POLITECNICO DI TORINO
Repository ISTITUZIONALE

Learning from humans how to grasp: A data-driven architecture for autonomous grasping with anthropomorphic soft hands

Original
Learning from humans how to grasp: A data-driven architecture for autonomous grasping with anthropomorphic soft hands / Santina, C. D.; Arapi, V.; Averta, G.; Damiani, F.; Fiore, G.; Settimi, A.; Catalano, M. G.; Bacciu, D.; Bicchi, A.; Bianchi, M.. - In: IEEE ROBOTICS AND AUTOMATION LETTERS. - ISSN 2377-3766. - 4:2(2019), pp. 1533-1540. [10.1109/LRA.2019.2896485]

Availability:
This version is available at: 11583/2970293 since: 2022-09-02T09:19:09Z

Publisher:
Institute of Electrical and Electronics Engineers Inc.

Published
DOI:10.1109/LRA.2019.2896485

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

24 November 2024

Learning from humans how to grasp: a data-driven architecture for autonomous grasping with anthropomorphic soft hands

Cosimo Della Santina[1,3], Visar Arapi[1], Giuseppe Averta[1,2,3], Francesca Damiani[1], Gaia Fiore[1], Alessandro Settimi[1], Manuel G. Catalano[2], Davide Bacciu[4], Antonio Bicchi[1,2,3], and Matteo Bianchi[1,3]

***Abstract*—Soft hands are robotic systems that embed compliant elements in their mechanical design. This enables an effective adaptation with the items and the environment, and, ultimately, an increase of their grasping performance. These hands come with clear advantages in terms of ease-to-use and robustness if compared with classic rigid hands, when operated by a human. However, their potential for autonomous grasping is still largely unexplored, due to the lack of suitable control strategies. To address this issue, in this work we propose an approach to enable soft hands to autonomously grasp objects, starting from the observations of human strategies. A classifier realized through a deep neural network takes as input the visual information on the object to be grasped, and predicts which action a human would perform to achieve the goal. This information is hence used to select one among a set of human-inspired primitives, which define the evolution of soft hand posture as a combination of anticipatory action and touch-based reactive grasp. The architecture is completed by the hardware component, which consists of a RGB camera to look at the scene, a 7-DoF manipulator, and a soft hand. The latter is equipped with IMUs at the fingernails for detecting contact with the object. We extensively tested the proposed architecture with 20 objects, achieving a success rate of 81.1% over 111 grasps.**

***Index Terms*—Natural Machine Motion; Deep Learning in Robotics and Automation; Modeling, Control, and Learning for Soft Robots; Grasping.**

I. Introduction

THE execution of reliable and stable grasping with artificial hands is a main challenge in the robotics field, due to its practical relevance and theoretical complexity. The *classic* approach used to grasp with rigid robotic hands generally favored object-centric analytical solutions. More specifically, a set of available contact points is hypothesized, while their position and contact forces are evaluated from the object knowledge [1]. Although very elegant and theoretically sound, this approach has not yet produced the desired outcomes in practice. To address these limitations, in soft artificial hands part of the control intelligence has been directly embedded in their mechanism, through the purposeful introduction of elastic elements and under-actuation patterns [2], [3]. Thanks to their intrinsic compliance, soft hands can mold around the external items and exploit their environment, thus multiplying their grasping capabilities.

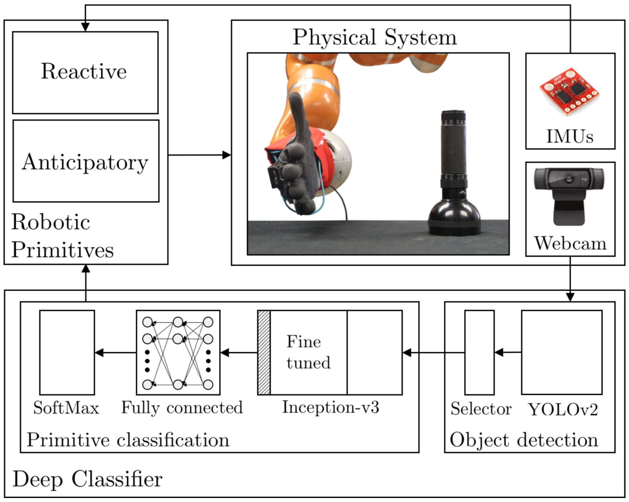

Fig. 1. High level organization of the proposed architecture, which combines anticipatory actions and reactive behavior. A deep classifier looks at the scene and predicts the strategy that a human operator would use to grasp the object. This output is employed to select the corresponding robotic primitive. 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.

Several papers have shown that soft end-effectors can achieve high-level grasping performance when operated by humans (see e.g. [4], [5]). However, such level of dexterity is still unmatched in autonomous grasp execution. Indeed, *classic* approaches cannot be applied to this kind of hands, which - by their own nature - do not allow fingertips placement with the required precision and relative independence. On the contrary, data driven approaches could be the key to learn from humans how to manage soft hands, towards higher levels of autonomous grasping capabilities.

Recently machine learning has become very popular for grasping generation, with positive results [6], [7], [8]. However, so far only few works in literature have applied learning methods in the control of soft hands. In [9] learning by demonstration is combined with reinforcement learning to

Manuscript received: September, 10, 2018; Revised November, 10, 2018; Accepted January, 11, 2019. This paper was recommended for publication by Editor Nikolaos G. Tsagarakis upon evaluation of Reviewers' comments. This work was supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement 688857 (SoftPro) and 645599 (Soma). The work was partially supported by the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab project (Departments of Excellence).

