Combining Force and Visual Feedback for Physical Interaction Tasks in Humanoid Robots

S. Wieland, D. Gonzalez-Aguirre, N. Vahrenkamp, T. Asfour and R. Dillmann

Institute for Anthropomatics, University of Karlsruhe,
Haid-und-Neu-Strasse 7, Karlsruhe-Germany.
$\{wieland, gonzalez, vahrenkamp, asfour, dillmann\}@ira.uka.de$

***Abstract*— In this paper we present a framework for combining force and visual feedback in the task space to deal with humanoid interaction tasks like open doors in a real kitchen environment. We present stereo-vision methods for markerless recognition and estimation of environmetal elements for applying force control strategies for compliant execution using a 6D force-torque sensor mounted in the humnaoids's wrist. The framework consists of components for model-based self-localization, visual planning, door recognition, grasping based on visual servoing, real time handle tracking and task execution considering force feedback during the physical interaction with the environment. Experimental results on the humanoid robot ARMAR-IIIa are presented.**

I. Introduction

The development of humanoid robots for human daily environments is an emerging research field of robotics and a challenging tasks. Recently, considerable results in this field have been achieved and several humanoid robots have been realized with various capabilities and skills. Recently, integrated humanoid robots for daily-life environment tasks has been successfully presented with various complex behaviors (see e.g. [1]). However, in order for humanoid robots to enter daily environments, it is indispensable to equip them with fundamental capabilities of grasping and manipulating objects encountered in the environment and of dealing with kitchen appliances and furniture such as fridge, dishwasher and doors.

The work presented in this paper is an extension of our previous work (see [2], [3], [4]) toward the realization of complex manipulations and grasping tasks in a kitchen environment. In this work, we are primarily interested in a frequently needed task in daily life: manipulating door tasks in a kitchen environment where the robot has to handle significant forces during the physical interaction with the environment.

Early work in humanoid manipulation considered body balancing [5] and collision-free motions to satisfy balance constraints [6], however they do not consider the physical interaction of the robot with the objects to be manipulated. In [7] the ZMP balance criterion is extended for the task of pushing an object with known dynamics and in [8] the physical interaction of a humanoid robot with unknown, large and heavy objects to learn the models of these objects.

Further work on physical interaction for door opening tasks has been proposed by Prats et al. in [9], [10] and [11]. In [9], only force feedback is used for task-oriented grasping and task frame estimation. In [10], artificial markers are added to visually estimate the task frame and in [11] both visual and force feedback are combined to solve door opening tasks. The approach we present in this paper is inspired by the work proposed by Prats et al. in [9], [10] and [11]. We investigate the integration of visual and force feedback to deal with similar task executed by a humanoid robot in a human centered environment, without using any artificial markers on the furniture.

In our work, we combine visual and force information to deal with the complexity of the underlying environment where the visual appearance, physical properties and dynamic behaviors of the elements are complex and time-varying. Therefore, we use both stereo vision task frame sensing and force feedback task frame estimation and their combination. In this manner, it is possible to simultaneously overcome the limitations of both sensing paradigms while generating synergy by exploiting the structure of the physical and modeled spaces.

The role of vision in our approach is to link the modeled world with the visual space. This is possible since image processing algorithms coupled with a model-based inference mechanism are effective for recognition without object contact and they provide information over a large area on a variety of conditions. On the other hand, the force feedback is responsible of the compliant interaction with the environments, delivering precise information upon contact. We present stereo-vision methods for marker-less recognition and estimation of the 6D task frame (the door and its axis) and the pulling direction (position and orientation of door handle) and force control strategies used for both, the task frame estimation and the compliant task execution, utilizing a 6D force-torque sensor at the humanoid robot wrist.

The remaining of the paper is organized as follows. In section II the general framework of our method is described. The task frame estimation methods, based on force- and vision-sensors, are explained in section III. In section IV the task execution on the humanoid robot ARMAR-IIIa along experimental results are presented. Conclusions and future work are given in section V.

II. General Framework

A. System overview

This work is performed on the humanoid robot ARMAR-IIIa [12], which consists of seven subsystems: head, left arm,

right arm, left hand, right hand, torso, and a mobile platform. The head has seven DOF and is equipped with two eyes. The eyes have a common tilt and can pan independently. Each eye is equipped with two 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 33 DOF: 14 DOF for the arms, 16 DOF for the hands and three DOF for the torso. Each arm is equipped with a five-fingered hand with eight DOF. Each joint of the arms is equipped with motor encoder and axis sensor to allow position and velocity control. In the wrists, 6D force-torque sensors are used for hybrid position and force control and for the realization of compliant arm movements as presented in [2].

B. Physical Interaction

Our framework for coupling visual- and force-sensor information is based on our previous work on physical interaction in household environments [2] and on the *Task Frame Formalism* (see [13] and [14]), because it is suitable for compliant task execution. In our work, the task frame $\mathcal{T}$ is defined as a Cartesian coordinate system aligned to the object to manipulate, in which the task is defined, using velocity or force references. However, the movement of the arms is controlled in the Cartesian coordinate system $\mathcal{E}$, attached to the end-effector, where the inverse jacobian matrix is used to transform the Cartesian into joint-velocities.

