Autonomous Acquisition of Visual Multi-View Object Representations for Object Recognition on a Humanoid Robot

Kai Welke, Jan Issac, David Schiebener, Tamim Asfour and Rüdiger Dillmann
Karlsruhe Institute of Technology, Institute for Anthropomatics, IAIM
P.O. Box 6980, 76128 Karlsruhe, Germany

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

***Abstract*— The autonomous acquisition of object representations which allow recognition, localization and grasping of objects in the environment is a challenging task, which has shown to be difficult. In this paper, we present a systems for autonomous acquisition of visual object representations, which endows a humanoid robot with the ability to enrich its internal object representation and allows the realization of complex visual tasks. More precisely, we present techniques for segmentation and modeling of objects held in the five-fingered robot hand. Multiple object views are generated by rotating the held objects in the robot's field of view. The acquired object representations are evaluated in the context of visual search and object recognition tasks in cluttered environments. Experimental results show successful implementation of the complete cycle from object exploration to object recognition on a humanoid robot.**

I. Introduction

For humanoid robots operating in human centered environments the ability of adaptation is a key issue. Autonomous adaptation to new tasks, domains, and situations is a necessary prerequisite for performing purposeful assistance functions in such environments. The combination of sophisticated sensor systems in humanoid platforms together with the ability to interact with the environment allows autonomous exploration in order to gain and consequently adapt knowledge about the surrounding world in a goal-oriented manner.

In the field of visual perception, an important aspect of world knowledge generation consists of the acquisition of internal representations of objects. While many available and published recognition systems rely on representations which have been acquired offline, exploration provides the opportunity to autonomously acquire internal representations. The benefits of such capabilities are two-fold: First, autonomous acquisition of internal representations of objects simplifies the generation of new object knowledge. Second, together with the ability of recognizing known and identifying unknown object instances in the environment, autonomous acquisition allows to adapt to changing environments as required in a human-centered world.

As a consequence, we follow an integrated perspective on visual object modelling and recognition. The goal is to equip the humanoid robot ARMAR-III [1] with the ability to acquire object representations autonomously which can then be used for object search and recognition in future tasks. In previous work [2], we proposed a system for active object search on a humanoid robot which is based on multi-view appearance-based representations. The representations were build from object images captured offline in an object modelling center and accurate segmentation could be achieved using a black background. In this work we focus on the autonomous generation of suitable object representations on the humanoid system. For this task, we propose an approach which combines object-hand and object-background segmentation in order to build a multi-view model of the object. Furthermore, we evaluate the applicability of these autonomously acquired representations in the active visual search task.

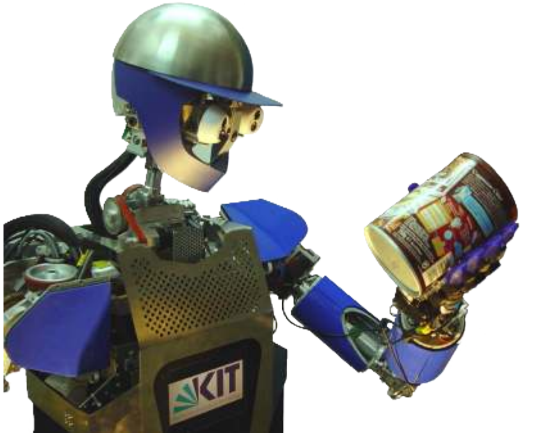


Fig. 1. The proposed approach allows to acquire representations of objects held in the hand of ARMAR-III.

Fig. 1 illustrates the acquisition of object representations on the humanoid robot platform ARMAR-III. During the acquisition process, unknown objects are held in the five-fingered hand. The focus is put on the acquisition of the appearance based part of object representations which can be used for a visual search task. The acquisition of geometric representations as required for grasping is difficult based on visual input only. Rather, the generation of the geometric part involves haptic as well as visual information which is beyond the scope of this paper.

Consequently, for this work, objects are put into the robot hand by a human assistant. The goal is then to generate different view angles of the object in order to cover as much visual information as possible. The views of the object captured with the robot's cameras contain major parts of the hand and the scene. In order to process the object views and to derive viable representations, object-background and object-hand segmentation has to be performed. Therefore, we propose a segmentation approach which allows to determine regions in the image which belong to the object based on a set of cues which are generated using visual as well as proprioceptive information. The segmented views are combined to form a multi-view object representation. As will be shown in the results section, the acquired representations are well-suited for active object search and recognition.

The remaining of this paper is organized as follows: The next paragraph gives an overview of related work in the field of humanoid robotics. In Section II, the humanoid platform ARMAR-IIIb is introduced and the movement generation is discussed. The approach for segmentation and generation of object representations is introduced in Section III, before we present experimental results in Section IV.

A. *Related Work*

Object recognition for humanoid robots has been subject to a vast amount of research. In most systems, the focus is put on reliable recognition and pose estimation as required for manipulation tasks. The underlying representations are usually generated offline, in many cases with the knowledge of exact 3D geometry.

