Visual Servoing for Dual Arm Motions on a Humanoid Robot

Nikolaus Vahrenkamp* Christian Böge*‡ Kai Welke* Tamim Asfour* Jürgen Walter‡ Rüdiger Dillmann*

*Institute for Anthropomatics
University of Karlsruhe
Haid-und-Neu Str. 7
76131 Karlsruhe, Germany
{vahrenkamp,welke,asfour,dillmann}@ira.uka.de

‡Department of Mechanical Engineering and Mechatronics
Karlsruhe University of Applied Sciences
Moltkestr. 30
76133 Karlsruhe, Germany
{christian.boege,juergen.walter}@hs-karlsruhe.de

***Abstract*— In this work we present a visual servoing approach that enables a humanoid robot to robustly execute dual arm grasping and manipulation tasks. Therefore the target object(s) and both hands are tracked alternately and a combined open- / closed-loop controller is used for positioning the hands with respect to the target(s). We address the perception system and how the observable workspace can be increased by using an active vision system on a humanoid head. Furthermore a control framework for reactive positioning of both hands using position based visual servoing is presented, where the sensor data streams coming from the vision system, the joint encoders and the force/torque sensors are fused and joint velocity values are generated. This framework can be used for bimanual grasping as well as for two handed manipulations which is demonstrated with the humanoid robot Armar-III that executes grasping and manipulation tasks in a kitchen environment.**

I. INTRODUCTION

Humanoid robots are developed to work in human-centered environments and to assist people in doing the housework, e.g. cleaning the dishes or serving a meal. To enable the robot operating in a safe and robust manner, a lot of components have to collude and to operate cooperatively. In this paper we show how dual arm tasks, like bimanual grasping or dual arm manipulations, can be executed with high accuracy. The fusion of multiple modalities combined with position based visual servoing allows an exact positioning of both arms in workspace and thus enables the robot to execute dexterous dual arm tasks. The proposed algorithms are implemented and evaluated on the humanoid robot Armar-III [1].

The perceptional components needed for the dual arm visual servoing approach are discussed in section II. In section III the general approach for visually controlled movements is described and the extensions for dual arm manipulation tasks are discussed in section IV. Finally, in section V the application of the proposed algorithms is attested by two experiments on the humanoid robot Armar-III.

II. PERCEPTION

Target positions of objects and the positions of both hands have to be made available to the visual servoing approach. Therefore, the images from the stereo camera pair are processed with appropriate recognition and localization algorithms. Both, the pose of the target objects and the position of the end effectors are determined in Cartesian space. For fast and robust object recognition and localization, the approach proposed in [2] is deployed. The marker-based end effector localization is performed similar to the approach presented in [3]. In the following, we will describe extensions made to this prior work for the scenario at hand.

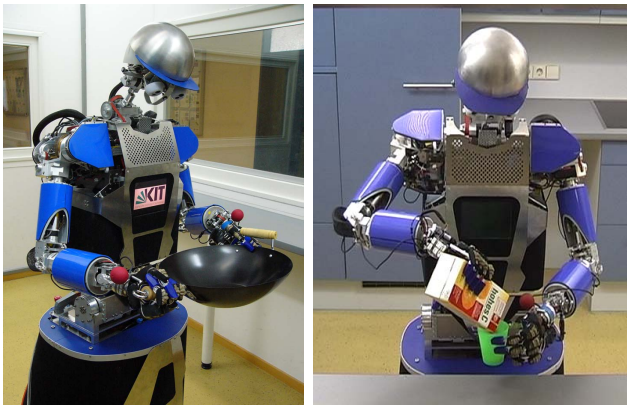


Fig. 1. Armar-III executing bimanual manipulations in the kitchen environment.

A. Object recognition and localization for Dual Arm Grasping Tasks

The vision framework of Armar-III offers methods for recognition and localization of every day kitchen objects like cups or cereal boxes [2]. The algorithms can handle uniformly colored and textured objects as long as they are fully visible in the stereo camera images. If the robot is supposed to handle large objects, like the wok that is used in the experiments (see Fig. 2(a)), the object is not always visible as a whole in both camera images during the task. Therefore, the wok is decomposed into two smaller objects (the handles) which are easy to track, in order to avoid loosing visual target information. The two handles are good features, since they mark the target of grasping actions and can be tracked independently.

