Modeling Human Motor Skills to Enhance Robots' Physical Interaction

Giuseppe Averta[1,2], Visar Arapi[1], Antonio Bicchi[1,2], Cosimo della Santina[3,4], and Matteo Bianchi[1]

[1] Research Center E. Piaggio and Department of Information Engineering, University of Pisa, Pisa, Italy
g.averta3@gmail.com,
[2] SoftBots lab, Istituto Italiano di Tecnologia, Genova, Italy
[3] Delft University of Technology (TU Delft), Delft, Netherlands
[4] German Aerospace Center (DLR), Oberpfaffenhofen, Germany

Abstract. The need for users' safety and technology acceptability has incredibly increased with the deployment of co-bots physically interacting with humans in industrial settings, and for people assistance. A well-studied approach to meet these requirements is to ensure human-like robot motions and interactions. In this manuscript, we present a research approach that moves from the understanding of human movements and derives usefull guidelines for the planning of arm movements and the learning of skills for physical interaction of robots with the surrounding environment.

Keywords: human motor control, Human-like Robotic movements, Machine Learning, Learning from Humans

1 Deriving a basis of Human movements

There are many examples in literature that have highlighted the importance of human-likeness (HL) to ensure a safe and effective Human-Robot Interaction (HRI) and Enviroinment-Robot Interaction [8]. This aspect has gained increasing attention, since it could open interesting perspectives for the control of artificial systems that closely interact with humans, as is the case of assistive, companion and rehabilitative robots. For the latter category, for example, human-inspired movement profiles could be used as reference trajectories for rehabilitation exoskeletons (see [10] for review), as an alternative to, and/or in association with, classic rehabilitation procedures [9]. Similarly, human likeliness of movements is of paramount importance for robots that interact with the surrounding environment in an unstructured scenario shared with humans.

Indeed, in these cases the motion of a robot can be more easily *predicted*, and hence accepted, by the user, if its movements are designed taking inspiration from actual human movements [14], leading to a general enhancement in terms of system usability and effectiveness. However, the design of control laws that effectively ensure human-like behavior in robotic systems is not straightforward, representing an important topic within the general framework of robot motion planning.

Fig. 1. Example of the 30 different tasks included in this study. Three blocks are considered, the first including gestures (intransitive tasks), the second actions which involve the contact with an object (transitive tasks), while the third actions that requires an object to interat with another object (tool-mediated tasks).

The solution we implemented to solve this problem exploits functional analysis to derive a basis of Eigenfunctions of human movements, which encode the characteristics of typical physiological motions.

To this end, we recorded the motion of 33 healthy subjects performing a list of 30 different actions of daily living (see Fig. 2).

Then, functional Principal Components Analysis (fPCA) was used to identify a basis of principal functions (or eigenfunctions), characterized by the fact that they are ordered in terms of importance. More specifically, let us assume, without any loss of generality, a 7 DoF kinematic model to represent upper limb joint trajectories $q(t) : \mathbb{R} \rightarrow \mathbb{R}^7$ where $t \in [0,1]$ is the normalized time. In these terms, generic upper limb motion $q(t)$ can be decomposed in terms of the weighted sum of base elements $S_i(t)$, or functional Principal Components (fPCs):

$$q(t) \simeq \bar{q}(t) + S_0(t) + \sum_{i=1}^{s_{\max}} \alpha_i \circ S_i(t) \,, \tag{1}$$

where $\alpha_i \in \mathbb{R}^{\mathrm{n}}$ is a vector of weights, $S_i(t) \in \mathbb{R}^{\mathrm{n}}$ - in this case n equals to 7 - is the i^{th} basis element or fPC and $s_{\max}$ is the number of basis elements. The operator $\circ$ is the element-wise product (Hadamard product), $\bar{q} \in \mathbb{R}^7$ is the average posture of q while $S_0 : \mathbb{R} \rightarrow \mathbb{R}^7$ is the average trajectory, also called *zero-order* fPC. The output of fPCA, which is calculated independently for each joint, is a basis of functions $\{S_1, \ldots, S_{s_{\max}}\}$ that maximizes the explained variance of the movements in the collected dataset. Given a dataset with N elements collecting the trajectories recorded in a given joint j, the first fPC $S_{j,1}(t)$ is the function that solves the following problem

$$\begin{aligned} \max_{S_{j,1}} \quad & \sum_{j=1}^{N} \left(\int S_{j,1}(t) q_j(t) dt \right)^2 \\ \text{subject to} \quad & ||S_{j,1}(t)||_2^2 = \int_0^1 S_{j,1}^2(t) dt = 1 \,. \end{aligned} \tag{2}$$

