Towards Grasp Learning in Virtual Humans by Imitation of Virtual Reality Users

Matthias Weber, Guido Heumer, Bernhard Jung

ISNM International School of New Media
University of Lübeck
Willy-Brandt-Allee 31c
23554 Lübeck
Tel.: +49 (0)451 2967 0
Fax: +49 (0)451 2967 40
E-Mail: {weber|gheumer|jung}@isnm.de

Abstract: Virtual humans capable of autonomously interacting with virtual objects could prove highly beneficial in virtual prototyping, e.g., for demonstration and verification operating, maintenance, assembly and other procedures. An attractive method for skill acquisition for such virtual humans would be to enable the virtual humans to imitate procedures first performed by Virtual Reality (VR) users on the virtual prototypes. This paper presents first steps towards such autonomous, learning virtual humans and describes methods for the analysis of grasps performed by VR users equipped with data-gloves as well as methods for autonomous, behavior-based grasping in virtual humans. The methods for grasp analysis and synthesis share a sensor-enriched hand model as well as an empirically founded grasp taxonomy which serve to compensate for imprecisely performed human grasps, e.g., due to lack of tactile feedback during object interactions, to enable a collision-free grasp behavior in virtual humans.

Keywords: Virtual Humans, Autonomous Grasping, Imitation Learning

1 Introduction

Animated virtual humans demonstrating the operation, maintenance, or other procedures on digital product models play an increasing role in virtual prototyping. Applications range from relatively simple animations that serve as visual communication means for marketing purposes or coordination within product development teams to more complex ergonomic verifications of virtual prototypes. The overall goal of our research is the development of a novel animation method for virtual humans in virtual prototyping applications: First, a human VR user performs a procedure on a virtual prototype. Then, through suitable recording and abstraction of that procedure, virtual humans of different sizes are enabled to perform the procedure themselves. Note that this approach differs from conventional motion capture that usually does not involve interactions with 3D objects; rather, it is related to methods known as Programming by Example or imitation learning in the field of robotics.

One benefit of the proposed approach is that it would significantly simplify the animation production process as animations are generated from natural 3D interactions in VR instead of complex WIMP interfaces characteristic of today's animation systems. Furthermore, value would be added to interactive VR systems in that prototype evaluations would not only be based on the experience of *one* VR user but on documentable performances of *many* virtual humans of different size, gender, and other anthropometric properties. An example of such a virtual human is "Vincent" as shown in figure 1.

Figure 1: Virtual human "Vincent"

This paper describes on-going work and first results towards the outlined goal of imitation learning in virtual humans. More concretely, it focuses on the analysis and synthesis of different types of one-handed grasping (rather than complete, possibly two-handed procedures performed on virtual prototypes). Related work in robotics and empirical sciences is described in section 2. To compensate for inaccurate sensor information when analyzing grasps of VR users as well as to support the autonomous grasping of virtual humans, a knowledge-based approach involving an empirically founded grasp taxonomy and a collision-sensor enriched hand model have been developed (section 3). The main output of the analysis phase of grasps performed by the VR user is their classification w.r.t. the grasp taxonomy; classification is a multi-stage process that involves the computation of features on several levels based on the contact points of the virtual hand with a 3D object (section 4). The second phase of our imitation learning approach is the synthesis of grasp animation in virtual humans; a behavior-based method has been implemented where grasps are generated from high-level descriptions and executed under continuous feedback from collision sensors (section 5). Section 6 presents current results and outlines further developments.

2 Related Work

Human grasping has been a subject of research for a long time. In the medical field a considerable amount of research has been carried out to learn how the hand works and how humans grasp objects (an overview of grasps is given in [EBMP02]). Additionally many achievements in grasping, particularly concerning algorithmic simulations, come from the robotics field; see, e.g., [BK00] for an overview. Research in robotics is however not restricted to the mere generation of grasps but also to learning from human instructors who demonstrate the grasp first, e.g., [ZR03]. This leads to the field of Programming by Demonstration (see, e.g., [ACR03]).