B. Tracking Multiple Objects

In the case of servoing two hands using visual feedback, tracking by turns of both hands is necessary since the area covered by vision is limited by the camera's field of view. In order to enlarge this visually observable area, we move the

gaze to relevant locations in the scene (see section IV-A). Using this behavior, a continuous tracking of objects alone is not possible since the field of view changes frequently. Therefore, a combined closed- and open-loop control scheme is proposed in order to gain as much information as possible.

If the marker of the hand is visible in frame t, the joint angles of the robot are used to compute the kinematic TCP pose p^t_{kin} and the offset to the visually retrieved TCP pose p^t_{vis} is computed with

$$\Delta p^t_{kin} = p^t_{kin} - p^t_{visual}. \quad (1)$$

The offset Δp^t_{kin} is only valid in the current configuration of the robot.

In cases where the hand marker is not visible, a pose estimate p_{est} can be calculated using the offset Δp^t_{kin} in the following way:

$$p^{t+1}_{est} = p^{t+1}_{visual} + \Delta p^t_{kin}. \quad (2)$$

For small movements of the hand, p^{t+1}_{est} is a good estimate for the TCP pose.

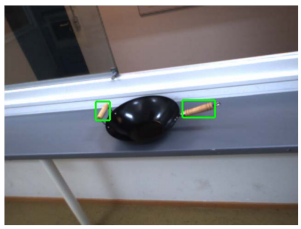
For target objects we use a similar approach if recognition in the current camera images is not possible. The last visually verified pose is used to derive the estimated object pose in succeeding frames. For moving object, the last known velocity is considered in order to derive the estimated object pose.

C. Perception with Active Cameras

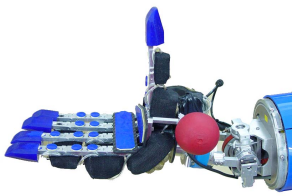
The gaze shifts required for adapting the camera's field of view to the needs of the task are implemented using the active head of ARMAR-III [4] which offers 7 DoF including 3 DoF for eye panning and common tilt. In order to allow object recognition and hand localization with actuated eyes, the kinematic model of the head eye system is determined offline using the visual calibration approach proposed in [5]. The calibration process is a necessary prerequisite since the optical centers of the cameras can not be determined based on CAD-models from the construction process. With the resulting kinematic model, stereo vision is possible with moving eyes. More details on the execution of gaze shifts are discussed in IV-A.

III. Visual Servoing

Hand-eye calibration techniques can be used to handle the errors between visual object localization and physical positioning of the end effectors. This requires a high repeatability of the execution which may not be guaranteed when using wire-driven lightweight arms, e.g. like the arms of humanoid robot Armar-III. Nevertheless the execution must handle error-prone sensor channels as well as the inaccurate execution of actions. The visually controlled execution offers a possibility to supervise the execution of grasping or manipulation tasks without the need for exact hand-eye calibration. Instead the robot operates on the relative distance between tracked objects and hands and thus a closed-loop controller for robust and secure execution of movements can be constructed.



(a) The pose of the wok is derived from the two poses of the handles.



(b) One hand of the humanoid robot Armar-III with attached marker.

Fig. 2. The perception system is used to track objects and the robot's hand.

A. Visual Servoing Approaches

Visual servoing approaches usually track the tool center point (TCP) and/or features of a target object and the TCP is moved to a desired position and orientation [6]. The term visual servoing was introduced by Hill and Park in 1979 to describe a visual-closed control loop [7] and over the years a lot of setups for perception (e.g. camera in hand or stereo camera systems) as well as different control strategies have been developed [6], [8]. The main strategies for visually controlling the actuators are the three concepts *Image-Based Visual Servoing (IBVS), Position-Based Visual Servoing (PBVS)* and *Hybrid Visual Servoing (2.5D VS)* which is a combination of the two approaches IBVS and PBVS [9].

In IBVS approaches the position of the TCP in the camera image(s) is derived by image features, and through the movements of these features a local Image-Jacobian can be retrieved [10], [11], [12], [13]. By computing the pseudo-inverse of this Jacobian, the desired 2D movements of the features in the image plane can be used to calculate the desired TCP movements. These approaches work without any stereo computations and thus with a single camera and with multi camera systems. Problems can arise, if the features leave the field of view and some problems like the *Chaumette Conundrum* cannot be solved [14].

Position Based Visual Servoing is used to control the pose of an end effector in workspace. If the object and the kinematic structure of the robot are known and if the pose of the end effector as well as the workspace pose of the target object can be retrieved visually, it is possible to build a control structure for moving the arm to a desired goal. The visual determined error between the current and the target pose are used as input for the control loop and the Pseudinoverse Jacobians are used to derive joint velocities.

