Grasp Recognition and Mapping on Humanoid Robots

Martin Do, Javier Romero, Hedvig Kjellström, Pedram Azad,
Tamim Asfour, Danica Kragic, Rüdiger Dillmann

Abstract— **In this paper, we present a system for vision-based grasp recognition, mapping and execution on a humanoid robot to provide an intuitive and natural communication channel between humans and humanoids. This channel enables a human user to teach a robot how to grasp an object. The system comprises three components: human upper body motion capture system which provides the approaching direction towards an object, hand pose estimation and grasp recognition system, which provides the grasp type performed by the human as well as a grasp mapping and execution system for grasp reproduction on a humanoid robot with five-fingered hands. All three components are real-time and markerless. Once an object is reached, the hand posture is estimated, including hand orientation and grasp type. For the execution on a robot, hand posture and approach movement are mapped and optimized according to the kinematic limitations of the robot. Experimental results are performed on the humanoid robot ARMAR-IIIb.**

I. Introduction

A humanoid robot's capability of autonomously adapting and acting in new and unstructured environments is very limited. In the majority of cases, a skilled and experienced user is needed for the programming in order to adapt an existing action to a new situation. To enable teaching of a robot by non-expert users, a natural intuitive interface is needed. Since imitation presents an obvious solution for tackling this problem, this field has received great interest in humanoid robotics. The benefit of exploiting demonstration is clearly revealed in [1], where an anthropomorphic arm is capable of balancing a pole in the first trial after observing a human.

A challenging problem where a robot could greatly benefit from a human demonstration is an object grasping task. Such a task involves the control of several degrees of freedom, visual servoing, tactile feedback, etc., turning it to a highly complex task. About the grasp action, a grasp can be divided in two stages: an approach stage and final grasp stage. Due to high object variety concerning shape, size, and mass, determining an adequate approach movement and selecting a suitable grasp type increase the chances that an object is successfully grasped. Instead of telling the robot explicitly which approach movement and which grasp type shall be used, it is desirable to have a system which enables the robot to observe a human during grasp execution and to imitate the demonstration. For the implementation of such a system, various problems have to be tackled, like observation of the human performing the grasp, the mapping of the grasp, and the final execution on the robot.

An important part of the grasp imitation system is the block in charge of getting information about the arm and hand movements. In order to provide this information, the approach movement of the arm as well as the hand pose have to be recognized. Aiming towards ease of use, markerless systems seem to be the most obvious solution for the observation of human grasps since, besides vision sensors, additional equipment is avoided and the preparation effort is kept to a minimum. However, markerless 3D motion capturing and reconstruction of hand pose based on image data are extremely difficult problems due to unstructured environments, the large self-occlusion, high dimensionality and non-linear motion of the arm and the fingers.

Besides the perception modules, another crucial part of an imitation system consists of the mapping and the execution of an observed human grasp on a humanoid robot. Due to severe constraints of mechanical systems and differences between the human and the robot's embodiment, a large number of requirements arise, which are difficult to be satisfied at once. Towards enabling a humanoid to imitate a human grasp, our system integrates several subsystems and methods. First, using a stereo camera setup human observation is initiated by capturing upper body motion and scanning the scene for known objects to attain information on the approach stage. Subsequently, grasp classification and hand orientation are provided through the estimation of the full hand pose in a non-parametric fashion. Finally, the motion data is gathered and mapped onto the robot for execution. The mapping is accomplished via a standardized interface and the ensuing execution is achieved by means of non-linear optimization.

II. Related Work

Several approaches have been made to create a markerless human motion capture system for humanoid robots. Especially, image-based approaches have been a major focus of this field. These approaches are either search-based ([2], [3]), utilize an optimization approach based on 2D-3D correspondences [4], [5], or are based on particle filtering. In [6], it was shown that human motion can be successfully tracked with particle filtering, using three cameras positioned around the scene of interest.

J. Romero, H. Kjellström and D. Kragic are with KTH - Royal Institute of Technology, Stockholm, Sweden, as members of the Computer Vision & Active Perception Lab., Centre for Autonomous Systems, e-mail: jrgn,hedvig,dani@kth.se

M. Do, P. Azad, T. Asfour and R. Dillmann are with the University of Karlsruhe (TH), Karlsruhe, Germany, as members of the Institute for Anthropomatics, e-mail: do,azad,asfour,dillmann@ira.uka.de

Towards imitation of human motion by a robot, the mapping and execution of motion capture data are issues whereas possible solutions pursue strategies which either make use of artificial markers and landmarks or which are based on the transfer and post-processing of joint angles. Marker-based approaches are presented in [7] and [8] where methods based on minimization of the mismatch between robot and human markers are introduced. However, in [9] and [10], joint angles of a demonstrators posture are determined and transferred to the robot for execution. Due to joint and velocity constraints, a scaling and transformation process must be performed in order to obtain a feasible joint angle configuration for the robot.

