag{3}$$

Finally, we fuse all proto-object hypotheses of one salient region and/or event using multimodal spatio-temporal fusion (see III-B.4) to create a new proto-object $o = (s_o, \boldsymbol{\mu}_o, \boldsymbol{\Sigma}_o)$.

2) Visual Saliency: The visual saliency defines which image regions are interesting, because they are likely to contain objects of interest. We use the image signatures of the quaternion DCT transformed image to calculate the visual saliency [13], [14], see Fig. 2, which provides state-of-the-art performance and can be calculated in less than $0.5\,\mathrm{ms}$ per image. As quaternion image we use a combined quaternion-based representation of the image intensity, red-green and blue-yellow color opponents and an additional motion component. The latter is calculated using difference images. In order to determine the salient object regions in the visual saliency map, we analyze the local isophote curvature and estimate the peaks including their spatial extent and saliency value (see [2]). Using the stereo-camera setup, we are able to approximate the parameters for a proto-object hypothesis $h_i = (s_{h_i}, \boldsymbol{\mu}_{h_i}, \boldsymbol{\Sigma}_{h_i})$ for each salient peak.

3) Auditory Saliency: We apply the well-known approach of Bayesian Surprise [19] to audio signals in order to detect acoustically salient events (see [2]). To this end, we use the Short-Term Fourier-Transform of the audio signal and incorporate the spectrogram of the windowed audio signal over time in order to calculate the auditory surprise $S_A(t)$. Subsequently, a localization of each acoustic event is performed with the well-known steered response power with phase transform (SRP-PHAT) approach [35], which uses the inter-microphone time difference of arrival. Afterwards, we perform a spatial clustering in order to remove outliers and improve the localization accuracy. Finally, we create a proto-object hypothesis h_i for each cluster with saliency $s_{h_i} = S_A(t)$, spatial mean $\boldsymbol{\mu}_{h_i}$ and spatial covariance $\boldsymbol{\Sigma}_{h_i}$.

4) Multimodal Spatio-Temporal Fusion: To fuse the information of the acquired proto-object hypotheses, reduce the influence of noise and create proto-objects as unique representations in the world model, we perform a spatio-temporal mean shift clustering [36]. Therefore, we interpret each cluster c as a (saliency-weighted) Gaussian mixture model, that consists of auditory and/or visual proto-object hypotheses $h_j = (s_{h_j}, \boldsymbol{\mu}_{h_j}, \boldsymbol{\Sigma}_{h_j}) \in c \subseteq \mathcal{H}$. This allows us

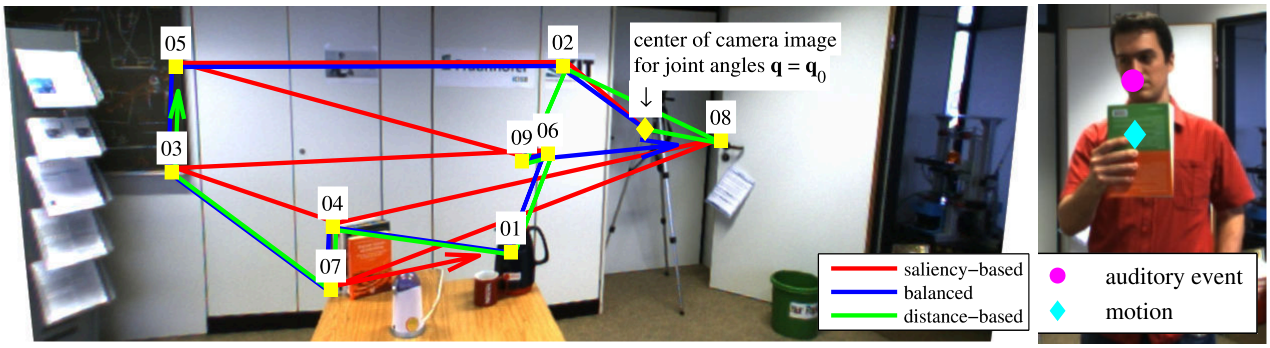

Fig. 3: Left: An example to illustrate the different focus of attention selection strategies. The attention shifts for each path are illustrated in the stitched image (only the nine most salient locations are shown and yellow squares mark the positions of the objects). Right: A person starts talking and consequently attracts the focus of attention. Thus, while the robot analyzes the proto-objects on the exploration path, an auditory salient event outside the visual field of view is detected and the gaze, i.e. overt attention, shifts towards the event location. Subsequently, the motion of the book increases the visual saliency.