Due to the fact that the task frame is always aligned to the object and not to the end-effector of the humanoid robot, the relative pose between the end-effector and the task frame must be computed, in order to assign the task frame jacobian matrix. Using this matrix, the desired task frame velocity $\dot{x}_d$ is transformed into the joint velocities $\dot{q}$. Thus, it is necessary to estimate the task frame online during task execution to be able to compute the relative pose between the two frames (see Sec.III-A and Sec.III-B).

The force control law introduced in [2] is running during the task execution, based on a Cartesian impedance control using a joint velocity controller. To cope with the joint redundancy of our humanoid robot, a joint limit avoidance secondary task is used. Additional redundant DOFs are provided in the task space, if one (or more) Cartesian DOFs are not required for task execution.

C. Model and Appearance-Based Object Recognition

The world-model and the available context acquired during self-localization (the associations between model elements and visual percepts) will not only make it possible to solve, complex visual assertion queries, but it will also dispatch them with a proficient performance. In our experiments, we have used this knowledge to switch between our components for door and handle recognition and pose estimation, i.e. a context-less component, as described in [15] deals with a wide distance range but with reduced tolerance to perturbations. Then, a middle-to-close distance algorithm [16] is used in closer range distances without occlusions tolerance mechanisms. Finally, the following technique uses context information, to easily ignore very intricated recognition outliers.

The pose estimation of the partial occluded door handle, when the robot has already grasped it, turns out to be a difficult task because there are many perturbation factors. No size rejection criteria may be assumed, since the robot hand is partially occluding the handle surface. Secondly, the hand slides during task execution, producing variation of the apparent size. No assumption about the background of the handle could be made, because when the door is partially open and the perspective view overlaps handlers from lower doors, the same chromatic distributions appear. On the top of that, the glittering of the metal surfaces on both, robot hand and door handle, produce very confusing phenomena, when using standard segmentation techniques [17]. The state of the art in parameterless robust segmentation techniques [15] deliver acceptable results, however they are not suitable for tracking purposes due to their performance speed. Their runnig time is approximately 1-6 seconds, depending on the selected (which is a problem itself) spatial and chromatical band widths.

In this context, we propose an application dependent, but very robust and fast technique (15-20 ms) in order to simultaneously segment the regions and erode the borders, producing non-connected regions which suits our desired preprocessing-filtering phase as follows. First, the raw *RGB*-color image $I_\chi(x,y) \in N^3$ is used to compute the *power image* using Eq. 1. These two images are illustrated in Fig.1.

$$I_\phi(x,y) = (I_\chi(x,y)^T I_\chi(x,y))^n \tag{1}$$

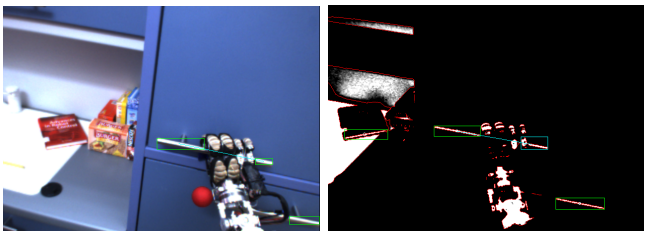

Fig. 1. a) Left Input Image b) The Power Image

Then a linear normalization and adaptive thresholding produces the binary image $I_B(x,y) \in \{0,1\}$ which is used to extract the blobs B_k and build feature vectors for rejection purposes.

The feature vector $F(B_k)$ (see Eq. 3) is formed by the blobs area $\omega(B_k)$, the energy density $\delta(B_k)$, and the elongation descriptor, i.e. the ratio of the eigenvalues $E_{\sigma_{i,j}}(B_k)$ of the energy covariance matrix M_{B_k} expressed by Eq. 2.

$$[\overrightarrow{E_{\sigma_{1,2}}} E_{\sigma_{1,2}}] = SVD(M_{B_k}) \tag{2}$$

$$F(B_k) := [\delta(B_k), \omega(B_k), E_{\sigma_1}(B_k)/E_{\sigma_2}(B_k)] \tag{3}$$

This characterization enables us to reject blobs when verifying the right-left cross matching by only allowing candidates in pairs (B_k, B_m) where the criterion $K(B_k, B_m) := \|E_{\sigma_1}(B_k)\cdot$

$E_{\sigma_1}(B_m)\| > K_{min}$ is fulfilled, i.e. the orientation of their axis shows a discrepancy less than $\arccos(K_{min})$ radians.

Subsequently, the interest point I_p is selected as the furthest pixel along the blobs main axis in opposed direction[1] of the vector Γ_R, i.e. unitary vector from the door center to the center of the line segment where the rotation axis is located (see Fig.4). This vector is obtained from the world model and localization context by using virtual cameras in the model. Moreover, the projected edges of a door within the kitchen aids the segmentation phase to extract the door pose and improves precision by avoiding to consider edges pixels close to the handle, because in this region a hough transformation coupled with a linear regression for finding the 2D-lines will fail, see Fig.2.