Analysis of human hand pose for the purpose of learning by demonstration (LbD), see [11] has been thoroughly investigated, almost exclusively with the help of markers and/or 3D sensors attached to the human hand. In the work by Oztop [12] motion capture, color segmentation with artificially colored hands, and active-marker capture systems were compared. Magnetic gloves have also been used extensively because of their accuracy [13]. Another input source for LbD systems is the passive joint measurements of the robot itself [14]. However, the methods shown above all use invasive devices. We envision a LbD scenario where the teaching process can be initiated without calibration and where the robot-user interaction is as natural as possible. For this reason, we want to reconstruct the hand posture in a visual markerless fashion.

Methods for hand pose estimation that are not constrained to a limited set of poses can largely be classified into two groups [15]: I) model based tracking and II) single frame pose estimation. Methods of type I) usually employ generative articulated models [16], [17], [18], [19]. Since the state space of a human hand is extremely high-dimensional, they are generally very computationally demanding, which currently makes this approach intractable for a robotics application. Methods of type II) are usually non-parametric [20], [21]. They are less computationally demanding and more suited for a real-time system, but also more brittle and sensitive to image noise, since there is no averaging over time. The method presented here falls into the second approach. However, it takes temporal continuity into account and it can be used for online real-time reconstruction.

III. Grasp Observation

As mentioned before, we assume that a grasp consists of an approaching stage and a final grasp stage. The observation of the whole grasping process involves recognition of the grasp type, estimation of the approach arm movement and object detection. Following the target of having an intuitive and natural programming interface for robots, we use a markerless human motion capture system for the observation of human motion using the stereo vision system of the robot's head [22]. The head has two eyes and each eye is equipped with two cameras, one with a wide-angle lens for peripheral vision and one with a narrow-angle lens for foveal vision.

First, the robot recognizes known objects in the scene and starts capturing human motion. The hand pose estimation system is triggered as soon as the human hand is in the vicinity of the object. To obtain a close-up of the hand, the foveal cameras are used. The grasp observation is finished with the classification of the observed human grasp.

A. Hand Pose Estimation

The input to the method is a sequence $[\mathbf{I}_t], t = 1, \ldots, n$ of monocular images of the human hand [21].

In each frame $\mathbf{I}_t$, the hand is segmented using skin color segmentation based on color thresholding in HSV space. The result is a segmented hand image $\mathbf{H}_t$.

The shape information contained in $\mathbf{H}_t$ is represented with a Histogram of Oriented Gradients (HOG). This feature has been frequently used for representation of human and hand shape [23], [24], [25]. It has the advantage of being robust to small differences in spatial location and proportions of the depicted hand, while capturing the shape information effectively.

1) Non-parametric Pose Reconstruction: In this section, we omit the time index and regard the problem of reconstructing a *single* pose $\mathbf{p}$ from a *single* HOG $\mathbf{x}$.

Our goal is to obtain the grasp class and orientation of the human hand. We can infer this information from the pose $\mathbf{p}$ of the hand, since all this information is stored for each entry of the database. Therefore, we want to find the mapping $\hat{\mathbf{p}} = \mathcal{M}(\mathbf{x})$, where $\hat{\mathbf{p}}$ is the estimated 31D hand pose in terms of global orientation (lower arm yaw, pitch, roll) and joint angles (3 wrist joint angles, 5 joint angles per finger) , and $\mathbf{x}$ is the observed 512D HOG representation of the hand view, described in Section III-A.

The mapping function $\mathcal{M}$ can be expected to be highly non-linear in the HOG space, with large discontinuities. Following [21], $\mathcal{M}$ is therefore represented non-parametrically, i.e., as a database of example tuples $\{\langle \mathbf{x}_i, \mathbf{p}_i \rangle\}, i \in [1, N]$. Due to the high dimensionality of both the HOG space (512D) and the state space (hereafter denoted JOINT space, 31D), the database needs to be of a considerable size to cover all hand poses to be expected; in our current implementation, $N = 90000$. This has two implications for our mapping method, as outlined in the subsections below.

2) Generation of Database Examples: Generating a database of 10^5 examples from real images is intractable.

(a) HOG $\mathbf{x}$, JOINT $\mathbf{p}$ (b) HOG $\mathbf{x}_1$, JOINT $\mathbf{p}_1$ (c) HOG $\mathbf{x}_2$, JOINT $\mathbf{p}_2$

Fig. 1. Ambiguity in mapping from HOG space to JOINT space. Even though it is visually apparent that $\|\mathbf{p} - \mathbf{p}_2\| \ll \|\mathbf{p} - \mathbf{p}_1\|$ in JOINT space, database instance 1 will be regarded as the nearest neighbor as $\|\mathbf{x} - \mathbf{x}_1\| < \|\mathbf{x} - \mathbf{x}_2\|$. *Note that the object in the hand just contributes with occlusion of the hand in HOG extraction, as it is then colored uniformly with background color.*

Instead, we used the graphics software Poser 7 to generate synthetic views $\mathbf{H}_i^{\text{synth}}$ of different poses. The database examples are chosen as frames from short sequences of different grasp types from different view points, different grasped objects, and different illuminations.

The grasp types are selected according to the taxonomy developed in the GRASP project[1], which integrates the Cutkosky [26], Kamakura [27], and Kang [28] taxonomies. The whole database is also available at the same place.

