2009 IEEE International Conference on Robotics and Automation
Kobe International Conference Center
Kobe, Japan, May 12-17, 2009

Active Multi-View Object Search on a Humanoid Head

Kai Welke, Tamim Asfour and Rüdiger Dillmann
University of Karlsruhe (TH), IAIM, Institute of Computer Science and Engineering (CSE)
P.O. Box 6980, 76128 Karlsruhe, Germany

{welke,asfour,dillmann}@ira.uka.de

Abstract— **Visual search is a common daily human activity and a prerequisite to the interaction with objects encountered in cluttered environments. Humanoid robots that are supposed to take part in human daily life should possess similar capabilities in terms of representing, attending to and recalling objects of interest in order to ensure robust perception in human-centered environments.**

In this paper, we present necessary processes, memories and representations which allow to identify and store locations of objects, encountered from different angles of view, in a visual search task. In particular, we introduce the so-called Feature Ego-Sphere (FES) as the scene memory for a humanoid robot. Experiments comprising different visual search tasks have been carried out on an active humanoid head equipped with perspective and foveal stereo camera systems. The scene is analyzed actively using both camera systems in order to find instances of searched objects in a consistent and persistent manner.

I. INTRODUCTION

The ability of humans to search for required objects is a prerequisite to interaction. Almost all actions that humans perform rely on specific items which support the action, e.g. as tools. For example, drinking requires a cup, eating a fork, and writing requires a pencil. While the task of searching for such objects is natural to humans it is still hard to implement on a technical system.

In the context of human visual perception, the pop-out effect is a well known phenomenon which supports the guidance of attention towards a specific object within a cluttered scene. According to [1] and [2], the pop-out is attributed to the interplay between dorsal and ventral pathways of the human visual system and is modelled using a blackboard architecture, which strongly relies on parallel processing and distributed representations of objects.

For technical systems, the visual search task has often been formulated as top-down attention guidance. Different approaches in the literature modulate the output of a bottom-up attention system, e.g. using the feature-gate technique [3]. While these approaches follow the line of biologically plausible systems it is hard to achieve good results for arbitrary objects due to the parallel and distributed nature of the problem.

In this work, we propose an approach which takes into account the difference between the ”wetware” used for processing in human brains and the hardware of technical systems. Instead of successively filtering the visual stimuli starting with low-level cues as in traditional attention systems [4], our approach starts with a search for object instances in the scene with coarse features. The resulting hypotheses are then verified with local, more descriptive features. This hypotheses and verify approach allows to decompose the required search space and to reduce the computational complexity of the problem. The target platform for our work is the Karlsruhe Humanoid Head [5]. As shown in Fig. 1, the head provides two pairs of active stereo cameras. One pair with wide angle lenses for perspective views and one pair using a small angle of view, which allows a more detailed visual inspection and mimics the fovea of the human visual system. The proposed approach uses both camera systems to actively analyze the current scene. Hypotheses of object locations are extracted in the perspective view. Based on these hypotheses, eye movements are executed to direct the gaze of the foveal cameras towards the corresponding location. Verification is performed using the foveal camera images. In order to store the object information from the current scene as collected during the object search process, a scene memory is required. In the following we propose a scene memory which assures the persistence and consistence of already acquired information about the scene. As will be shown, the scene memory allows for the integration of multiple hypotheses based on spatial coherence, which makes the search task more robust.

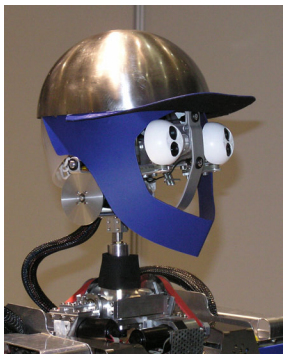


Fig. 1. The Karlsruhe Humanoid Head is equipped with a 3 DoF active camera system and offers one perspective and one foveal camera pair.