[1] C. Della Santina, V. Arapi, G. Averta, F. Damiani, G. Fiore, A. Settimi, A. Bicchi, and M. Bianchi are with the Centro di Ricerca "Enrico Piaggio", Università di Pisa, Largo Lucio Lazzarino 1, 56126 Pisa, Italy {cosimodellasantina, visararapi22, ale.settimi}@gmail.com, matteo.bianchi@centropiaggio.unipi.it

[2] G. Averta, M. G. Catalano, and A. Bicchi are with the Soft Robotics for Human Cooperation and Rehabilitation, Fondazione Istituto Italiano di Tecnologia, via Morego, 30, 16163 Genova, Italy {Manuel.Catalano, antonio.bicchi}@iit.it

[3] C. Della Santina, G. Averta, A. Bicchi, and M. Bianchi are with the Dipartimento di Ingegneria dell'Informazione, Università di Pisa, via G. Caruso 16, 56126 Pisa, Italy giuseppe.averta@ing.unipi.it

[4] D. Bacciu is with the Dipartimento di Informatica, Universita' di Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy bacciu@di.unipi.it

The authors want to thank Gaspare Santaera for his really valuable support in setting up the experimental setup.

Digital Object Identifier (DOI): see top of this page.

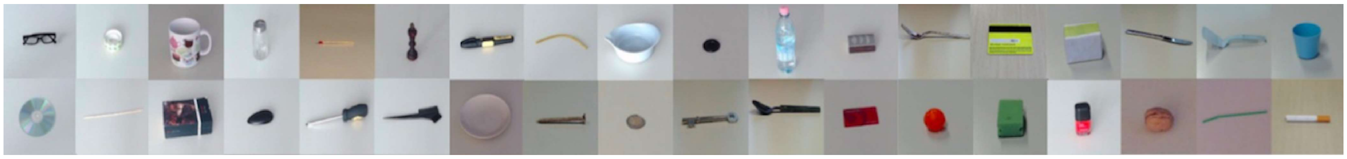

Fig. 2. The set of 36 objects used during human videos acquisition. Photos are not in scale. From left to right and from top to bottom we have: a pair of glasses, a tape, a coffee mug, a salt shaker, a match, a piece of chess, a marker, a rubber, a colander, a button, a bottle, a match box, a fork, a credit card, a box, a knife, a spatula, a glass, a CD, a shashlik, a book, a shell, a screwdriver, a comb, a plate, a screw, a coin, a key, a spoon, a game card, a ball, a sponge, a nail polish container, a walnut, a straw and a cigarette.

transfer grasping capabilities of known objects from a human operator to the robotic system. In [10] autoencoders and generalized regression neural networks are used to learn from examples generated by a human operator, how to manipulate previously unseen thin objects with a soft gripper. While very promising, both works show no generalization capabilities in terms of objects to be grasped. Two works recently attempted to go beyond this limitation. In [11] a library of reactive strategies was collected from a subject operating a soft hand, and successfully translated for robotic grasping of new items, in a human-robot handover scenario. In [12] a 3D convolutional neural network is trained with tens of thousands of labeled images. The network output provides the control input for the hand approaching direction.

These works represent an important step forward, not only in terms of performance, but more fundamentally for recognizing the complementary yet intertwined nature of machine learning methods and soft hands. Indeed, 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.

The aim of this work is to build upon this principle, and fully exploit hardware adaptability to grasp a vast range of very different objects. To this end we propose a human inspired multi-modal, multi-layer architecture (Fig. 1) that combines feedforward components with reactive sensor-triggered actions. We extensively tested the proposed architecture on a set of 20 objects previously unseen by the network. The orientation of the objects placed on a table also varied. We performed three repetitions for each condition, for a total of 111 tests, reaching an overall percentage of grasp success of 81.1%.

II. Proposed Approach

Humans are able to accomplish very complex grasps by employing a vast range of different strategies [13]. 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 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 [14].

Rather than learning a monolithic end-to-end map, we built the proposed architecture as combination of interpretable basic elements organized as in Fig. 1. 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.

We realized the classifier through a deep neural network. This was trained to predict the object-directed grasp action chosen among nine human-labeled strategies, using as input only a first-person RBG image of the scene. 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 [11]. This component take as input the accelerations coming from six IMUs placed on the soft hand to generate the desired 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. Without loss of generality, we use here the Pisa/IIT SoftHand [15]. We report in the next sections the detailed description of these components.

To conclude, the main contributions of this work are:

- A deep neural network, which is able to predict with high accuracy the strategy that a human would adopt to grasp an object, using a first-person RGB image of the scene. This result is then used to plan a suitable primitive execution on the robotic side;
- A set of reactive primitives that reproduce human grasps, substantially extending [11];
- The definition and extensive experimental validation of an autonomous grasping system, which combines these two blocks with the adaptability of a soft hand.

III. 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. 1: one for detecting the object, and the second one to perform the actual association with the required motion. Before going through the details of these two components, we briefly describe the phases of primitive extraction and labeling from human videos.

a) Dataset creation and human primitive labeling: We collected 6336 first person RGB videos (single-object, table-top scenario), from 11 right-handed subjects grasping the 36 objects in Fig. 2. The list of objects was chosen to span a wide range of possible grasps, taking inspiration from [16]. 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

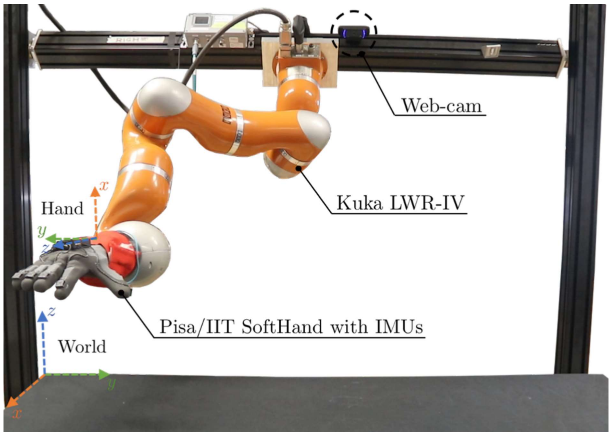

Fig. 3. Experimental setup. The Pisa/IIT SoftHand is mounted as end effector of a Kuka LWR. A web-cam records the scene from a first-person point of view. The hand is equipped with inertial measurement units used to detect contacts. Reference frames are reported.