From each example view $\mathbf{H}_i^{\text{synth}}$, the tuple $\langle \mathbf{x}_i, \mathbf{p}_i \rangle$ is extracted, where $\mathbf{x}_i$ is generated from $\mathbf{H}_i^{\text{synth}}$ as described in Section III-A, and $\mathbf{p}_i$ is the pose used to generate the view $\mathbf{H}_i^{\text{synth}}$ in Poser 7.

3) Approximate Nearest Neighbor Extraction: Given an observed HOG $\mathbf{x}$, the goal is to find an estimated pose $\hat{\mathbf{p}} = \mathcal{M}(\mathbf{x})$. With the non-parametric mapping approach, the mapping task $\hat{\mathbf{p}} = \mathcal{M}(\mathbf{x})$ is one of searching the database for examples $\langle \mathbf{x}_i, \mathbf{p}_i \rangle$ such that $\mathbf{x}_i \approx \mathbf{x}$. More formally, X_k, the set of k nearest neighbors to $\mathbf{x}$ in terms of Euclidean distance in HOG space, $d_i = \|\mathbf{x} - \mathbf{x}_i\|$ are retrieved.

As an exact kNN search would put serious limitations on the size of the database, an approximate kNN search method, Locality Sensitive Hashing (LSH) [29] is employed. LSH is a method for efficient ϵ-nearest neighbor (ϵNN) search, i.e. the problem of finding a neighbor $\mathbf{x}_{\epsilon\text{NN}}$ for a query $\mathbf{x}$ such that

$$\|\mathbf{x} - \mathbf{x}_{\epsilon\text{NN}}\| \leq (1+\epsilon)\|\mathbf{x} - \mathbf{x}_{\text{NN}}\| \tag{1}$$

where $\mathbf{x}_{\text{NN}}$ is the true nearest neighbor of $\mathbf{x}$. The computational complexity of ϵNN retrieval with LSH [29] is $\mathcal{O}(DN^{\frac{1}{1+\epsilon}})$ which gives sublinear performance for any $\epsilon > 0$.

4) The Mapping $\mathcal{M}$ is Ambiguous: The database retrieval described above constitutes an approximation to the true mapping $\hat{\mathbf{p}} = \mathcal{M}(\mathbf{x})$, robust to singularities and discontinuities in the mapping function $\mathcal{M}$.

However, it can be shown empirically that $\mathcal{M}$ is inherently ambiguous (one-to-many); substantially different poses $\mathbf{p}$ can give rise to the similar HOGs $\mathbf{x}$ [23]. An example of this is shown in Figure 1.

Thus, the true pose $\mathbf{p}$ can not be fully estimated from a single HOG $\mathbf{x}$ (using any regression or mapping method); additional information is needed. In the next section, we describe how temporal continuity assumptions can be employed to disambiguate the mapping from HOG to hand pose.

5) Time Continuity Enforcement in JOINT Space: We now describe how temporal smoothness in hand motion can be exploited to disambiguate the mapping $\mathcal{M}$.

Consider a sequence of hand poses $[\mathbf{p}_t], t = 1, \ldots, n$, that have given rise to a sequence of views, represented as HOGs $[\mathbf{x}_t], t = 1, \ldots, n$. Since the mapping $\mathcal{M}$ is ambiguous, the k nearest neighbors to $\mathbf{x}_t$ in the database, i.e. the members of the set X_k, are all similar to $\mathbf{x}_t$ but not necessarily corresponding to hand poses similar to p_t. An important implication of this is that a sequence of hand poses $[\mathbf{p}_t], t = 1, \ldots, n$ does not necessarily give rise to a sequence of HOGs $[\mathbf{x}_t], t = 1, \ldots, n$ continuous in the HOG space.

[1] www.grasp-project.eu.

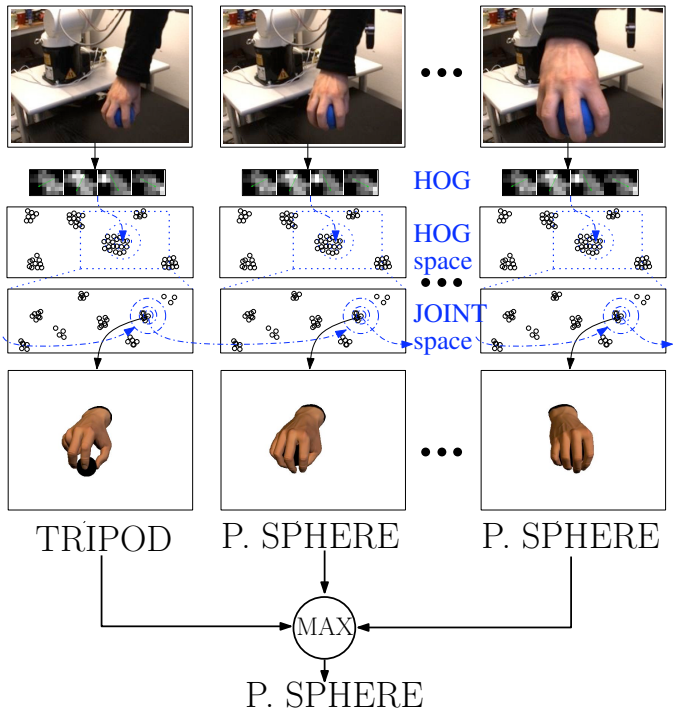

Fig. 2. Grasp Classification with continuity enforcement in JOINT space

However, due to the physics of the human body, the speed of the hand articulation change is limited. Thus, the sequence of hand poses $[\mathbf{p}_t], t = 1, \ldots, n$, i.e. the *hidden variables*, display a certain continuity in the JOINT space. This is illustrated in Figure 2.