With the availability of the necessary technical systems, a large number of capable vision systems for humanoid robots has been presented in the last years. In the following we discuss state-of-the-art systems which make use of foveated vision and address the problem of visual search. In [6], the authors present a vision system which integrates foveal and perspective cameras on a humanoid robot. Object detection

978-1-4244-2789-5/09/$25.00 ©2009 IEEE

is performed in the perspective image. Once a known object is detected, the gaze of the foveal camera is directed towards the object for recognition. The system proposed in [7] makes use of the perspective cameras to calculate hypothetical locations in the scene for a given object using its 3D size and hue cues. The gaze of the foveal camera is directed towards the hypotheses in order to perform recognition using SIFT features. The system works in real-time and takes into account multiple canonical views of objects. In [8] and [9] the authors propose a system which comprises the acquisition of object representations and view-based object recognition on a humanoid robot. The proposed work focusses on interaction. Recognition and acquisition is performed on objects in the hand of an assistant. More recent work has been presented in the context of the Semantic Robot Vision Challenge (SRVC) [10]. In [11] the authors describe a system which combines bottom-up attention and SLAM in order to perform robust recognition of objects.

While most of these systems provide solutions for the task of visual search using active camera systems with foveal and peripheral cameras, the underlying problem of figure-ground segmentation is still not solved for cluttered environments. Most systems make use of disparity maps in order to determine salient regions in the visual field ([8], [9], [11], [7]). However, in cluttered scenes, the segmentation based on disparity maps is not applicable. In [6] the background is represented using Gaussian mixture models which tend to fail for complex backgrounds and in the presence of clutter.

Unlike the systems described above, our approach constructs a consistent and persistent scene memory during the visual search task, which is constantly verified and can be used for successive visual tasks. Making use of a scene memory to collect evidence for searched objects allows to identify instances in a cluttered scene without the need of segmentation. The provided experimental results comprise complex visual search tasks and show that object instances can be identified even in the presence of clutter using our approach.

II. Active Multi-View Object Search

Fig. 2 illustrates the memories and processes involved in the object search task. The input of the system consists of the foveal and perspective camera image pairs as provided by the Karlsruhe Humanoid Head and the ID of an object to search for. The search process generates hypotheses for locations that correspond to the provided object ID and updates the scene memory accordingly. The attention process serializes the verification process by guiding the gaze of the foveal camera pair to salient locations in the scene. Each new gaze initiates a verification process. Hypotheses in the scene memory are verified using the more detailed images from the foveal camera pair. The scene memory is updated in order to obtain consistent and persistent locations of the searched object in the scene. The information from the scene memory can then be used for further visual or interaction tasks.

In the following, we describe the different parts of the system depicted in Fig. 2.

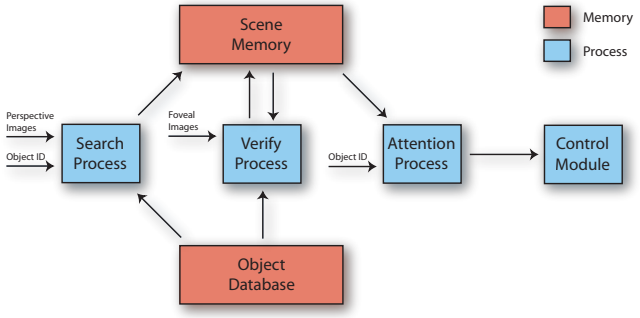


Fig. 2. Overview of the memories and processes involved in the proposed system. The search process generates hypotheses of object locations. From these hypotheses, the attention and control processes generate a gaze sequence for the foveal camera pair which is used for subsequent verification steps. The object database contains appearance based multi-view representations of acquired objects. The system iteratively assures consistent and persistent information in the scene memory.

A. Object Database

The object database contains all object specific information that is required during the object search procedure. The visual search is performed on the basis of multiple views of objects - no volumetric information is used. This facilitates the online acquisition of representations in the sensor space defined by the robotic system.