to split each cluster c into an auditory ($c_A = c \cap \mathcal{H}_A$) and/or visual ($c_V = c \cap \mathcal{H}_V$) sub-cluster and estimate the saliency for each modality separately. Subsequently, we consider a linear combination[1] to integrate the audio-visual saliency

$$s_o = \frac{1}{2}\Big(\sum_{h_k \in c_A} w^A_{h_k} f^G_{h_k}(\boldsymbol{\mu}_o) + \sum_{h_l \in c_V} w^V_{h_l} f^G_{h_l}(\boldsymbol{\mu}_o) \Big) \quad , \quad (4)$$

using the modality specific weights $w^A_{h_j}$ and $w^V_{h_j}$ (analogous to Eq. 5). Consequently, we use the spatial mean of every proto-object hypotheses of the cluster to estimate the position

$$\boldsymbol{\mu}_o = \mathbb{E}[c] = \sum_{h_j \in c} \boldsymbol{\mu}_{h_j} w_{h_j} \quad \text{with } w_{h_j} = \frac{s_{h_j}}{\sum_{h_i \in c} s_{h_i}} \ . \quad (5)$$

Finally, we determine the spatial variance of the cluster $\boldsymbol{\Sigma}_o$ by iteratively fusing the variance of the hypotheses

$$\boldsymbol{C}_j = \boldsymbol{C}_{j-1} - \boldsymbol{C}_{j-1} \left(\boldsymbol{C}_{j-1} + \boldsymbol{\Sigma}_{h_j} \right)^{-1} \boldsymbol{C}_{j-1}, \ \forall_{j=2,\dots,H} \quad (6)$$

with $\boldsymbol{C}_1 = \boldsymbol{\Sigma}_{h_1}$ and $\boldsymbol{\Sigma}_o = \boldsymbol{C}_H$. Accordingly, we are able to build a new proto-object $o = (s_o, \boldsymbol{\mu}_o, \boldsymbol{\Sigma}_o)$ with integrated saliency s_o, spatial mean $\boldsymbol{\mu_o}$ as well as spatial variance $\boldsymbol{\Sigma}_o$.

Another important feature of the spatio-temporal fusion in combination with the employed world model (see Sec. III-A) is the unique object-based representation of salient regions and events over time. To this end, we use the euclidean distance metric to relate the current proto-object hypotheses with already existing entities in the world model in order to decide whether to create a new proto-object as world model entity or fuse it with an already existing entity.

C. *Focus of Attention Selection: The Exploration Gaze Path*

After we determined the salient proto-objects, we need to decide which of these objects should be attended in which order. To this end, we determine an initial object exploration order, i.e. the exploration path. We call it "initial", because in our implementation the analysis of an object as well as the exploration path can always be interrupted and changed by newly detected salient regions, sudden changes in the saliency of tracked objects, or higher-level processes.

1) Exploration Path Strategies: Iteratively attending the most salient region in combination with inhibition-of-return is the classical approach to saliency-based overt and covert attention. However, this bottom-up approach has some drawbacks as overt attention scheme for active sensing robots. Most important, for application in robotics it is also important to provide smooth exploration paths that reduce the amount of necessary ego-motion in order to save energy and reduce wear-and-tear (i.e. to consider a certain form of necessary motion "laziness"), to minimize the time to focus the next and/or all relevant objects, and to provide a more human-like scan path and consequently head movement in unexplored environments. Accordingly, it is also necessary to take into account the joint angle configurations that are necessary to focus specific objects.