Fitzpatrick et al. [3] were among the first to follow a different approach. In their work visual information is extracted through autonomous exploration of the environment. Therefore, the humanoid robot Cog moves its manipulator over a planar surface. If the robot hits an object placed on that surface, the movement of the object is exploited in order to perform object-background segmentation. The authors highlight that following the causal chain from the robots action allows to develop visual competence. While the theoretical impact of this work has been immense, it does not focus on generating object representations suitable for online object search and recognition.

Goerick et al. [4] follow a different approach and focus on object representations and their application in recognition. In their system, the acquisition of object representations is performed online, while a human assistant presents an unknown, or partly unknown object in his hand in front of a camera. The input image is decomposed in a disjunct set of segments using the adaptive scene dependent filters (ASDF) [5]. In order to select segments which belong to the object, disparity information and the location of the object in the center of the camera are used. In contrast to their work and as aforementioned, in our approach the ability to interact with the environment is exploited. This allows us to move the object out of the visual field of the robot and create a background representation. Thus, the segmentation in cluttered environments is facilitated.

The work of Orabona et al. [6] deals with attentional mechanisms which are used in combination with the movement of the end-effector in order to compose an object of parts which move in a homogeneous manner. While the presented work is promising, the necessity of translational motions of the end-effector in order to group object parts would slow down the acquisition process significantly, especially when rotation is used to generate different view angles as is the case in our system.

In contrast to the aforementioned research, Stasse et al. take an integrated view on object learning and visual search [7]. The goal of the Treasure Hunt Project is to perform object modelling online and use the resulting visual representations in search tasks on the humanoid platform HRP-2. For the acquisition of models, unknown objects are placed on a table. The robot captures object images at different viewpoints in order to develop an object representation valid for multiple viewpoints. Segmentation is performed using dense disparity maps and texture information. Furthermore, the motion of the robot is exploited in order to discard spurious matches of features between two object views. The resulting representation, composed of collected 3D features, is then used for complex visual search tasks which involve determining salient parts of the scene in the perspective cameras of the robot and matching based on SIFT descriptors in the foveal cameras [8]. To our knowledge this is the first system that combines autonomous object modelling and visual search on a humanoid platform. In contrast to the work of Stasse et al. our approach acquires views of an object in the hand of the robot. We believe that multi-sensory object representations, as required e.g. for grasping, can only be generated through a direct physical interaction with the objects.

The proposed approach makes use sensori-motor primitives introduced in our earlier work [9]. In contrast to the segmentation and object learning proposed in [10] the focus is put on general probabilistic methods that support the integration of different segmentation techniques and on hand-object segmentation. We develop a sensor fusion scheme, based on Bayesian methods, which allows to perform segmentation based on different cues, such as background subtraction, disparity and hand localization exploiting proprioceptive sensors. In combination with the movement generation proposed in [9], multiple views of objects are revealed. Based on the generated views, multi-view object representations are constructed. In the experiments we will demonstrate the feasibility of the resulting representations in an active object search task on a humanoid system.

II. Platform and Movement Generation

A. *The Robot Platform*

The system for autonomous object representation and active object search is developed for the humanoid robot ARMAR-IIIb, which is a copy of the humanoid robot ARMAR-III (see [1], [11]). The underlying embodiment is a crucial factor for the design of active approaches. Thus, in the following, we give a brief overview of the structure of the ARMAR-IIIb humanoid robot.

From the kinematics point of view, the robot consists of seven subsystems: head, left arm, right arm, left hand, right hand, torso, and a mobile platform. The head (see [12]) has seven DoF and is equipped with two eyes, which have a

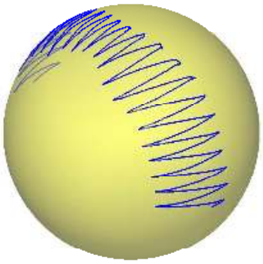


Fig. 2. Trajectory on the 3-D view sphere generated by the utilized control scheme, similar to [9].

common tilt joint and can pan independently. Each eye is equipped with two digital color cameras, one with a wide-angle lens for peripheral vision and one with a narrow-angle lens for foveal vision. The upper body of the robot provides 17 DoF: 2×7 DoF for the arms and three DoF for the torso. The arms are designed in an anthropomorphic way: three DoF for each shoulder, two DoF in each elbow and two DoF in each wrist.

Each arm is equipped with a five-fingered hand [13] with 11 DoF (two for each finger and one for the palm) which are actuated with fluidic actuators. Each DoF is equipped with position and pressure sensors. A position and force control schema is realized based on a simplified model of the fluidic actuator [14].

B. Movement Generation

The goal of the movement generation consists of revealing as many different view directions of the object in the hand of the robot as possible. For this tasks, we resort to the control scheme proposed in [9], which introduces a systematic method to control a robot in order to achieve a maximum range of motion across the 3-D view sphere.

The movement is controlled in the velocity space of position and orientation of the robot hand, both given in the camera coordinate system. For the acquisition of object representations, the rotation of the object in the view plane is dispensable and can be ignored. Consequently, the task of controlling the object position and orientation has 5 DoF. As aforementioned, the arm of ARMAR-III has 7 DoF which leaves 2 DoF of redundancy for our task. As already shown in [9], the redundancy of the system can be utilized to avoid joint limits which results in a higher range of motion across the view sphere.