The hand pose recognition for a certain frame t is therefore divided into two stages; I) retrieval of a set of k nearest neighbors X_k using single frame non-parametric mapping, as described in Section III-A.1; II) weighting of the members of X_k according to their time continuity in the JOINT space.

Let P_k be the set of poses corresponding to the kNN set X_k found in stage I). Moreover, let $\hat{\mathbf{p}}_{t-1}$ be the estimated pose in the previous time step. In stage II), the members $\mathbf{p}_j, j \in [1, k]$ of P_k are weighted as

$$\omega_j = e^{-\frac{\|\mathbf{p}_j - \hat{\mathbf{p}}_{t-1}\|}{2\sigma^2}} . \tag{2}$$

where σ^2 is the variance of the distance from each entry pose $\mathbf{p}_j$ to the previous estimated pose p_{t-1}.

The pose estimate at time t is computed as the weighted mean of P_k:

$$\hat{\mathbf{p}}_t = (\sum_{j=1}^{k} \omega_j \mathbf{p}_j)/(\sum_{j=1}^{k} \omega_j) . \tag{3}$$

The grasp class estimation G_t is obtained through a majority voting process within the N_p poses with the highest weight ω_j (for our experiments $N_p = 15$). G_t is then smoothed temporally taking the majority vote in a temporal window of N_f frames ($N_f = 10$ in our experiments). This

can be seen in Figure 2. The whole system runs at 10 Hz on a 1.8 GHz single core CPU.

B. Object Recognition

For the robust recognition and accurate 6D pose estimation of single-colored objects, in our previous work, we have developed a model-based approach based on a combination of stereo triangulation, matching of global object views and online projection of a 3D model of the object [30]. The requirement for the approach is global segmentation of the objects, which is accomplished by color segmentation. For training, a 3D model of the object is used to generate views with different object orientations in simulation. Each view is stored along with its corresponding orientation. For recognition, each region candidate obtained by the segmentation routine is matched against the database. An initial orientation estimate is given by the stored orientation information with the matched view. An initial position estimate is given by the stereo triangulation result of the segmented regions in the left and right camera image. The triangulation result of the centroids depends on the view of the object and thus cannot serve as a constant reference point. In order to solve these problems, a pose correction algorithm is applied, which make use of online projection of the 3D model. This pose correction algorithm is an iterative procedure, which in each iteration corrects the position vector by computing the triangulation error in simulation and correcting the orientation estimate on the basis of the updated position estimate.

C. Markerless Motion Capture

In the following, our real-time stereo-based human motion capture system presented in [31] will be summarized briefly. The input to the system is a stereo color image sequence, captured with the built-in wide-angle stereo pair of the humanoid robot ARMAR-IIIb, which can be seen in Figure 5. The input images are preprocessed, generating output for an edge cue and a so-called distance cue, as introduced in [32]. The image processing pipeline for this purpose is illustrated in Figure 3. Based on the output of the image processing pipeline, a particle filter is used for tracking the movements in joint angle space. For tracking the movements, a 3D upper body model with 14 DoF (6 DoF for the base transformation, 2·3 for the shoulders, and 2·1 for the elbows) consisting of rigid body parts is used, which provides a simplified description of the kinematic structure of the human upper body. The model configuration is determined by the body properties like the limbs length of the observed human subject. The core of the particle filter is the likelihood function that evaluates how well a given model configuration matches the current observations, i.e. stereo image pair. For this purpose, an edge cue compares the projected model contours to the edges in the image. On the basis of an additional 3D hand/head tracker, the distance cue evaluates the distance between the measured positions and the corresponding positions inferred by the forward kinematics of the model. Various extensions are necessary for robust real-time application such as a prioritized fusion

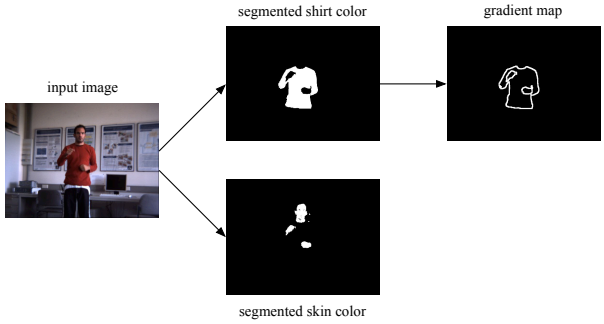

Fig. 3. Illustration of the image processing pipeline.

method, adaptive shoulder positions, and the incorporation of the solutions of the redundant arm kinematics. The system is capable of online tracking of upper body movements with a frame rate of 15 Hz on a 3 GHz single core CPU. Details are given in [31].