B. Visually Controlled Movements of one Hand

The work presented in this paper is based on the single arm approaches of [3] where we described a robust system for visually controlled grasping with one hand. The proposed position based visual servoing approach is used to control an Cartesian error between the hand and a target object to zero. The position of the hand is tracked by an artificial marker (see Fig. 2(b)) and the orientational component of the hand's pose is derived by forward kinematics methods. The Cartesian error between the current and the desired hand

pose is used for computing joint velocity values which are calculated via the Pseudionverse of the robot's Jacobian. The system was successfully implemented on Armar-III for grasping cups, boxes and door handles.

When two hands are needed to grasp an object the complexity of the visual servoing framework increases since two kinematic chains have to be moved and due to the limited view of the robot an alternating observation of multiple workspace positions is required.

C. Positioning the Robot via Visual Servoing

Dual arm grasping and manipulation tasks require a good robot position with respect to the manipulation object in order to be able to reach the target poses with both hands. Visual servoing can be used to control the platform of the humanoid robot Armar-III to guide it to a pose from where the manipulation task can be executed. In our experiments we move to a position offset relative to the center position of both dual arm target positions. The orientation of the platform is set to be parallel to the table the object is placed on and does not depend on the object's orientation. In a later stage, the redundancy of the arms is used to compensate any orientational differences. The difference between the current and the target position is used as input for the velocity controller of the platform.

IV. Multi Target Visual Servoing

Han et. al. use Image Based Visual Servoing techniques with a camera-in-hand system for controlling a dual arm assembly robot [15], [16]. In [17] also IBVS methods are used to control the pose of two instruments for minimally invasive surgery. The work presented in [18] deals with IBVS methods to realize an automated object capture system with a two-arm flexible manipulator. Contrary to these approaches we propose a Position Based Visual Servoing system which is able to track and control two end effectors of a humanoid robot. With the dual-arm Visual Servoing algorithms an exact positioning of both arms relative to each other or to an object is possible and thus the robot is enabled to execute dual arm grasping and manipulation tasks.

A. Adapting the Gaze

During dual arm manipulation tasks, almost the complete workspace of both arms has to be observed within the robot's camera system. Therefore, five DoF of the active head are used in order to adapt the gaze to regions of interest such as hands and target objects. Two DoF of the neck are not necessary to cover the workspace and remain fixed.

The calibrated kinematic model of the head-eye system (see II-C) is used to solve the inverse kinematics problem and direct the gaze of the system in order to fixate target objects. The neck DoF of the head are only used, if the fixation point is not reachable with the 3 DoF of the eye system.

While performing gaze shifts, the visually determined positions are not updated in order to avoid motion blur. Updates only take place if the angular velocity of the gaze is below a certain threshold.

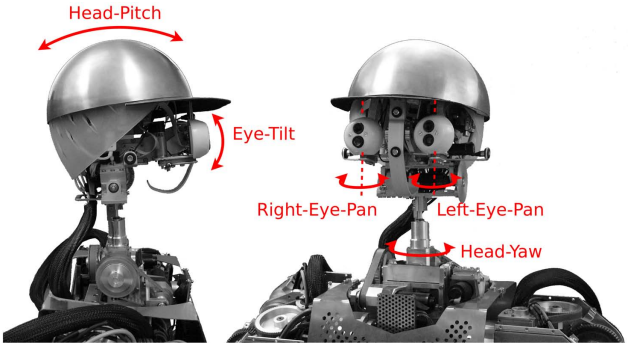


Fig. 3. The actuated joints of the head: Two DoF of the neck (Head-Pitch and Head-Yaw) and three DoF of the eyes (Eye-Tilt, Left-Eye-Pan and Right-Eye-Pan).

B. Allocation of Hand Markers

In case of tracking the markers of the hands, the determined markers have to be registered to the left or right hand. Therefor estimated hand marker positions are calculated for both hands in the same way it is done with the tcp positions as described in II. The allocation is done by calculating the Cartesian distance of the estimated positions and the visually determined marker positions. Markers within a given threshold around the kinematic hand pose are allocated to the nearest hand and their 3D position is used for further hand pose calculations. The final TCP hand pose is then calculated by adding a fixed Cartesian offset in the hand's coordinate system, describing the distance between the marker and the center of the TCP, to the marker pose and the orientational component is derived by forward kinematics methods.

