Gaze Selection during Manipulation Tasks

Kai Welke, David Schiebener, Tamim Asfour and Rüdiger Dillmann
Karlsruhe Institute of Technology (KIT)
Institute for Anthropomatics (IFA)

{welke,schiebener,asfour,dillmann}@kit.edu

***Abstract*— A major strength of humanoid robotics platforms consists in their potential to perform a wide range of manipulation tasks in human-centered environments thanks to their anthropomorphic design. Further, they offer active head-eye systems which allow to extend the observable workspace by employing active gaze control. In this work, we address the question where to look during manipulation tasks while exploiting these two key capabilities of humanoid robots.**

We present a solution to the gaze selection problem, which takes into account constraints derived from manipulation tasks. Thereby, three different subproblems are addressed: the representation of the acquired visual input, the calculation of saliency based on this representation, and the selection of the most suitable gaze direction. As representation of the visual input, a probabilistic environmental model is discussed, which allows to take into account the dynamic nature of manipulation tasks. At the core of the gaze selection mechanism, a novel saliency measure is proposed that includes accuracy requirements from the manipulation task in the saliency calculation. Finally, an iterative procedure based on spherical graphs is developed in order to decide for the best gaze direction. The feasibility of the approach is experimentally evaluated in the context of bimanual manipulation tasks on the humanoid robot ARMAR-III.

I. INTRODUCTION

The anthropomorphic design of humanoid robots makes these platforms most suitable for manipulation tasks in human-centered environments and facilitates human-robot interaction. The integration of active head-eye systems in such platforms is a direct consequence of the anthropomorphic design. While the application of head and eye movements to fixate, to saccade, or to perform smooth pursuit plays an important role in interaction, there is also a technical benefit in using active systems. In contrast to passive camera systems where an increase in the field-of-view leads to a loss in the resolution of details, the application of active systems allows to increase the observable area while keeping the details. During manipulation this behavior is desirable, especially if the task involves multiple objects which are spatially distributed. Fig. 1 illustrates such a setup, where the goal of a bimanual manipulation task is pouring of juice into a glass. By fixating the objects sequentially using active gaze control, almost the full camera resolution is available to perform reliable object recognition and pose estimation.

The classical approach to solving complex manipulation tasks involves the sense-plan-act scheme. Thereby, entities of the world are visually captured in an internal representation which is then used as basis for action sequencing and motion planning followed by the execution of the resulting trajectories. On humanoid robots such an approach bears several problems that render its applicability in real world scenarios difficult. First, the complexity of humanoid platforms leads to divergence between the kinematic and dynamic model and the real execution. Thus, the planned trajectories are usually not executed with the required accuracy. Second, the internal representation of the scene is also affected by inaccuracies. These stem on the one hand from noisy measurements of the perception and on the other hand from unpredictable behavior in unconstrained environments. Consequently, in order to achieve robust execution of manipulation tasks, a continuous adaptation of the internal representations is favorable over a sense-plan-act approach.

Fig. 1. ARMAR-IIIa [1] performing bimanual visual servoing in order to pour juice into a cup. During this task, both target objects as well as the hands of the robot need to be observed by the visual system. Due to the large workspace, the selection of appropriate gaze directions is required in order to cover all task relevant objects and thus allow successful execution of the task.

In order to take into account the inaccuracies of perception and execution, a common approach consists in formulating the processes involved in the overall task in a probabilistic fashion [2]. The derivation of an internal representation then becomes a probabilistic inference problem where appropriate models for uncertainty in the perception and execution processes need to be provided. In our approach, the environment is represented using a spatial environmental model of all entities involved in the manipulation task. Each entity corresponds to an object in the real world, i.e. cups, juice boxes, or hands of the robot, and is accompanied with uncertainties

about its current state. Appropriate observation models for the visual perception are introduced and applied in a data fusion scheme thus reducing the amount of uncertainty over time. Further, the motion of all objects is predicted in order to account for the dynamic nature of manipulation tasks. This is essential since e.g. the arms of the robot move during most manipulation tasks.

During a manipulation task, the different world entities compete for being fixated by the active camera system. Each fixation of an entity allows a more accurate representation of its state within the environmental model by fusing the new observation with the past sensor data. Depending on the manipulation task the requirements on the accuracy of the environmental model might differ significantly. While e.g. transportation tasks usually do not require a precise estimation of the object's position, other tasks such as grasping will fail if the estimated poses of the object and the robot hand are not accurate enough. Consequently, we propose to include these accuracy requirements in the gaze selection mechanism by fixating objects accordingly. In order to include this task specific guidance in the gaze selection strategy, we introduce the *task acuity* in the calculation of saliencies. By implementing a gaze selection mechanism on top of the task acuity, the active perceptual process can be configured in order to guarantee a specific accuracy for each element in the environmental model as suitable for the manipulation task.

This paper is organized as follows: In the next section, related work is discussed and the novel aspects of the proposed work in the paper are highlighted. Subsequently, in Section III, the proposed gaze selection mechanism is introduced including the environmental model, the saliency measure based on the task acuity and the decision for the most feasible gaze direction. The proposed mechanism is then put into the context of a bimanual manipulation task, and appropriate motion models are defined in Section IV. The achieved results are discussed in Section V, before the contribution of the proposed work is summarized in Section VI.