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, we first visually inspected the video and identified ten main primitives

- Top: the object is approached from the top with the palm down parallel to the table. Object center is approximatively at the level of the middle phalanx. When contact is established, subjects close simultaneously all their fingers, achieving a firm power-like grasp.
- Top left: same for the top grasp, but with the palm rotated clockwise of at least $\pi/9$ radians.
- Top right: as for the top grasp, but with the palm rotated counter-clockwise of at least $\pi/9$ radians.
- Bottom: the object is approached from its right side. The palm is roughly perpendicular to the table, but slightly tilted so that the fingertips are more close to the object than the wrist. When the contact is reached, the hand closes with the thumb opposing the four long fingers. This primitive is used to grasp large and concave objects, e.g. a salad bowl.
- Pinch: same as for the top, but the primitive concludes with a pinch grasp.
- Pinch left: same as for the top left, but the primitive concludes with a pinch grasp.
- Pinch right: same as for the top right, but the primitive concludes with a pinch grasp.
- Slide: the hand is placed on the object from above as to push it toward the surface. Maintaining this hand posture, the object is moved towards the edge of the table until it partially protrudes. A grasp is then achieved by moving the thumb below the object, and opposing it to the long fingers. This strategy is used to grasp objects whose thickness is smaller compared to the other dimensions, such as a book or a compact disk.
- Flip: the thumb is used together with the environment on one side, and the index and/or the middle on the opposite one, to pivot the object. The item rotates of about $\pi/2$ and then it is grasped with a pinch. This strategy is used to grasp small and thin objects, as a coin.
- Lateral: the same as for the top grasp, but the palm is perpendicular to the object during the approaching phase. This strategy is used to grasp tall objects, like a bottle.

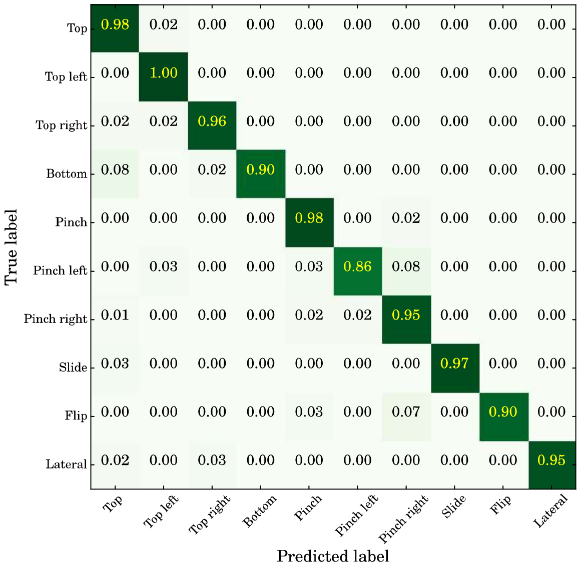

Fig. 4. Confusion matrix summarizing the performance of the proposed deep classifier on the test set. Each entry shows the rate at which the primitives identified by the row labels are classified as the ones identified by the column labels. Rate is also color coded, from low rate coded with white to high rate coded with dark green.

The choice of these primitives was done taking inspiration from literature [16], [13], and to provide a representative yet concise description of human behavior, without any claim of exhaustiveness. Note that the selection of the action primitive is not only object-dependent but also configuration dependent. This is clear for the left/right modifier. Consider for example a bottle; if placed on its base it triggers a lateral grasp, while when laying down on its side induces a top grasp.

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.

A. *Object detection*

Object detection is implemented using the state of the art detector YOLOv2 [17]. Given the RGB input image, YOLOv2 produces as output a set of labeled bounding boxes containing all the objects in the scene. We first discard all the boxes labeled as person. We assume that the target is localized close to the center of the image. Hence, we select the bounding box closest to the scene center. Once the object of interest is identified, the image is automatically cropped around the bounding box, and resized to 416×416 pixels (size expected by the subsequent layer). The result is fed into the following block to be classified.

B. *Primitive Classification*

a) Architecture: Instead of building from scratch a completely new architecture, we follow a transfer learning approach. The idea is to exploit the existing knowledge learned from one environment to solve a new problem, which is different yet related. In this way, a smaller amount of data is sufficient to train the model, and achieve high accuracy with a short training time. We select as starting point Inception-v3 [18], trained on the ImageNet data set to classify objects from images. We keep the early and middle layers and remove the softmax layer. In this way, we have direct access to the highly refined and informative set of neural features that Inception-v3 uses to perform its classification. It is important to note that the object signature is not one-to-one but it aims at extracting high level semantic descriptions that can be applied to objects with similar characteristics. On the top of the original architecture we add two fully connected layers containing 2048 neurons each (with ReLU activation function). These layers operate an adaptive non-linear combination of the high-level features discovered by the convolutional and pooling layers, further refining the information. In this way, the geometric features are implicitly linked each other to serve as the base for the classification. The output of the last fully-connected layer is thus fed into the softmax, which produces a probability distribution over the considered set of motion primitives. We chose the one with maximum probability as output of the network.

b) Training and validation: We use the labeled dataset described above to train the network. The parameters of the two fully connected layers at the top of the Inception-v3 architecture are trained from scratch, while the original parameters of the network are fine-tuned. To this end we impose layer-specific learning rates. More specifically, we freeze the weights in the first 172 layers (over the total 249) of the pre-trained network. These layers capture indeed universal features like curves and edges that are also relevant to our problem. We instead use the subsequent 77 layers to capture dataset-specific features. However, we expect the pre-trained weights to be already good if compared to randomly initialized ones. Hence, we avoid to abruptly change them using a relatively small learning rate λ_{ft}. Finally, given that the weights of the two last fully connected layers are trained from scratch, we randomly initialize them and use a higher learning rate λ_{tr} w.r.t. the one we use in previous layers. We further reduce the risk of over-fitting by using dropout; before presenting a training sample to the network, we randomly disconnect neurons from its structure (actually, this is implemented by masking their activation). Each neuron is removed with probability p_{drop}. In this way, a new topology is produced each time the network is trained, introducing variability and reducing the production of pathological co-adaptation of weights. We use Keras library for network design and training. All the procedures were executed trough an NVIDIA Tesla M40 GPU with 12GB of on-board memory.