C. Object-Relative Approach Poses

For grasping two taught target poses p_{pre}^{obj} and p_{grasp}^{obj} are used for each hand. These poses are defined relatively to the pose of the object and thus the final target poses are calculated by combining them with the object's pose (see eq. 3).

The grasping action is performed by first moving the TCP toward the pre-pose p_{pre}^{world} and then toward the final grasping-pose p_{grasp}^{world} (see Fig. 4). This procedure enables movements with a defined approaching direction $v_{grasp} = p_{grasp}^{world} - p_{pre}^{world}$ in order to avoid collisions with the target object during approaching.

$$\begin{aligned} p_{pre}^{world} &= p_{obj}^{world} \cdot p_{pre}^{obj} \\ p_{grasp}^{world} &= p_{obj}^{world} \cdot p_{grasp}^{obj} \end{aligned} \tag{3}$$

The pre- and grasp-poses are taught using a zero-force controller to move the hands to the desired poses relatively to the object. After that, the translational and rotational offset between hand and object(s) are determined visually (see eq. 4), which means that we use the same algorithms for localizing that are used later during visual servoing.

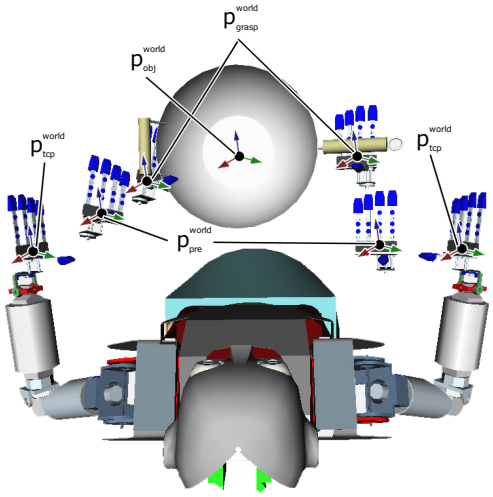


Fig. 4. The wok with associated pre- and grasping-poses for each hand.

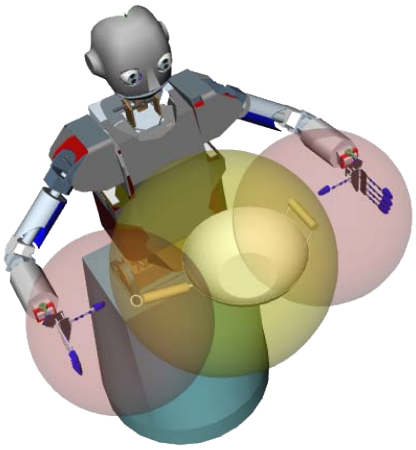


Fig. 5. Areas of interest for visual servo control of two hands relative to one object for grasping

$$p_{grasp}^{obj} = (p_{obj}^{world})^{-1} \cdot p_{tcp}^{world} \tag{4}$$

D. Bimanual Visual Servoing For Grasping

Exact positioning by visual servoing for moving one hand toward a relative pose with respect to an object requires a determination of the positions of the object and the hand in the same stereo image pair since accuracy problems can come up when using different pairs of images for determining these positions. These problems can be caused by imprecise encoders, slightly inexact kinematic models or bad calibrations.
So for using two hands, e.g. for grasping, an exact positioning of both hands with respect to one or two objects is needed. Therefor the relative relation (relative pose?) of the hand to the object is more important than the absolute pose in the world and thus the visually determined hand and object pose should be performed on the same image data.

1) Initialization: At the beginning of a dual arm grasping task the orientation of the object is determined by looking to the central area of interest as shown in illustration 5. After that, only the red areas of interest between the hands and the object borders are used to update the positions of the hand with respect to the object. In case of grasping two objects the orientations of both objects are determined.