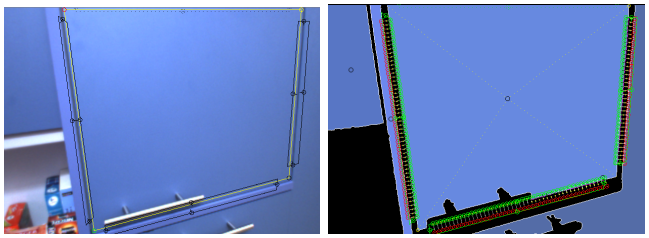

Fig. 2. Model and Appearance-Based Object Recognition

The key factor of this vision-to-model coupling relies on the fact that very general information is used, i.e. from the projected lines and blobs using the virtual camera and the scene graph, only their direction is used (plugged into a noise-tolerant criterion K_{min}) and not the position itself, which differs to the real one, due to the discretization, quantization, noise and uncertainty accumulated starting with the self-localization, the camera calibrations up to the joint-encoders deviations.

III. Online Task Frame Estimation

A. Force-based Estimation

In our previous work [2], we use only the force sensor information, provided by two 6D force/torque sensors, mounted at both wrists of the robot, to estimate the task frame. The open direction, which is the z-axis of the task frame, is initially set to a normal vector on the kitchen furniture, because we assume a fully closed door. The x-axis is a normal on the floor and the y-axis is their cross-product. The task frame origin is calculated by adding a translational displacement to the end-effector frame position.

During execution, the door rotates and therewith the task frame changes. Our solution is to record a history of the task frame positions during task execution in order to estimate the new frame adjustment.

The manipulated furniture and the error in the frame estimation generate small forces in the hand. The robot tries to minimize this forces, by updating its hand position. Furthermore the robot aligns the task frame z-axis with the vector tangent to the saved task frame positions, plotted as red dots in Fig.3.

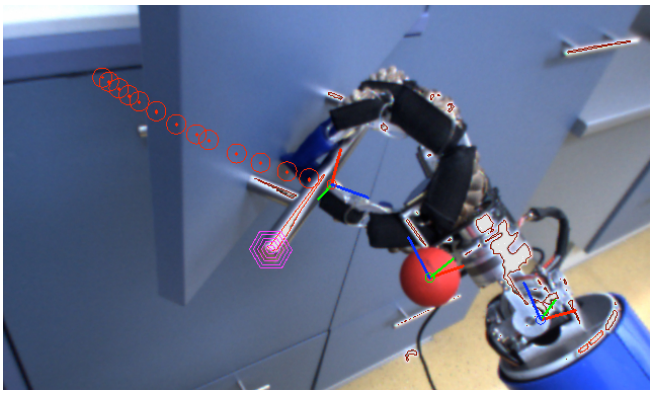

Fig. 3. Trajectory of the task frame positions, plotted as red dots. Furthermore the z-axis (blue) of the task frame is displayed, aligned to the trajectory vector tangent

In summary, the task frame is aligned to the hand movement and the knowledge of the particular mechanism of the furniture is not required during task execution.

B. Vision- and Model-based Estimation

One advantage of our approach is the usage of the vision-to-model coupling dealing with limited visibility (3D reasoning). In order to provide the required information for the interaction module, it is necessary to estimate the interest point I_p, and the normal vector N_p of the grasping element (see Fig.4), e.g. the door handle. Translating the interest point along the handle direction with a offset, acquired from the model, the midpoint M_p of the handle is calculated, which is used as origin of the task frame $\mathcal{T}$.

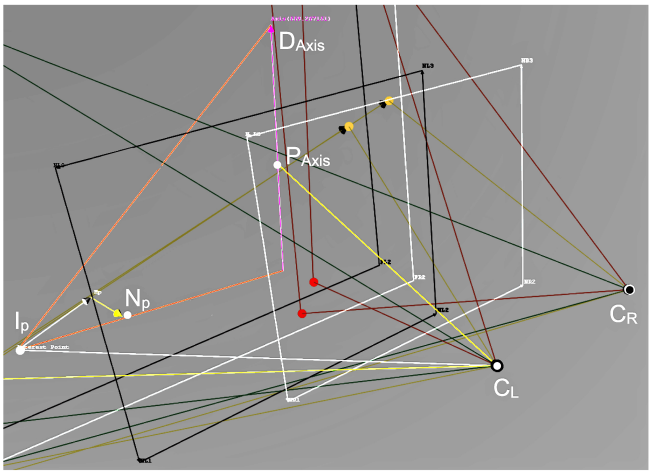

Fig. 4. Door and handle recognition

Because of the sizes of both, the door and the 3D field of view, it can be easily corroborated that the minimal distance within the subspace Ψ (where the robot must be located for

[1]The context opposed direction means where the scalar product evaluates to maximal negative value.

the complete door to be contained inside the robots 3D field of view) may lie outside of the reachable space of the humanoid robot. In this situation, the geometric definition of the door (a rectangular prism) allows the planner to switch from pure data driven algorithm to the following recognition approach which only requires three partially visible edges of the door and uses the context (robots pose) and model to assert the orientation of the door's normal vector and as an aftereffect the door's angle of aperture.

The door axis recognition uses the following facts: first, a 2D-line Υ_i on an image and the center of its capturing camera C_j define a 3D-space plane $\Phi_{(i,j)}$, hence two such planes $\Phi_{(L,L)}$ and $\Phi_{(\mu(\Upsilon_L,\Upsilon_R),R)}$, resulting from the matching $\mu(\Upsilon_L, \Upsilon_R)$ of two lines in left and right images in a stereo system define an intersection[2] subspace: $\Lambda_i = \Phi_{(L,L)} \wedge \Phi_{(\mu(\Upsilon_L,\Upsilon_R),R)}$,i.e. a 3D-line. These 3D-lines Λ_i are subject to noise and calibration artifacts. Thus, they are not suitable to compute 3D intersections. However, their direction is robust enough for our purposes. Next, the left image 2D points $H_{(L,i)}$ resulting from the intersection of 2D-lines Υ_i are matched against those in the right image $H_{(R,j)}$ producing 3D points $X_{(R,j)}$ by means of triangulation in a minimal square fashion [18].