To verify the generalization and robustness of primitive classification, we use hold out validation. The goal is to estimate the expected level of model predictive accuracy independently from the data used to train the model. We split our data set in: 70% objects for training, 20% objects for validation and 10% for testing. We maintained a balanced number of objects per class among over the three data sets. We trained 30 different network configurations using the *cross entropy* cost function to adjust the weights by calculating the error between the output of the softmax layer and the label vector of the given sample category. Each configuration was obtained by varying the most relevant model learning hyper-parameters, i.e. learning rates $\lambda_{\text{ft}} \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ and $\lambda_{\text{tr}} \in \{10^{-2}, 10^{-3}, 10^{-4}\}$, dropout probability $p_{\text{drop}} \in \{0.4, 0.5, 0.6\}$, number of epochs in $\{10, 20, 30, 40\}$, and the batch size in $\{10, 20, 30, 40\}$. The training time for each network ranged from 1 to 5 hours. We selected the configuration that provided the highest $f1$-score accuracy [19] on the validation data set - which is 97%. The selected hyper-parameters are $\lambda_{\text{ft}} = 10^{-5}$, $\lambda_{\text{tr}} = 10^{-3}$, $p_{\text{drop}} = 0.5$, 30 epochs, and batches size 20.

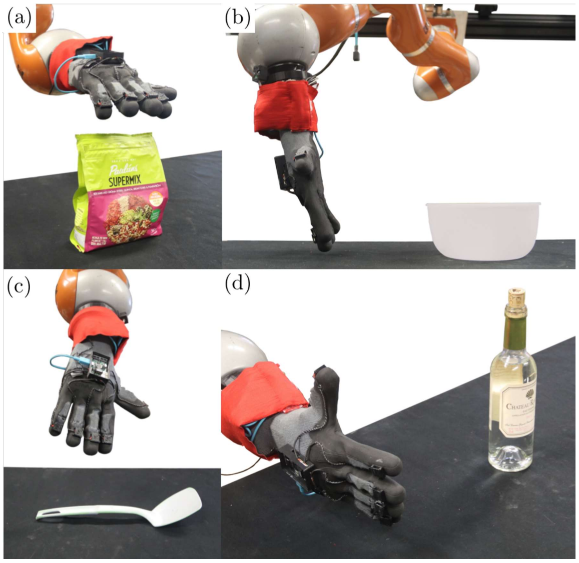

Fig. 5. Four significant relative object-hand postures assumed by the hand during the approaching phase. Starting from these initial configurations, the hand translates until a contact is detected by the IMUs. Directions of translation are perpendicular to the table for top and pinch primitives, and parallel to it for lateral and bottom.

With such parameters, the network is able to classify the primitives in the test set with an accuracy ranging from 86% to 100%, depending on the primitive, and 95% on average. Fig. 4 shows the normalized accuracy of the classifier for all ten classes. Visually inspecting the results reveals two main causes behind the occasional failures of the network. The first one is a limitation in the problem formulation itself, which makes intrinsically not possible to achieve 100% classification accuracy. Indeed, it seldom occurs that the same object in the same configuration is grasped in two different ways by two subjects. This happens for example for the coin, which is often grasped through a flip, while sometimes slide primitive is used instead. The second cause is connected to the fact that using only a single RGB image, the network sometimes misinterprets the object size. This could, for example, lead to predict a top grasp rather than a bottom grasp for a bowl, since this object may be interpreted as a ball-like item. In future work we will consider the use of a stereo camera to prevent this issue.

IV. Robotic Grasping Primitives

In [20], Johansson and Edin affirm that the Central Nervous System “*monitors specific, more-or-less expected, peripheral sensory events and use these to directly apply control signals*

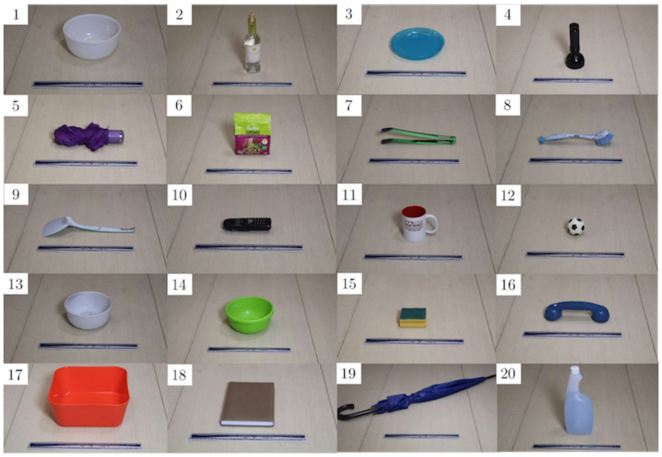

Fig. 6. Set of objects used in the experimental validation. None of them was part of the set used during training. A 30cm ruler is present in all the photos to help in qualitatively understanding object sizes.

that are appropriate for the current task and its phase". These signals are largely precomputed (i.e. anticipatory, or feedforward). Driven by this observation, we decided to implement the robotic grasping strategies relying mostly on anticipatory actions. To do this, we took inspiration from the visual inspection of the videos described in the previous section, and decided to trigger primitive execution by specific events. The first event is generated by the detection of an object and scene classification. This triggers one primitive among all the available ones. We do not consider here flip, which can not be implemented by the soft hand that we use in this work. As a trade-off between performance and complexity, we divide all primitives in two phases: i) approach and ii) reactive grasp. 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).