In summary, the controller is defined by the equation

$$\dot{q} = \begin{pmatrix} J_{pos} \\ J_{rot} \end{pmatrix}^{\dagger} \begin{pmatrix} \dot{x} \\ \dot{\theta} \end{pmatrix} + P_n \, \dot{q}_n,$$

where $\dot{x}$ and $\dot{\theta}$ are the desired position and rotation velocities in the camera coordinate system, J_{pos} and J_{rot} denote the positional and rotational part of the arm Jacobian and $\dot{q}$ is the vector of joint velocities. For joint limit avoidance, a secondary task $\dot{q}_n$ is defined and projected into the null space of the Jacobian using P_n.

For our experiments we used a fixed position x_0 in the camera coordinate system which is optimal in terms of stereo processing and allows to fit typical household objects within the camera images. The rotation of the object is performed around the two relevant axes in the camera coordinate frame. Fig. 2 visualizes the resulting reachable orientations of the hand during our experiments.

III. Autonomous Object Acquisition

Fig. 3 illustrates the components involved in the autonomous acquisition of object representations. The object in the hand of the robot is observed with one stereo camera pair, which is kept static during the procedure. The five-fingered hand as well as the robot arm offer proprioceptive sensor information in terms of joint angles. Both, camera images and joint angles are made available for three different sensor modules which together constitute the object segmentation in the fusion step. For each view captured along the trajectory, one segmented object view is calculated. The segmented views are accumulated and processed in the modelling step in order to derive a multi-view object representation.

To obtain a segmentation of an unknown object in cluttered environments we select different types of sensors and deploy a segmentation fusion yielding the final object segmentation. In particular, our system uses three segmentations generated by three different probabilistic sensor models.

The background sensor performs background subtraction based on the eigenbackground approach in order to determine the area covered by the object and the robot arm in the camera images. Since a completely static head cannot be guaranteed during execution of the trajectory, the eigenbackground subtraction method produces false-positive foreground in areas with high intensity gradient in the background. To compensate for these false-positives, a disparity sensor is deployed in order to detect background pixels based on their distance to the robot. Finally, a hand localization sensor is used to perform object-hand segmentation. Proprioceptive information as well as camera images are used in a particle filter approach to subtract the robot arm and hand from the object view, thus maintaining only the object in the final segmentation.

A. Sensor Models

All sensor models share a common probabilistic concept which accommodates uncertainty that arises in the perception of the robot. The general approach of the sensor models is to generate occupancy probability grids based on Bayes filters with static state assumption [15]. More precisely, we use binary Bayes filters since the segmentation in our case estimates a fixed binary quantity. The underlying idea of occupancy grids is to represent the field of vision of the robot as a field of binary random variables in a grid. A sensor model calculates the posterior estimate over the binary variables conditioned on the measurement data such as camera images up to time t.

Each sensor model possesses an occupancy grid which is updated recursively by applying an inverse measurement

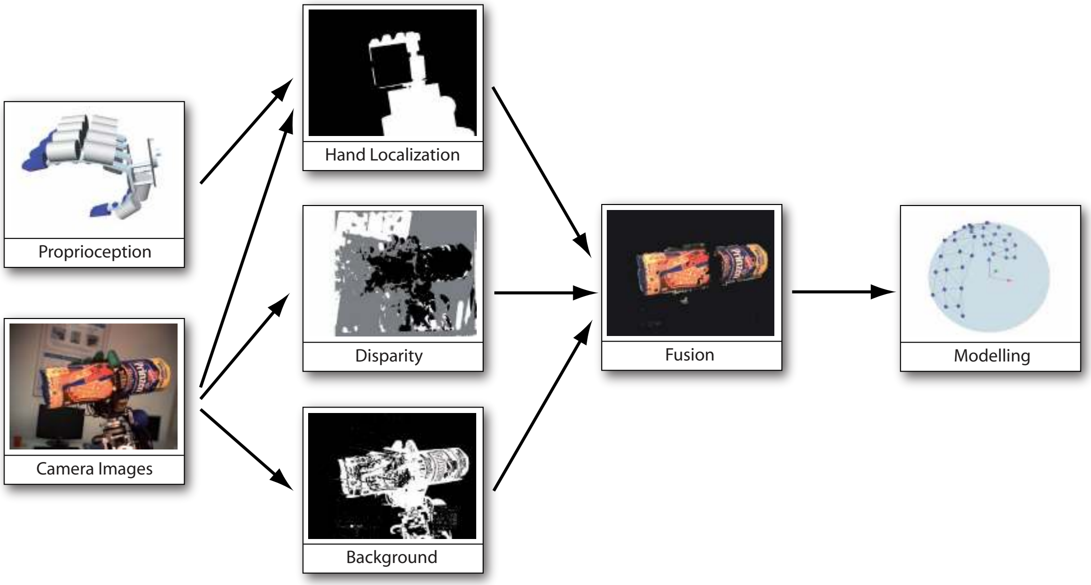


Fig. 3. The system takes input from the camera images and the joint angle sensors in the arm and the hand of the robot. The segmentation is composed from the fusion of background, disparity and hand localization sensors. From all segmented object views, the object representation is modelled.