Subsequent fPCs $S_{j,i}(t)$ are the functions that solve the following:

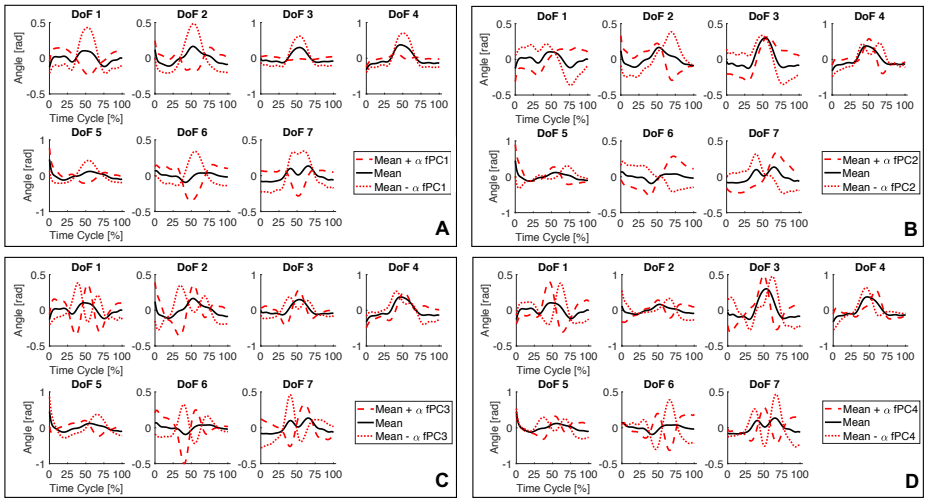

Fig. 2. Plots of the first four orthogonal functional Principal Components extracted 33 healthy subjects while performing a list of 30 different daily living activities. Image adapted from [3].

$$\begin{aligned} \max_{S_{j,i}} \quad & \sum_{j=1}^{N} \left(\int S_{j,i}(t) q_j(t) dt \right)^2 \\ \text{subject to} \quad & ||S_{j,i}(t)||_2^2 = 1 \\ & \int_0^1 S_{j,i}(t) S_p(t) dt = 0 \,, \; \forall p \in \{1, \ldots, i-1\} \,. \end{aligned} \tag{3}$$

A detailed implementation of this method - which bypasses the solution of the minimization problem - is presented in [2]. The core idea is that the output of this process is an ordered list of functions that are organised following the importance that each function has in reconstructing the whole dataset. (See Fig. **??**

Note that this formalization of human trajectories is very compact and full of information, and can enable several practical implementations. For example, one can observe that the higher in the number of functional PCs required to reconstruct one specific movement, the more complex (or jerky) the motion is. This can have a direct impact for the evaluation of motion impairment, for example as a consequence of a stroke event. Indeed, pathological movements are typically characterized by jerky movements, and an assessment of the level of impairment can rely on the fPCA characterization, as we proposed in [15].

Moreover, the hierarchy in the definition of subsequent fPCs is a key characteristics, since it can be exploited to design incremental algorithms [1] of motion planning, as presented in the following of this manuscript.