II. RELATED WORK

In the context of human visual processing, the problem of gaze selection is often referred to as overt visual attention. The most prominent computational model for visual attention has been proposed by Itti et al. ([3]) followed by several extensions (e.g. [4]) and implementations on robotic platforms (e.g. [5]). An extensive review of such approaches can be found in [6]. In contrast to this line of research, where the goal consists in mimicking the human visual attention processes, our work focuses on establishing a technically motivated approach which allows to support manipulation in a real world environment while making use of active camera systems.

Another line of research deals with active visual search, where the goal consists in detecting and recognizing objects in the extended observable area. Thereby, models for search targets are usually made available as cue for the search task ([7], [8]). Recently, active visual search has been extended by means of integrating a spatial memory that allows to fuse visual information over several gaze directions ([9], [10]). Further, in [11], the active visual search task is extended to a treasure hunting task involving not only gaze selection but also locomotion of the robot in order to detect the object. It has been shown that such systems already can be applied in manipulation tasks ([12]). Nevertheless, constraints arising from manipulation tasks are not taken into account. This applies to the requirement of a dynamic environmental model as well as to the inclusion of accuracy requirements in the saliency measure as proposed in our approach.

The competition for the limited resources of the visual perception system of humanoid robots stands at the core of several gaze selection approaches in the context of humanoid locomotion. During locomotion, usually at least two different perceptual tasks compete: the self-localization of the robot and the obstacle avoidance. In [13], a gaze selection mechanism is proposed which minimizes the self-localization uncertainty as well as the obstacle avoidance uncertainty. In [14], the authors approach the gaze selection problem in a RoboCup scenario, where self-localization, obstacle avoidance, and ball detection compete for the limited resources. Thereby, the environmental model allows for dynamic entities using occupancy grid mapping techniques. The gaze selection mechanism aims to reduce the uncertainty within the grid based representation. For this purpose, a saliency measure is proposed based on the Shannon entropy. Being tailored for locomotion, these approaches do not propose any mechanism to include constraints from manipulation tasks in the gaze selection.

Another possibility of calculating gaze sequences for a manipulation task consists in using the knowledge from a motion planning step. In [15], the knowledge from planning is used to determine the position of objects in the scene and thus adapt the gaze accordingly. In [16], the gaze direction is planned together with the robot motion under consideration of visibility constraints.

In contrast to these approaches and as motivated in the introduction, we seek to establish a gaze selection mechanism which can handle inaccurate or incomplete world knowledge. Thus, we propose an online approach to gaze selection in contrast to the offline calculation of gaze sequences based on the a-priori model during motion planning. Therefore, we establish an environmental model which is updated online and can handle dynamic world entities. This update is formulated as probabilistic inference process. Similar to [13] and [14], the goal of redirecting the gaze is then the reduction of uncertainty in this model. The main contribution of this work consists in the generation of a task specific gaze sequence by introducing the task acuity as saliency measure. The task acuity allows to configure the required accuracy for entities in the environmental model. Based on this saliency measure, an approach for selecting the optimal gaze direction during manipulation is introduced.

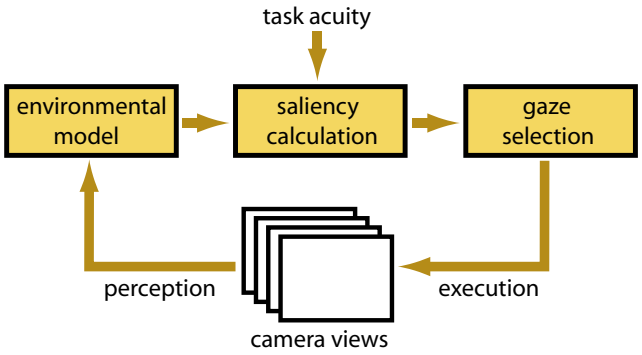

Fig. 2. The proposed approach generates gaze sequences in a perception-action loop. The processing chain includes fusion in the environmental model, calculation of the saliency under consideration of task constraints, and selection of the best gaze direction.

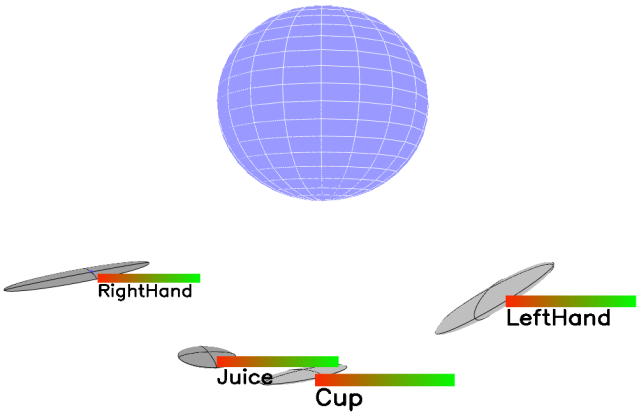

Fig. 3. The environmental model represents the current state estimation of the pose of objects relevant to the manipulation task in a fixed ego-centric reference frame. Each object is associated with a label, a pose estimate, and a recognition certainty. The figure illustrates the environmental model during a bimanual manipulation task involving both hands and two objects. The estimated position uncertainties are indicated by ellipsoids corresponding to the covariance matrix. For each object the recognition certainty is visualized with a bar, where green denotes certainty close to one.