model $p(m \mid z_{1:t})$. Here $z_{1:t}$ are the measurements up to time t and $m = \{m_i\}_{i=1}^N$ is the grid corresponding to the field of view. Each element m_i denotes the cell with the index i and corresponds to the i-th pixel in the field of vision. To avoid numerical instabilities for probabilities near zero or one, we use the log odds representation of occupancy:

$$l_{t,i} = \log \frac{p(m_i \mid z_{1:t})}{1 - p(m_i \mid z_{1:t})} \tag{1}$$

The recursive occupancy grid update for the i-th cell is then given by

$$l_{t,i} = l_{t-1,i} + \log \frac{p(m_i \mid z_t)}{1 - p(m_i \mid z_t)} - l_0, \tag{2}$$

where l_0 is the prior occupancy probability in the log odds ratio.

In the following we describe the implementation of the three sensor models. A brief introduction to the underlying algorithms is given and the inverse measurement model $p(m_i \mid z_t)$ is derived.

1) Background Sensor: The main segmentation cue is the background subtraction (Fig. 4). Therefore, the eigenbackground approach [16] is deployed which models the background variation based on eigenvalue decomposition by applying principal component analysis (PCA) on a sample of N images. In this way, the background can be represented by the mean image μ_b and the M eigenvectors corresponding to the M largest eigenvalues. Let $\tilde{\phi}_i$ denote the i-th sample image vector and $\phi_i = \tilde{\phi}_i - \mu_b$ is the i-th mean normalized image vector. To perform PCA the covariance matrix C of $A = [\phi_1 \ \phi_2 \ldots \phi_N]$ has to be determined. Since the dimension of C is large, the computation of eigenvalues and eigenvectors is impracticable. Instead, the eigenvalues $\tilde{\lambda}_i$ and eigenvectors $\tilde{\upsilon}_i$ of $C^T = A^T A$ are computed. The eigenvalues λ_i and eigenvectors υ_i of C can then be recovered using

$$\lambda_i = \tilde{\lambda}_i, \quad \upsilon_i = \frac{A\tilde{\upsilon}_i}{\sqrt{\tilde{\lambda}_i}}. \tag{3}$$

Assuming the eigenvalues are in descending order we determine the number of used eigenvectors M by a ratio weight γ using

$$M = \min_k \{k \in [1, N] \mid \frac{\sum_{i=1}^k \lambda_i}{\sum_{j=1}^N \lambda_j} > \gamma\}. \tag{4}$$

Once the eigenbackground model is computed, each mean normalized input image $I_t - \mu_b$ is projected onto the eigenspace expanded by the eigenbackground vectors:

$$\tilde{I}_t = \Phi_M^T (I_t - \mu_b) \tag{5}$$

where $\Phi_M = [\upsilon_1 \ \upsilon_2 \ldots \upsilon_M]$ is the eigenbackground matrix. $\tilde{I}_t$ is then backprojected onto the image space to determine the static parts pertaining to the background

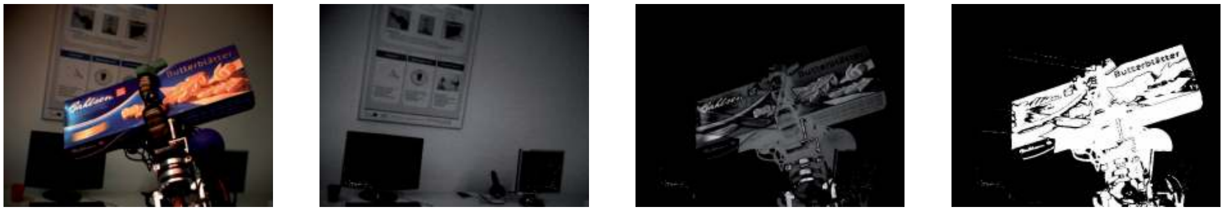
$$\psi_t = \Phi_M \tilde{I}_t + \mu_b \tag{6}$$

and the reconstructed background ψ_t (see Fig. 4) is subsequently subtracted from the input image to obtain the distance image

$$\Delta_t = |I_t - \psi_t|, \tag{7}$$

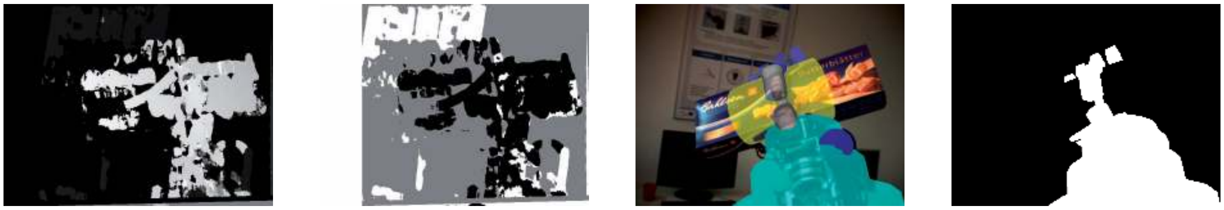
as depicted in Fig. 4(c).