2 Planning Robots' Movements with fPCs

As previously discussed, typical approaches used in literature to achieve human likeness [12] in robotic motions rely on the strong assumption that human movements are generated by optimizing a known cost function $J_{\text{hl}}(q) : \text{C}_7^1[0,1) \rightarrow \mathbb{R}^+$, where $\text{C}_7^1[0,1)$ is the space of smooth functions going from $[0,1)$ to the joint space $\mathbb{R}^7$, and 1 stands for the final normalized time. The function J_{hl} is used to produce artificial natural motions by solving the problem

$$\min_{q \in \text{C}_7^1[0,1)} \quad J_{\text{hl}}(q) \; . \tag{4}$$

How to choose J_{hl} is not obvious, and it is indeed a very debated topic in literature. However only achieving human likeness is meaningless without specifying also a task to be accomplished. For this reason also a model of the task should be added to (4). The latter point can be formulated in terms of the minimization of an additional cost function $J_{\text{task}} : \text{C}_7^1 \rightarrow \mathbb{R}^+$. As soon as the need for minimizing J_{task} is introduced, (4) becomes a multi-objective optimization, which is of very difficult formulation and solution, except for very simple cases [12].

The solution we proposed is able to by-pass this issue. Indeed, instead of using data to guess a reasonable $J_{\text{hl}}(\cdot)$, and then explicitly optimize it, our solution directly embeds human likeness in the choice of the functional subspace where the optimization occurs. More specifically, the problem move from the infinite dimensional functional space $\text{C}_7^1[0,1)$, to its finite dimensional subspace containing all the functions so constructed:

$$q(t) = \bar{q} + S_0(t) + \sum_{i=1}^{M} \alpha_i \circ S_i(t) \tag{5}$$

with $\bar{q}, S_i, \alpha_i$ defined as in the previous section. In this way the principal components can be used to generate motions happening within any time horizon $[0, t'_{\text{fin}})$.

$M \leq s_{\max}$ is the number of functional Principal Components considered in the optimization (with $s_{\max}$ as in previous Section). According to the preliminary results presented in [2] and further extended in [3], it is plausible to expect that a low number of functional Principal Components should be sufficient to implement most of the human-like motions at the joint level. Therefore the multi-object and unconstrained optimization can be formulated as the following constrained optimization problem:

$$\begin{aligned} \min_{\bar{q}, \alpha_1, \ldots, \alpha_M} \quad & J_{\text{task}}(q) \\ \text{subject to} \quad & q(t) = \bar{q} + S_0(t) + \sum_{i=1}^{M} \alpha_i \circ S_i(t) \; . \end{aligned} \tag{6}$$

In this manner, the search space is narrowed, with the twofold purpose of ensuring human likeness, and strongly simplifying the control problem (indeed, the search space is now of dimension $M+1$).

Point-to-Point Free Motions Point-to-point motion can be generated by solving the following optimization problem, instance of the more general formulation (6)

$$\begin{aligned} \min_{\bar{q},\alpha_1,\ldots,\alpha_M} \quad & ||q(0)-q_0||_2^2 + ||q(1)-q_{\text{fin}}||_2^2 \\ \text{subject to} \quad & q(t) = \bar{q} + S_0(t) + \sum_{i=1}^{M} \alpha_i \circ S_i(t) \ , \end{aligned} \tag{7}$$

where $q(0)$ and $q(1)$ are the initial and final poses of the calculated trajectory, while q_0 and q_{fin} are the desired initial and final poses respectively. In this simple case, a single functional Principal Component (i.e. $M = 1$) is already sufficient to solve (7) with zero error, and the solution can be written in closed form (see [3]).

Obstacle avoidance Let us consider the case in which we also need to avoid one or more obstacles, while performing the point-to-point motion. The problem can be generalized as:

$$\begin{aligned} \min_{\bar{q},\alpha_1,\ldots,\alpha_M} \quad & \left|\left|\begin{bmatrix} q(0)-q_0 \\ q(1)-q_{\text{fin}} \end{bmatrix}\right|\right|_2^2 + \rho P(q,P_O) \\ \text{subject to} \quad & q(t) = \bar{q} + S_0(t) + \sum_{i=1}^{M} \alpha_i \circ S_i(t) \ . \end{aligned} \tag{8}$$

Two terms can be distinguished in this cost function. The first contribution guarantees that the desired initial and final poses are achieved, as for the free motion case (7). The second term takes into account the distance w.r.t. obstacles. For the sake of conciseness, and without any loss of generality, we considered here N_{O} spherical obstacles. Given $P_{\text{O}} = \{P_{\text{O}_1}, \ldots, P_{\text{O}_{N_{\text{O}}}}\}$ the set containing the Cartesian coordinates of all the centers of these obstacles, $P(q,P_O)$ is a potential-based function that sums up, for each obstacle, a term inversely proportional to the minimum distance between the obstacle and the closest joint trajectory, i.e.

$$P(q,P_O) = \sum_{i=1}^{N_O} \frac{1}{m_i(q([0,1]),P_{\text{O}_i})})^2 \tag{9}$$

where m_i is the distance between the arm and the $i-th$ obstacle, defined as $m_i(q([0,1]),P_{\text{O}_i}) = \min_k\{d(h_k(q([0,1])),P_{\text{O}_i})\}$.