A. Experimental setup

While the proposed techniques are not specifically tailored on this specific setup, it is convenient to introduce it here to simplify the description of the next subsections (see Sec. II). The robotic architecture is composed of two main components: a KUKA LWR-IV arm, and a Pisa/IIT SoftHand [15] as end effector. This anthropomorphic soft hand has 19 degrees of freedom, and only one degree of actuation. The *intelligence embodied* in the hand mechanics is to be considered as an integral part of the control architecture itself, rather than as a simple effector to act onto the environment. A RGB camera is placed on the top of the manipulator to simulate a first-person point-of-view. The robotic hand is equipped with IMUs for contact detection, triggering reactive strategies for grasping. The principal reference frames used in our control framework are depicted in Fig. 3.

B. Approach phase

During the approach phase, human hand tends to follow straight lines connecting the starting position and the target [21]. We reproduce this behavior through the simple trajectory

$$x(t) = x_0 + d\,t\ , \quad Q(t) = Q_0\ , \tag{1}$$

where $x \in \mathbb{R}^3$ is the hand base frame position in Cartesian coordinates, and $Q \in \mathbb{R}^4$ its orientation as quaternion, both expressed in global coordinates. $x_0 \in \mathbb{R}^3$ and $Q_0 \in \mathbb{R}^4$ are the initial position and orientation, while $d \in \mathbb{R}^3$ is the direction of approach. All these three quantities are defined by the selected primitive, and dictated by the aim of heuristically reproducing as close as possible the human behavior observed in the videos. Fig. 5 shows photos of the hand in $t = 0$ for top, pinch, lateral and bottom grasps. Tab. I summarizes directions of approach and initial orientations for all the primitives.

TABLE I
INITIAL ORIENTATION Q_0 AND NORMALIZED DIRECTION OF APPROACH $\hat{d}$ FOR EACH PRIMITIVE.

Strategy	Q_0^{T}	$\hat{d}^{\mathrm{T}}$
Top	$[0.0\ 0.711\ 0.0\ 0.703]$	$[0\ 0\ -1]$
Top left	$[0.269\ 0.6570\ -0.2721\ 0.6496]$	$[0\ 0\ -1]$
Top right	$[0.269\ -0.657\ -0.272\ -0.649]$	$[0\ 0\ -1]$
Bottom	$[0.145\ -0.696\ 0.701\ 0.030]$	$[0\ 1\ 0]$
Pinch	$[0.084\ 0.816\ 0.17\ 0.458]$	$[0\ 0\ -1]$
Pinch left	$[0.116\ 0.733\ 0.483\ 0.463]$	$[0\ 0\ -1]$
Pinch right	$[0.186\ 0.890\ -0.110\ 0.400]$	$[0\ 0\ -1]$
Slide	$[0.0\ 0.711\ 0.0\ 0.703]$	$[0\ 0\ -1]$
Lateral	$[0\ -1\ 0\ 0]$	$[0\ 1\ 0]$

C. Grasp phase

The grasp phase is when the grasp actually happens, and thus where the primitives differentiate more from each others. When not differently specified, translations and rotations are here expressed in hand coordinates.

a) Top and lateral grasps: The reactive grasp framework leverages on a dataset of 13 prototypical rearrangements of the hand, extracted from human movements. In [11], 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 $\Delta_i : [0, T] \to \mathbb{R}^7$ describes the rearrangement performed by the subject to grasp the object. Acceleration signals $\alpha_1 \ldots \alpha_{13} : [0, T] \to \mathbb{R}^5$ were measured too through the IMUs. To transform these recordings into a grasping 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 $a : [0, T] \to \mathbb{R}^5$. The triggered sub-strategy is defined by the local rearrangement Δ_j, with

$$j = \arg\max_i \int_0^T a^{\mathrm{T}}(\tau)\alpha_i(\tau)\mathrm{d}\tau\ . \tag{2}$$

When this motion is completely executed, the hand starts closing until the object is grasped. This procedure proved its effectiveness in preliminary power grasp experiments on objects approached similarly as specified here by the top primitive [11]. We extend here its use to top left, top right and lateral strategies.