In the current state of the work, the object views are generated off-line in the object modelling center [12]. The object modelling center offers a camera pair mounted on a robotic arm and a rotating plate. The target positions for recording object views are calculated by subdividing an icosahedron in two stages. Due to the limits of the robotic arm, the zenith covers angles between $\pm 75°$. With this limit, 58 object views are recorded for each object. The collected views are then stored in an aspect graph representation ([13], [14]). In our system, the aspect graph is modelled as a bidirectional spherical graph which contains one node per stored view. The edges between nodes are generated using Delaunay triangulation and thus express the neighborship of views.

The aspect graph serves as basis for the feature extraction process. For each view, one global and a set of local descriptors are extracted and associated to the corresponding node. In the current implementation of the system we make use of color cooccurrence histograms (CCH) [15] as global descriptors. CCHs offer a description of the object that is invariant to rotation in the viewing plane and robust to scaling. Furthermore, they combine texture information (in terms of information about pairs of neighbored pixels) as well as color information. We currently use histograms which cover the hue channel of the HSV image. As local descriptors we use the scale invariant feature transform approach (SIFT)[16]. Each SIFT descriptor is stored together with a reference vector to the origin of the image.

In order to reduce the size of required memory, features are clustered into similar groups using the BIRCH [17] clustering approach for feature quantization. For this work,

we compared the performance of the BIRCH algorithm with the Growing Neural Gas (GNG) method which we used in our earlier work [18] and its incremental version IGNG. We observed that the BIRCH algorithm produces similar clustering results as the IGNG with superior efficiency. Both algorithms support incremental clustering, which is required to allow the incremental acquisition of object representations. In contrast to the GNG and IGNG, where the number of generated clusters depends on the maximum accumulated error per cluster (see [18]), the BIRCH algorithm produces a clustering of the feature space which fits into a given amount of memory.

After feature quantization all cluster centroids are stored in the feature pool. Furthermore, for each object, a feature graph is generated which has identical structure as the aspect graph. The nodes of the feature graph contain references to the corresponding clusters in the feature pool. The object database then consists of one feature graph per object and one common feature pool.

The feature pool itself is implemented as a two-level hierarchical memory. All features are held on disk, while the memory only contains a limited amount of features. Features are cached in memory during instantiation and removed from memory following the least recently used (LRU) strategy.

B. Scene Memory

A visual scene memory is necessary to provide a consistent visual model of the observed scene. It has been shown that human perception accumulates such a scene memory "across separate glances and over time" [19]. In our work, the scene memory contains information about matches between searched objects and the current scene associated with spatial information. These matches are successively verified by moving the foveal cameras to salient locations in the scene. Together with the processes specified in the Sections II-C, II-D and II-E, the scene memory provides consistent information about objects and their locations accumulated over time. The information is constantly verified and is persistently made available for further tasks.

The scene memory proposed in this work is constructed as ego-sphere. The application of ego-spheres as sensory memory is usually called Sensory Ego-Sphere (SES). In [20], the authors introduce the SES as sphere around the so-called ego-center, which is typically located in the base coordinate frame of the robot. The entries in the SES correspond to sensory stimuli and are stored with $2\frac{1}{2}D$ information using their spherical polar coordinates (ϕ, θ, r), thus forming an ego-centric representation of the current scene. The SES has been used in a number of different applications such as multi-modal bottom-up attention [21] and image mapping and visual attention [22].

In contrast to the SES, where usually sensory information is stored, we introduce the Feature Ego-Sphere (FES) as scene memory. The FES is implemented as ego-sphere. However, instead of sensory stimuli as typically stored in the SES, information about matches between features from the object database and the current scene are stored as nodes. Thus, the FES contains the knowledge gathered so far by comparing stored object representations with the current scene. Particularly, the FES does not only contain information about positive matches, but also retains information about the falsification of hypotheses.

Despite its function as scene memory, the FES supports the proposed hypotheses and verify approach in different ways. First it allows the integration of different hypotheses on the basis of spatial coherence. Neighbored entries in the FES which describe the same object can be combined to a common node and thus increase the certainty of the corresponding match. Second, the FES can be deployed to generate the necessary attentional shifts required to verify the hypotheses (see Section II-E).