2) Dual Arm Visual Servoing: The bimanual visual grasp controller continuously determines the object and hand poses as explained in section II and in case the tracking targets are too far away from each other the gaze is moved to alternately focus the left and the right points of interest (see section IV-A). Depending on the current state the Cartesian distance Δx to p_{pre} or p_{grasp} is determined for the left and the right hand and via the Pseudoinverse Jacobian the joint velocities are calculated (see eq. 5). In case Δx falls below a threshold, the next pose is selected or if the hands are already at the final grasping position the fingers are closed.

$$\begin{aligned} \dot{\boldsymbol{q}}_{left} &= k * J^{+}(\boldsymbol{q}_{left}) * \Delta \boldsymbol{x}_{left} \\ \dot{\boldsymbol{q}}_{right} &= k * J^{+}(\boldsymbol{q}_{right}) * \Delta \boldsymbol{x}_{right} \end{aligned} \tag{5}$$

For security reasons the impacting forces at the wrist are measured and in case a threshold is exceeded, Δx is adjusted to counteract these forces. This security behavior can result in a situation where the desired goal is not reachable because of the measured forces, but since there are no forces expected, maybe the environmental situation has changed or a human operator is guiding the robot in order to avoid an obstacle. In these cases the controller is able produce an intuitive reaction on interactions with the environment.

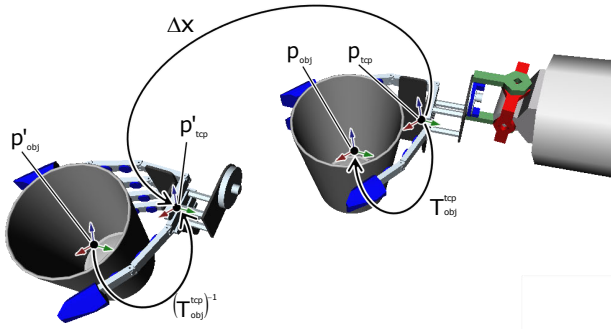


Fig. 6. The current and goal coordinate systems for visually controlled manipulation tasks.

E. Visually Controlled Dual Arm Manipulations

If one or two objects are grasped with the described visual servoing approaches, the target object(s) are difficult to track in general. But since the hand tracking is independent from the hand posture p_{tcp} it is possible to derive current object pose(s) p_{obj} from the visually observed hand positions (see eq. 6).

$$p_{obj} = p_{tcp} \cdot T_{obj}^{tcp} \tag{6}$$

The relation T_{obj}^{tcp} between hand and object can be stored during the grasp process or, if the discussed algorithms are used for grasping the object, T_{obj}^{tcp} is the inverse of p_{grasp}^{obj}, the relation between object and TCP when applying a grasp. To manipulate the object's pose, the algorithms used for grasping can be slightly modified in order to apply a coupled or a decoupled dual arm manipulation.

In case of a coupled manipulation (e.g. when a large object, like the wok used in section V-A, is grasped) the target poses for both TCPs are calculated in order to retrieve the translational and rotational relationship to each other. Through this relationship a virtual object pose can be derived and if this pose is modified, the corresponding hand positions can be calculated and used as new targets for the visual servoing algorithms. Since the resulting grasp forms a closed kinematic chain, forces and torques can appear during the execution of the manipulation task. To deal with that we counteract by slightly changing the hand poses in a way that the forces are reduced in case a threshold is exceeded (see [3]).

When considering decoupled manipulations (e.g. when two independent objects are grasped) the targets for visual servoing are generated by deriving a target pose p'_{tcp} from the desired object target pose p'_{obj} (see eq. 7). Since the current object pose p_{obj} is known (eq. 6), $\Delta \boldsymbol{x}$ can be calculated easily for both hands and used as input for the visual servoing controller.

$$p'_{tcp} = p'_{obj} \cdot (T_{obj}^{tcp})^{-1} \tag{7}$$

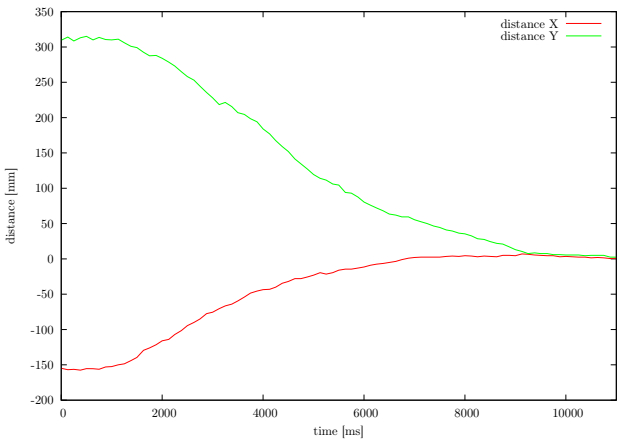


Fig. 7. Visual controlled positioning of the robot's platform toward the position were the object is graspable with both hands. The servoing distances for platform control are depicted.