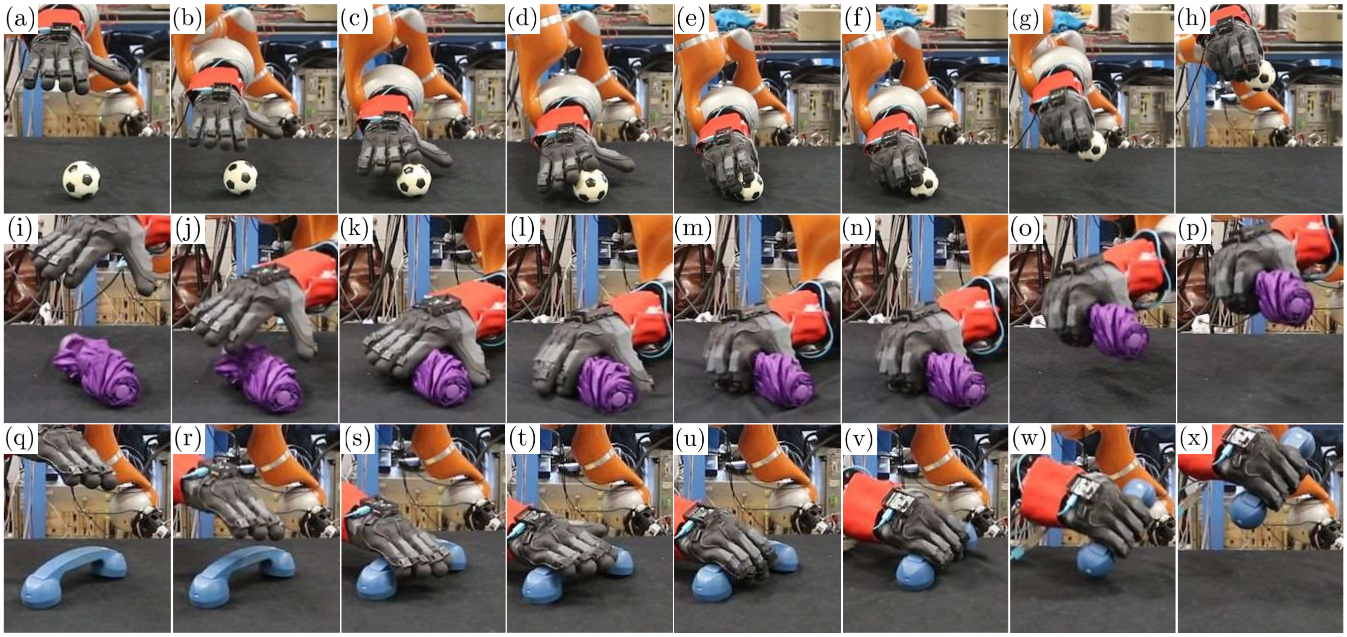

Fig. 7. Photosequences of grasps produced by the proposed architecture during validation: panels (a-h) present a top grasp of object 12, panels (i-p) a top-left grasp of object 5, and panels (q-x) a top-right grasp of object 16. Panels (a-b) depicts the approach phase. In (c) the contact is detected and classified using (2). In panels (c-f) the hand finely changes its relative position w.r.t. the object, as prescribed by the reactive routine, and grasps it. In (g) and (h) the item is firmly lifted.

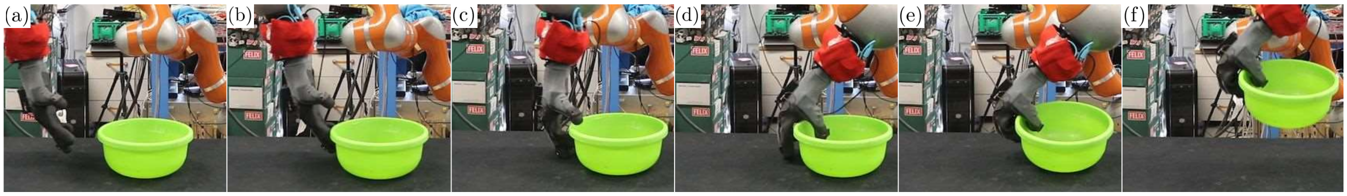

Fig. 8. Photosequence of a grasp produced by the proposed architecture during validation: Bottom grasp of object 14. The hand starts from the initial configuration of the primitive in panel (a). The contact happens in panel (b), triggering the reactive routine. In panel (f) the object is firmly lifted.

b) Bottom: To mimic human behavior described in previous section, when a contact is detected we rotate the hand along x of $\pi/3$ and translate 300mm along y. In this way the palm base moves over, and the thumb can enter into the concave part of the object during hand closure.

c) Pinches: In pinch, left pinch and right pinch strategies the hand just closes without changing its pose.

d) Slide: To mimic the human behavior we realized an anticipatory routine composed of the following sub-phases, triggered by the initial contact with the object and the environment: i) apply a force on the object along x axis to maintain the contact during sliding, by commanding a reference position to the hand 10 mm below the contact position; ii) slide the object towards the edge of the table, iii) unload the contact to avoid pushing the object out of the table, by translating 10 mm along x, iv) rearrange the hand to favor the grasp, by translating 100mm along x and 50mm along z, and rotating along y of $\pi/12$ radians, v) close the hand.

D. Control

A Jacobian based inverse kinematic algorithm is performed to obtain desired joint positions q_r from the prescribed end effector evolution. A joint-level impedance control is used to realize the motion, with $K = 10^3 \frac{\text{Nm}}{\text{rad}}$ as stiffness and $D = 0.7 \frac{\text{Nms}}{\text{rad}}$ as damping for each joint. The control law is $\tau(t) = Ke(t) + D\dot{e}(t) + \mathbb{D}(q, \dot{q})$, where τ are the applied joint torques, $e = q_r - q$ and $\dot{e} = \dot{q}$ are the error at joint level and its derivate. $\mathbb{D}$ is a compensation of the robot dynamics evaluated by the KUKA embedded controller. All the control and sub-strategies implementation were performed in ROS.

V. Experimental Results

We test the effectiveness of the proposed architecture by performing table-top object grasping experiments. A table is placed in front of the system, as depicted in Fig. 3. The object is placed by an operator approximatively in the center of the table. RGB information from the web-cam triggers scene classification through the proposed deep neural network, which is followed by primitive execution. The task is repeated three times. The exact position of the object and its orientation vary each time, the first in a circle of radius ~ 100mm, the second in the full angle range. All the process is repeated for each of the 20 objects depicted in Fig. 6, chosen so as to elicit different grasping strategies. Objects number 5,6,7,8,9,10,16, and 19 are classified with a different strategy depending on their positioning and orientation. We consider three tests for each possible classification. The total amount of grasp tested is 111. None of the selected objects was used during the network training phase.

TABLE II
STRATEGY USED, SUCCESSES AND FAILURES FOR EACH GRASP.

Object	Strategy	Successes	Failures	Object	Strategy	Successes	Failure	Object	Strategy	Successes	Failure
1	bottom	3	0	7	pinch right	3	0	14	bottom	3	0
2	lateral	2	1	8	pinch	2	1	15	top	2	1
3	slide	2	1		pinch left	2	1	16	top	2	1
4	lateral	3	0		pinch right	2	1		top left	3	0
5	top	3	0	9	pinch	0	3		top right	3	0
	top left	2	1		pinch left	1	2	17	bottom	3	0
	top right	3	0		pinch right	1	2	18	slide	3	0
6	lateral	3	0	10	top	3	0	19	top	3	0
	top	3	0		top left	2	1		top left	2	1
	top left	2	1		top right	3	0		top right	3	0
	top right	3	0	11	lateral	3	0	20	lateral	2	1
7	pinch	3	0	12	top	2	1				
	pinch left	2	1	13	bottom	3	0				
								Total	-	90	21