IV. Grasp Mapping

Before the execution on the robot, the approach movement in the form of joint angle configurations and the recognized grasp type are mapped onto the robot. In order to map motion onto the robot, we proposed in our previous work (see [33]) the Master Motor Map (MMM), a standardized interface which features a high level of flexibility and compatibility, since it allows mapping from various motion capture systems to different robot embodiments. The MMM provides a reference kinematic model of the human body by defining the maximum number of DoF, currently 58, that can be used by a human motion capture module and a robot. Trajectories in the MMM file format can be represented in joint angle space as well as in Cartesian space. Concerning movements in Cartesian space, in order to enable grasping and manipulation tasks, the MMM provides mapping of the desired 6D pose and the grasp type on the robot's end effector. A proper connection via the MMM of a motion capture module to a robot requires the implementation of a conversion module which transforms module specific data into the MMM file format and vice versa for overcoming different Euler conventions, active joint sets and orders of the joint angle values between the modules. As depicted in Figure 4, in the current system one conversion module has been implemented for each human motion capture system, converting the motion capture data to the MMM format. A third conversion module is implemented for mapping the MMM data to the kinematics of ARMAR-IIIb.

Along with the approach movement in the form of joint angle values the grasp type and the estimated hand orientation are passed from the hand pose estimation system to the robot through the MMM interface. According this data, from a set of preimplemented grasp the corresponding one is selected to be executed. To complete the grasp mapping, the grasp type to be performed is adjusted regarding the extent of the object shape. For this purpose, a rudimentary grasp type adjustment is implemented, which projects the object

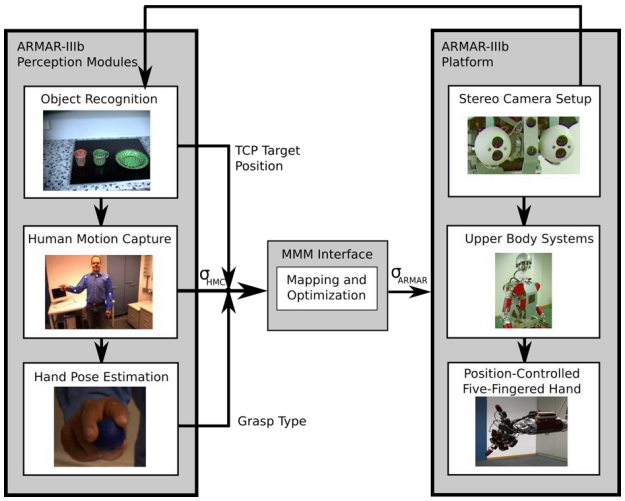

Fig. 4. Structure of the entire framework.

shape onto the thumbs position such that the thumbs tip lies on the shapes margin. The aperture of the fingers is scaled in a way that the positions of the remaining finger tips also approximately meet the margin of the shape. This method works on objects with simple shape properties.

A. Grasp Execution

The grasp reproduction of ARMAR-IIIb is performed in three different stages. The first stage describes the approach movement of the end effector towards the object based on the observed movement, while in the second stage the end effector is placed at the final grasp pose. The reproduction concludes with the execution of the recognized grasp type. Regarding the approach stage, by mapping these joint angle movements onto the robot, through forward kinematics one obtains a trajectory of the TCP in Cartesian space. The resulting trajectory is not sufficient for a goal-directed reproduction due to differences in the kinematic structure between the embodiments of the robot and a human e.g. mechanical joint constraints, differing joints and limb measurements. Therefore, the TCP trajectory for movements such as grasping is stretched and directed towards the object position to be reached. In order to attain a goal-directed reproduction, which additionally should feature a high similarity to the demonstrated human movement, in each frame, joint angles as well as desired TCP position of the modified trajectory have to be considered during execution. In [34], we developed an approach, which supports reproduction of observed human motion on the robot using non-linear optimization methods. In order to formulate an optimization problem which comprises displacements in Cartesian space regarding the TCP position as well as in joint angle space, a similarity measure is defined as follows:

$$S(\boldsymbol{\sigma}) = 2 - \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\sigma_i}^t - \sigma_i\right)^2}{\pi^2} - \frac{\frac{1}{3}\sum_{k=1}^{3}\left(\hat{p_k}^t - p_k\right)^2}{\left(2 \cdot l_{arm}\right)^2} \quad (4)$$

with n representing the number of joints, $\sigma_i, \hat{\sigma_i}^t \in [0, \pi]$ and $p_k, \hat{p_k}^t \in [-l_{arm}, l_{arm}]$, whereas l_{arm} describes the robot's arm length. The reference joint angle configuration is denoted by $\hat{\boldsymbol{\sigma}} \in \mathbb{R}^n$, while $\hat{\mathbf{p}} \in \mathbb{R}^3$ stands for the desired TCP position. The current TCP position $\mathbf{p}$ can be determined by applying the forward kinematics of the robot to the joint angle configuration $\boldsymbol{\sigma}$. Based on Equation 4 and the joint constraints $\{(C_{min}, C_{max})\}$ of a robot with n joints, one obtains following constrained optimization problem:

$$\min S'(\boldsymbol{\sigma}) = 2 - S(\boldsymbol{\sigma}) \quad (5)$$

$$\text{subject to} \quad C_{i_{min}} \leq \hat{\sigma}_i \leq C_{i_{max}} \quad (6)$$