The distance between the $k-th$ point of contact with forward kinematics h_k, and the $i-th$ sphere is

$$d(h_k(q([0,1])),P_{\text{O}_i}) = \max\left\{\min_{x\in h_k(q([0,1]))} ||P_{\text{O}_i}-x||_2, R_{\text{O}_i}\right\} , \tag{10}$$

with R_{O_i} radius of the sphere.

Incremental optimization procedure The problem of motion generation with obstacle avoidance does not have a closed-form solution, hence the optimal trajectory is calculated

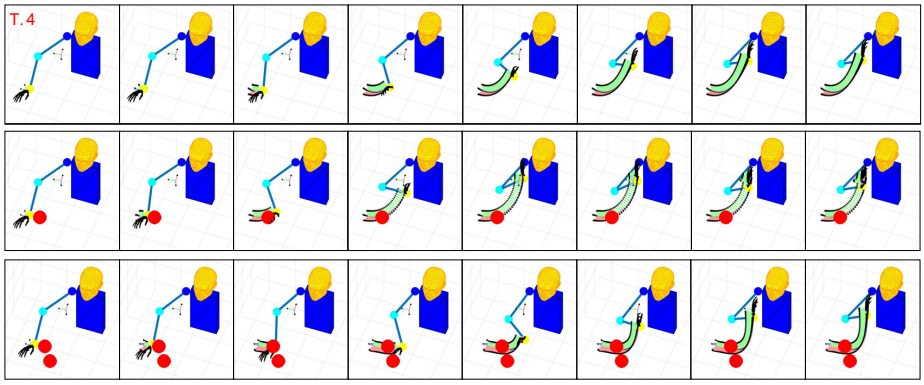

Fig. 3. In this example our approach is used to generate a ”drinking” task, with and without obstacles along the trajectory.

via numerical optimization. One solution to do this is to exploit the hierarchy of fPCs basis elements, according to a descending amount of the associated explained variance, and implemented an incremental procedure (see [3] for the implementation of the Algorithm). The proposed approach calculates, given a fixed number of fPCs enrolled, the optimal trajectory that minimizes the error in starting and final position while maximizing the distance from the obstacles. If the corresponding solution is sufficiently far from the obstacles, this choice already defines the globally optimal solution. If the obstacles are not very close to the aforementioned trajectory, then solving (8) with $M = 1$ would fine tune the initial guess, achieving good results. In case of obstacles very close to or even intercepting the free-motion trajectory, at least one more fPC should be enrolled to suitably solve the problem. The more are the basis elements enrolled, the more complex are the final trajectories that can be implemented (see e.g. **??** for the generation of a ”drinking” task).

3 Learning from Humans How to Grasp: Enhancing the Reaching Strategy

Deriving useful information from Humans can be pushed even further through the usage of machine learning techniques. However, it is important to recall that learning based techniques can only achieve solutions that are *close enough* to the desired ones, rather than exact. This uncertainty can be naturally compensated by the ability of soft hands to locally adapt to unknown environments. Following this approach, part of our effort has been devoted to the development of a human inspired multi-modal, multi-layer architecture that combines feedforward components, predicted by a Deep Neural Network, with reactive sensor-triggered actions (more details in [5]).