The FES contains two different types of nodes, which are motivated by the hypotheses and verify approach for object search:

- *hypothesis node:* The position of the hypothesis and information about the match between searched features and object is stored. Hypothesis nodes can contain pointers to verify nodes. They are made persistent in the scene memory in order to allow the search module to detect changes in the scene.
- *verify node:* Verify nodes result from the verification of a node (either verify or hypothesis node). They contain the verified position and information about the match between verified features and the scene.

The content of the FES is manipulated by two basic operations:

- *addEntry:* Adds a hypothesis node to the FES. The node is only added if there is no similar node already present in the proximity. If there is a node present that contains different data, change is detected and the hypothesis node is adapted.
- *verifyEntry:* This operation is called once a node of the FES has been verified. If the verified node is a hypothesis node, a new verify node is created and linked to the hypothesis node. If the verified node is a verify node, its position and match is updated.

Hypothesis nodes are generated by the search process (see Section II-C) using the perspective view of the cameras. The gaze is directed towards salient hypotheses and the verify process invalidates or verifies the corresponding nodes using the foveal camera views. The verification process successively generates verify nodes, if not already present, with corrected positions and creates links to corresponding hypotheses. Multiple verify nodes are combined to one node if they represent the same object and similar positions. In the course of the verification process, verify nodes are moved towards valid object positions in the scene.

C. Search Process

The search process is responsible for the generation of possible locations of object instances enriched by the quality of the match given a specific object. As input, the perspective image pair from the robot's camera system and the object ID

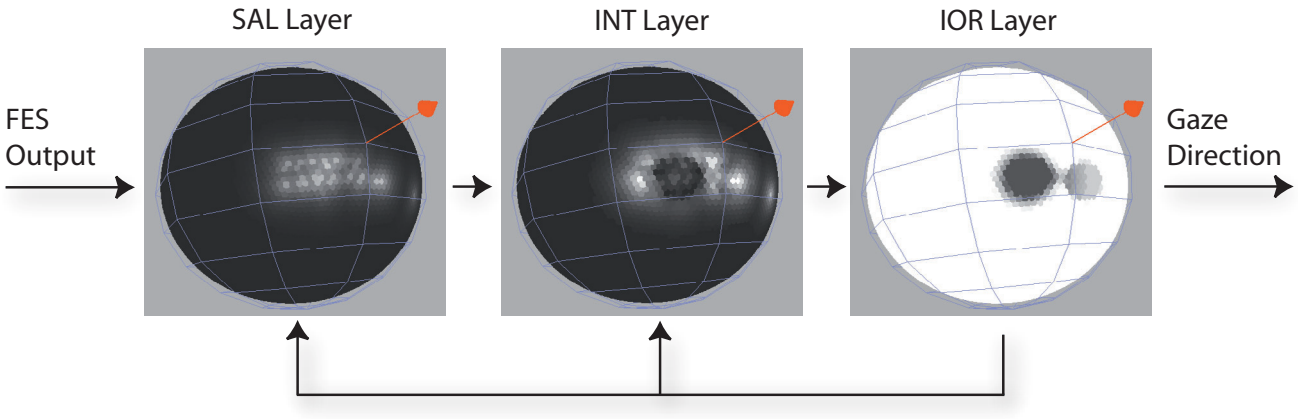


Fig. 3. The spherical winner-take-all (WTA) network as used to direct the gaze of the robot head. The network consists of the saliency layer (SAL), integration layer (INT) and inhibition-of-return layer (IOR). The input saliency is generated from nodes of the FES that correspond to the attended object. Leaky integrate-and-fire neurons in the INT layer initiate an attentional shift once a threshold of activation is reached.

to search for are provided. In order to determine hypotheses about object locations, the search process requests all CCH clusters as stored in the feature graph of the corresponding object. The search is accomplished using an integral image approach. Object positions are accepted on the basis of the histogram intersection with the database features. 3D positions are generated using the disparity map calculated on the perspective image pair.