III. Gaze Selection during Manipulation Tasks

The proposed gaze selection mechanism adapts the gaze of the active camera system online in a perception-action loop. The processing steps are illustrated in Fig. 2. First, the processed camera views are fused in the environmental model. Then, the saliency of all entities in the environmental model is calculated. Thereby, the task acuity allows to configure the accuracy of the active perceptual process. Finally, a selection mechanism steers the gaze redirection according to the saliency measure. All three steps of the processing chain are discussed in detail in the following sections.

A. Environmental Model Representation

In order to support manipulation, the environmental model needs to cover task relevant objects such as the hands of the robot or manipulation targets as well as their relevant properties. While the selection of these objects and properties is task specific and thus varies, all manipulation tasks share the common goal of physically interacting with the environment. Consequently, all entities stored within the environmental model need to provide at least means to direct interaction toward them.

To support the physical interaction, the environmental model is organized as a spatial memory covering 6D pose information for each entity. For most manipulation tasks, the number of objects that need to be considered and represented in the model is limited. Consequently, we choose a sparse landmark-based approach to represent the environmental model. For each landmark, its 3D position x accompanied with the location uncertainty Σ_x and its orientation in quaternion representation q is stored. The resulting environmental memory is a collection of the N task relevant entities:

$$M = (m_1, \cdots, m_N)\,,$$

where each entity m_i is represented as:

$$m_i = (x_i, \Sigma_{x_i}, q_i)\,.$$

In order to update the content of the environmental model, stereo-based object localization is performed in the current view of the cameras. For this purpose, we make use of the approach proposed in [17] for textured objects and in [18] for uniformly colored objects. The localization process provides the position z and the orientation q_z for each object. The uncertainty in the localization process is modeled as additive Gaussian noise in the position domain with the covariance matrix Σ_z. In order to approximate the 3D localization uncertainty, we assume 2D additive Gaussian noise of localization in each stereo image which is passed through the epipolar geometry using the unscented transform [19]. Further, we calculate a scalar value $\epsilon \in [0, 1]$ that quantifies the confidence of the object recognition and localization process.

The update of the environmental model based on the object localization result is implemented as probabilistic inference process. The correspondence between localized object and memory entity is solved on the spatial domain using the maximum a-posteriori estimate. Since only normally distributed random variables are involved, the update process is realized using Kalman filtering. The prediction step of the Kalman filter incorporates the motion and motion uncertainty of the memory entity, while the update step fuses the predicted estimate and the current observation. For the prediction step, we provide a motion model for each entity in the task. These motion models are task specific and will be further defined for the application in bimanual visual servoing in Section IV.

In the update step we incorporate the confidence of the current object localization ϵ as proposed in [20]. Instead of using just the Kalman gain matrix K, the position and uncertainty estimation is updated using $\epsilon \cdot K$. The resulting position estimates are illustrated in Fig. 3. The orientation of

each entity is updated by applying spherical linear interpolation to the stored and observed quaternions. The interpolation parameter κ is derived according to the predicted variance of the stored entity position ($\overline{\Sigma}_x$) and observed position variance (Σ_z). Thereby, we use the radius of a sphere with the same volume as the uncertainty ellipsoid as quantification for the amount of uncertainty:

$$\kappa = \frac{|\overline{\Sigma}_x|^{\frac{1}{6}}}{|\overline{\Sigma}_x|^{\frac{1}{6}} + |\Sigma_z|^{\frac{1}{6}}}.$$

As for the position, we incorporate the confidence of correct recognition and thus interpolate the orientation with the factor $\epsilon \cdot \kappa$.

B. Saliency Calculation

The saliency measure in our work encodes the necessity to fixate a location in the observable area. It forms the basis for deciding the optimal gaze in the gaze selection step. Thus, the definition of the saliency measure is the most crucial element in implementing the gaze selection strategy.

As already discussed in the introduction, each manipulation task has specific requirements on the perceptual process. In order to allow the inclusion of constraints from manipulation tasks, we propose a saliency measure which can be configured for specific tasks. For this purpose, we introduce the task acuity a_i which allows to specify the required accuracy of an entity m_i within the environmental model. More precisely, the task acuity is interpreted as desired upper bound for the uncertainty in the localization of a memory entity. As such, the accuracy of the localization estimate resulting from the active perceptual process becomes the main driving force for gaze selection. In the following, we will derive a consistent way to embed the task acuity in a saliency measure for gaze selection.