Kang and Ikeuchi [KI92] discuss the analysis and classification of human grasps based on a number of contact points between hand and object arranged in a 3D graphical representation – the *Contact Web*. Each finger segment has one contact point associated with it. Ekvall

and Kragic [EKed] present a hybrid approach of grasp recognition using Hidden-Markov-Model-based fingertip position evaluation and arm trajectory evaluation. However, only three fingertip positions are evaluated as hand configuration, which are furthermore tracked by rather imprecise magnetic trackers. Taking the whole joint configuration of the hand into account, tracked with a data-glove, could probably considerably improve recognition reliability.

Concerning the grasp generation in virtual humans, one type of distinguishing the methods for grasping is dividing them into semi-automatic and automatic methods [RBH+95]. Semi-automatic grasps are first performed by a user with a data-glove and then mapped to a virtual hand. Multi-sensor collision detection methods avoid penetration with the virtual object. Automatic grasp methods do not need the user's input via data-glove as they can execute the grasp themselves. In either way some kind of collision detection is needed. To accomplish grasping an object without penetrating it, sensors can be used to efficiently detect collision with the object to grasp [HBTT95].

In the Smart Object approach, virtual objects are annotated with information of how to grasp or otherwise interact with the object [KT99, Kal04]. As there are many possible ways of grasping different or even the same object, grasp taxonomies have also been considered in such research. Some of these taxonomies, particularly our own grasp taxonomy are described next.

3 Grasp Representation

The transfer of object grasping performed in VR to virtual humans requires robustness against inaccurate sensor data from VR input devices, also due to missing tactile feedback with conventional data-gloves. To compensate for the vagueness of the input data, a knowledge-based approach involving an empirically founded grasp taxonomy as well as a collision-sensor enriched hand model have been developed.

3.1 Grasp Taxonomy

To ensure independence of hand and object geometry, a high level representation of the grasp is required. Grasp taxonomies which categorize different grasp types provide such high level representations. Several categorizations of grasps have been proposed in the literature. This began with research in the medical field where grasp sequences of humans have been studied by Schlesinger [Sch19]. His classification is based on the shape of the object to grasp and includes six different grasp types: "cylindrical grasp", "tip grasp", "hook grasp", "palmar grasp", "spherical grasp" and "lateral grasp". Grasps can also be categorized by the stability of the grasp, as, e.g., in the work of Napier [Nap56] and Mishra and Silver [MS89]. This line of research differentiates between two basic grasp types, i.e., the power grip which holds an object firmly and the precision grip where the thumb and