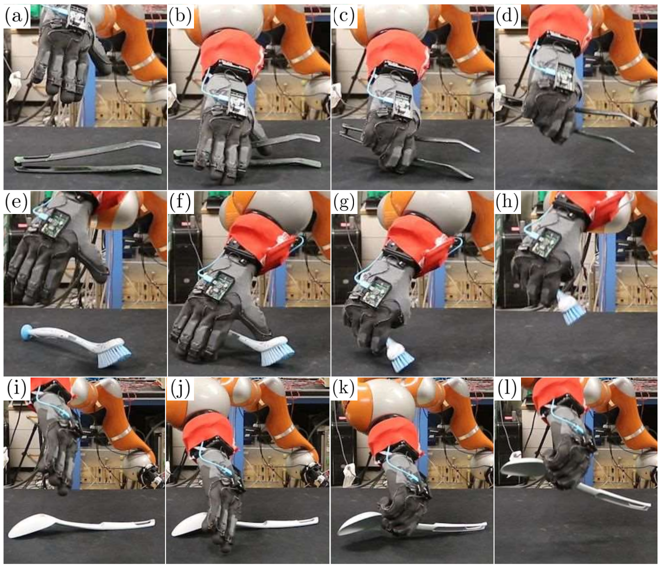

Fig. 9. Photosequences of grasps produced by the proposed architecture during validation: panels (a-d) present a pinch grasp of object 7, panels (e-h) a pinch-left grasp of object 8, and panels (i-l) a pinch-right grasp of object 9. Panel (a) shows the hand initial configuration. The contact is established in panel (b) through interaction with the environment, which also guides the hand towards the grasping achieved in panel (c). In (d) the object is firmly lifted.

Tab. II summarizes the results in terms of the primitive used, successes, and failures for each object. The overall grasping success rate is 81.1%. A grasp was considered successful if the robot maintained it for 5 seconds (after which the hand automatically opens). Note that objects 12 and 15 elicit only the top grasp primitive, independently from their orientation. They are indeed both (almost-)rotationally symmetric, so the classifier does not take in account their orientation to select the grasp. Looking instead at primitive-specific success rates we obtain: Top 85.7% (Fig. 7 (a-h)), Top left 73.3% (Fig. 7 (i-p)), Top right 100% (Fig. 7 (q-x)), Bottom 100% (Fig. 8), Pinch 55.6% (Fig. 9 (a-d)), Pinch left 55.6% (Fig. 9 (e-h)), Pinch right 66.7% (Fig. 9 (i-e)), Slide 83.3% (Fig. 10), Lateral 86.7% (Fig. 11).

VI. Discussion

This work represents a substantial improvement w.r.t. [11], where a similar success rate was obtained for human-robot handover, while only exploratory tests were performed on autonomous grasping. It is worth mentioning that this paper represents - together with [12] - the first work that validates over a large set of objects a combination of deep learning techniques and soft hands. Any formal comparison between the two works is prevented by the fact that neither this nor the other paper used a standardized object set and protocol [22], [23]. With this as premise, it is worth noticing that our success rate is only fairly lower than the one in [12] (which reports 87% of successes, versus the 81% reported here). However, in our work, we considered a higher number of objects for the testing phase (20 versus 10), spanning a wider range of shapes, and with larger differences w.r.t. the learning set. Another interesting consideration arises from a more in-depth analysis of the results. If we remove from the statistics the three objects that would require a pinch grasps (i.e. 7,8,9) the success rate jumps over 88%. This can be explained by an intrinsic feature of the soft hand we used, which was designed to perform power grasp. Nonetheless, using the environment as an enabling constraint, the end-effector can still partially overcome this limitation. We are sure - and we will test it in the future - that using other versions of the SoftHand that can execute both pinch and power grasping see e.g. [24], the success rate will increase.

VII. Conclusions

In this work, we proposed and validated a data-driven human-inspired architecture for autonomous grasping with soft hands. We achieve this goal by: i) introducing a novel deep neural network that processes the visual scene and predicts which action a human would perform to grasp the target object, ii) formulating and implementing an artificial counterpart of the strategies that we observed in humans, iii) combining them together in a integrated robotic platform, iv) extensively testing the proposed architecture in the execution of 111 autonomous grasps, achieving an overall success rate of 81.1%. Future work will be devoted to testing the use of other anthropomorphic and soft hands within this framework, as e.g. SoftHand 2 [24], RBO hand [25].

References

[1] A. Bicchi and V. Kumar, "Robotic grasping and contact: A review," in *ICRA*, vol. 348. Citeseer, 2000, p. 353.

[2] L. Birglen, T. Laliberté, and C. M. Gosselin, *Underactuated robotic hands*. Springer, 2007, vol. 40.

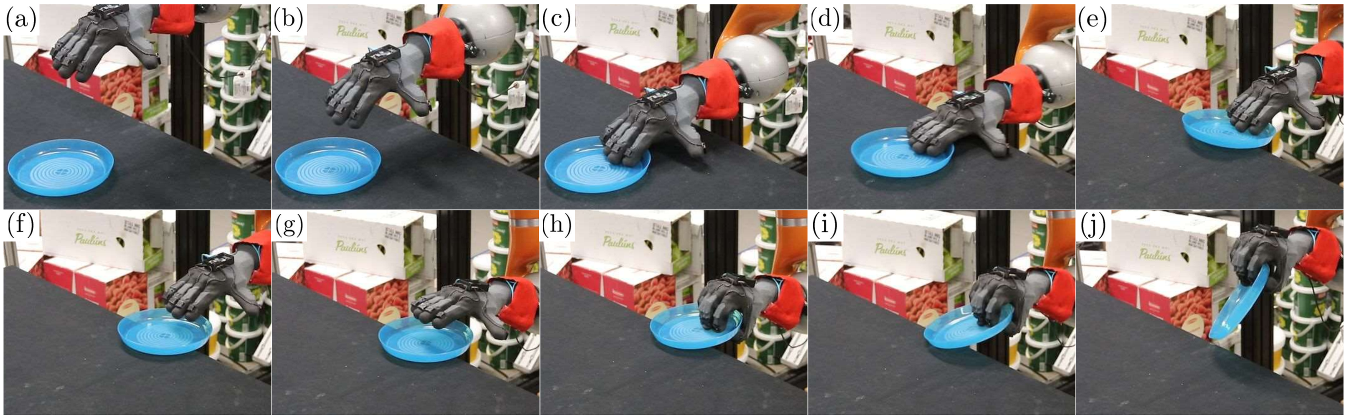

Fig. 10. Photosequence of a grasp produced by the proposed architecture during validation: slide grasp of object 3. Panels (a-c) depicts the approaching phase. In panels (d-e) the environment is exploited to guide the object to the table edge. In panels (f-g) the hand changes its relative position w.r.t. the object so to favor the grasp, which is established in panels (h-i). In panel (j) the item is firmly lifted.