The gaze selection strategy aims at reducing the overall localization uncertainty within the environmental model. For this purpose, the saliency is calculated for each memory entity m_i stored in the environmental model. A memory entity with high localization uncertainty Σ_{x_i} should thereby be assigned with a high saliency value in order to express the necessity for revalidation. In order to quantify the amount of uncertainty, the covariance matrix Σ_{x_i} needs to be mapped to a scalar value. A natural quantification of the amount of uncertainty is provided by the differential entropy, which is a generalization of the Shannon entropy to continuous probability distributions. Given that the localization uncertainty in our work is normally distributed, its differential entropy can be calculated in closed form using

$$u_i(t) = \frac{1}{2} \log \left[(2\pi e)^3 |\Sigma_{x_i}(t)| \right], \tag{1}$$

where $\Sigma_{x_i}(t)$ is the location uncertainty corresponding to the memory entity m_i.

Using $u_i(t)$ as saliency measure would result in a gaze selection strategy which aims at reducing the uncertainty of all memory entities irrespective of the manipulation task's requirements. In order to include these requirements, we express the task acuity $a_i(t)$ in the same domain as $u_i(t)$ using again the differential entropy

$$b_i(t) = \frac{1}{2} \log \left[(2\pi e a_i(t)^2)^3 \right]. \tag{2}$$

The resulting measure $b_i(t)$ encodes the desired upper bound for the entropy of the position estimate.

Putting equations (1) and (2) together yields the final saliency measure $s_i(t)$, where saliency is defined as difference between entropy resulting from the localization uncertainty and minimal entropy desired by the manipulation task

$$s_i(t) = u_i(t) - b_i(t). \tag{3}$$

The saliency measure $s_i(t)$ in relation to the localization uncertainty and the effect of a fixed task acuity a are illustrated in Fig. 4. The plot was generated using a univariate localization uncertainty with standard deviation σ for two different task acuities $a_1 = 10mm$ and $a_2 = 30mm$. The logarithmic shape of the entropy measure is desirable, since it results in a smaller validation effect for entities with higher localization uncertainty. The task acuity acts as shift and cut-off for the saliency and assures that once the localization uncertainty reaches the requested task acuity the saliency drops to zero. For values smaller zero we set $s_i(t) = 0$ in order to avoid negative saliency.

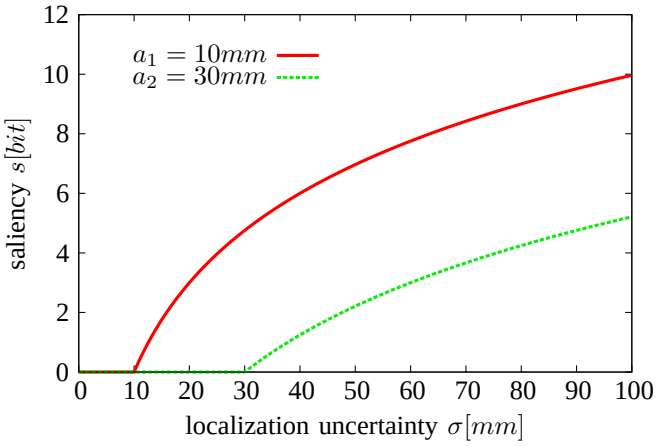

Fig. 4. Saliency in relation to localization uncertainty σ and task acuity a. The saliency measure drops to zero once the uncertainty reaches the requested task acuity.

In summary, the combination of differential entropy and task acuity yields a consistent integration of accuracy requirements in the saliency calculation. Moreover, the introduction of the task acuity renders the differential entropy usable at all for a saliency measure. A major drawback of the differential entropy, its definition on the interval $(-\infty, \infty)$, is compensated with the inclusion of the task acuity in equation (3). The task acuity limits the differential entropy to $[0, \infty)$ making it suitable for saliency calculation.

C. Gaze Selection

On a humanoid robot, a gaze can generally be realized by specifying the 6D pose of the camera system plus the version

and vergence parameters of the active cameras. Selecting gazes in this space would require to initiate full-body motions of the robot in order to achieve the optimal gaze. During a manipulation task, it is obviously not suitable to realize a gaze direction in this way, since its execution would interfere with the execution of the task. Rather, in order to not affect the manipulation task, we only consider the active head-eye system of the robot for realizing the optimal gaze.

For the selection of gaze directions, we further simplify the head-eye system in order to achieve a representation of gaze directions which allows computationally feasible online performance. For this reason, the gaze of the system is represented on a unit sphere with an origin at the center between both active cameras. The sphere representation allows to encode a gaze direction with the zenith θ and azimuth ϕ of the corresponding spherical polar coordinates. We omit the rotation around the tangential plane to the sphere, since this degree of freedom is hard to realize when using only the head-eye system. The unit sphere is represented using a spherical graph as illustrated in Fig. 5, where each of the equidistantly distributed nodes corresponds to one viewing direction of the active head-eye system.