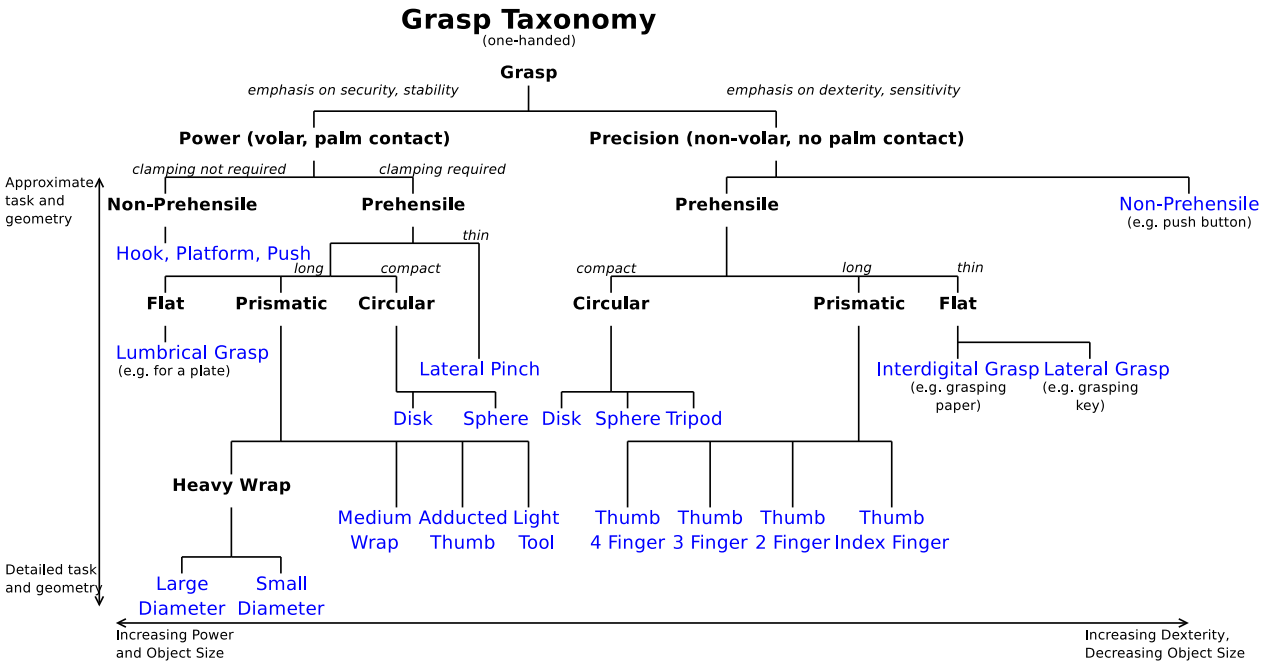

Figure 2: Grasp taxonomy (an extension of Cutkosky's taxonomy [Cut89])

other fingers hold the object. Cutkosky developed a taxonomy on the basis of research about the work of mechanics to achieve optimal grasp operations in factories [Cut89]. Ehrenmann et al. [ERZD02] distinguish static and dynamic grasps; in static grips the fingers remain unchanged, while in dynamic grips the finger positions vary to keep the grasped object stable.

The grasp taxonomy introduced by Kang and Ikeuchi [KI92] is strongly based on the contact web (see section 2) and thus facilitates grasp classification based on observed sensory data. On the highest hierarchy level a distinction is made between volar (palm contact) and non-volar (no palm contact) grasps. The non-volar grasps are subdivided into fingertip and composite non-volar grasps, while the more complex subdivision of volar grasps is based on the relative locations of contact points in space.

In our work we extended Cutkosky's grasp taxonomy [Cut89] and integrated the work of [KI92] and [EBMP02] to add some missing grasps. Mainly these additional grasps provide a broader distinction between flat-shaped grasps, like, e.g., the platform push, and non-prehensile grasps, like pushing a button. Figure 2 shows our taxonomy.

3.2 Hand Model

The skeleton structure of our hand model is based on the H-ANIM standard (www.h-anim.org), featuring 15 finger joints (three for each finger). The metacarpophalangeal joint of each finger – i.e., the joint attaching the finger to the palm – has two degrees of freedom (DOF): flexion and pivot. In contrast, the two subsequent joints only have one DOF: flexion. In addition to the finger joints, the hand model has a wrist joint with three DOFs, with its

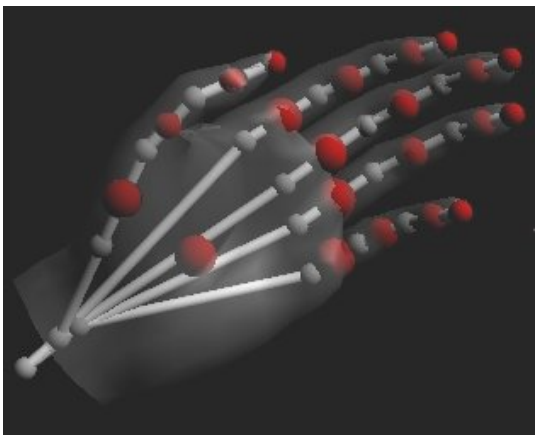

Figure 3: Skeleton model of the hand. The larger spheres denote sensors while the light grey spheres represent joints.

rotational center forming the origin of the hand coordinate system. This results in a total number of 23 DOFs. Joint angle constraints are modeled according to [ST94].

To perform collision detection and contact point determination, the hand model is fitted with sphere sensors. This approach has also been taken in [RBH+95], but differing from that approach, our sensors are placed in the center of each segment instead of in the joints. Additionally we have placed one sensor in the palm center to determine palm contact (volar/non-volar) of a grasp, which is essential for the identification of power vs. precision grasps. Both modifications provide a more accurate mapping to the contact web set of contact points [KI92], which are also situated in the segment centers. Furthermore, by also providing small sphere sensors in the fingertips, we extend the contact web structure to represent a broader range of grasps, which emphasize fingertip contact, like, e.g., button press. Figure 3 illustrates our hand model with attached sphere sensors.