Humans are able to accomplish very complex grasps by employing a vast range of different strategies [7]. This comes with the challenging problem of finding the right strategy to use for a given scenario. It is commonly suggested that the animal brain

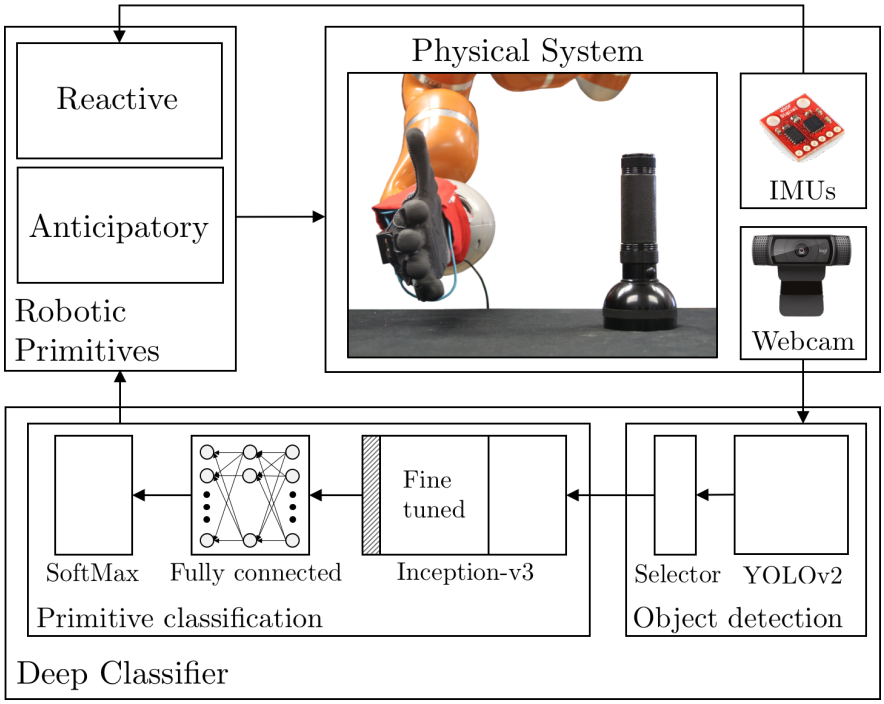

Fig. 4. High level organization of the proposed architecture, which combines anticipatory actions and reactive behavior. A deep classifier looks at the scene and predicts the strategy that a human operator would use to grasp the object. This output is employed to select the corresponding robotic primitive. These primitives define the posture of the hand over time, to produce a natural, human-like motion. The IMUs placed on the fingers of the hand detect the contact with the items and triggers a suitable reactive grasp behavior.

addresses this challenge by first constructing representations of the world, which are used to make a decision, and then by computing and executing an action plan [11]. Rather than learning a monolithic end-to-end map, we built the proposed architecture as combination of interpretable basic elements organized as in Fig. 4. The *intelligence* is here distributed on three levels of abstractions; i) high level: a classifier which plans the correct action among all the available ones, ii) medium level: a set of human-inspired low level strategies implementing both the approaching phase and the sensor-triggered reaction, iii) low level: a soft hand whose *embodied intelligence* mechanically manages local uncertainties. All the three levels are human-inspired.

The classifier was realized through a deep neural network, trained to predict the object-directed grasp action chosen among nine human-labeled strategies, using as input only a first-person RBG image extracted from a video. These actions were implemented on the robotic side to reproduce the motions observed in the videos. A reactive component was then introduced, following the philosophy of [4]. This component take as input the accelerations coming from six IMUs placed on the soft hand to generate the desired

Fig. 5. Photosequence of a grasp produced by the proposed architecture during validation. The hand starts from the initial configuration of the primitive in panel (a). The contact happens in panel (b), triggering the reactive routine. In panel (f) the object is firmly lifted.

evolution of the hand pose. The lower level of intelligence consists of the soft hand itself, which can take care of local uncertainties relying on its intrinsic compliance. Any robotic hand being soft and anthropomorphic both in its motions and in its kinematics can serve to the scope (as for example the Pisa/IIT SoftHand).

3.1 Deep Classifier

The aim of this deep neural network is to associate to an object detected from the scene the correct primitive (i.e. hand pose evolution) humans would perform to grasp it. The deep learning model consists of two stages, depicted in Fig. 4: one for detecting the object, and the second one to perform the actual association with the required motion.