In order to determine the optimal gaze, each node of the graph is assigned a rating based on the saliency measure introduced in the last section. The rating for a node with coordinates (θ, ϕ) is calculated as a weighted sum of saliencies:

$$r_{(\theta,\phi)} = \sum_{i \in 1 \ldots N} v_i(\theta, \phi) \cdot s_i. \tag{4}$$

The weight $v_i(\theta, \phi)$ encodes the visibility in the cameras of each environmental memory entity with the current gaze direction. In order to determine this visibility, a simplified camera model is used which approximates the view frustum of the cameras by a single cone. For each memory entity, such a cone is intersected with the sphere of gaze directions. Entities which are situated close to the limits of the cone are likely to be partially occluded and subject to lens distortion effects. Consequently, the visibility is attenuated with increasing distance to the cone center since localization performance decreases towards the limits. An example of resulting spherical graphs is illustrated in Fig. 5. For both cases, four memory entities were involved in the task. In the first case, two entities are situated at the same position in the center, resulting in an increased rating of the region. The second case illustrates how overlapping visibility regions can generate maxima on the sphere by considering the sum of saliencies. A gaze direction towards such a maximum allows to fixate multiple objects at the same time.

The representation of possible viewing directions as nodes in a spherical graph allows straightforward and efficient determination of optimal gazes by detecting the maximum on the spherical graph. However, as pointed out earlier, it is also a simplification of the active head-eye system since it assumes a fixed reference frame for the camera system during calculation of the rating. In reality however, the cameras move resulting in inaccurate approximations of the real saliency distribution. In order to compensate this effect, we use an iterative procedure for calculating the corrected optimal gaze direction: First, the rating in equation (4) is performed using the current posture of the head-eye system. Then, a candidate gaze direction is determined by searching the node with maximum rating on the graph. Using this gaze direction as input, the new posture of the head-eye system is calculated using inverse kinematics. With this new posture, the rating procedure is repeated with a sphere centered at the new reference frame. The iterative procedure stops when the posture of the head-eye system does not change anymore. The optimal gaze direction is then approximated by the node corresponding to the maximum peak on the spherical graph. In practice, it never occurred that more than one additional iteration was required, as the posture of the head-eye system does not change significantly during the iterations.

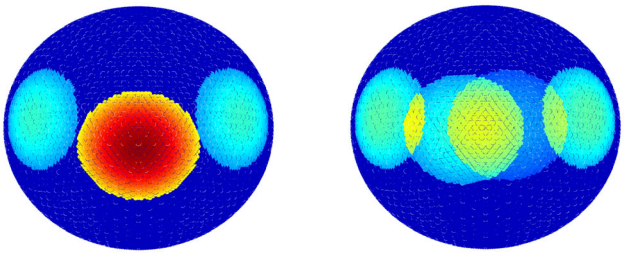

Fig. 5. Gaze directions are represented as a spherical graph with equidistantly distributed nodes. The nodes are rated according to the saliency of visible memory entities when taking the corresponding gaze direction. Thereby, the visibility of the objects in the cameras is approximated by viewing cones with decreasing localization reliability towards the limits. The node with the maximum rating is chosen as optimal gaze direction. Both graphs illustrate an example task involving four objects, where in the left graph two objects completely overlap in the center.

IV. Application in Bimanual Manipulation

In this section, the proposed method for gaze selection is applied in a bimanual manipulation task on the humanoid robot ARMAR-III. In previous work, we demonstrated the execution of bimanual tasks using visual servoing techniques [21]. Thereby, the wide operational space necessitates head-eye movements in order to observe all objects involved in the task. In the previous work, the gaze selection was accomplished in a manner specific for the task. Based on such an application we will demonstrate, how the proposed gaze selection mechanism allows to produce gaze sequences for a given task in a more consistent and general way.

In the following, a brief introduction to the implementation of the bimanual manipulation task is given. Subsequently, the motion models required to complete the definition of the gaze selection mechanism for this task are introduced.

A. Bimanual Visual Servoing

In our previous work we solved bimanual manipulation tasks such as pouring or carrying big objects using position-based visual servoing. The benefits of applying visual ser-

voing techniques lie in their robustness towards inaccuracies in the kinematic model of the system. Thus, task execution based on visual servoing shares the goal with our approach in being applicable in the presence of inaccuracies, making it well suited for complex integrated platforms such as humanoid robots. In contrast to planned motions, trajectories resulting from visual servoing are not guaranteed to be collision-free, thus limiting its applicability to tasks which do not include possible collision with obstacles. Nevertheless, the application of visual servoing provides a feasible test-bed for the proposed gaze selection strategy for two reasons: First, it allows the execution of manipulation tasks including multiple objects. Second, the feedback from the perceptual processes can easily be integrated in the execution by directly using the content of the environmental model as input. For the integration with motion planning, a suitable plan monitoring and re-planning step would need to be implemented which goes beyond the scope of this paper.

For bimanual manipulation we observe the two robot hands and two target objects with the proposed gaze selection mechanism. Their position and orientation from the environmental model are then used in the position-based visual servoing approach. The trajectory is generated by successively reducing the distance of robot hands and target objects using differential inverse kinematics as discussed in [21].

To realize the desired gaze directions, we use the active head of ARMAR-III offering 3 DoF in the neck, a common tilt and a separate pan for both cameras. The joint angles for these 6 DoF are calculated by solving the inverse kinematics problem using optimization. We use an objective function that assures the correct gaze direction and generates natural looking postures. The inverse kinematics solution is calculated using gradient-free local optimization. The kinematic model for the head-eye system is calibrated offline using the approach proposed in [22]. The same model is used to retain stereo perception while the extrinsic camera parameters change.