4 Grasp Analysis

In order for a virtual human to learn from a user's grasp, the grasp has to be analyzed and the defining features of the grasp have to be determined. On the hardware level, a tracking device is needed to acquire the user's hand posture and position. Contact points of the user's (virtual) hand and the virtual object are determined in a subsequent software collision-detection step. From these low-level grasp features, higher-level features can be deduced, that provide a basis for classification of the grasp according to the grasp taxonomy. Finally, further inferences about the grasp can be drawn in a post-processing step, e.g., about the grasp purpose. This section describes this process of feature extraction and grasp classification as conceived in our current research work.

4.1 Basic Features

At the basic level a human grasp consists of a number of finger (and hand) joint angles, which define the hand posture. Furthermore in interaction with an object, a grasp consists of a number of *contact points* – points, where the hand touches the grasped object. These low-level features are mere mechanical facts, which need to be determined as exactly as possible by a combination of hardware tracking and software. Currently we are using an 18-sensor Cyberglove (by Immersion Corp.) to track the user's hand posture. The sensor data of this type of data-glove does not provide a complete representation of the hand posture. Flexion angles of the distal finger joints are not tracked and pivot movements of the finger are only determined as relative angles between the fingers.

Since we deal with virtual objects, no real contact between these objects and the user's hand occurs. Therefore, a virtual hand model (see section 3.2) is added to the virtual scene, representing the user's hand in the virtual world. In a first processing step, the glove sensor input is mapped to joint rotation angles of the virtual hand model. The mapping of the 18 Cyberglove sensor values to the 15 finger joints (with a total of 20 DOFs) is based on a heuristics that estimates the missing information.

While the user performs the grasp, the sensors of the hand model provide contact point information. This simple model only provides touch information on a per-segment basis, but does this in a quick and efficient way. As shown in [KI92] it is sufficient to regard one contact point per finger segment for the purpose of grasp classification. The exact position (or area) of contact is not necessary to uniquely classify a grasp within a grasp taxonomy. As only approximate contact positions are required, the feature extraction process becomes robust against inaccuracies introduced by tracking and contact point calculation.

After joint rotations and contact points have been determined, the grasp posture of the virtual hand is corrected, so that no intersections of hand and object occur, and to guarantee that joint rotations stay within given constraints.

4.2 Medium-Level Features

From the basic features, several medium-level features can be extracted or calculated, such as *virtual fingers*, *opposition space* and *grasp cohesive index*. Virtual fingers, introduced by Arbib et al. [AIL85] describe a functional unit of one or more real fingers. Real fingers comprising one virtual finger exert a force in unison, opposing the object or other virtual fingers in the grasp. The mapping from real to virtual fingers can be determined, based on the contact web.

Related to virtual fingers is the concept of *opposition space* as defined by Iberall et al. [IBA86] as: "the area within coordinates of the hand, where opposing forces can be exerted between virtual finger surfaces in effecting a stable grasp". Mainly important for prehensile grasps,

three different forms of opposition are identified, with which an object can be clamped. The type of opposition present in a grasp proves helpful for its characterization.

Lastly, Kang and Ikeuchi [KI92] define the concept of *grasp cohesive index*, a numerical value, indicating the overall similarity of action of fingers within the given virtual finger mapping of a grasp. Grasp classification within the contact web grasp taxonomy is strongly based on this feature.

4.3 Concept-Level Features

Based on its basic and medium-level features, a grasp can be classified according to the grasp taxonomy. First a broad classification is performed with regard to the contact web taxonomy based on the number and position of detected contact points. This classification is refined further based on medium-level features to reflect the full depth of our taxonomy. Furthermore certain tool or special purpose grasps can be identified based on particular finger configurations, like, e.g., scissors, chopsticks etc.

After classification, the grasp category yields a high-level representation of the grasp involving features on the concept-level, such as whether or not the hand clamps the object (*prehensile / non-prehensile*), whether the focus lies on exerting as much force as possible on the object or to manipulate the object as precisely as possible (*power / precision grasp*) etc.