We define an exploration path EP $\in S_{\mathcal{O}}$ as a permutation of the proto-objects $\{o_1, o_2, \dots, o_N\} = \mathcal{O}$ in the world model that have to be attended, where $S_{\mathcal{O}}$ is the permutation group of $\mathcal{O}$ with $|\mathcal{O}| = N!$. For example, the exploration path $\text{EP}_{\text{example}} = (o_1, o_3, o_2, o_4)$ would first attend object o_1, then o_3 followed by o_2, and finally o_4. In the following, we denote s_{o_i} as the saliency of object o_i and $\boldsymbol{q}_{o_i}$ represents the robot's joint angle configuration needed to focus object o_i.

a) Saliency-based Exploration Path: All perceived proto-objects in the scene (i.e., all proto-objects in the world model, which also includes objects that are currently outside the field of view) are sorted in descending order by their saliency s_{o_i} and analyzed correspondingly

$$\text{EP}_{\text{saliency}} = (o_{i_1}, o_{i_2}, \dots, o_{i_N}) \text{ with } s_{o_{i_1}} \geq, \dots, \geq s_{o_{i_N}} \ . \quad (7)$$

This is equivalent to the classical approach.

b) Distance-based Exploration Path: Alternatively, we can also neglect the saliency and minimize the accumulated joint angle distances that are necessary to attend all objects

[1] Note that there exist other biological plausible audio-visual integration schemes (see [37]) that can be implemented in our model as well.

on the exploration path in the specified order

$$\mathrm{EP}_{\text{distance}} = \underset{\mathrm{EP}\in S_{\mathcal{O}}}{\arg\min}\left\{\sum_{k=1}^{N}\left\|\boldsymbol{q}_{o_{i_k}} - \boldsymbol{q}_{o_{i_{k-1}}}\right\|\right\} \;, \qquad (8)$$

where $\boldsymbol{q}_{o_{i_k}}$ represents the joint angles needed to focus the kth object on the exploration path, and accordingly $\boldsymbol{q}_{o_{i_{k-1}}}$ is the joint angle configuration for the preceding object. We define $\boldsymbol{q}_{o_{i_0}}$ as the initial joint angle configuration at which we start the exploration using the calculated exploration path. Here, the norm of the joint angle differences $d_{m,n} = \|\boldsymbol{q}_m - \boldsymbol{q}_n\|$ measures the angular distance between two joint configurations and is used as a measure for the amount of necessary ego-motion. However, the underlying problem of determining the minimal accumulated distance to attend all objects equates to the traveling salesman problem (TSP) and is consequently NP-complete[2]. In consequence, we have to limit the computation to K local neighbors of the currently focused object that were not already attended, where K is chosen depending on run-time constraints. In our current real-time implementation, we use $K = 10$, which – in our experience – provides good reasults at acceptable computational costs. By reducing the required amount of ego-motion this strategy leads to more time and energy efficient paths, but it does not take the saliency into account.

c) Balanced Exploration Path: We consider the exploration path estimation as a multi-objective optimization problem, which allows us to combine the saliency-based and distance-based approaches. Therefore, we can minimize a single aggregate objective function

$$\mathrm{EP}_{\text{balanced}} = \underset{\mathrm{EP}\in S_{\mathcal{O}}}{\arg\min}\left\{\sum_{k=1}^{N} f_d(\|\boldsymbol{q}_{o_{i_k}} - \boldsymbol{q}_{o_{i_{k-1}}}\|)\cdot f_s(s_{o_{i_k}})\right\}, \qquad (9)$$

where $s_{o_{i_k}}$ is the saliency value of the proto-object o_{i_k}, f_d is a distance transformation function, and f_s is a saliency transformation function. In our current implementation, f_d is defined as identity function and $f_s(s;\alpha) = s^{-\alpha}$, i.e.

$$\mathrm{EP}_{\text{balanced}}(\alpha) = \underset{\mathrm{EP}\in S_{\mathcal{O}}}{\arg\min}\left\{\sum_{k=1}^{N}\|\boldsymbol{q}_{o_{i_k}} - \boldsymbol{q}_{o_{i_{k-1}}}\|\cdot s_{o_{i_k}}^{-\alpha}\right\}. \quad (10)$$

This combined optimization function tries to balance the tradeoff between far away proto-objects with a high saliency and nearby proto-objects with a lower saliency, where the choice of α influences the objective priorities. However, this definition equates to an asymmetric TSP (i.e., due to the object dependent saliency term the aggregate function's distance between two joint configurations is not identical in each direction) and consequently we have to limit the search for the next object to attend at each step to K local neighbors of the currently attended and focused object.