V. Experiment on a Humanoid Robot.

In this section the results of two experiments with the humanoid robot Armar-III are presented. The discussed algorithms are fully integrated in the software framework of Armar-III and the experiments show how dual arm grasping and manipulation actions can be executed robustly.

A. Grasping a Wok with Both Hands

In this experiment a visually guided dual arm grasping task is realized. The humanoid robot is supposed to locate the wok in a kitchen environment and then to use the visual servo platform control (see section III-C) for guiding the robot to a position that allows to grasp the wok with both arms. In figure 7 the Cartesian servoing distances for platform control during approaching the wok, which is located at the sideboard (see Fig. 2(a)), are illustrated. Finally the dual arm visual servoing techniques of section IV are used for grasping the wok robustly with both hands (see Fig. 10). The servoing distances measured during grasping are shown in figure 8 and 9.

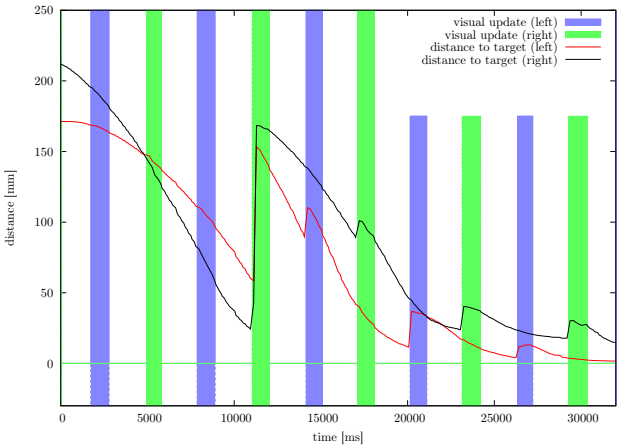


Fig. 8. The Cartesian distances between target and actual TCP position of left and right hand. The distances are estimated alternately and the visual update phases are marked by green and blue bars. The abrupt rising of the distances at $t = 11s$ is caused by changing the target from pre- to grasping poses.

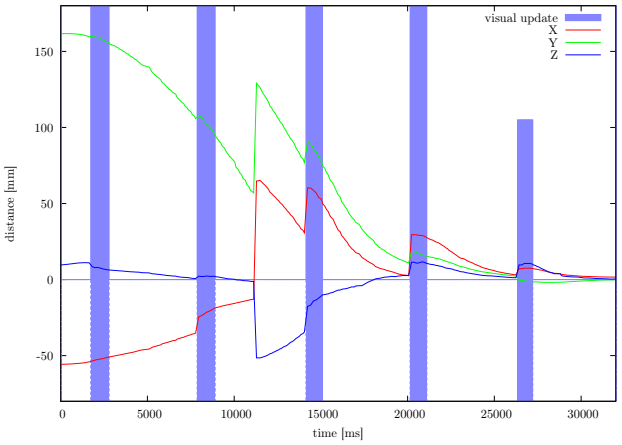


Fig. 9. The x,y and z distances between target and current position of the left hand during execution of dual arm visual servoing. The abrupt rising of the distances at $t = 11s$ is caused by changing the target from pre- to grasping poses.

B. A Visually Controlled Dual Arm Manipulation Task

In this experiment Armar-III should grasp a cup and a beverage carton in order to pour a drink. Both objects are