B. Motion models

The prediction of motion within the environmental model is necessary since not all objects are visible to the cameras all of the time during the manipulation task. Thereby, two kinds of motion need to be considered: The motion of the head-eye system and the motion of entities physically controlled by the robot such as its hands. Both motions can be approximated by reading the joint encoders of the robot. Due to remaining inaccuracies in the positioning and in the kinematic model of the system, these measurements are not entirely correct, thus necessitating the inclusion of motion uncertainty in the motion model.

In order to calculate uncertainties implied by the head motion, we use the frame of the left camera as reference. This reference frame changes during head-eye movements. In order to cope with the inaccuracies in the kinematic model and in the positioning, we assume additive Gaussian noise in the joint angles of the head-eye system. Using the unscented transform, this noise is passed through the system in order to retrieve an estimate of the uncertainty implied by the head motion to the position of an entity. The resulting covariance matrix is used in the Kalman filter prediction step for the entities.

In order to define motion models for objects in the scene, we differentiate between the objects which are controlled by the robot, as e.g. its hands, and objects which are target of the manipulation. In the current setup, the objects the robot wants to manipulate are assumed to be static, i.e. they do not move on their own. Consequently, the pose and the associated uncertainty do not change over time and no additional uncertainty needs to be considered in the motion model. In contrast, for the hands of the robot again the inaccuracies in the kinematic model need to be considered with respect to the joint encoder readings of the arm. As for the head-eye system, we make use of the unscented transform in order to calculate the uncertainty of motion implied by the model inaccuracies. The pose of the objects is then predicted using the resulting covariance matrix within the Kalman filter prediction step.

V. Experimental Evaluation

A. Experimental Setup

In the following, the proposed gaze selection mechanism is evaluated in a typical kitchen environment task on the humanoid robot ARMAR-IIIa. A complex task in this environment which involves multiple objects and requires two arms is the pouring scenario, where the robot pours juice from a container in one hand into a cup held in the other hand. In the context of gaze selection, the first phase of this task, the approach and grasping of both objects with the five-fingered hand is the most demanding part, since four distinct objects need to be observed: the cup, the juice, and both hands. Consequently, we restrict the experiments to this first phase.

For all experiments, we used the green cup and the vitamin juice placed on a table in front of ARMAR-IIIa as shown in Fig. 1. The positions of cup and juice were varied within the workspace of the robot. The desired grasps for both objects and thus the target poses for visual servoing were predefined relative to the objects' local coordinate frames.

In order to initiate the task execution, an estimate of the position of both hands and both objects needs to be provided as prior in the environmental model. For the hands, we use the pose from the kinematic model as initial estimate. For both objects, a position on the table in front of the robot is provided as initial estimate. We choose a conservative initial localization uncertainty with a standard deviation of $500\,mm$ in all directions for the hands whereas the objects are assigned with a higher uncertainty corresponding to a standard deviation of $1000\,mm$. The execution itself is started once the localization uncertainty of all objects drops below a standard deviation of $d_{max} = 50\,mm$.

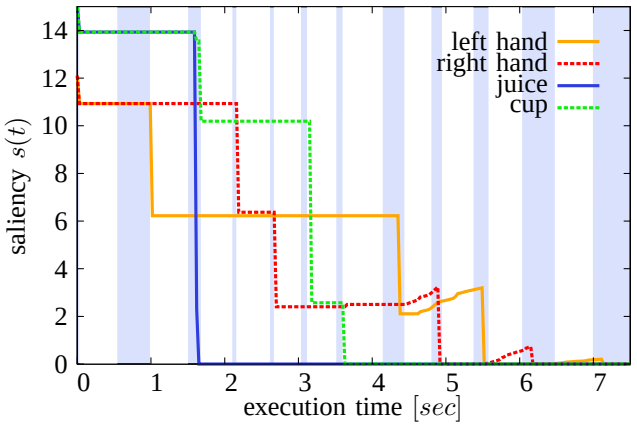

Fig. 6. The bimanual visual servoing task requires pose estimates of both hands and both involved objects with the necessary accuracy. The saliency measure $s(t)$ encodes the necessity to perform a localization of an element. For a task acuity of $a = 5\,mm$, the plot shows the development of the saliency over a complete approach and grasp phase. Blue regions indicate phases where localizations are performed while white regions denote head movements to selected gaze directions.

B. Saliency during Manipulation

Fig. 6 illustrates the course of the saliency measure $s_i(t)$ for all objects involved in the task over one complete approach and grasp phase. The task duration from the initial localization of all objects until successful grasp execution is about seven seconds, where the first four seconds are required in order to successively reduce the uncertainty of all objects under the limit d_{max}. Once the uncertainty drops below this limit, the visual servoing procedure is started until the target is reached and the grasp is executed.

During the execution of the manipulation task, successive fixations of the involved objects are performed according to the gaze selection mechanism. After each redirection of the gaze, the object localization modules are triggered in order to determine the pose of all visible objects and to update the environmental model. The localization is stopped once a new gaze direction is requested by the gaze selection mechanism. The time intervals, when localization is performed are marked with blue background in Fig. 6. The update of the environmental model is performed delayed, once the object localization processes finish the computation of the pose.