For solving Equation 5, we apply the Levenberg-Marquardt algorithm, since it features numerical stability and more robust convergence compared to other optimization algorithms such as the Gauss-Newton and the steepest descent method. Following this optimization approach a trade-off is attained, which on the one hand results in an accurate TCP positioning with small displacement error while it provides on the other hand a feasible robot joint angle configuration resembling the observed human configuration. This way goal-directed imitation of the approach movement is achieved. For further details, the reader is referred to [34]. For the execution of the final grasp phase, due to errors and inaccuracies originating from the object localization and the robot's mechanical elements, a displacement error arises between the TCP and the object that has to be diminished. To achieve exact alignment of the end effector and the robot, we make use of visual servoing methods as presented in [35]. Within this approach the hand and object are tracked. The resulting distance between both is reduced and the hand orientation is controlled. The hand orientation estimate coming from the grasp recognition module is used to determine if the grasp should be executed from the top or from the side. Therefore, the hand is placed over the object if the palm orientation was similar to the table plane, or next to the object otherwise.

V. Experiments

A. Experimental Setup

The humanoid platform ARMAR-IIIb, a copy the humanoid robot ARMAR-IIIa [36], serves as the experimental platform in this work. From the kinematics point of view, the robot consists of seven subsystems: head, left arm, right arm, left hand, right hand, torso, and a mobile platform. The head has seven DoF and is equipped with two eyes, which have a common tilt and independent pan. Each eye is equipped with two digital color cameras, one with a wide-angle lens for peripheral vision and one with a narrow-angle lens for foveal vision. The upper body of the robot provides 33 DoF: 2·7 DoF for the arms and three DoF for the torso. The arms are designed in an anthropomorphic way: three DoF for each shoulder, two DoF in each elbow and two DoF in each wrist. Each arm is equipped with a five-fingered hand with eight DoF. The locomotion of the robot is realized using a wheel-based holonomic platform.

The proposed approach was integrated on the humanoid platform ARMAR-IIIb and was successfully applied. For

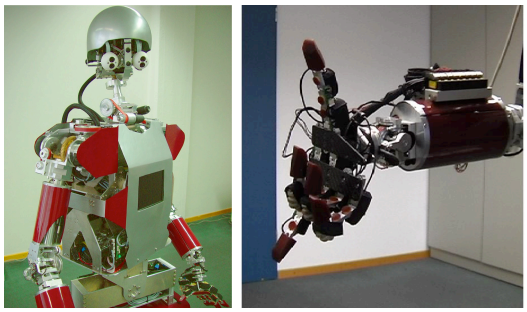

Fig. 5. Left: The humanoid robot ARMAR-IIIb. Right: Position-controlled right hand with 8 DoF.

the experiments, objects were used which can be easily identified such as single-colored cups. The experimental setup stipulates that demonstration of the grasp is performed in front of the robot. Observation is initiated by scanning the scene for known objects. Once an object is found, tracking of the human upper body is triggered leading to the capturing process of movements in the approach stage. This process is finished once the hand is positioned within a tolerated distance to a specific object. At this point, observation is switched to the hand pose estimation whereby its classification and the outcoming orientation complete the motion data of the grasp. As described in Section IV, the data is mapped onto robot, optimized to its embodiment and executed. In the execution phase, the robot searches for the same object which was grasped in the demonstration and approaches it. Based on the classification of the grasp type, an adequate instance is selected from the set of implemented grasp on the robot which is modified to the objects appearance. The hand pose recognition system was running on an external computer, while the rest of the system was running on ARMAR-IIIb. The communication between the two systems was performed through UDP sockets. It is possible to run the whole system on the robot, but this setup was more preferable for debugging purposes. Two sets of experiments were performed: in the first one, the whole system (grasp observation, mapping and execution) was tested with a reduced set of grasps: power grasp from top, power grasp from side, and pinch grasp(see Figure 6). In the second one, the set of grasps was extended to five of them (power sphere, prismatic wrap, parallel extension, tripod, and pinch). However, the execution of the grasp was reduced to the hand pose, keeping the arm still (see Figure 7).

B. Experimental Results

As depicted in Figures 6 and 7 the robot successfully imitated the demonstrated grasp including approach and grasp type. Since a non-linear optimization method is applied during approaching, we attained a trade-off between the similarity of the reproduced movement concerning the demonstration and accuracy in terms of positioning of the end effector regarding goal-directed tasks. Furthermore, the applied method provided a unique solution in terms of joint angles, which standard inverse kinematics methods fail to do due to singularities and redundancies. Nevertheless, in the approach phase, we experienced a displacement error of max $65mm$ caused by kinematic inaccuracies which varies depending on the cups distance regarding the end effector. The displacement could be recovered by using visual servoing. In order to test the grasp classification module, each grasp was executed 20 times for the Experiment 2. The results are shown in Table I. An overall classification accuracy of 72% was achieved, clearly over the human baseline for grasp recognition with similar grasps [21], with four out of five grasp types with accuracies over 80%. The differences between human model and synthetic had a stronger effect in the parallel extension grasp, lowering the accuracy for that particular grasp. Results of the grasp recognition, mapping and execution on the humanoids robot ARMAR-IIIb are shown in the accompanying video submission, which is also available under `wwwiaim.ira.uka.de/users/do/GraspRecognitionDivx.avi`.