Dataset creation and human primitive labeling The network was trained on 6336 first person RGB videos (single-object, table-top scenario), from 11 right-handed subjects grasping the 36 objects. The list of objects was chosen to span a wide range of possible grasps, taking inspiration from [6]. During the experiments, subjects were comfortably seated in front of a table, where the object was placed. They were asked to grasp the object starting from a rest position (hand on the table, palm down). Each task was repeated 4 times from 4 points of view (the four central points of the table edges). To extract and label the strategies, videos were visually inspected to identify ten main primitives (power, pinch, sliding, lateral and flip grasps in different relative orientations). The choice of these primitives was done taking inspiration from literature [6,7], and to provide a representative yet concise description of human behavior, without any claim of exhaustiveness. The first frame of each video showing only the object in the environment was extracted, and elaborated through the object detection part of the network (see next subsection). The cropped image was then labeled with the strategy used by the subject in the remaining part of the video. This is the dataset that we used to train the network.

Object detection and Primitive classification Object detection is implemented using the state of the art detector YOLOv2 [13]. Given the RGB input image, YOLOv2 produces as output a set of labeled bounding boxes containing all the objects in the scene. Assuming that the target is localized close to the center of the image, we select the bounding box closest to the scene center. Then, a modification of Inception-v3 [16], trained on the ImageNet data set, was used to classify objects from images and extract high level semantic descriptions that can be applied to objects with similar characteristics.

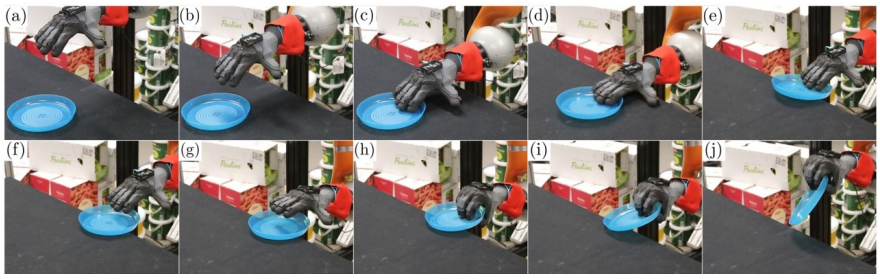

Fig. 6. Photosequence of a grasp produced by the proposed architecture during validation: slide grasp of a flat plate. Panels (a-c) depicts the approaching phase. In panels (d-e) the environment is exploited to guide the object to the table edge. In panels (f-g) the hand changes its relative position w.r.t. the object so to favor the grasp, which is established in panels (h-i). In panel (j) the item is firmly lifted.

Technical details on training and validation are here omitted, the interested reader is invited to refer to [5].

3.2 Robotic Grasping Primitives

The output of the network introduced in the previous section is a direction of approach, described in terms of an high level description of the human preference for the specific object shape and orientation. For each primitive, a Human-Like approaching trajectory needs to be planned (following, for example, the approach presented in section II).

As a trade-off between performance and complexity, the approaching phase is associated with an additional reactive behavior. The role of the latter is to introduce a feedback control leveraging on measures recorded through IMUs at the fingertip level, with the ultimate goal of locally precisely arrange the relative configuration between hand and object (see [4]). The transition between the first and the second phase is triggered by a contact event, detected as an abrupt acceleration of the fingertips (as read by IMUs). In [4], a subject was asked to reach and grasp a tennis ball while maneuvering a Pisa/IIT SoftHand. The grasp was repeated 13 times, from different approaching directions. The user was instructed to move the hand until the contact with the object, and then to react by adapting the hand/wrist pose w.r.t. the object. Poses of the hand were recorded through a PhaseSpace motion tracking system. We subtract from the hand evolution recorded between the contact and the grasp (T represents the time between them) the posture of the hand during the contact. The resulting function $\Delta_i : [0,T] \rightarrow \mathbb{R}^7$ describes the rearrangement performed by the subject to grasp the object. Acceleration signals $\alpha_1 \dots \alpha_{13} : [0,T] \rightarrow \mathbb{R}^5$ were measured too through the IMUs. To transform these recordings into a local adaptation strategy, we considered the acceleration patterns as a characteristic feature of the interaction with the object. When the Pisa/IIT SoftHand touches the object, IMUs read an acceleration profile $a : [0,T] \rightarrow \mathbb{R}^5$. The triggered sub-strategy is defined by the local rearrangement Δ_j, with

$$j = \arg\max_i \int_0^T a^{\mathrm{T}}(\tau)\alpha_i(\tau)\mathrm{d}\tau \,. \tag{11}$$

When this motion is completely executed, the hand starts closing until the object is grasped.