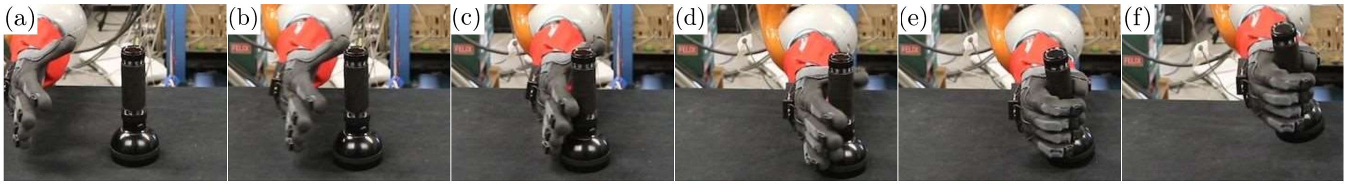

Fig. 11. Photosequence of a grasp produced by the proposed architecture during validation: lateral grasp of object 4. Panels (a-c) present the approaching phase. In panel (c) contact is detected, and in (e) the grasp is established. The object is lifted in panel (f).

[3] C. Piazza, G. Grioli, M. Catalano, and A. Bicchi, "A century of robotic hands," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. (In Press), 2019.

[4] C. Erdogan, A. Schröder, and O. Brock, "Coordination of intrinsic and extrinsic degrees of freedom in soft robotic grasping," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–6.

[5] M. Haas, W. Friedl, G. Stillfried, and H. Höppner, "Human-robotic variable-stiffness grasps of small-fruit containers are successful even under severely impaired sensory feedback," *Frontiers in neurorobotics*, vol. 12, p. 70, 2018.

[6] X. Yan, J. Hsu, M. Khansari, Y. Bai, A. Pathak, A. Gupta, J. Davidson, and H. Lee, "Learning 6-dof grasping interaction via deep geometry-aware 3d representations," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–9.

[7] P. Schmidt, N. Vahrenkamp, M. Wächter, and T. Asfour, "Grasping of unknown objects using deep convolutional neural networks based on depth images," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 6831–6838.

[8] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo *et al.*, "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–8.

[9] A. Gupta, C. Eppner, S. Levine, and P. Abbeel, "Learning dexterous manipulation for a soft robotic hand from human demonstrations," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 3786–3793.

[10] T. Nishimura, K. Mizushima, Y. Suzuki, T. Tsuji, and T. Watanabe, "Thin plate manipulation by an under-actuated robotic soft gripper utilizing the environment," in *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*. IEEE, 2017, pp. 1236–1243.

[11] M. Bianchi, G. Averta, E. Battaglia, C. Rosales, A. Tondo, M. Poggiani, G. Santaera, S. Ciotti, M. G. Catalano, and A. Bicchi, "Tactile-based grasp primitives for soft hands: Applications to human-to-robot handover tasks and beyond," in *Robotics and Automation (ICRA), 2018 IEEE International Conference on*. IEEE, 2019.

[12] C. Choi, W. Schwarting, J. DelPreto, and D. Rus, "Learning object grasping for soft robot hands," *IEEE Robotics and Automation Letters*, 2018.

[13] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The grasp taxonomy of human grasp types," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 66–77, 2016.

[14] G. A. Miller, E. Galanter, and K. H. Pribram, *Plans and the structure of behavior.* Adams Bannister Cox, 1986.

[15] C. Della Santina, C. Piazza, G. M. Gasparri, M. Bonilla, M. G. Catalano, G. Grioli, M. Garabini, and A. Bicchi, "The quest for natural machine motion: An open platform to fast-prototyping articulated soft robots," *IEEE Robotics & Automation Magazine*, vol. 24, no. 1, pp. 48–56, 2017.

[16] C. Eppner, R. Deimel, J. Alvarez-Ruiz, M. Maertens, and O. Brock, "Exploitation of environmental constraints in human and robotic grasping," *The International Journal of Robotics Research*, vol. 34, no. 7, pp. 1021–1038, 2015.

[17] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," *arXiv preprint*, 2017.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[19] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 42–49.

[20] R. S. Johansson and B. B. Edin, "Predictive feed-forward sensory control during grasping and manipulation in man," *BIOMEDICAL RESEARCH-TOKYO-*, vol. 14, pp. 95–95, 1993.

[21] T. Flash, "The control of hand equilibrium trajectories in multi-joint arm movements," *Biological cybernetics*, vol. 57, no. 4-5, pp. 257–274, 1987.

[22] J. Leitner, A. W. Tow, N. Sünderhauf, J. E. Dean, J. W. Durham, M. Cooper, M. Eich, C. Lehnert, R. Mangels, C. McCool *et al.*, "The acrv picking benchmark: A robotic shelf picking benchmark to foster reproducible research," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4705–4712.

[23] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Yale-cmu-berkeley dataset for robotic manipulation research," *The International Journal of Robotics Research*, vol. 36, no. 3, pp. 261–268, 2017.

[24] C. Della Santina, C. Piazza, G. Grioli, M. G. Catalano, and A. Bicchi, "Toward dexterous manipulation with augmented adaptive synergies: The pisa/iit softhand 2," *IEEE Transactions on Robotics*, no. 99, pp. 1–16, 2018.

[25] R. Deimel and O. Brock, "A novel type of compliant and underactuated robotic hand for dexterous grasping," *The International Journal of Robotics Research*, vol. 35, no. 1-3, pp. 161–185, 2016.