Grasp Type	Illustration	Correct Classification Rate
Power Sphere		80 %
Prismatic Wrap		95 %
Parallel extension		50 %
Tripod		85 %
Pinch		80 %

TABLE I
GRASP TYPE CLASSIFICATION RESULTS.

VI. Conclusions

In this paper, we presented a system for grasp recognition, mapping and execution on a humanoid robot. Human grasping activities are captured using markerless motion capture system and mapped to the humanoid robot ARMAR-IIIb. Human upper body tracking, object tracking and hand pose estimation techniques are applied to perceive human object grasping movements. The recognized grasps are mapped and executed on a humanoid robot with a five-fingered hand.

VII. Acknowledgments

The work described in this paper was partially conducted within the EU Cognitive Systems projects GRASP (FP7-215821) and PACO-PLUS (FP6-027657) funded by the European Commission.

References

[1] C. G. Atkeson and S. Schaal, "Learning tasks from a single demonstration," in *IEEE International Conference on Robotics and Automation (ICRA97)*, 1997, pp. 1706–1712.

[2] D. Gavrila and L. Davis, "3-D Model-based tracking of humans in action: a multi-view approach," in *Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, USA, 1996, pp. 73–80.

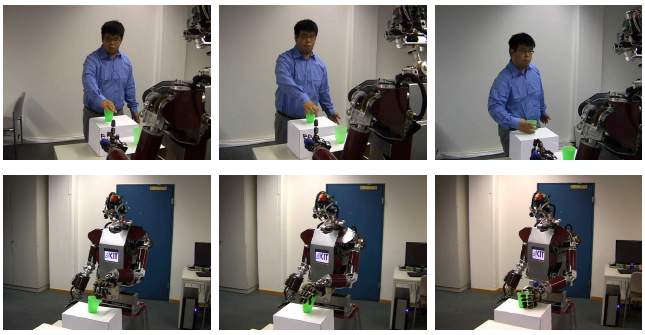

Fig. 6. Image Samples of Experiment 1. Top: human performing pinch grasp, top power grasp, and side power (left to right). Bottom: reproduction of the approach movement and the grasp type to grasp an object.

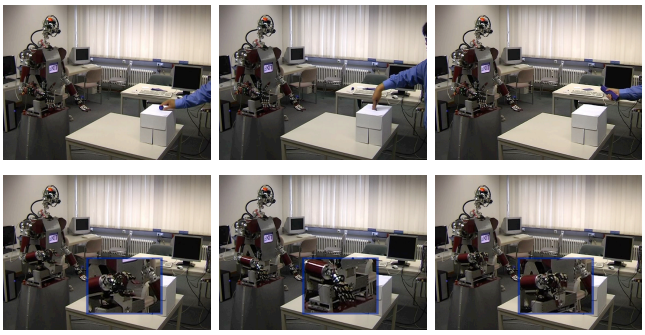

Fig. 7. Image Samples of Experiment 2. Top: human performing tripod grasp, parallel extended finger grasp, and prismatic wrap (left to right). Bottom: reproduction of the grasp type.

[3] K. Rohr, "Human Movement Analysis based on Explicit Motion Models," in *Motion-Based Recognition*, 1997, pp. 171–198.

[4] C. Bregler and J. Malik, "Tracking People with Twists and Exponential Maps," in *Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, Santa Barbara, USA, 1998, pp. 8–15.

[5] D. Grest, D. Herzog, and R. Koch, "Monocular Body Pose Estimation by Color Histograms and Point Tracking," in *DAGM-Symposium*, Berlin, Germany, 2006, pp. 576–586.

[6] J. Deutscher, A. Blake, and I. Reid, "Articulated Body Motion Capture by Annealed Particle Filtering," in *Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, Hilton Head, USA, 2000, pp. 2126–2133.

[7] C. Kim, D. Kim, and Y. Oh, "Adaption of Human Motion Capture Data to Humanoid Robots for Motion Imitation using Optimization," *Integrated Computer-Aided Engineering*, vol. 13, no. 4, pp. 377–389, 2006.

[8] A. Ude, C. Atkeson, and M. Riley, "Programming Full-Body Movements for Humanoid Robots by Observation," *Robotics and Autonomous Systems*, vol. 47, no. 2-3, pp. 93–108, 2004.

[9] X. Zhao, Q. Huang, Z. Peng, and K. Li, "Humanoid Kinematics Mapping and Similiarity Evaluation based on Human Motion Capture," in *IEEE International Conference on Information Acqusition*, Hefei, China, June 2004, pp. 426–431.