Using this facts, it is possible to acquire corners of the door and directions of the lines connecting them, even when only partial edges are visible. Herein, the direction of the vector Γ_R is the long-term memory clue simultaneously used to select 3D line edge (direction D_{Axis} and its point P_{Axis}, see Fig.4) corresponding to the door rotation axis and ensuring that the direction vector of the axis points upwards, avoiding that the normal vector calculations may be confused with its twisted pair.

Furthermore, the online handle tracking uses the previous valid position of the interest point I_p, to restrict the search region, using the predicted trajectory (only to limit the search region in the image) and a fixed radius, see Fig.1. In this way the power image computation performs fast calculation by searching the region of interest. Using the new tracked interest point the midpoint M_P of the door handle is updated. Afterwards, the normal vector N_p on the door, see Fig.4, is computed using the axis of rotation $R := [D_{Axis}, P_{Axis}]$, shown in Eq. 4.

$$N_p = \frac{(P_{Axis} - I_p) \times D_{Axis}}{\|(P_{Axis} - I_p) \times D_{Axis}\|} \tag{4}$$

Subsequently, this calculated normal vector N_p is used to adjust the task frame z-axis, in order to provide a normal pulling direction. Simultaneously the knowledge about the handle is used to update the x-axis, which is parallel to the handle, and furthermore their cross-product is calculated, resulting in the y-axis of the task frame $\mathcal{T}$.

[2]The plane-plane (in Hessian normal form) intersection operator $\wedge$ provides a point on the line and its direction.

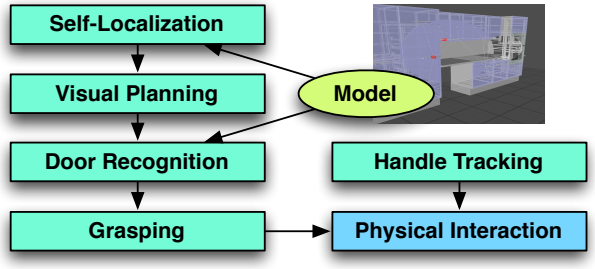

Fig. 5. The main components of the system: self-localization, visual planning, door recognition, handle tracking, grasping, and physical interaction.

IV. Task Execution and Results

A. Assignment of Tasks

The experiment of door opening in a regular kitchen environment with the humanoid robot ARMAR-IIIa is performed several times with different modalities during the physical interaction phase:

1) During the first execution we only use the force sensor channel, according to our previous work [2].
2) The novel stereo vision tracking methods are used to perform the task in our second experiment. The force sensor information is used for compliant movement of the robot hand, but not for the task frame estimation.
3) Furthermore the results of the task space sensor fusion, using the visual- and the force-based estimation of the task frame, are shown.

However, the opening velocity, applied in the direction of the task frame z-axis (normal on the door, in order to minimize the external forces), is set to $20\,mm/s$ during all three experiments. The complete task of door opening, illustrated in Fig.6, is split into the following different modules (see Fig.5). The task assignment to the robot is done by using speech commands, where a human has to specify the door, which the robot should open.

B. Preliminary Task Execution

1) Self-Localization: In this first phase of the task execution the robot localizes itself in the kitchen. The visual self-localization procedure consists of three elements: a collection of active visual *perception-recognition* components (see Sec.II-C), a world model and a statistical hypotheses generation-validation apparatus. The multiple recognized elements (*Percepts*) are fused, filtered and mapped into the ego center frame of the humanoid robot. These ego-percept subgraphs are carefully matched against the model by considering the noise in the relative position and orientation among them. Ideally, the match provides enough information to compute the relative pose of the ego center frame $\mathcal{E}$ as a kinematic chain by using the inverse transformation of the perception frame $\mathcal{P}$ and the deducted frame transformation from the world-model $\mathcal{W}_M$, like $\mathcal{E} = [\mathcal{W}_M][\mathcal{P}]^{-1}$. The runtime for this module is 15-20 seconds using 20 real stereo images. For detailed explanations and experimental results see [19].