In order to calculate the conditional probability and update the occupancy grid of the background sensor model, we map the distances Δ_t onto probabilities by means of the Gaussian



(a) Input image (b) Reconstructed background image ψ_t (c) Distance image Δ_t (d) Occupancy grid after 10 iterations

Fig. 4. Processing steps and output of the background sensor using eigenbackgrounds



(a) Disparity map (b) Disparity occupancy grid after 10 iterations (c) Localization of the hand (d) Occupancy grid of hand localization

Fig. 5. Results of the disparity and hand localization sensors

cumulative distribution sigmoid function

$$p(m_i \mid \Delta_t^i) = \frac{1}{\sigma_{\Delta_t}\sqrt{2\pi}} \int_{-\infty}^{\Delta_t^i} \exp\left\{ -\frac{(t-\mu_{\Delta_t})^2}{2\sigma_{\Delta_t}^2} \right\} dt, \quad (8)$$

where μ_{Δ_t} is the mean and σ_{Δ_t} the standard deviation of the distances in Δ_t. Thus, the sigmoid function is determined adaptively and outliers maintain a lower probability. The resulting occupancy grid after 10 updates is illustrated in Fig. 4(d).

2) Disparity Sensor: To eliminate remaining background parts which result from small movements of the cameras, we construct the disparity map D_t (see Fig. 5(a)) given the calibrated stereo images. A fixed threshold δ is used to set the margin between foreground and background depth. The conditional probability of the inverse measurement model is then easily obtained by

$$p(m_i \mid D_t^i) = \begin{cases} p_B & \text{if } D_t^i < \delta \\ 1 - p_B & \text{otherwise} \end{cases}, \quad (9)$$

where p_B is the background probability with $p_B > p(m_i)$ and $p(m_i)$ is the prior probability of the cell. Fig. 5(b) illustrates the resulting occupancy grid.

3) Hand Localization Sensor: While the first two sensors perform foreground-background segmentation, object-hand segmentation is accomplished using the localization of the robot hand in the camera images. The goal is to identify the area covered by the robot hand and arm. The robot hand is localized using a particle filter approach [17] to estimate position, orientation and finger joint configuration of the hand. A reduced model of the hand with 6 DoF is used, thus the dimension of the configuration space is 12.

As cues for the particle filter, ratings for the color q_c, the edges q_e and the edge directions q_d of the finger tips and the color of a marker q_m attached to the wrist of the robot are calculated. The conditional probability $p(z|s)$ of an image z given the particle configuration s is calculated according to

$$p(z) \;\propto\; e^{w_c q_c \,+\, w_e q_e \,+\, w_d q_d \,+\, w_m q_m},$$

where w_c, w_e, w_d and w_m are weighting factors for the different cues. In order to derive a precise estimate of the configuration and pose of the hand, simulated annealing is deployed which supports convergence to a local optimum [18]. For the initialization of the particles we use the pose and configuration of the hand as determined from the proprioceptive sensors as well as the result of the last estimation.

In order to detect cases where the fingertips are not clearly visible in the image, the reliability r_{pf} of the particle filter estimation s_{pf} is calculated using heuristics which determine the visible parts of the hand. To handle cases with low reliability a new pose is predicted using the last estimation and the proprioceptive sensors. The rating r_{pred} of the predicted configuration s_{pred} is the product of the rating of the last particle filter estimation and a factor β that incorporates the uncertainty of the prediction.

The configuration s_{used} that provides the localization result is calculated as a weighted mean of the particle filter estimation s_{pf} and the prediction s_{pred}. From the calculated ratings the weight w is determined as

$$w \;=\; \frac{r_{pf}}{r_{pf} + r_{pred}}.$$

The combination of predicted configuration and particle filter estimation is calculated using

$$s_{i,used} \;=\; f(s_{i,pf},\; s_{i,pred},\; w).$$

For elements s_i corresponding to the translation of the end-effector and the joint angles of the fingers f denotes the linear interpolation. The orientation of the end-effector is calculated using spherical linear interpolation. Another

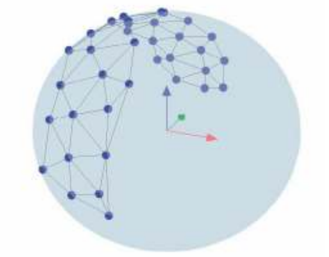


Fig. 6. Object representation using an aspect graph with 36 views as generated with the proposed approach

particle filter iteration is performed in order to find a local optimum in the proximity of the interpolated configuration.

After localization, the hand model is projected into the input image to determine the area covered by the hand as shown in Fig. 5(c). In order to model occlusion of the hand by the object, a cylinder approximates the extent of the object within the robot hand. The diameter and orientation of the cylinder are calculated using the positions of characteristic points of the hand. With the resulting image, the occupancy grid update is carried out using the following conditional probability:

$$p(m_i \mid H_t^i) = \begin{cases} p_H & \text{if } H_t^i \text{ belongs to the hand} \\ 1 - p_H & \text{otherwise} \end{cases} \quad (10)$$

with the probability $p_H > p(m_i)$ of the hand area. The resulting occupancy grid is depicted in Fig. 5(d).