2) Focus of Attention and Inhibition of Return: In principle, the proto-objects on the exploration path are focused and analyzed successively in the specified order. To this end, the spatio-temporal clustering and tracking (see Sec. III-B.4) in combination with the integrated mean-shift tracking of world model entities allow us to keep track of already attended entities in the environment and thus directly enable object-based inhibition of return. However, the exploration sequence is interrupted and updated if one of the following events occurs: a new proto-object with a high saliency has been detected, a sudden increase of the saliency of an already attended object entity occurred, or a higher-level process prioritizes another proto-object than the currently analyzed (see [2], [28]). In these cases, the corresponding proto-object is attended and analysed directly in order to swiftly react on salient events and allow top-down control of the overt attention.

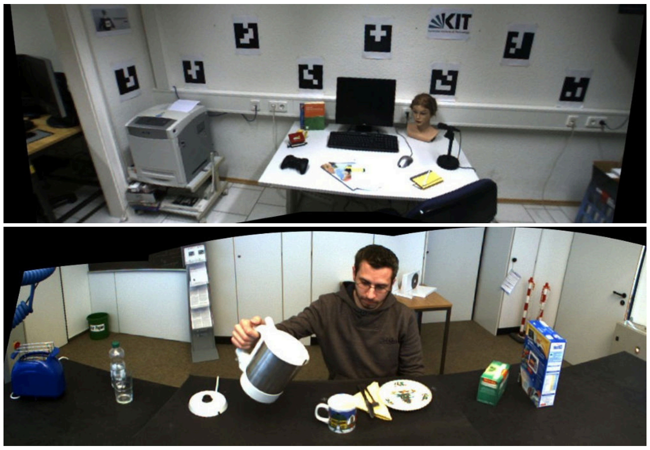

Fig. 4: Sample image stitches of the recordings in room 1 with the stereo camera PTU head (top) and in room 2 with the ARMAR-III head (bottom).

IV. EXPERIMENTAL EVALUATION

In our previous work [2], we demonstrated an improvement of the visual as well as the auditory perception when directing the sensors towards the object of interest. Furthermore, we evaluated the focus of attention selection and object-based inhibition of return in several scenarios. However, the focus of attention selection strategy was solely based on the saliency and did not take into account other factors such as, e.g., the necessary head movement – which costs time and energy – to attend the next object of interest. Here, we want to show that using the proposed balanced exploration path (see Sec. III-C), we can reduce the necessary head movement while still focusing objects in an order that strongly favors salient objects. For this purpose and to "visually" compare the behavior of attention systems, we collected a new data set that we will make publicly available.

A. Data Set

1) Description: Due to the active, scene-dependent nature of overt attention, a quantitative evaluation method of the behavior of active systems hardly exists (see [38]). This is due to the fact that it seems impossible to reproduce all factors such as, e.g., the environment, timing, stimuli,

[2] Please note that the additional requirement of the TSP to return to the starting city does not change the computational complexity.

	Number of Recordings		
scene	PTU stereo setup	ARMAR-III head	$\sum$
breakfast	15	15	30
office	10	10	20
neutral	5	5	10
	30	30	60

TABLE I: Structure of the data set

and implementation details that result in a specific (active) behavior. However, when we focus on the problem of evaluating intelligent exploration strategies, we can approach the problem in two steps: First, we scan the whole room in a scan sweep and calculate the locations of salient (proto-)objects in the room. This is related to the human behavior of taking a quick, initial glance around the room to get an overview of an unknown environment. Second, given a starting configuration, we can use the 3-D locations of the objects to plan the head movement. The first step makes it possible to assess how good our methods are able to determine salient regions. Then, the second step makes it possible to investigate the generated active behavior. This makes it also possible to exchange methods in each step and investigate the resulting behavior. For example, we can add noise to the salient (proto-)object detections or exchange the underlying definition of what is salient. Or, we can implement different exploration path strategies and investigate their effectiveness, which is what we will use our data set for in this paper.