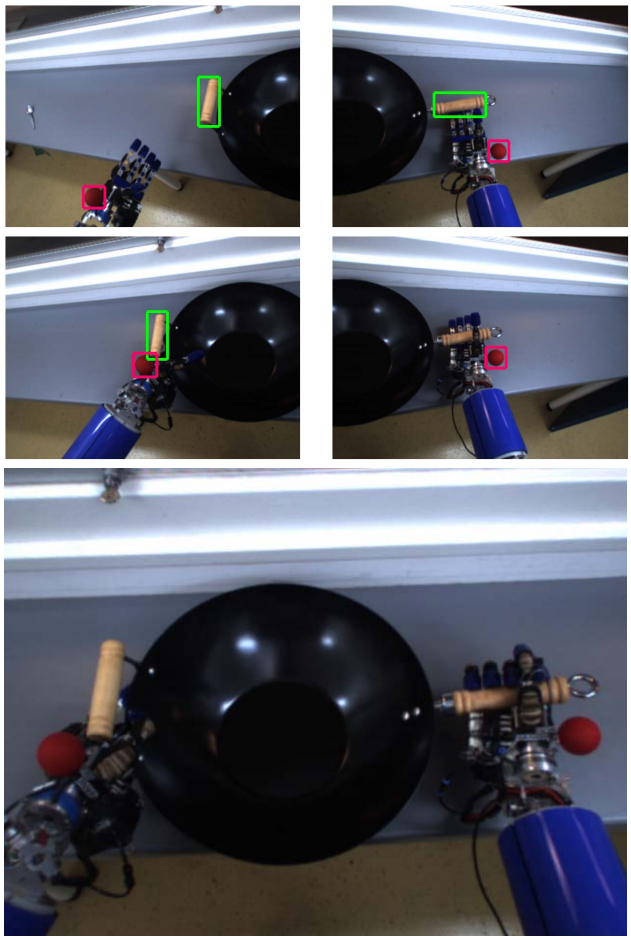


Fig. 10. Procedure of grasping the wok where the pose of the wok is derived from the two handles (marked in green) and the hand is tracked via the red markers. The final grasping pose (relative to the object's pose) was taught-in by a zero-force controller.

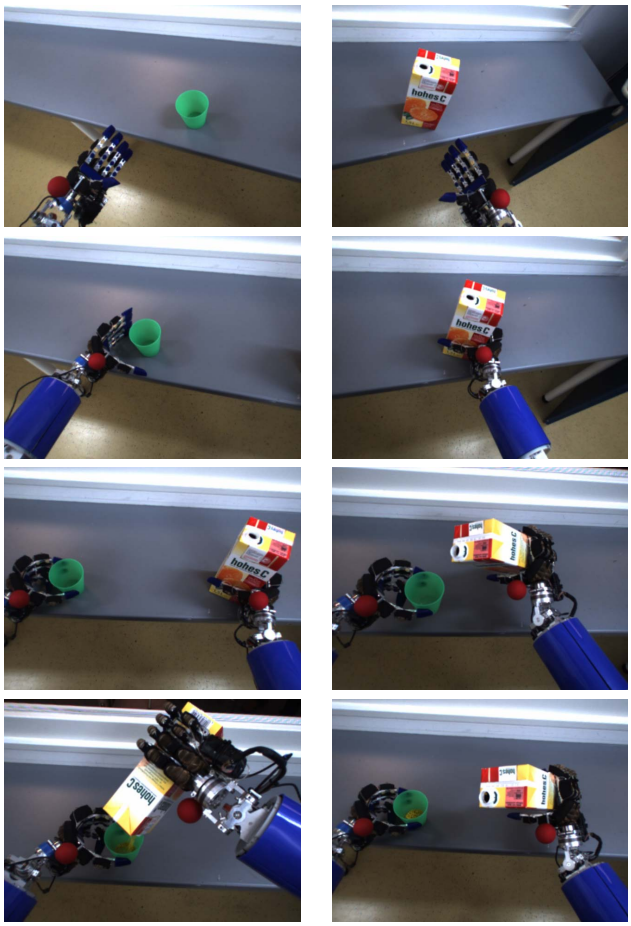


Fig. 11. Armar-III grasps a cup and an orange juice simultaneously with both hands and pours the juice in the cup. The grasping as well as the manipulation action are executed by the proposed dual arm visual servoing controller.

grasped in parallel and then a dual arm manipulation action is performed by moving the hands to a taught in pouring pose. All movements are executed with high accuracy since they are supported by the proposed dual arm visual servoing techniques described in section IV (see Fig. 11).

VI. CONCLUSIONS AND FUTURE WORKS

In this work a visual servoing controller for dual arm grasping and manipulation tasks was presented. The control framework allows it to track multiple targets and by estimating the poses of the objects and the robot's hands that are currently not in the field of view a large area of operation is covered. The use of an active vision system allows a non parallel arrangement of the eyes without loosing the ability for Cartesian object localization and thus the observable workspace can be increased additionally. The implementation on the humanoid robot Armar-III shows how dual arm manipulations can be robustly executed in spite of noisy sensor data, inaccurate kinematic models and without any hand eye calibration.

In the future we want to improve the hand tracking to be able to visually retrieve the hand orientation and position. The orientational accuracy should benefit from this additional information channel and more dexterous operations can be executed. Furthermore additional dual arm operations with decoupled targets could be investigated, e.g. the robot could open a door of a cabinet with one hand while grasping an object with the other hand. These dual arm operations could support a human like behavior of the robot since the sequential execution of single arm manipulation tasks is avoided.