We extensively tested the proposed architecture with 20 objects, different than the ones used for the training of the network. Results demonstrated that this approach is very reliable, achieving a success rate of 81.1 % over 111 grasps tested, thus demonstrating that taking inspiration from humans can provide very interesting solutions for classic and novel problems toward a new generation of anthropomorphic robots.

References

1. Averta, G., Angelini, F., Bonilla, M., Bianchi, M., Bicchi, A.: Incrementality and hierarchies in the enrollment of multiple synergies for grasp planning. IEEE Robotics and Automation Letters **3**(3), 2686–2693 (2018)
2. Averta, G., Della Santina, C., Battaglia, E., Felici, F., Bianchi, M., Bicchi, A.: Unvealing the principal modes of human upper limb movements through functional analysis. Frontiers in Robotics and AI **4**, 37 (2017)
3. Averta, G., Della Santina, C., Valenza, G., Bianchi, M., Bicchi, A.: Exploiting upper-limb functional synergies for human-like motion generation of anthropomorphic robots. Journal of NeuroEngineering and Rehabilitation
4. Bianchi, M., Averta, G., Battaglia, E., Rosales, C., Bonilla, M., Tondo, A., Poggiani, M., Santaera, G., Ciotti, S., Catalano, M.G., et al.: Touch-based grasp primitives for soft hands: Applications to human-to-robot handover tasks and beyond. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 7794–7801. IEEE (2018)
5. Della Santina, C., Arapi, V., Averta, G., Damiani, F., Fiore, G., Settimi, A., Catalano, M.G., Bacciu, D., Bicchi, A., Bianchi, M.: Learning from humans how to grasp: a data-driven architecture for autonomous grasping with anthropomorphic soft hands. IEEE Robotics and Automation Letters **4**(2), 1533–1540 (2019)
6. Eppner, C., Deimel, R., Alvarez-Ruiz, J., Maertens, M., Brock, O.: Exploitation of environmental constraints in human and robotic grasping. The International Journal of Robotics Research **34**(7), 1021–1038 (2015)
7. Feix, T., Romero, J., Schmiedmayer, H.B., Dollar, A.M., Kragic, D.: The grasp taxonomy of human grasp types. IEEE Transactions on Human-Machine Systems **46**(1), 66–77 (2016)
8. Fink, J.: Anthropomorphism and human likeness in the design of robots and human-robot interaction. In: International Conference on Social Robotics, pp. 199–208. Springer (2012)
9. Krebs, H.I., Hogan, N.: Robotic therapy: the tipping point. American journal of physical medicine & rehabilitation/Association of Academic Physiatrists **91**(11 0 3), S290 (2012)
10. Maciejasz, P., Eschweiler, J., Gerlach-Hahn, K., Jansen-Troy, A., Leonhardt, S.: A survey on robotic devices for upper limb rehabilitation. Journal of neuroengineering and rehabilitation **11**(1), 3 (2014)
11. Miller, G.A., Galanter, E., Pribram, K.H.: Plans and the structure of behavior. Adams Bannister Cox (1986)
12. Piazzi, A., Visioli, A.: Global minimum-jerk trajectory planning of robot manipulators. IEEE transactions on industrial electronics **47**(1), 140–149 (2000)
13. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. arXiv preprint (2017)
14. Riek, L.D., Rabinowitch, T.C., Chakrabarti, B., Robinson, P.: How anthropomorphism affects empathy toward robots. In: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction, pp. 245–246. ACM (2009)

15. Schwarz, A., Averta, G., Veerbeek, J.M., Luft, A.R., Held, J.P., Valenza, G., Biechi, A., Bianchi, M.: A functional analysis-based approach to quantify upper limb impairment level in chronic stroke patients: A pilot study. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4198–4204. IEEE (2019)
16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826 (2016)