The resulting hypotheses comprise the locations of hypotheses and the result of the histogram intersection as the quality of the match. Using the *addEntry* operation of the FES (see Section II-B) new hypothesis nodes are added to the scene memory if not already present.

D. Attention Process

The attention process determines the sequence for the verification of the FES content. Two factors influence the decision which FES nodes to verify next: the quality of the corresponding match and the elapsed time since the last verification. Such problems of selective attention can be solved using a winner-take-all (WTA) network as introduced in [4].

In order to provide the necessary input for the WTA network, the FES content is filtered using the object ID of the currently attended object. From the content of the FES, a spherical saliency map is generated. Each leaf node from the FES, which corresponds to the attended object ID, generates a stimulus on the saliency sphere represented as a 2D Gaussian with an amplitude proportional to the stored match quality. Multiple stimuli are combined using a MAX operator.

The saliency sphere is then used as input for a spherical implementation of the WTA network. Fig. 3 shows the three layers of the network. The saliency sphere is modulated with the feedback from the inhibition of return (IOR) layer in the saliency (SAL) layer. The resulting activations are integrated using leaky integrate-and-fire neurons in the integration (INT) layer. Once the activation of a node in the INT layer exceeds a threshold, the neuron fires and generates activation in the IOR layer with 2D Gaussian shape. The inverse ouput of the IOR layer is used as feedback to the SAL and INT layers.

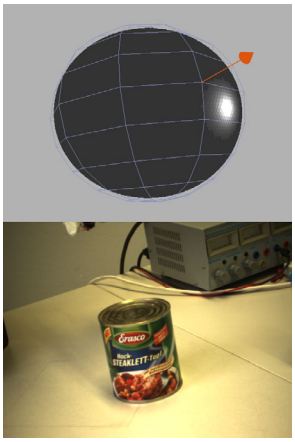
Each time a neuron in the INT layer fires, a new saccadic eye movement is initiated in order to direct the gaze of the foveal cameras towards the corresponding FES entries. For this purpose, all FES nodes which have similar positions on the sphere are determined and the gaze is directed towards the closest node.

E. Verify Process

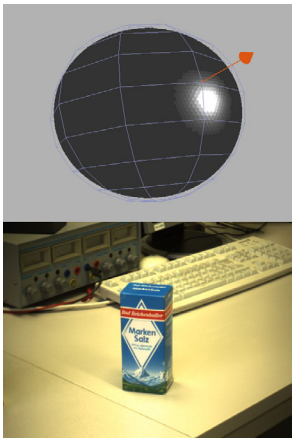
The verify process is responsible for the constant verification of the FES content. Each time the gaze of the robot is adapted by the attention process, a new verification cycle is initiated using the foveal camera images. The verification process requests all nodes from the FES that are visible within the current field of view. Using the match and object IDs stored with the nodes, the SIFT features for all associated object views are determined using the corresponding feature graph. The object's presence is verified by filtering the SIFT matches using a 2D Hough space and voting for the center of the object (see [23]). The corresponding match is thresholded and used to modulate the quality previously associated with the node.

In order to refine the estimated location of the object encoded in the node, the distance in the image planes between the object's position and the principal point of the foveal cameras is determined. Since no stereo calibration is available for the foveal cameras, the target position for the inverse kinematics of the cameras (see Section II-F) is updated in order to move the optical axes of the foveal cameras closer to the location of the verified instance. Note that using this approach the nodes adapt to positions which allow to bring the object instances into the center of the foveal cameras.

(a) Example scene setup used for the object search experiments viewed from the left perspective camera. Two objects were presented to the system.



(b) Resulting saliency sphere and foveal view for the soup can search task.



(c) Resulting saliency sphere and foveal view for the salt box search task.

Fig. 4. Results of the object search experiments for two objects.

For each processed node the operation *verifyEntry* of the FES is called which updates the content of a verify node or generates a new one.