The plot clearly illustrates how the proposed approach allows to reduce the localization uncertainty by actively redirecting the gaze appropriately. Each localization results in a reduction of the uncertainty of the observed entity. Once the desired accuracy, defined by the task acuity a is reached, the saliency $s(t)$ drops to zero. For the cup and the juice box, the saliency drops to zero after a few localizations and remains there. For the two robot hands, the uncertainty in the pose estimate increases due to the movement of the robot arms, accompanied by an increase of the associated saliency. As expected, the gaze selection mechanism compensates this increase by initiating additional localizations of the robot hands.

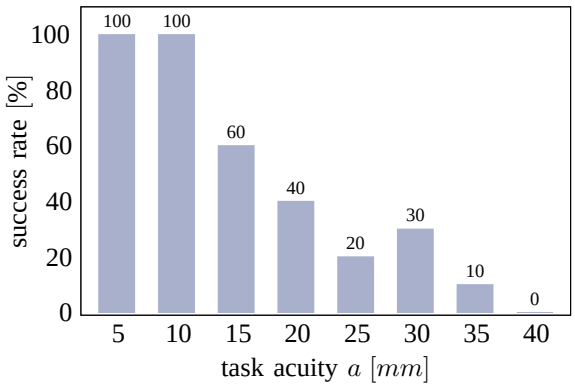

Fig. 7. Success rate of the bimanual visual servoing and grasping task in relation to the selected task acuity. For each setting of the task acuity 10 trials were executed. For a task acuity of $5\,mm$ and $10\,mm$ all trials could be carried out successfully. With increasing task acuity, the success rate drops.

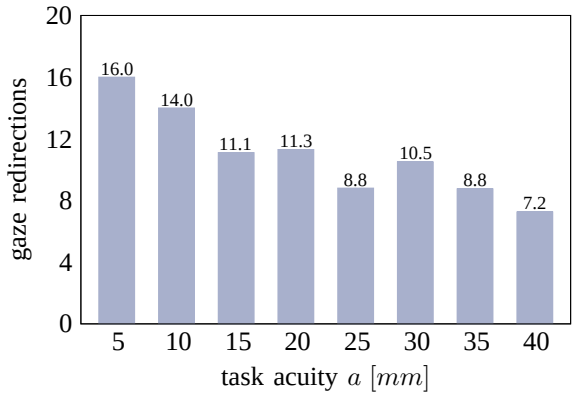

Fig. 8. Number of head movements in relation to the selected task acuity. For lower task acuity, more head movements need to be executed in order to achieve the required acuity of the environmental model.

C. Influence of the Task Acuity

In order to evaluate the feasibility of the task acuity as means of integrating manipulation task constraints with the gaze selection approach, the bimanual task was performed several times with varying task acuity settings. Thereby, a task acuity from the range $a_{min} = 5\,mm$ to $a_{max} = 40\,mm$ was used with an increment of $5\,mm$. For each setting from this range, the manipulation task was executed ten times. After each execution, the success was assessed by lifting both objects. The resulting success rate in relation to the tested task acuity is illustrated in Fig. 7. For the task acuity settings of $5\,mm$ and $10\,mm$ the execution succeeded in all ten trials. The success rate drops with increasing task acuity until no successful execution is possible with the maximum tested task acuity of $40\,mm$. These results are into accordance with our expectations, since a minimal required localization accuracy of $10\,mm$ for both – robot hand and object – seems to be feasible in order to produce a stable grasp of the object.

In addition to the success rate we investigated the number of gaze redirections required to perform the manipulation task in relation to the task acuity. While higher task acuity

obviously lead to better pose estimates, it also necessitate the execution of more gaze shifts. An appropriate selection of the task acuity should minimize the number of gaze redirections required but still retain the ability to successfully accomplish the task. As illustrated in Fig. 8 the number of required gaze redirections drops with increasing task acuity. Considering the number of redirections, the optimal choice for the task acuity in the bimanual visual servoing and grasping task amounts to $a = 10\,mm$.

VI. Conclusion

In this work, we introduce a gaze selection approach tailored for manipulation tasks on humanoid robots. The applicability in a manipulation task influences the proposed approach in several ways: First, the proposed environmental model allows for dynamic entities by the inclusion of motion models. Second, the saliency calculation includes accuracy constraints from the manipulation task by means of the task acuity. Finally, the gaze selection and redirection is implemented using only the DoF of the head-eye system in order to not interfere with the manipulation task.

The gaze selection approach was evaluated in a bimanual visual servoing task involving four objects that would fail without the application of active gaze control. Using the proposed gaze selection approach, the task could be accomplished with a success rate of 100%. Further, we could demonstrate that the inclusion of the task acuity allows to intuitively configure the perceptual processes. The optimal trade-off between accuracy and number of required gaze redirections was achieved for a task acuity of $a = 10\,mm$. Being able to perform the task with this accuracy is feasible as well as intuitive.