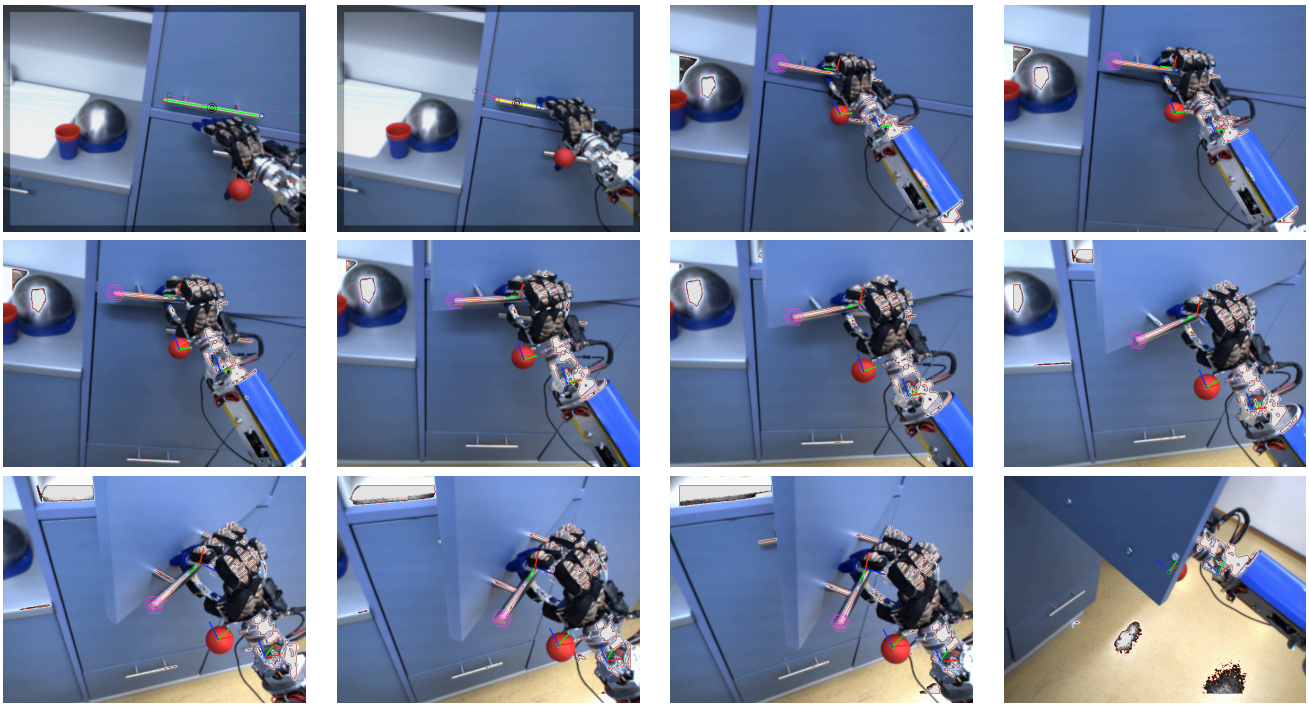

Fig. 6. The humanoid robot ARMAR-IIIa interacting with a cupboard in the kitchen environment. The first two pictures in the first line show the door recognition and grasping. The following illustrate the physical interaction on the cupboard. At this the magenta circles represent the tracked interest point I_p. Using I_p, the midpoint M_p of the handle is calculated and parallel the normal vector on the door N_p is assigned. The result of this calculations is shown as the task frame $\mathcal{T}$, tagged with a red origin (M_p). The x-axis is drawn in red, the y-axis in green and the z-axis (N_p) in blue. Additionally the pose aligned hand tracking and end-effector frame are displayed with the same axis colors. In the last picture the tracker cannot find the handle, because the door covers it. The whole door opening by the humanoid robot ARMAR-IIIa is available in the accompanied video.

2) Visual Planning: Visual planning for physical interaction involves three fundamental aspects: First, once the visual target has been established, the corresponding target-node provides a frame and the definition of a subspace Ψ where the robot has to be located, therewith the target-node can be robustly recognized. Subsequently, the visual planner uses the restriction subspace and target node frame to generate a transformation from the current pose to a set of valid poses. These poses are submitted to the navigation layer [12] to be unfolded and executed into a safe trajectory. Once the robot has reached his desired position, the planner uses the description of the node to predict parametric transformations and properties (how the image content should look like). The general and practical usage of this concept requires specific tailored knowledge for image processing (*2D Reasoning*), space inference (*3D Reasoning*) and the expertise to implement natural compelling functionality, depicting elaborated dynamic states of the nodes. These are the complex long-term memory and attention components of the cognitive architecture [20], e.g. door opening, faucet flowing [1] etc.

3) Door Recognition: This module provides an initial recognition of the door, using the methods introduced in section II-C, even if the visibility is limited. The interest point I_P and the axis of rotation are assigned, which are used later by the handle tracking module, to restrict the handle search region, in order to increase the performance.

4) Grasping: The first part of this module is the grasp planning. Depending on the recognized door, different strategies are chosen, see section IV-B.3. Doors with a left sided rotation axis are opened with the left hand, the other mechanisms with the right one. Grasp types used in the task are determined in an offline manner. For reaching the handle, we used a position-based visual servoing approach as described in [4], which is based on [21] and [22]. The control loop of the visual servoing minimizes the Cartesian error between the destination, which is the task frame $\mathcal{T}$ and the actual end-effector frame $\mathcal{E}$ to zero. The hand orientation is calculated using the arms direct kinematic.

During the execution of the servoing task, the error between the task- and the end-effector-frame mentioned above is minimized. If the error falls below a defined threshold, the 6D force sensor information is taken into account, to indicate a contact with the handle. Once this contact event occurs, the grasping procedure starts. To react on minor inaccuracies, the control of the robot arms is switched to zero force mode, assisting the alignment of the hand and thus improving the robustness of the grasp.

5) Handle Tracking: The handle tracking module provides the handle midpoint M_P, the normal vector on the door N_P and finally the task frame $\mathcal{T}$, as explained in section III-B. Robustness and reliability of the tracker are one of the key achievements of our approach. The figure 7 shows the tracked Cartesian position of the handle midpoint M_P, related

to the ego center frame $\mathcal{E}$. The tracked position follows the movement of the cupboard handle until iteration 165, because the door covers the handle after this point of task execution. This case is shown in the last picture of Fig.6.