Our data set consists of 60 recordings with a length of 30 seconds each, see Tab. I. We recorded the data set on two hardware systems: a PTU stereo setup (see [2]) and the ARMAR-III humanoid head (see [39]). We considered three evaluation scenarios: office environments, breakfast scenes, and neutral scenes. Neutral scenes are reference scenes that are recorded in the same environment, but with a drastically reduced amount of salient objects. The recordings have been made on different days and at different times of day in order to vary the lighting conditions, ranging from varying natural to artificial lighting. The data set includes the stereo images, the head joint angles, the depth maps, and further information. The total volume of the data set is 36 GigaByte.

2) Sensor Setup: Both hardware platforms, i.e. the PTU sensor setup and the ARMAR-III humanoid head (see Fig. 5), share essentially the same sensor configuration. The wide angle and foveal cameras have a focal length of 4 mm and 12 mm, respectively. The stereo baseline separation between each camera pair is 90 mm. The camera sensors provide a resolution of 640×480 pixels at a frame rate of 30 Hz. In the evaluation, we only use the front and side omni directional microphones. The distance between the side microphones is approximately 190 mm and the vertical distance between the front microphones is approximately 55 mm. The audio data is processed at a sampling rate of 48 kHz. The ARMAR-III head provides 7 degrees of freedom (DoF) and uses a cycle time of T=30 ms to control them (see [3]). The pan-tilt unit has 2 DoF and 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. The wide angle and foveal cameras of the PTU setup have a focal length of 3.5 mm and 6 mm, respectively.

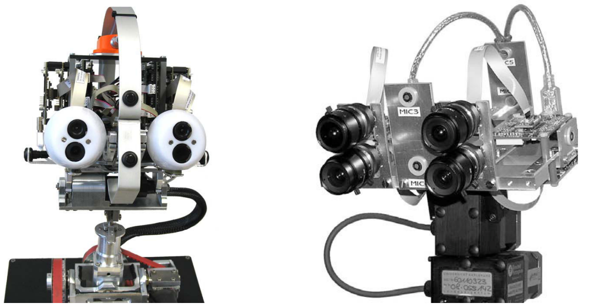
(a) ARMAR-III robot head (b) PTU stereo setup

Fig. 5: The ARMAR-III humanoid robot head (a) and PTU stereo setup (b) provide 7 and 2 degrees of freedom, respectively. Both setups perceive their environment with 6 omnidirectional microphones (1 left, 1 right, 2 front, 2 rear) and 2 stereo camera pairs (coarse and fine view, respectively).

B. Evaluation Procedure and Measures

1) Audio-Visual Exploration and Motion: First, we will discuss the influence of motion onto the visual saliency as continuation of our last experiments [2]. To this end, we performed a couple of additional experiments. To reliably verify the behavior of the proposed system, we repeated all experiments several times with varying external influences like lighting, number of objects and clutter.

2) Exploration Path: To investigate the presented exploration path strategies (see Sec. III-C), we use the presented data set (see Sec. IV-A) and introduce two evaluation measures. First, we use the cumulated joint angles distances (CJAD) as measure of ego-motion

$$\mathrm{CJAD}(\mathrm{EP}) = \frac{1}{N} \sum_{j=1}^{N} \left\| \boldsymbol{q}_{\mathrm{EP}_j} - \boldsymbol{q}_{\mathrm{EP}_{j\text{-}1}} \right\| , \qquad (11)$$

where EP_j is the index of the jth attended object of exploration path EP, and $\boldsymbol{q}_{o_i}$ represents the joint angle configuration that focuses object o_i, see Sec. III-C.1.b. Since we want to reduce the amount of necessary head motion, we want to minimize the CJAD. To investigate the influence of saliency on the exploration order, we use the cumulated saliency (CS) of already attended objects

$$\mathrm{CS}(i; \mathrm{EP}) = \sum_{j=1}^{i} s_{\mathrm{EP}_j} \quad , i \in \{1, 2, \ldots, N\} \qquad (12)$$