F. Head Control Module

The head control module is responsible for the generation of target values for the head-eye system which correspond to the gazes generated by the attention module. There are essentially two possible strategies to execute the required movements: closed-loop control and open-loop control. In closed-loop control, usually visual feedback is used in order to derive the position error of the eyes iteratively. In contrast open-loop control does not depend on visual feedback but uses the kinematic model of the system to determine the desired posture. Since the target posture in the context of our work is defined by the spike of a single neuron in the WTA network, the necessary visual feedback for closed-loop control cannot be provided.

In order to control the head using the open-loop strategy, the kinematic model of the head-eye system has to be determined. Therefore a kinematic calibration process is performed. We use the approach introduced in [24], which yields accurate results since it avoids methodical errors which are usually introduced with the assumption of two intersecting rotation axes.

The inverse kinematics problem is solved on the basis of the calibrated kinematic model. Since only eye movements are used in the system, the problem can be formulated as non-redundant mapping from 3D Cartesian space to 3D joint angle space (for more details see [5]). We use the inverse Jacobian approach to solve for the joint angles of the camera system. Furthermore, the stereo calibration of the perspective camera system is made available in order to provide the disparity map required for the generation of 3D positions in the search module.

III. Experimental results

A. Setup

For the experiments presented in this section, five objects were stored in the object database. The object view acquisition generated 58 views per object covering equidistant angles in the range of $\theta = [-75°; 75°]$ and $\phi = [0°; 360°[$. The resulting 290 CCH descriptors used in the search module were quantized to 75 cluster centers.

The Karlsruhe Humanoid Head was equipped with a pair of 4 mm lenses for the perspective cameras and 12 mm lenses for the foveal cameras. For the experiments the objects were positioned at about 1 m distance to the head.

B. Experiment I: Object Search Task

In the first experiments, a simple object search task was performed. Two objects from the acquired set were presented in front of a non-uniform background. An example input scene is depicted in Fig. 4(a).

In separate search tasks each of the two objects was enquired. Fig. 4(b) shows the results of the search for the soup can. The upper image illustrates the saliency sphere after 22 verifications. All incorrect hypotheses have been eliminated and the correct hypotheses have been fused by the FES to form a combined position estimate on the saliency sphere. The lower picture in Fig. 4(b) shows the foveal image of the left camera. Similar results could be achieved for the salt box (see Fig. 4(c)). The search task for the salt box could be completed within nine verification steps.

The number of required verification steps depends on the number of hypotheses generated by the search process. Considering the input scene in Fig. 4(a), the salt box has an outstanding color signature and could be brought into focus on the first saccade. The remaining eight verifications adapted the positions of the verify nodes to a single estimate in the FES. In contrast, the CCH of the soup can was found in multiple incorrect scene parts as e.g. the robot arm and

(a) Scene setup used for the complex search task. Two instances of one object are presented to the system in a cluttered scene.

(b) Resulting saliency sphere and foveal views of the left foveal camera. In the final state, the system focusses alternately on the position of both object instances.

Fig. 5. Results of the object search task for two instances in a complex scene.

the teach box. These spurious hypotheses could be invalidated by performing additional saccadic eye movements and verification steps.

The same procedure was performed for all objects in the object database. Similar results could be achieved with a mean number of required verifications steps of 17 in order to retrieve a saliency sphere similar to Fig. 4(b) and Fig. 4(c).

C. Experiment II: Complex Search Task

The goal of the second experiment was to evaluate the performance of the propose approach in cluttered environments. The scene shown in Fig. 5(a), which contains a large amount of distractor objects was presented to the system. The task of this experiment was to find both instances of the cereal box among the distractor objects.

The system performed a saccade containing 28 verification steps in order to retrieve the results depicted in Fig. 5(b). Two locations of high intensity are visible on the saliency sphere, which correspond to the locations of both object instances. After the 28 iterations, the gaze alternately focussed the two instances of the cereal box. Despite the two peaks on the saliency sphere, other local intensity maxima are visible. The local maxima result from unverified hypothesis which lie in the proximity of highly activated areas. If the activation of such hypotheses is below a threshold they can be dominated by strong stimuli nearby. The local maxima can be removed by reducing the IOR size which results in a prolonged verification procedure.