B. Object Segmentation

The object segmentation is composed of the occupancy grids resulting from the three proposed sensor models:

- Background sensor model occupancy grid m^Δ
- Disparity sensor model occupancy grid m^D
- Hand localization sensor model occupancy grid m^H

The probabilities m_i^Δ, m_i^D and m_i^H are recovered from the log odd ratio in (1). Since the three resulting probabilites reside in different observation spaces, a fusion based on the occupancy grids is not possible. Hence, we apply a logic operation to fuse the probabilities and thus obtain the final object segmentation as follows:

$$S_t^i = \begin{cases} I_t^i & (m_i^{\Delta_t} > \delta_S) \wedge \neg(m_i^{D_t} > \delta_S) \wedge \neg(m_i^{H_t} > \delta_S) \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

where $\delta_S \in [0.5, 1]$ is a probability threshold. $m_i^{\Delta_t} > \delta_S$ specifies whether a cell is a foreground cell or not, $m_i^{D_t} > \delta_S$ states if the i-th cell belongs to the background and $m_i^{H_t} > \delta_S$ indicates if the cell belongs to the hand or not. The resulting segmentation mask S_t is post-processed using morphological operations to erase scattered noise pixels.

C. Object Modelling

As stated above, the goal of the segmentation of unknown objects is to generate an object representation which is suitable for object search on a humanoid robot as presented in [2]. The modelling step comprises all necessary steps to build such a representation from the segmented object views. In our case, objects are represented using a multi-view appearance-based representation, called aspect graph [19]. Each node in the aspect graph corresponds to a specific view of the object. From the set of segmentations generated along the executed trajectory, we select object views in such a way that an equidistant distribution over the view sphere is approximated. Fig. 6 shows an example of a view sphere generated by the proposed approach. Each view is then processed using feature extraction methods. For the object search procedure, Color Cooccurrence Histograms [20] and SIFT features [21] are extracted for each object view. To reduce the amount of required features and to derive prototypical views, clustering in feature space is performed using vector quantization methods.

IV. Experimental Results

A. Setup

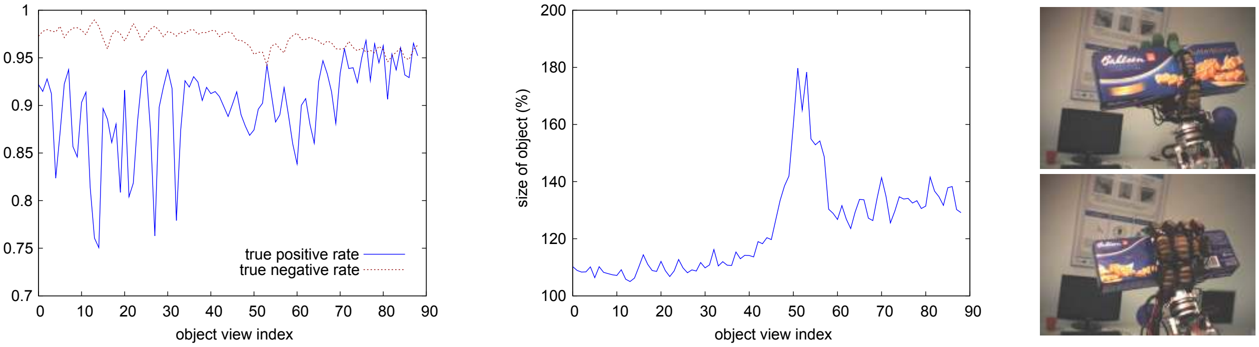
All experiments on object segmentation were carried out on the humanoid robot platform ARMAR-IIIb. In order to show the robustness of the approach we chose a common background which contains a typical amount of clutter. Unknown objects were put into the hand of the robot by a human assistant. Segmentation was performed each 8 seconds along the trajectory (see Section II-B). Thus, about 90 object views for each object were generated. For image capturing we used the perspective stereo camera pair equipped with 6mm lenses.

The parameters for the sensors were chosen as follows. The eigenbackground model was calculated from 60 sample images and we used $\gamma = 0.98$ as a ratio weight of the eigenvalues in order to determine the M best eigenvectors expanding the eigenbackground space. Disparity was extracted in a range of 0 to 200 pixels. The number of particles for hand localization was set to 3000. Four particle filter iterations with decreasing variance were performed for each localization. All occupancy grids were updated with 10 iteration for each object segmentation cycle. The general probability threshold δ_S of the occupancy grid was set to 0.78. The probability constants p_H and p_B were set to 0.75.

B. Object View Acquisition

In order to evaluate the proposed approach we manually segmented all views of one object. From the manual segmentation and the autonomous segmentation, the number of true-positive and true-negative object pixels were determined. Fig. 7(a) illustrates the resulting rates over all views of the object. In average, 90% of the object and 97% of the background could be recovered.