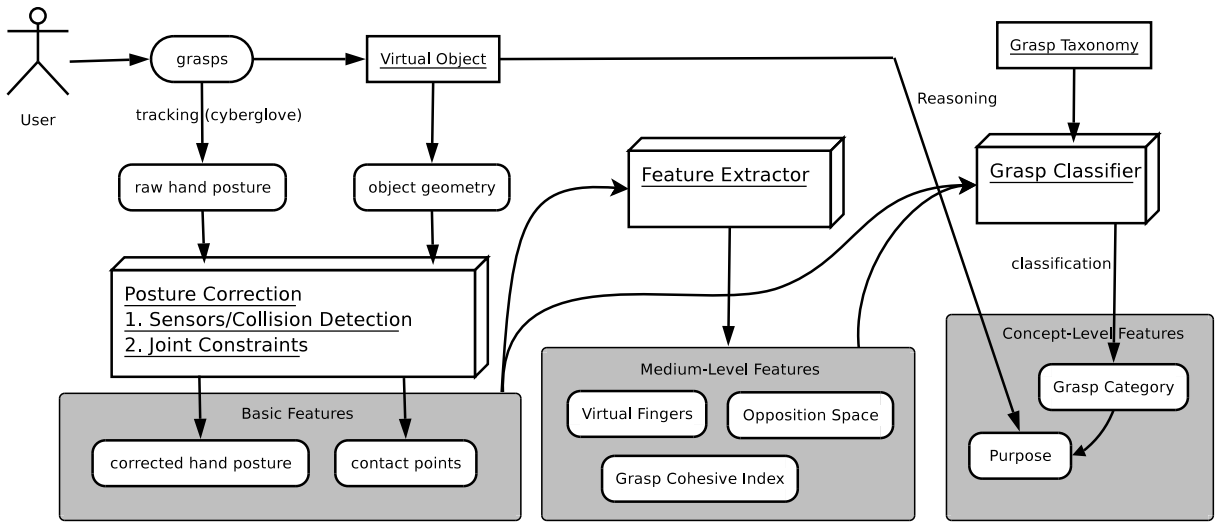

Figure 4: The grasp analysis process.

If the grasp falls into the category of tool or special-purpose grasps, additional statements can be concluded about its purpose. Purpose information can also be concluded from the information, where an object has been grasped. For instance a distinction can be made

between use and displacement grasps. In the former case a knife would be firmly grasped by its grip to cut with, while in the latter case it would probably be carefully grasped by its blade to, e.g., hand it to another person. This type of purpose information can provide additional cues in the grasp synthesis functions in virtual humans.

Figure 4 illustrates the complete grasp analysis process. The different levels of grasp features together form a representation of the user's grasp with high-level features being the most abstract and low-level features being the most concrete. This representation enables a virtual human to imitate or reproduce the grasp, while being independent of exact object or hand geometries. This process of grasp synthesis is described in the following section.

5 Grasp Synthesis

To close the circle of learning and imitation, a virtual human not only has to analyze manipulation tasks performed by a human, but also has to manipulate virtual objects himself. A behavior-based approach to grasp synthesis in virtual humans has been implemented, where grasping is performed under continuous feedback from collision sensors.