The scene memory content generated during the search task is depicted in Fig. 6. From the initially large amount of hypothesis nodes only a small amount could be verified and has been associated to verify nodes. All other hypotheses nodes are either invalidated or dominated by a stronger stimulus in the proximity. The system produced one unique verify node per instance of the object in the scene.

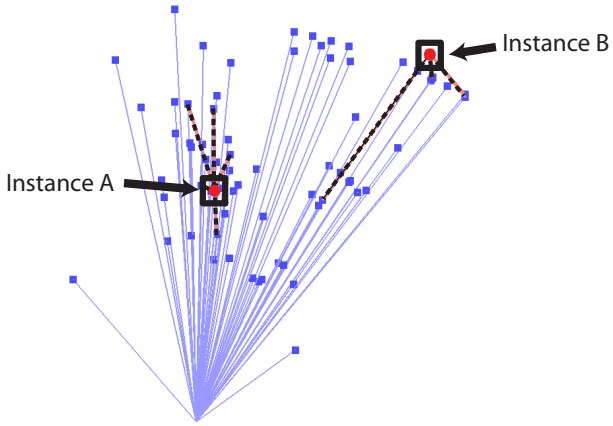


Fig. 6. Content of the FES after 28 eye movements and verifications. Both instances of the searched object have one unique verify node (marked by the black box), which is supported by multiple hypotheses nodes (associations marked with dotted line). For each hypotheses node the connection to the ego-center is illustrated.

D. Discussion

For the experiments we abstain from giving recognition rates of the system since the object search performance directly depends on the object to find in the search task. Because the approach relies on CCH and SIFT features, it is limited to objects and object views which exhibit properties that allow a robust representation with the considered descriptors.

The CCH implementation using the hue channel cannot handle object views that contain black and white in major parts. This results from the fact that black and white do not have a well-defined representation in HSV color space. Furthermore, lighting is an issue when using color descriptors. The experiments were carried out using natural lighting conditions with variations during the day. Reducing the threshold for CCH matching allowed to account for

small changes in ambient lighting, but increased the number of invalid hypotheses produced during the search process. Since the verification process is not that critical concerning lighting, good results could still be achieved. To provide good verification performance using SIFT features, enough texture has to be present in the object view in order to provide the necessary number of features. While logos and pictures printed on the objects provide good features, small written text is usually not covered by the SIFT approach. In our database the short side views of the objects usually contained text and large white areas and thus could not be consistently identified. For other views, e.g. as presented in the previous sections, the results could be reproduced consistently.

The proposed system currently runs on a single core 3.0 GHz linux PC. Each verification step took about 20 seconds. We did not use an optimized implementation of the feature matching process. The approach is intended and already prepared to run on our vision cluster which comprises 6 IBM eServer connected via Ethernet. We expect that by means of optimization and cluster implementation we will reach a verification run-time of less then 1 second.

IV. CONCLUSIONS

In this work we presented an approach which provides persistent and consistent information about object locations resulting from an object search task. The FES datastructure and associated processes were introduced as scene memory. The necessary processes and modules required to perform a visual object search task were presented and discussed. Experiments comprising the search for one object at a time and the search for multiple instances of an object in cluttered scenes were carried out and discussed.

The results show that even for complex tasks the proposed hypotheses and verify approach is able to identify the object locations by actively analyzing the scene.

V. ACKNOWLEDGMENTS

The work described in this paper was conducted within the EU Cognitive Systems project PACO-PLUS (FP6-2004-IST-4-027657) funded by the European Commission.

REFERENCES

[1] F. van der Velde and M. de Kamps, "A model of visual working memory in PFC," *Neurocomputing*, vol. 52-54, pp. 419–424, 2003.