Here, a steep growing curve is desired, because it indicates that objects with higher saliency are attended first. Since the number of attended salient objects may vary depending on the saliency distribution in the scene, we denote the percentage of already attended objects as p, which makes it possible to integrate over the curves of different scenes. This way, we can calculate the area under the CS curve as a compact evaluation measure, i.e.

$$\mathrm{ICS}(\mathrm{EP}) = \int \mathrm{CS}(p; \mathrm{EP})\, \mathrm{d}p \quad . \qquad (13)$$

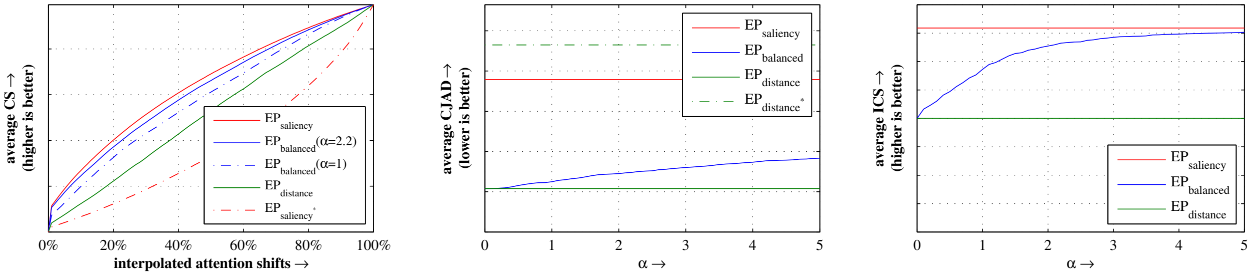

Fig. 6: Left-to-right: The average cumulated saliency (left), the average cumulated joint angle distances (middle), and the average area under the cumulated saliency curve (right) over all recordings in the database.

Additionally, we introduce two normalized versions of CJAD and CS, i.e. NCJAD and NCS, respectively, that take into account the spatial distribution of objects in the scene as well as their saliency distribution:

$$\mathrm{NCJAD(EP)} = \frac{\mathrm{CJAD(EP)} - \mathrm{CJAD(EP_{distance})}}{\mathrm{CJAD(EP_{saliency})} - \mathrm{CJAD(EP_{distance})}} \quad (14)$$

$$\mathrm{NCS(EP)} = \frac{\mathrm{ICS(EP_{saliency})} - \mathrm{ICS(EP)}}{\mathrm{ICS(EP_{saliency})} - \mathrm{ICS(EP_{distance})}} \quad (15)$$

Here, we use two facts and two observations for normalization: The saliency-based exploration strategy will lead to the fastest growth of CS, but is likely to have a high CJAD. In contrast, the distance-based strategy will lead to the smallest CJAD, but is likely to exhibit a slow growth of CS.

Additionally, to serve as a lower boundary for CS, we calculate $\mathrm{EP_{saliency^*}}$ which is the opposite strategy to $\mathrm{EP_{saliency}}$ that selects the least salient unattended object at each shift. Analogously, we calculate $\mathrm{EP_{distance^*}}$ which greedily selects the object with the highest distance at each step and is an approximate (greedy) strategy opposite to $\mathrm{EP_{distance}}$.

C. Results and Discussion

1) Audio-Visual Exploration – Motion: We introduced the motion component (see Sec. III-B.2) to increase the saliency of moving objects (see [15]), which is in accordance with the human biological model (see [4]). Therefore, we evaluated several sequences that involve motion, e.g., a person that pours coffee into a cup or moves a book (see Fig. 2 and 3). As already shown in [15], the chosen visual saliency model proves to be a computationally efficient and reliable method in practice. We always observed a drastic increase of the saliency of moving objects, which consequently increases the saliency of the corresponding proto-object and finally results in a higher prioritization during the exploration path estimation and an earlier attraction of the attentional focus.