5.1 Grasp generation overview

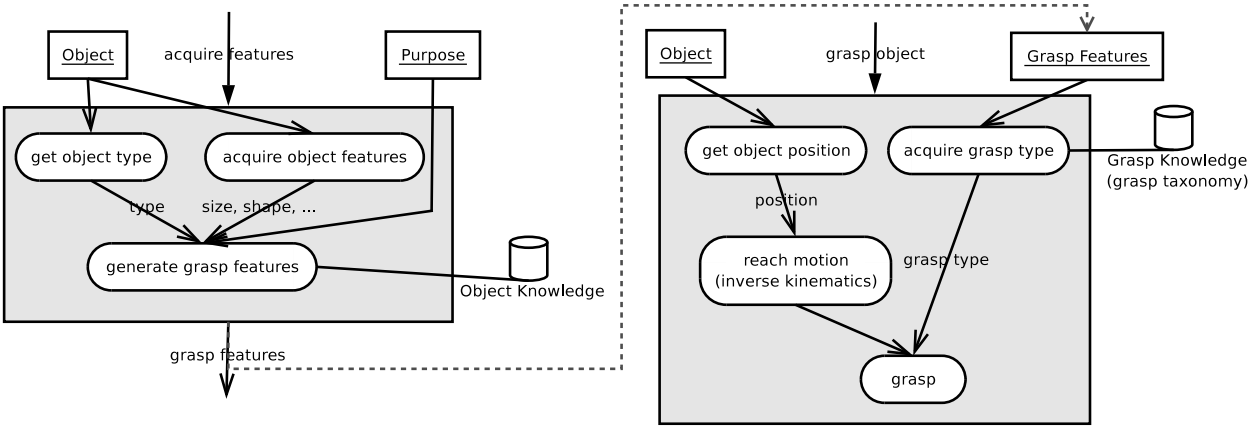

Figure 5: Obtaining grasp features and generating the grasp

The virtual humans in our approach are capable of autonomously performing grasps (and in the future more complex procedures), based on high-level descriptions and plans. They learn from a VR user on the basis of specific examples but can later apply their skills to a range of similar tasks, involving, e.g., different, similarly shaped objects. This means, in typical mode operation, the grasp synthesis module will just receive as input the object to grasp and possibly also the purpose (see section 4.3) of the grasp.

The first step of grasp planning thus involves an analysis of the object to grasp (see left side of figure 5). Relevant object features include, among others, the generic object type (like hammer, cup, etc.), size and shape. Based on these features and the purpose of the grasp,

grasp features can be generated. These grasp features are based on the contact web (see sections 2 and 3). The next step is to use these grasp and object features to generate an appropriate grasp.

Based on the computed grasp and object features, the grasp type is determined. This is achieved through a mapping of features to grasp types. This mapping is contained in a grasp knowledge-base. The grasp types in this knowledge-base correspond to our grasp taxonomy (see section 3.1).

The execution of the actual grasp action is preceded by a reach motion. During the reach motion, the hand is moved towards the object, while at the same time already shaping the hand to get a good starting position for the grasp itself. The target hand position of the reach motion is calculated, using a specific type of inverse kinematics that is based on forces applied to the joints in the kinematic chain. The inverse kinematics is computed using an iterative method suitable for the 7 DOF arms of our virtual humans. After finishing the reach motion, the object is grasped according to the already established grasp type. Figure 5 shows the complete process of grasp planning, reach motion, and grasping.

5.2 Sensor-based grasping

The internal process of moving the body for reaching and grasping is shown in figure 6. The animation system includes a motor control component that controls the state of the virtual human's skeleton and all underlying motor programs. Motor programs are primitive parts of the system that generate simple movements, like moving joints to a given end rotation