[2] F. van der Velde, M. de Kamps, and G. T. van der Voort van der Kleij, "CLAM: Closed-loop attention model for visual search," *Neurocomputing*, vol. 58-60, pp. 607–612, 2004.

[3] M. Wright, J. Chodzko, and D. Luk, *Biologically Motivated Computer Vision*. Springer Berlin / Heidelberg, 2000, ch. Development of a Biologically Inspired Real-Time Visual Attention System, pp. 779–785.

[4] 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.

[5] T. Asfour, K. Welke, P. Azad, A. Ude, and R. Dillmann, "The Karlsruhe Humanoid Head," in *Proc. IEEE/RAS International Conference on Humanoid Robots (HUMANOIDS)*, 2008.

[6] A. Ude, C. G. Atkeson, and G. Cheng, "Combining peripheral and foveal humanoid vision to detect, pursue, recognize and act," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2003, pp. 2173–2178.

[7] M. Björkman and J.-O. Eklundh, "Vision in the real world: Finding, attending and recognizing objects," *International Journal of Imaging Systems and Technology*, vol. 16, pp. 189–208, 2003.

[8] C. Goerick, H. Wersing, I. Mikhailova, and M. Dunn, "Peripersonal space and object recognition for humanoids," in *Humanoid Robots, 2005 5th IEEE-RAS International Conference on*, 2005, pp. 387–392.

[9] C. Goerick, I. Mikhailova, H. Wersing, and S. Kirstein, "Biologically motivated visual behaviors for humanoids: Learning to interact and learning in interaction," *Humanoid Robots, 2006 6th IEEE-RAS International Conference on*, pp. 48–55, 2006.

[10] "The Semantic Robot Vision Challenge (SRVC)," http://www.semantic-robot-vision-challenge.org.

[11] D. Meger, P.-E. Forssén, K. Lai, S. Helmer, S. McCann, T. Southey, M. Baumann, J. J. Little, and D. G. Lowe, "Curious George: An attentive semantic robot," *Robotics and Autonomous Systems*, vol. 56, no. 6, pp. 503–511, 2008.

[12] R. Becher, P. Steinhaus, R. Zöllner, and R. Dillmann, "Design and implementation of an interactive object modelling system," in *Proc. Conference Robotik/ISR 2006*, 2006.

[13] J. Koenderink and A. van Doorn, "The singularities of the visual mapping," *Biological Cybernetics*, vol. 24, no. 1, pp. 51–59, 1976.

[14] ——, "The internal representation of solid shape with respect to vision," *Biological Cybernetics*, vol. 32, pp. 211–216, 1979.

[15] P. Chang and J. Krumm, "Object recognition with color cooccurrence histogram," in *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1999.

[16] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 1999, p. 11501157.

[17] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in *Proc. ACM-SIGMOD International Conference on Management of Data*, 1996, pp. 103–114.

[18] K. Welke, E. Oztop, G. Cheng, and R. Dillmann, "Exploiting similarities for robot perception," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2007, pp. 3237–3242.

[19] D. Melcher, "Persistence of visual memory for scenes," *Nature*, vol. 26, 2001.

[20] R. A. Peters II, K. E. Hambuchen, K. Kawamura, and D. M. Wilkes, "The sensory ego-sphere as a short-term memory for humanoids," in *Proc. IEEE/RAS International Conference on Humanoid Robots (HUMANOIDS)*, 2001.

[21] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer, "Multimodal saliency-based bottom-up attention. A framework for the humanoid robot iCub," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2008, pp. 962–967.

[22] K. A. Fleming, R. A. Peters, and R. E. Bodenheimer, "Image mapping and visual attention on a sensory ego-sphere," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2006, pp. 241–246.

[23] P. Azad, T. Asfour, and R. Dillmann, "Stereo-based 6D Object Localization for Grasping with Humanoid Robot Systems," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, San Diego, USA, 2007.

[24] K. Welke, M. Przybylski, T. Asfour, and R. Dillmann, "Kinematic calibration for saccadic eye movements," in *Robotics: Science and Systems Conference*, 2009, (submitted to).