Subsequently, the second experiment, see Fig.7, shows the robustness of the handle tracker against external disturbances. Between iteration 65 and 123 the two cameras of the humanoid robot are covered, causing an interruption of the handle tracker. During this part of execution no new interest point is estimated and the tracker provides the last predicted one to the interaction module, see Fig.5.

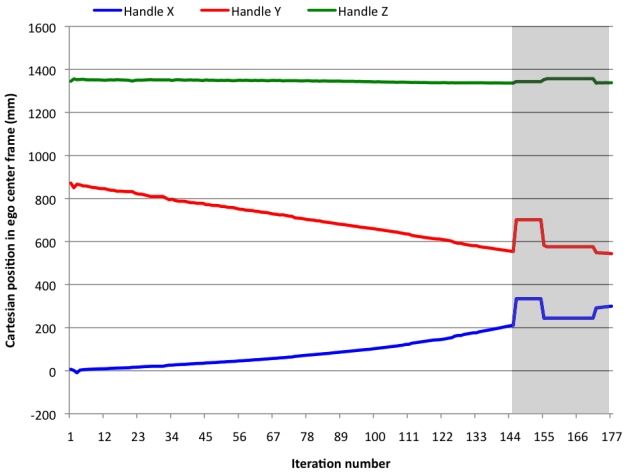

Fig. 7. Cartesian position of the handle midpoint, related to the ego center frame. After iteration 143 the door itself covers the handle and thus the position estimation is not reliable any more. Between iteration 65 and 123 the cameras of the humanoid robot are covered, which causes an interruption of the handle tracker, because no new interest point could be estimated.

C. Force Based Physical Interaction

The results in this section are based on the methods of our previous work [2]. After the handle is grasped, the robot uses the above introduced online force sensor based task frame estimation method, to open the cupboard, see section III-A. Because of the alignment of the task frame z-axis (normal on the door), where the opening velocity is assigned, the external forces are reduced. However during this experiment only the force sensor information is used, which causes external forces at the robot hand, displayed in Fig.8 (red curve).

In summary, the door opening task could be successfully completed with our proposed method, but the errors in the robot kinematics and the inaccuracy of the task frame position tangent vector leads to considerable external forces appeared at the robot hand.

D. Stereo Vision Based Physical Interaction

To reduce the external forces, which appeared during the first experiment, described in Sec.IV-C, online stereo vision-based methods are used to estimate the task frame during task execution, according to section III-B. This methods lead to a more precisely estimation of the opening direction.

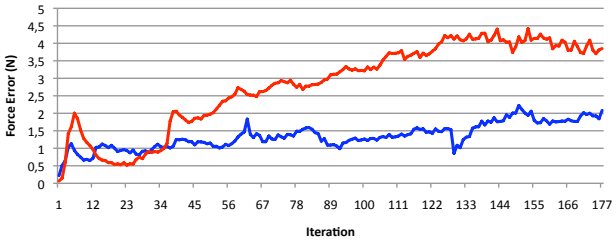

Fig. 8. Comparative plots of the forces appeared at the task frame, without the force in the door normal (pulling) direction. The blue curve shows our improved results, which uses a vision estimated task frame and force feedback. Our previous work is illustrated in the red curve, wherein the task frame is estimated by the history of the task frame position.

The complete task execution is shown in Fig.6, including the task frame $\mathcal{T}$ and the end-effector frame $\mathcal{E}$ with a red origin and the blue z-axis, the green y-axis and the red x-axis. Fig.8 compares the summarized Cartesian force error, except the force in the normal direction on the door, between this experiment (blue curve) and our previous experiment (red curve), see Sec.IV-C. It can easily be seen, that the visual estimation of the task frame is leading to smaller external forces on the hand, produced by the handle and the door mechanism.

Combining stereo vision and force control provides the advantage of real-time task frame estimation by vision, which avoids the errors of the robots kinematics and adjustment of actions by the impedance control. Some inaccuracies concerning the vision modules are caused by environmental influences, e.g. people crossing, changing light conditions. Therefore the force component manages to balance these external forces and torques, which is visualized in Fig.8.

E. Physical Interaction by Coupling both Sensor Information

In conclusion, the first two experiments show, that neither the force-based estimation nor the vision-based estimation of the task frame are satisfying on their own, see table I. Furthermore, the vision estimation does not work, if the handle is completely covered or abruptly changing light conditions disturb the recognition, which leads to a wrong task frame alignment.

TABLE I
COMPARISON BETWEEN THE TASK FRAME ESTIMATION REALIZATIONS

	Force Sensor	Stereo Vision	Combined
Pro	reliable	accurate	reliable accurate
Contra	inaccurate	susceptible	-

For this reason, the two sensor channels for task frame estimation are combined by introducing a vision-gain V_g and a force-gain F_g, connected by $F_g = V_g - 1$. At the beginning of the interaction phase, V_g equals 1. This indicates, that the stereo vision based estimation is used exclusively, because the last two experiments show that the vision estimation is more

precisely than the force estimation. During task execution, this gain could be decreased, depending on the quality of the stereo vision task frame estimation, due to the fact that the estimation by force is more reliable.