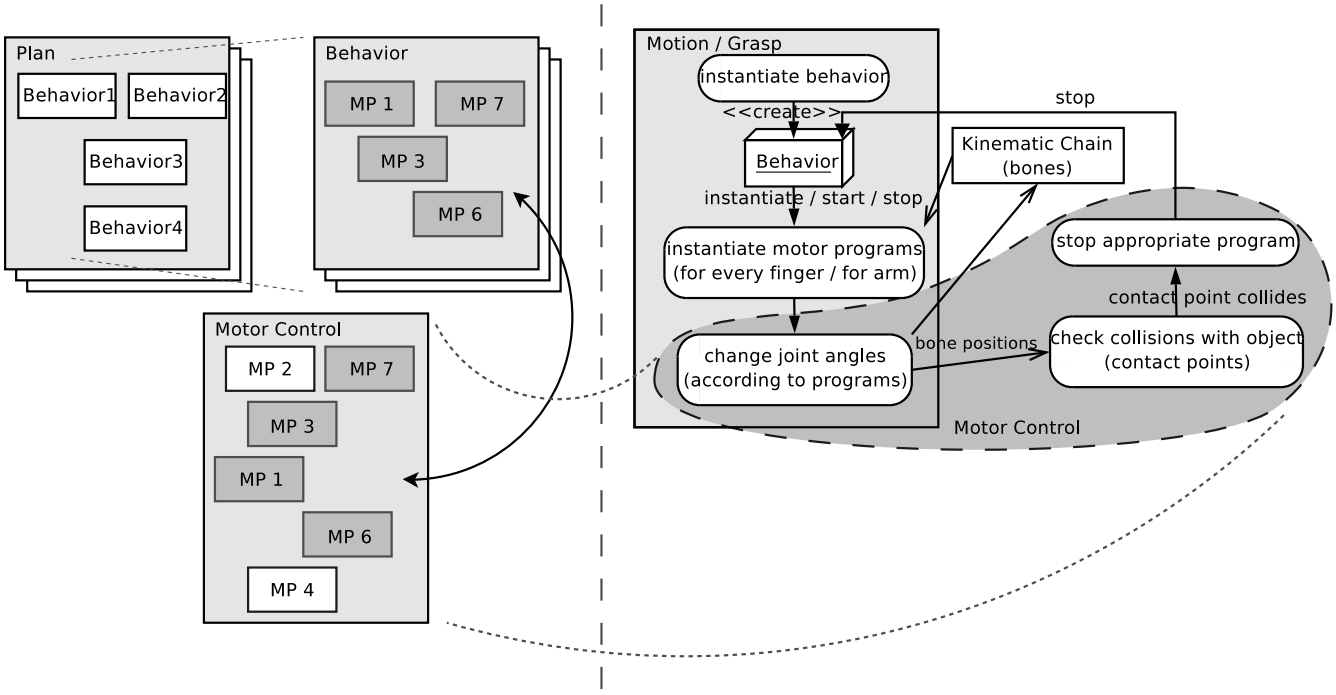

Figure 6: Plans, behaviors and motor control

or to move the end effector of a kinematic chain to a given end position. One level above, behaviors are scheduling these motor programs to achieve certain movements. Behaviors have a specific goal, e.g., closing the hand with a given grasp type, opening the hand, or moving it to a given position. Behaviors instantiate motor programs and can also stop them, e.g., when collisions are detected for a finger involved in a motor program. Furthermore, behaviors can be grouped into plans. At the moment, these plans typically consist of a reach motion followed by a grasp action. In general, plans specify the consecutive or concurrent execution of behaviors. Plan parameters can be passed to behaviors, such as start and end time of a movement or the object to grasp. Plans are described in an XML-based language.

Details of the movement simulation loop are shown on the right side of figure 6. All movement plans, behaviors, and motor programs are goal directed, such as achieving a grasp type or reaching an end position of the end effector. Triggered by plans, the behaviors themselves start motor programs that change angles of joints they were assigned to. The motor programs are executed in every simulation step; priority values are used to handle conflicts when different motor programs attempt to update the same joint angles. The motor programs are informed when a collision occurs and therefore might stop their movement. If not stopped by collision detection they will stop their movement, when they reach their goal. Similarly, behaviors are informed when their motor programs finish their movement. Behaviors terminate when they reach their goal which usually is the termination of all its motor programs. In a plan this can lead to the instantiation of new behaviors. Finally, the movement process stops when all behaviors, either instantiated by hand or by plan execution, are finished.

6 Results and Future Work

We have presented a concept that aims at enabling virtual humans to imitate grasps performed by a human VR user. In the current stage of the work, several components have been implemented, including a sensor-enriched virtual hand model and a grasp taxonomy shared by both grasp analysis and synthesis processes. Further, on the analysis side, glove sensor data are mapped to joint angles of the virtual hand to provide one half of features from the basic feature set. From the virtual sensors on the hand model, the contact points are calclulated to provide the other part of the basic feature set. On the grasp synthesis

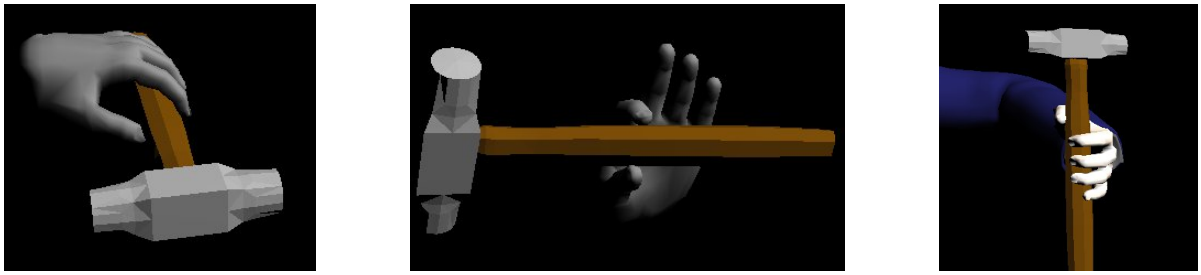

Figure 7: Different ways of grasping a hammer

side, simple precision and power grasps can currently be generated by commands, based on a partial implementation of plans, behaviors, and motor programs as described in section 5.2. Figure 7 shows examples how an object can be grasped in different ways using this approach. We have further implemented an extension to the Avango / Performer VR software by integration of the Cal3D skeletal character animation library (cal3d.sourceforge.net) to allow for the inclusion of deformable virtual human hand and body models. The goal of future work is to enable virtual humans to learn and execute longer virtual prototype operation and assembly procedures by imitation of VR users, as outlined in the introduction.