2) Exploration Path I – saliency-based: First, we examine the traditional saliency-based exploration path approach (see Fig. 3, red; see Sec. III-C.1.a). As expected, this strategy leads to the highest amount of head movement of all strategies and highest growth of the cumulated saliency (see Fig. 6). Accordingly, the larger amount of necessary head movements leads to a slower exploration of the scene, but fast analysis of the most salient objects.

3) Exploration Path II – distance-based: Secondly, we analyze the exploration path which minimizes the angular distances (see Fig. 3, green; see Sec. III-C.1.b). In the comparison to the other strategies, this path does not take the saliency into consideration and thus leads to the slowest growth of the cumulated saliency (see Fig. 6). However, as can be clearly seen in Fig. 6, the necessary angular distances and thus the time for exploration is minimized. Please note that the computational limitation of only using K local neighbors for the TSP optimization (see Sec. III-C.1.b) leads to a $25\,\%$ longer distance in average (see [40]).

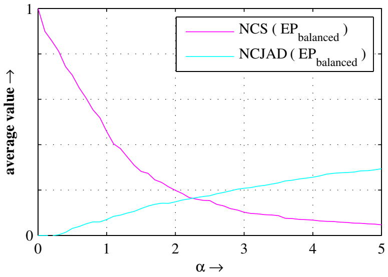

Fig. 7: NCS and NCJAD over all recordings in the data set using the balanced exploration path with varying α.

4) Exploration Path III – balanced: Finally, we consider the balanced method, which tries to balance the tradeoff between a small cumulated joint angle distance and an efficient saliency-based exploration with a steep growth of the cumulated saliency (see Fig. 3, blue; see Sec. III-C.1.c). We can adjust the influence of the two aspects by changing the operating parameters α. As can be seen, even a relatively high α can already drastically reduce the CJAD while providing a high ICS, see Fig. 6.

One question remains: What is an appropriate operating parameter α^*? In our opinion, an optimal α^* should be the point where a further decrease of NCS no longer outweighs a further increase of NCJAD, see Fig. 7. This is related to the point where $\mathrm{ICS}(\mathrm{EP_{balanced}}(\alpha))$ begins to approach a constant value, i.e. where the growths $\mathrm{ICS}(\mathrm{EP_{balanced}}(\alpha))$ starts to approach 0, see Fig. 6. In our experience, a good heuristic to determine a good α^* candidate is to select the value of α where the NCS and NCJAD curves intersect, i.e.

$$\mathrm{NCJAD}(\mathrm{EP_{balanced}}(\alpha^*)) = \mathrm{NCS}(\mathrm{EP_{balanced}}(\alpha^*)) \; . \quad (16)$$

Accordingly, we currently operate the presented system using $\alpha = 2.2$, see Fig. 7 and Tab. II.

	CJAD	ICS
$\mathrm{EP}_{\text{distance}}$	0.2157 (72.6 %)	130.0 (83.1 %)
$\mathrm{EP}_{\text{balanced}}$	0.2972 (100.0 %)	156.5 (100.0 %)
$\mathrm{EP}_{\text{saliency}}$	0.7576 (254.9 %)	161.8 (103.4 %)

TABLE II: Comparison of the results for different exploration path strategies ($\alpha = 2.2$).

When α is set to 2.2, we achieve an average CJAD of 0.2972. For comparison the distance-based and saliency-based strategy achieve a CJAD of 0.2157 (72.6 %) and 0.7576 (254.9 %), respectively. At the same time, we achieve an average ICS of 156.5. Here, the distance-based and saliency-based strategy achieve an average ICS of 130.0 (83.1 %) and 161.8 (103.4 %), respectively. Thus, we provide an exploration strategy that effectively balances between favoring highly salient objects and efficient head movements.

V. CONCLUSION AND FUTURE WORK

In our attention system, we combine adaptive predictive gaze control, detection and localization of salient audio-visual proto-objects, saliency-driven exploration with object-based inhibition of return, and hierarchical knowledge-driven object analysis. We presented a multi-criterial optimization approach to generate efficient head motion for scene exploration that combines the advantages of distance minimization and a winner-take-all-like salient proto-object selection. To evaluate our approach, we collected a novel data set and demonstrated that we are able to achieve a good trade-off between necessary head motion and attending the proto-objects with the highest saliency first.