To determine the quality of the stereo vision estimation, the distance between the actual estimated handle midpoint M_P and the actual force sensor estimated task frame position, is calculated, see section III-A. The vision gain is inverse proportional to the estimation error, which is only considered between 35 and 70 mm, see Fig.9 and 10. Using the gains, both estimated task frames are weighted and added. In detail, the axis- and the position-vectors of the task frame, the vision-based estimation as well as the force-based estimation, are multiplied by the associated gain and subsequently added.

The results of our first experiment utilizing the fused task frame estimation are shown in Fig.9, 10 and 11.

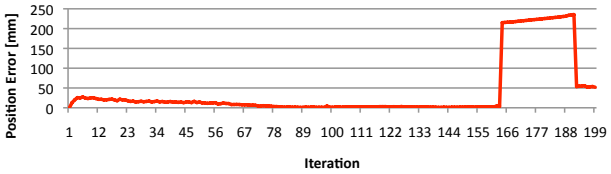

Fig. 9. Error between the visually- and kinematically-estimated task frame position. This value influences the force-vision gain, see Fig.10. After iteration 165 the door itself covers the handle, cf. last picture of Fig.6, and thus the position estimation is not reliable any more.

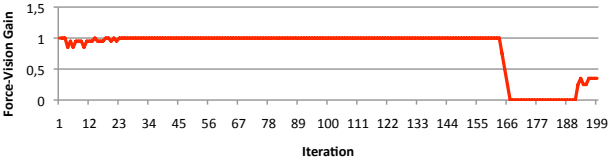

Fig. 10. Gain between visual- and force-sensor estimated task frame. If the vision gain equals one, the interaction module only uses the visually estimated interest point for task execution. The information appreciated by the force sensor is taken into account, if the vision gain decreases ($F_g = V_g - 1$). After iteration 165 the door itself covers the handle and thus the visual estimation is not reliable any more, which leads to a decreasing vision-gain.

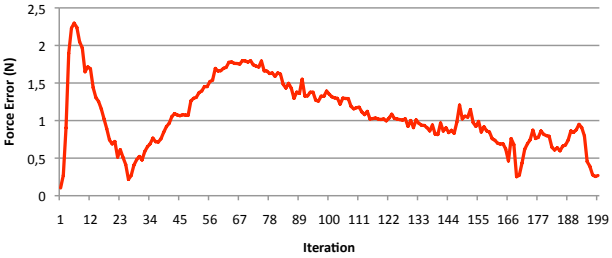

Fig. 11. Comparative plot of the forces appeared at the task frame, without the force in the door normal (pulling) direction.

To demonstrate the robustness of our approach, the stereo cameras of the humanoid robot are covered between iteration 67 and 122 during the second experiment. The figure 12 shows that the robot cannot find a new interest point, because of the covered cameras, which results in an increasing estimation error. Therefore, the force gain increases over this period of time, see Fig.14. After the removal of the coverage at iteration 122, the robot directly recognizes the interest point and the force gain is decreased. The growing position error after iteration 144 has the same reason as in the first experiment, namely the handle covering by the door itself.

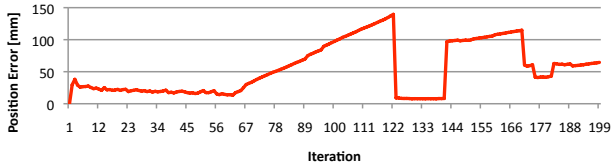

Fig. 12. Error between the visually- and kinematically-estimated task frame position. This value manipulates the force-vision gain, see Fig.13. After iteration 143 the door itself covers the handle, cf. last picture of Fig.6, and thus the position estimation is not reliable any more. Between iteration 65 and 123 the cameras of the humanoid robot are covered, which causes a increasing position error, because no new interest point could be estimated.

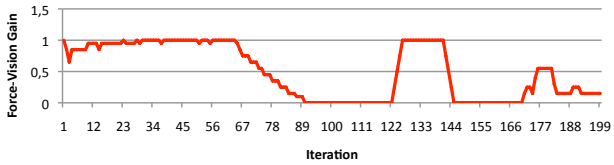

Fig. 13. Gain between visual- and force-sensor estimated task frame. If the vision gain equals one, the interaction module only uses the visually estimated interest point for task execution. The information appreciated by the force sensor is taken into account, if the vision gain decreases ($F_g = V_g - 1$). After iteration 143 the door itself covers the handle and between iteration 65 and 123 the cameras of the humanoid robot are covered, which leads to a decreasing vision-gain, because the visual estimation is not reliable any more.

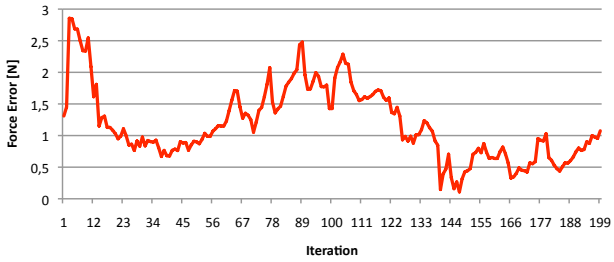

Fig. 14. Comparative plot of the forces at the task frame, without the force in the door normal (pulling) direction.