[10] N. Pollard, J. Hodgins, M. Riley, and C. Atkeson, "Adapting Human Motion for the Control of a Humanoid Robot," in *IEEE International Conference on Robotics and Automation*, Washington, DC, USA, May 2002, pp. 1390–1397.

[11] M. J. Matarić, "Sensory-motor primitives as a basis for imitation: Linking perception to action and biology to robotics," in *Imitation in Animals and Artifacts*, K. Dautenhahn and C. Nehaniv, Eds., 2000.

[12] E. Oztop, J. Babic, J. G. Hale, G. Cheng, and M. Kawato, "From biologically realistic imitation to robot teaching via human motor learning," in *ICONIP (2)*, vol. 4985, 2007, pp. 214–221.

[13] M. C. Lopes and J. S. Victor, "Motor representations for hand gesture recognition and imitation," in *IROS Workshop on Robotic Programming by Demonstration*, 2003.

[14] S. Calinon, A. Billard, and F. Guenter, "Discriminative and adaptative imitation in uni-manual and bi-manual tasks," in *Robotics and Autonomous Systems*, vol. 54, no. 5, 2005, pp. 370–384.

[15] A. Erol, G. N. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "A review on vision-based full DOF hand motion estimation," in *Vision for Human-Computer Interaction*, 2005, pp. III: 75–75.

[16] Q. Delamarre and O. D. Faugeras, "3D articulated models and multiview tracking with physical forces," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 328–357, 2001.

[17] B. D. R. Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla, "Filtering using a tree-based estimator," in *IEEE International Conference on Computer Vision*, 2003.

[18] E. Sudderth, M. I. Mandel, W. T. Freeman, and A. S. Willsky, "Visual hand tracking using non-parametric belief propagation," in *IEEE Workshop on Generative Model Based Vision*, 2004.

[19] M. de la Gorce, N. Paragios, and D. J. Fleet, "Model-based hand tracking with texture, shading and self-occlusions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[20] R. Rosales, V. Athitsos, L. Sigal, and S. Sclaroff, "3D hand pose reconstruction using specialized mappings," in *IEEE International Conference on Computer Vision*, 2001, pp. I: 378–385.

[21] H. Kjellström, J. Romero, and D. Kragić, "Visual recognition of grasps for human-to-robot mapping," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008.

[22] T. Asfour, K. Welke, P. Azad, A. Ude, and R. Dillmann, "The karlsruhe humanoid head," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, Daejeon, Korea, December 2008.

[23] W. T. Freeman and M. Roth, "Orientational histograms for hand gesture recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 1995.

[24] G. Shakhnarovich, P. Viola, and T. Darrell, "Fast pose estimation with parameter sensitive hashing," in *IEEE International Conference on Computer Vision*, vol. 2, 2003, pp. 750–757.

[25] H. N. Zhou, D. J. Lin, and T. S. Huang, "Static hand gesture recognition based on local orientation histogram feature distribution model," in *Vision for Human-Computer Interaction*, 2004, p. 161.

[26] M. Cutkosky, "On grasp choice, grasp models and the design of hands for manufacturing tasks," *IEEE Transactions on Robotics and Automation*, vol. 5, no. 3, pp. 269–279, 1989.

[27] N. Kamakura, M. Matsuo, H. Ishi, F. Mitsuboshi, and Y. Miura, "Patterns of static prehension in normal hands," *Am J Occup Ther*, vol. 7, no. 34, pp. 437–45, 1980.

[28] S. B. Kang and K. Ikeuchi, "Grasp recognition using the contact web," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1992, pp. 194–201.

[29] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," *Communications of the ACM*, vol. 51, 2008.

[30] P. Azad, T. Asfour, and R. Dillmann, "Stereo-based 6d object localization for grasping with humanoid robot systems," in *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, 29 2007-Nov. 2 2007, pp. 919–924.

[31] P. Azad, T. Asfour, and R.Dillmann, "Robust real-time stereo-based markerless human motion capture," in *IEEE International Conference on Humanoid Robots*, Daejeon, Korea, December 2006, pp. 700–707.

[32] P. Azad, A. Ude, T. Asfour, and R. Dillmann, "Stereo-based Markerless Human Motion Capture for Humanoid Robot Systems," in *IEEE International Conference on Robotics and Automation (ICRA)*, Roma, Italy, 2007, pp. 3951–3956.

[33] P. Azad, T. Asfour, and R. Dillmann, "Toward an Unified Representation for Imitation of Human Motion on Humanoids," in *IEEE International Conference on Robotics and Automation*, Rome, Italy, April 2007.

[34] M. Do, P. Azad, T. Asfour, and R. Dillmann, "Imitation of Human Motion on a Humanoid Robot using Non-Linear Optimization," in *IEEE International Conference on Humanoid Robots*, Daejeon, Korea, December 2008, pp. 545–552.

[35] 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.

[36] T. Asfour, K. Regenstein, P. Azad, J. Schröder, N. Vahrenkamp, and R. Dillmann, "ARMAR-III: An Integrated Humanoid Platform for Sensory-Motor Control," in *IEEE/RAS International Conference on Humanoid Robots*, 2006.