REFERENCES

[1] 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.

[2] B. Schauerte, B. Kühn, K. Kroschel, and R. Stiefelhagen, "Multimodal saliency-based attention for object-based scene analysis," in *IROS*, 2011.

[3] G. Milighetti, A. De Luca, and L. Vallone, "Adaptive predictive gaze control of a redundant humanoid robot head," in *IROS*, 2011.

[4] D. Mahapatra, S. Winkler, and S.-C. Yen, "Motion saliency outweighs other low-level features while watching videos," in *Proc. SPIE Human Vision and Electronic Imaging XIII*, vol. 6806, 2008.

[5] N. Butko, L. Zhang, *et al.*, "Visual saliency model for robot cameras," in *ICRA*, 2008.

[6] 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.

[7] J. Ruesch, M. Lopes, A. Bernardino, *et al.*, "Multimodal saliency-based bottom-up attention: A framework for the humanoid robot iCub," in *ICRA*, 2008.

[8] F. Orabona, G. Metta, and G. Sandini, "A proto-object based visual attention model," in *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, 2008, pp. 198–215.

[9] D. Figueira, M. Lopes, R. Ventura, and J. Ruesch, "From pixels to objects: Enabling a spatial model for humanoid social robots," in *ICRA*, 2009.

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

[11] 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.

[12] J. K. Tsotsos, *A Computational Perspective on Visual Attention*. The MIT Press, 2011.

[13] B. Schauerte and R. Stiefelhagen, "Predicting human gaze using quaternion DCT image signature saliency and face detection," in *Proc. IEEE Workshop Applications of Computer Vision (WACV)*, 2012.

[14] B. Schauerte and R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction," in *Proc. European Conf. on Computer Vision*, 2012.

[15] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, 2008.

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

[17] 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.

[18] 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.

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

[20] K. A. Fleming, R. A. Peters II, and R. E. Bodenheimer, "Image mapping and visual attention on a sensory ego-sphere," in *IROS*, 2006.

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

[22] T. Machmer, A. Swerdlow, B. Kühn, and K. Kroschel, "Hierarchical, knowledge-oriented opto-acoustic scene analysis for humanoid robots and man-machine interaction," in *ICRA*, 2010.

[23] 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.

[24] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vis. Res.*, vol. 40, pp. 1489–1506, 2000.

[25] A. Dankers, N. Barnes, and A. Zelinsky, "A reactive vision system: Active-dynamic saliency," in *Proc. Int. Conf. on Computer Vision Systems*, 2007.

[26] F. Saidi, O. Stasse, K. Yokoi, and F. Kanehiro, "Online object search with a humanoid robot," in *IROS*, 2007.

[27] A. Andreopoulos, S. Hasler, H. Wersing, *et al.*, "Active 3D object localization using a humanoid robot," *IEEE Trans. Robotics*, pp. 47–64, 2010.

[28] B. Kühn, B. Schauerte, R. Stiefelhagen, and K. Kroschel, "A modular audio-visual scene analysis and attention system for humanoid robots," in *Proc. 43rd Int. Symp. Robotics (ISR)*, 2012.

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

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

[31] H. Wersing, S. Kirstein, M. Götting, *et al.*, "Online learning of objects in a biologically motivated visual architecture," *Int. Journal of Neural Systems*, pp. 219–230, 2007.

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

[33] J. M. Henderson, "Human gaze control during real-world scene perception," *Trends in Cog. Science*, pp. 498–504, 2003.

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

[35] 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 *EUSIPCO*, 2009.

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

[38] F. Shic and B. Scassellati, "A behavioral analysis of computational models of visual attention," *Int. Journal of Computer Vision*, vol. 73, pp. 159–177, 2007.

[39] T. Asfour, K. Regenstein, P. Azad, *et al.*, "ARMAR-III: An integrated humanoid platform for sensory-motor control," in *Humanoids*, 2006.

[40] D. Johnson and L. McGeoch, *Local search in combinatorial optimization*, 1997, ch. The traveling salesman problem: A case study in local optimization, pp. 215–310.