In summary, the figures 8, 11 and 14 show, that the external forces are reduced to a minimum during the whole task execution, due to the combination of both very different sensor channels in task space. However, interferences can always appear, which makes the integrated Cartesian impedance control necessary during the physical interaction. The entire door

opening task by the humanoid robot ARMAR-IIIa is available in the accompanied video.

V. Conclusion and Future Work

In this work, we have presented a framework for coupling force and vision information in the task space, to accomplish door opening tasks in a real kitchen environment with a humanoid robot. First, the humanoid robot localizes itself in the environment using model-based vision. In the next step the door is recognized. Later, using position-based visual servoing the system is able to grasp the handle. Furthermore, the implemented force control law and the real-time task frame tracking, combined both, vision- and force-sensors, is used to interact on the furniture and adapt the direction of the opening force depending on the opening angle of the door.

Future work will concentrate on the extension of the presented work towards the recognition of different door handles and the association of different grasps to them. Furthermore, we will investigate the integration of the tactile sensor information provided by the hand in this framework.

VI. Acknowledgements

The work described in this paper was partially conducted within the the German Humanoid Research project SFB588 funded by the German Research Foundation (DFG: Deutsche Forschungsgemeinschaft).

References

[1] K. Okada, M. Kojima, Y. Sagawa, T. Ichino, K. Sato, and M. Inaba, "Vision based behavior verification system of humanoid robot for daily environment tasks," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids 2006)*, 2006, pp. 7–12.

[2] M. Prats, S. Wieland, T. Asfour, A. P. del Pobil, and R. Dillmann, "Compliant interaction in household environments by the armar-iii humanoid robot," *Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on*, pp. 475–480, Dec. 2008.

[3] 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.

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

[5] K. Inoue, H. Yoshida, T. Arai, and Y. Mae, "Mobile manipulation of humanoids-real-time control based on manipulability and stability," *Robotics and Automation, 2000. Proceedings. ICRA '00. IEEE International Conference on*, vol. 3, pp. 2217–2222 vol.3, 2000.

[6] J. J. Kuffner, S. Kagami, K. Nishiwaki, M. Inaba, and H. Inoue, "Dynamically-stable motion planning for humanoid robots," *Auton. Robots*, vol. 12, no. 1, pp. 105–118, 2002.

[7] K. Harada, S. Kajita, K. Kaneko, and H. Hirukawa, "Pushing manipulation by humanoid considering two-kinds of zmps," *Robotics and Automation, 2003. Proceedings. ICRA '03. IEEE International Conference on*, vol. 2, pp. 1627–1632 vol.2, Sept. 2003.

[8] M. Stilman, K. Nishiwaki, and S. Kagami, "Learning object models for whole body manipulation," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids 2007)*, 2007.

[9] M. Prats, P. Sanz, and A. del Pobil, "Task-oriented grasping using hand preshapes and task frames," *Robotics and Automation, 2007 IEEE International Conference on*, pp. 1794–1799, April 2007.

[10] M. Prats, P. Martinet, A. del Pobil, and S. Lee, "Vision/force control in task-oriented grasping and manipulation," *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, San Diego, USA*, 2007.

[11] M. Prats, P. Sanz, and A. del Pobil, "A sensor-based approach for physical interaction based on hand, grasp and task frames," *Science and Systems 2008, workshop on robot manipulation: intelligence in human environments, Zrich*, 2008.

[12] T. Asfour, K. Regenstein, P. Azad, J. Schroder, A. Bierbaum, N. Vahrenkamp, and R. Dillmann, "Armar-iii: An integrated humanoid platform for sensory-motor control," *6th IEEE-RAS International Conference on Humanoid Robots*, pp. 169–175, 2006.

[13] M. T. Mason, "Compliance and force control for computer controlled manipulators." *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 11, p. 418432, 1981.

[14] H. Bruyninckx and J. D. Schutter, "Specification of force-controlled actions in the 'task frame formalism': A synthesis," *IEEE Trans. on Robotics and Automation*, vol. 12(5), pp. 581–589, 1996.

[15] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, 2002.

[16] D. I. Gonzalez-Aguirre, T. Asfour, E. Bayro-Corrochano, and R. Dillmann, "Improving model-based visual self-localization using gaussian spheres," *To Appear, Applications of Geometric Algebras in Computer Science and Engineering 2008., Proceedings of the 3rd International Conference on*, 2008.

[17] ——, "Model-based visual self-localization using conformal geometric algebra," *To Appear, Pattern Recognition, 2008., Proceedings of the 19th International Conference on*, 2008.

[18] A. M. Andrew, "Multiple view geometry in computer vision," *Robotica*, vol. 19, no. 2, pp. 233–236, 2001.

[19] D. Gonzalez-Aguirre, T. Asfour, E. Bayro-Corrochano, and R. Dillmann, "Model-based visual self-localization using geometry and graphs," *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pp. 1–5, Dec. 2008.

[20] S. Patnaik, *Robot Cognition and Navigation: An Experiment with Mobile Robots (Cognitive Technologies)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[21] S. Hutchinson, G. Hager, and P. Corke, "A tutorial on visual servo control," *IEEE Trans. on Robotics and Automation*, vol. 12(5), pp. 651–670, 1996.

[22] W. Wilson, C. W. Hulls, and G. Bell, "Relative end-effector control using cartesian position based visual servoing," *IEEE Transactions on Robotics and Automation*, vol. 12, pp. 684–696, 1996.