In Fig. 7(b) a more critical measure was used in order to judge the quality of the segmented images. Therefore, the number of segmented object pixels were set into relation to the number of expected pixels from the ground truth. Thus, the size of the autonomously segmented areas relative to the object size is measured. From the beginning of the trajectory up to view 30 a stable performance of about 110% size



(a) True-positive and true-negative rates for all generated rotations of one object. In average, 90% of the object and 97% of the background could be recovered.

(b) Number of segmented pixels in relation to manually segmented object. With increasing rotation the fingertips become less visible which results in inaccurate hand-object segmentation.

(c) Orientations with advantageous fingertip visibility (top) and hidden fingertips (bottom)

Fig. 7. Segmentation results for all generated rotations of one object

relative to the ground truth is achieved. Starting with view 40 the smallest side of the object becomes visible. Thus, the expected segmented area is smaller which results in a peak at view 51. In the second half of views, starting from index 60, the extent of the object increases. As illustrated in Fig. 7(c) the fingertips are not clearly visible in these cases. Consequently, the hand localization relies mostly on prediction, which is affected by inaccuracies in the kinematic model of the arm. Hence, the segmentation of the hand is not optimal which results in more false-positive object pixels.

In order to show the generality of the proposed approach, 20 objects were segmented autonomously. In this experiment we did not perform the complete movement but used exemplary orientations of the end-effector from the beginning of the trajectory. As can be seen in Fig. 8, for all objects good segmentation results could be achieved. Compared to the manual segmentation, a true-positive rate of 87% and a true-negative rate of 96% could be achieved in average. As a consequence of the application of eigenbackgrounds, object parts which share the same intensity with the background cannot be recovered.

C. Object Search

In order to demonstrate the feasibility of the multi-view representations generated from the segmented object view for object recognition, we apply one exemplary representation in an active visual search procedure. As described in [2] the goal of our visual search system consists in filling a scene memory with occurrences of searched object instances in the scene while performing saccadic eye movements using the active stereo camera system. A successful search for an object results in focusing the correct object instance in the foveal cameras after several saccadic eye movements.

The aspect graph for the test object resulting from the modelling step comprised 36 distinct viewpoints which were clustered to 3 prototypical views. For the search experiment we used a cluttered scene as depicted in Fig. 9(a). The searched object was presented in different view angles which were covered by the generated multi-view representation.

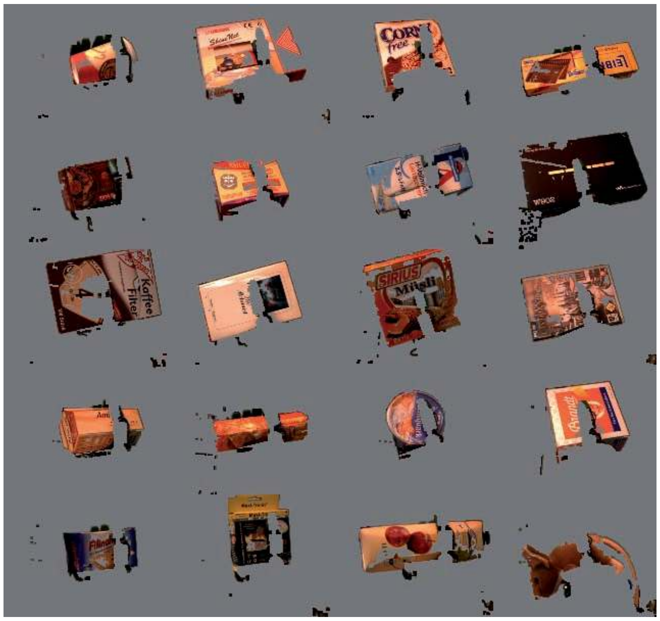


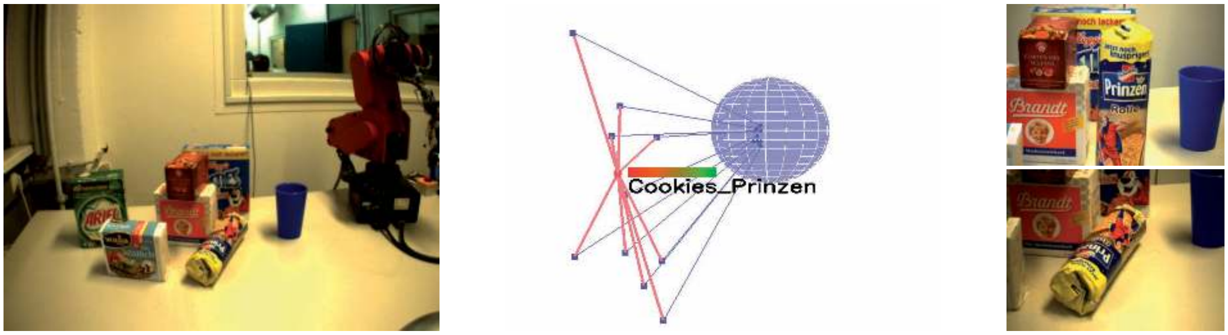
Fig. 8. Segmentation result for 20 test objects

Fig. 9(b) illustrates the scene memory content after 22 saccadic eye movements. As depicted, the instance of the test object could be found and stored. In Fig. 9(c) the final focus of one foveal camera is depicted for two object views.