While the evaluation in the context of a bimanual visual servoing task is suitable to demonstrate the feasibility of the approach, it only covers a fraction of possible applications for the proposed gaze selection mechanism. For complex manipulations involving obstacles and dexterous abilities, motion planning is required in order to achieve an executable and collision-free trajectory. Having performed motion planning, the motion models as well as the required task acuity could be directly derived from the resulting trajectory and its relation to the world model.

In summary, the described gaze selection approach enables the robot to exploit two of its key capabilities, manipulation and active gaze control, in an integrated fashion. The inclusion of constraints based on the task acuity allows the adaptation of the generated gaze sequence in order to support successful task execution. Thus, the proposed approach substantially contributes in increasing the autonomy of humanoid platforms.

Acknowledgment

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement №270273 (Xperience).

References

[1] T. Asfour, P. Azad, N. Vahrenkamp, K. Regenstein, A. Bierbaum, K. Welke, J. Schröder, and R. Dillmann, "Toward humanoid manipulation in human-centred environments," *Robotics and Autonomous Systems*, vol. 56, no. 1, pp. 54–65, 2008.

[2] S. Thrun, W. Burgard, and D. Fox, *Probabilistic robotics*, ser. Intelligent Robotics and Autonomous Agents. MIT Press, 2005.

[3] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *pami*, vol. 20, no. 11, pp. 1254–1259, 1998.

[4] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Proc. SPIE 48th Annual International Symposium on Optical Science and Technology*. SPIE Press, 2003, pp. 64–78.

[5] A. Ude, V. Wyart, L.-H. Lin, and G. Cheng, "Distributed visual attention on a humanoid robot," in *humanoids*, 2005, pp. 381–386.

[6] S. Frintrop, E. Rome, and H. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," *ACM Transactions on Applied Perception*, vol. 7, no. 1, pp. 1–39, 2010.

[7] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," in *International Joint Conference on Artificial Intelligence*, 1999, pp. 1146–1153.

[8] M. Bjorkman and D. Kragic, "Combination of foveal and peripheral vision for object recognition and pose estimation," in *humanoids*, 2004, pp. 5135–5140.

[9] D. Figueira, M. Lopes, R. Ventura, and J. Ruesch, "Towards a spatial model for humanoid social robots," in *Progress in Artificial Intelligence*. Springer Berlin / Heidelberg, 2009, pp. 287–298.

[10] K. Welke, "Memory-based active visual search for humanoid robots," Ph.D. dissertation, Karlsruhe Institute of Technology (KIT), Computer Science Faculty, Institute for Anthropomatics (IFA), 2011.

[11] O. Stasse, T. Foissotte, D. Larlus, A. Kheddar, and K. Yokoi, "Treasure hunting for humanoid robots," in *Workshop on Cognitive Humanoids Vision,*, 2009.

[12] B. Rasolzadeh, M. Björkman, K. Huebner, and D. Kragic, "An active vision system for detecting, fixating and manipulating objects in real world," *The International Journal of Robotics Research*, vol. 29, pp. 133–154, 2009.

[13] F. Seara, J. Strobl, K. H. Martin, and E. Schmidt, "Task-oriented and situation-dependent gaze control for vision guided autonomous walking," in *humanoids*, 2003.

[14] S. Kohlbrecher, A. Stumpf, and O. von Stryk, "Grid-based occupancy mapping and automatic gaze control for soccer playing humanoid robots," in *Proc. 6th Workshop on Humanoid Soccer Robots,*, 2011.

[15] K. Okada, M. Kojima, S. Tokutsu, Y. Mori, T. Maki, and M. Inaba, "Task guided attention control and visual verification in tea serving by the daily assistive humanoid HRP2JSK," in *iros*, sept. 2008, pp. 1551 –1557.

[16] P. Michel, C. Scheurer, J. Kuffner, N. Vahrenkamp, and R. Dillmann, "Planning for robust execution of humanoid motions using future perceptive capability," in *Proceedings of the IEEE/RSJ IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'07)*, October 2007, pp. 3223–3228.

[17] P. Azad, T. Asfour, and R. Dillmann, "Stereo-based 6D Object Localization for Grasping with Humanoid Robot Systems," in *iros*, San Diego, USA, October 2007, pp. 919–924.

[18] ——, "Accurate Shape-based 6-DoF Pose Estimation of Single-colored Objects," in *iros*, St. Louis, USA, October 2009, pp. 2690–2695.

[19] S. J. Julier and J. K. Uhlmann, "A new extension of the kalman filter to nonlinear systems," 1997, pp. 182–193.

[20] T. Kirubarajan and B. Y. Shalom, "Probabilistic data association techniques for target tracking in clutter," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 536–557, 2004.

[21] N. Vahrenkamp, C. Böge, K. Welke, T. Asfour, J. Walter, and R. Dillmann, "Visual Servoing for Dual Arm Motions on a Humanoid Robot," in *humanoids*, Paris, France, Dec. 2009, pp. 208–214.

[22] K. Welke, M. Przybylski, T. Asfour, and R. Dillmann, "Kinematic calibration for saccadic eye movements," Institute for Anthropomatics, Universität Karlsruhe, Tech. Rep., 2008.