VII. ACKNOWLEDGMENTS

The work described in this paper was partially conducted within the German Humanoid Research project SFB588 funded by the German Research Foundation (DFG: Deutsche Forschungsgemeinschaft) and the EU Cognitive Systems projects PACO-PLUS (FP6-2004-IST-4-027657) and GRASP (IST-FP7-IP-215821) 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] P. Azad, T. Asfour, and R. Dillmann, "Stereo-based 6D Object Localization for Grasping with Humanoid Robot Systems," in *International Conference on Intelligent Robots and Systems (IROS)*, San Diego, USA, 2007.

[3] N. Vahrenkamp, S. Wieland, P. Azad, D. Gonzalez, T. Asfour, and R. Dillmann, "Visual servoing for humanoid grasping and manipulation tasks," in *Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on*, Dec. 2008, pp. 406–412.

[4] T. Asfour, K. Welke, P. Azad, A. Ude, and R. Dillmann, "The karlsruhe humanoid head," in *Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on*, Dec. 2008, pp. 447–453.

[5] K. Welke, M. Przyblyski, T. Asfour, and R. Dillmann, "Kinematic calibration for saccadic eye movements," Universität Karlsruhe (TH), Fak. f. Informatik, Institute for Anthropomatics, Tech. Rep., 2008.

[6] S. Hutchinson, G. Hager, and P. Corke, "A tutorial on visual servo control," *IEEE Trans. on Robotics and Automation*, vol. 12, no. 5, pp. 651–670, oct 1996.

[7] J. Hill and W. Park, "Real time control of a robot with a mobile camera," in *The 9th International Symposium on Industrial Robots*, 1979, pp. 233–246.

[8] F.Chaumette and S.Hutchinson, "Visual servo control, part I: Basic approaches," *IEEE Robotics and Automation Magazine*, vol. 13, no. 4, pp. 82–90, dec 2006.

[9] E. Malis, F. Chaumette, and S. Boudet, "2 1/2 d visual servoing," *IEEE Transaction on Robotics and Automation*, vol. 15, no. 2, pp. 234–246, April 1999.

[10] L. Weiss, A. C. Sanderson, and C. P. Neuman, "Dynamic sensor-based control of robots with visual feedback," *IEEE Journal on Robotics and Automation*, vol. RA-3, no. 5, October 1987.

[11] K. Hosoda and M. Asada, "Versatile visual servoing without knowledge of true jacobian," in *Intelligent Robots and Systems '94. 'Advanced Robotic Systems and the Real World', IROS '94.*, vol. 1, September 1994, pp. 186–193.

[12] F. Chaumette, "Visual servoing using image features defined upon geometrical primitives," in *33rd IEEE Conf. on Decision and Control*, vol. 4, Orlando, Florida, December 1994, pp. 3782–3787.

[13] K. HoWon, C. JaeSeung, and K. InSo, "A novel image-based control-law for the visual servoing system under large pose error," in *Intelligent Robots and Systems, 2000. (IROS 2000).*, vol. 1, 2000, pp. 263–268.

[14] F. Chaumette, "Potential problems of stability and convergence in image-based and position-based visual servoing," in *The Confluence of Vision and Control*, D. Kriegman, G. Hager, and A. Morse, Eds. LNCIS Series, No 237, Springer-Verlag, 1998, pp. 66–78.

[15] S. Han, W. See, J. Lee, M. Lee, and H. Hashimoto, "Image-based visual servoing control of a scara type dual-arm robot," in *Industrial Electronics, 2000. ISIE 2000. Proceedings of the 2000 IEEE International Symposium on*, vol. 2, 2000, pp. 517–522 vol.2.

[16] S. H. Han and H. Hashimoto, "A study on feature-based visual servoing control of robot system by utilizing redundant feature," *The Korean Society of Mechanical Engineers*, vol. 16, no. 6, pp. 762–769, June 2002.

[17] P. Hynes, G. Dodds, and A. Wilkinson, "Uncalibrated visual-servoing of a dual-arm robot for mis suturing," in *Biomedical Robotics and Biomechatronics, 2006. BioRob 2006. The First IEEE/RAS-EMBS International Conference on*, Feb. 2006, pp. 420–425.

[18] T. Miyabe, A. Konno, M. Uchiyama, and M. Yamano, "An approach toward an automated object retrieval operation with a two-arm flexible manipulator," in *The International Journal of Robotics Research 2004*, vol. 23, 2004, pp. 275–291.