V. Conclusion

In the proposed work, we demonstrated autonomous acquisition of multi-view representations on a humanoid platform. The presented approach exploits the ability of a humanoid robot to actively change the perceived world in order to generate knowledge about previously unknown objects. An object segmentation procedure is presented which allows to segment views of objects held in the hand of the robot even in cluttered environments.

The probabilistic formulation of the different sensor models together with the application of occupancy grids results



(a) Scene setup used for the visual search task

(b) Content of the scene memory after 22 saccadic eye movements

(c) Resulting focus of the system for two different orientations of the object

Fig. 9. Results of the visual search task using an autonomously acquired object representation

in a extensible fusion scheme for segmentation. Based on the segmented object views, an object representation suitable for object recognition is generated.

The experiments show that the autonomous segmentation is very accurate. Inaccuracies occur in cases where the fingertips are not visible or the object itself is too similar to the background. Finally, we could demonstrate the complete cycle from visual object exploration to recognition in an active visual search task.

VI. Acknowledgements

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] T. Asfour, K. Regenstein, P. Azad, J. Schröder, A. Bierbaum, N. Vahrenkamp, and R. Dillmann, “Armar-III: An integrated humanoid platform for sensory-motor control.” in *IEEE-RAS International Conference on Humanoid Robots (Humanoids 2006)*, December 2006, pp. 169–175.

[2] K. Welke, T. Asfour, and R. Dillmann, “Active multi-view object search on a humanoid head,” in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2009)*, 2009.

[3] P. Fitzpatrick and G. Metta, “Grounding vision through experimental manipulation,” *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, vol. 361, no. 1811, pp. 2165–2185, 2003.

[4] 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.

[5] J. J. Steil, M. Götting, H. Wersing, E. Körner, and H. Ritter, “Adaptive scene dependent filters for segmentation and online learning of visual objects,” *Neurocomput.*, vol. 70, no. 7-9, pp. 1235–1246, 2007.

[6] F. Orabona, G. Metta, and G. Sandini, *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*. Springer Berlin / Heidelberg, 2007, ch. A Proto-object Based Visual Attention Model, pp. 198–215.

[7] O. Stasse, T. Foissotte, D. Larlus, A. Kheddar, and K. Yokoi, “Treasure hunting for humanoid robots,” in *Workshop on Cognitive Humanoids Vision,8th IEEE-RAS International Conference on Humanoid Robots*, 2008.

[8] O. Stasse, D. Larlus, B. Lagarde, A. Escande, F. Saidi, A. Kheddar, K. Yokoi, and F. Jurie, “Towards autonomous object reconstruction for visual search by the humanoid robot hrp-2.” in *Proc. IEEE RAS/RSJ Conference on Humanoids Robots*, 2007.

[9] D. Omrcen, A. Ude, K. Welke, T. Asfour, and R. Dillmann, “Sensorimotor processes for learning object representations,” in *IEEE-RAS 7th International Conference on Humanoid Robots*, 2007.

[10] A. Ude, D. Omrcen, and G. Cheng, “Making object recognition an active process,” *International Journal of Humanoid Robotics (IJHR)*, vol. 5, pp. 267–286, 2008.

[11] T. Asfour, P. Azad, N. Vahrenkamp, K. Regenstein, A. Bierbaum, K. Welke, J. Schröder, and R. Dillmann, “Toward humanoid manipulation in human-centred environments,” *Robot. Auton. Syst.*, vol. 56, no. 1, pp. 54–65, 2008.

[12] 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.

[13] I. Gaiser, S. Schulz, A. Kargov, H. Klosek, A. Bierbaum, C. Pylatiuk, R. O. T. Werner, T. Asfour, G. Bretthauer, and R. Dillmann, “A new anthropomorphic robotic hand,” in *Proc. IEEE/RAS International Conference on Humanoid Robots (HUMANOIDS)*, 2008, pp. 418–422.

[14] A. Bierbaum, J. Schill, T. Asfour, G. Bretthauer, and R. Dillmann, “Hybrid force position control for a pneumatic anthropomorphic hand,” in *Proc. IEEE/RAS International Conference on Humanoid Robots (HUMANOIDS)*, 2008, to appear).

[15] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, September 2005.

[16] N. M. Oliver, B. Rosario, and A. P. Pentland, “A bayesian computer vision system for modeling human interactions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 831–843, 2000.

[17] M. Isard and A. Blake, “Condensation - conditional density propagation for visual tracking,” *International Journal of Computer Vision*, vol. 29, pp. 5–28, 1998.

[18] J. Deutscher, A. Blake, and I. Reid, “Articulated body motion capture by annealed particle filtering,” in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 2, 2000, pp. 126–133 vol.2.

[19] J. Koenderink and A. van Doorn, “The internal representation of solid shape with respect to vision,” *Biological Cybernetics*, vol. 32, pp. 211–216, 1979.

[20] P. Chang and J. Krumm, “Object recognition with color cooccurrence histograms,” *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 2, p. 2498, 1999.

[21] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1999, pp. 1150–1157.