7 Acknowledgments

This research is partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the project ,,Virtual Workers".

References

[ACR03] J. Aleotti, S. Caselli, and M. Reggiani. Toward Programming of Assembly Tasks by Demonstration in Virtual Environments. *12th IEEE Int. Workshop on Robot and Human Interactive Communication*, 2003.

[AIL85] M.A. Arbib, T. Iberall, and D.M. Lyons. *Coordinated control programs for movements of the hand*, pages 111–129. Springer-Verlag, 1985.

[BK00] A. Bicchi and V. Kumar. Robotic Grasping and Contact: A Review. In *IEEE Int. Conf. on Robotics and Automation*, 2000.

[Cut89] M.R. Cutkosky. On grasp choice, grasp models and the design of hands for manufacturing tasks. *IEEE Trans. on Robotics and Automation*, 5(3), 1989.

[EBMP02] S.J. Edwards, D.J. Buckland, and J.D. McCoy-Powlen. *Developmental & Functional Hand Grasps.* SLACK Incorporated, Thorofare, NJ USA, 2002.

[EKed] S. Ekvall and D. Kragic. Grasp Recognition for Programming by Demonstration. *IEEE/RSJ International Conference on Advanced Robotics*, 2005 (to be published???).

[ERZD02] M. Ehrenmann, O. Rogalla, R. Zöllner, and R. Dillmann. Analyse der Instrumentarien zur Belehrung und Kommandierung von Robotern. *1. SFB-Aussprachetag, Human Centered Robotic Systems, HCRS*, 2002.

[HBTT95] Zhiyong Huang, Ronan Boulic, Nadia Magnenat Thalmann, and Daniel Thalmann. A Multi-sensor Approach for Grasping and 3D Iinteraction. In *Computer graphics: developments in virtual environments*, pages 235–253, London, UK, 1995. Academic Press Ltd.

[IBA86] T. Iberall, G. Bingham, and M.A. Arbib. *Opposition space as a structuring concept for the analysis of skilled hand movements*, pages 158–173. Number 15 in Experimental Brain Research Series. Springer-Verlag, 1986.

[Kal04] M. Kallmann. *Interaction with 3-D Objects*, pages 303–322. John Wiley & Sons Ltd., Chichester, West Sussex, England, 2004.

[KI92] S.B. Kang and K. Ikeuchi. Grasp Recognition Using the Contact Web. In *Proc. IEEE/RSJ Conference on Intelligent Robots and Systems*, 1992.

[KT99] M. Kallmann and D. Thalmann. A Behavioral Interface to Simulate Agent-Object Interactions in Real-Time. In *Proc. Computer Animation 99*, pages 138–146. IEEE Computer Society Press, 1999.

[MS89] B. Mishra and N. Silver. Some discussion of static gripping and its stability. *IEEE Transactions on Systems, Man and Cybernetics*, 19:783–796, 1989.

[Nap56] J. Napier. The Prehensile Movements of the Human Hand. *The Journal of Bone and Joint Surgery*, 38b(4):902–913, 1956.

[RBH+95] S. Rezzonico, R. Boulic, Z. Huang, N. Magnenat-Thalmann, and D. Thalmann. Consistent Grasping Interactions with Virtual Actors Based on the Multi-sensor Hand Model. In *Proc. 2nd Eurographics workshop on Virtual Environments*, 1995.

[Sch19] G. Schlesinger. Der Mechanische Aufbau der Künstlichen Glieder. In *Ersatzglieder und Arbeitshilfen für Kriegsbeschädigte und Unfallverletzte*, pages 321–699. Springer-Verlag: Berlin, Germany, 1919.

[ST94] R.M. Sanso and D. Thalmann. A Hand Control and Automatic Grasping System for Synthetic Actors. *Computer Graphics Forum*, 13(3):167–177, 1994.

[ZR03] J. Zhang and B. Rössler. Self-Valuing Learning and Generalization of Visually Guided Grasping. *IROS-2003 Workshop on Robot Programming by Demonstration*, 2003.