Combining Sensory Modalities and Exploratory Procedures to Improve Haptic Object Recognition in Robotics*

Bertrand Higy[1,2], Carlo Ciliberto[1,3], Lorenzo Rosasco[2,3] and Lorenzo Natale[1]

***Abstract*— In this paper we tackle the problem of object recognition using haptic feedback from a robot holding and manipulating different objects. One of the main challenges in this setting is to understand the role of different sensory modalities (namely proprioception, object weight from F/T sensors and touch) and how to combine them to correctly discriminate different objects.**

We investigated these aspects by considering multiple sensory channels and different exploratory strategies to gather meaningful information regarding the object's physical properties. We propose a novel strategy to train a learning machine able to efficiently combine sensory modalities by first learning individual object features and then combine them in a single classifier. To evaluate our approach and compare it with previous methods we collected a dataset for haptic object recognition, comprising 11 objects that were held in the hands of the iCub robot while performing different exploration strategies. Results show that our strategy consistently outperforms previous approaches [17].

I. INTRODUCTION

The literature on object recognition focuses primarily on vision. However there are many properties of objects that are difficult to infer from vision alone, but are directly observable using haptic information. Exploiting haptic information for object recognition is, however, a difficult task. This is because haptic information includes data from multiple channels (namely proprioception, touch and force) which are affected by the various object properties (weight, texture, size and volume) in different ways. Also, haptic object recognition is intrinsically active: appropriate exploratory procedures are required to extract the information related to different properties and channels. To tackle this complexity, experiments are usually performed in simplified settings in which objects are manipulated in a controlled, repetitive way. On the same line, existing work usually rely on carefully designed features and machine learning techniques.

In this paper we consider a more realistic and less controlled scenario, in which objects are given to the robot while varying their orientation and position in the hand. Drawing our inspiration from humans, two compatible strategies can be adopted to compensate for the additional complexity: 1) exploit the complementary sensory modalities available to haptic perception and 2) use different object-exploration strategies. Indeed, as pointed out in Lederman and Klatzky [10] humans make use of stereotypical exploratory procedures (EPs) to assess object properties (and also increase the information available by repeating these actions multiple times).

A main challenge for haptic recognition in robotics is how to deal with the multiple sensory modalities available during this process. To this end, we propose a novel method that learns how different sensory modalities (acquired during different EPs) can be combined for haptic recognition. To evaluate the effectiveness of our approach we collected a dataset comprising 11 objects held in the hand of the iCub robot while it performed different movements to explore them. Empirical evidence shows that our strategy for haptic object recognition consistently outperforms previous methods. We will make the dataset available for the community as a benchmark for haptic object recognition.

The paper is organized as follows. The state-of-the-art is described in Sec. II, followed by a description of the experimental setup, Sec. III, and the tools for data analysis, Sec. IV. Sec. V-A provides a comparison of several techniques for combining data from different sensory modalities or EPs. Sec V-B and Sec.V-C discusses pros and cons of combining multiple EPs together or several trials of the same EP. Finally, Sec. VI draws the conclusions of the paper.

II. STATE-OF-THE-ART

One way to cope with the complexity of the haptic modality is to use controlled exploration schemes to get reliable data. A good example is [2] where a carefully designed protocol is used to gather the data. However, this is not how human usually interact with their environment and not how we expect robots to behave. If one wants a robot to be able to work effectively in the same environment as humans, they should be able to face uncertainty and deal with less structured situations. In this work, we explore how well a robot can recognize objects in such conditions, leveraging on several human-inspired strategies that have already been successfully applied in the literature.

Humans are particularly good at combining different haptic cues to recognize objects, and the different haptic channels have been extensively used in robotic applications, confirming that they can provide complementary information about objects: joints positions is really effective for global shape and size estimation [8], [13], while tactile data gives essential information about textures and material [3], [7], [9] or precise shape [11]. Finally, force/torque (F/T) sensors

*This research has received funding from the European Unions FP7 for research, technological development and demonstration under grant agreement No. 610967 (TACMAN) and No. 270273 (Xperience).

[1]iCub Facility, Istituto Italiano di Tecnologia, Via Morego, 30, 16100, Genova, Italy, (bertrand.higy@iit.it, lorenzo.natale@iit.it).

[2]Università di Genova, Via All'Opera Pia, 13, 16145 Genova.

[3]Laboratory for Computational and Statistical Learning, Istituto Italiano di Tecnologia, Via Morego, 30, 16100, Genova, Italy, (cciliber@mit.edu, lrosasco@iit.it).

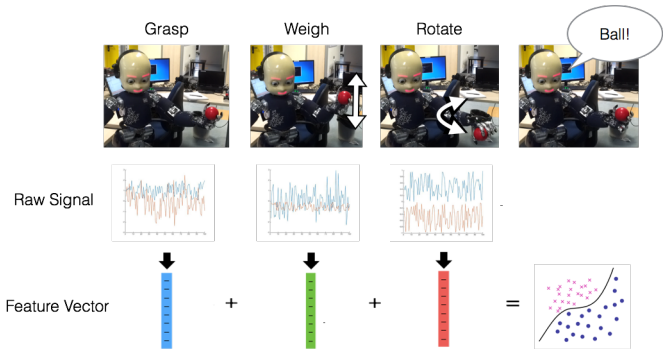

Fig. 1. Object recognition pipeline: (Top) iCub performs different exploratory procedures while holding an object. (middle) the raw sensory signal is coded into feature vectors (bottom) which are used to train haptic classifiers.

are useful to infer weight [12]. It has further been shown that object recognition is improved when several sources of information are available [14], [17].

Haptic object recognition is an active process. In humans, Lederman and Klatzky [10] have well described how stereotypical exploratory procedures are used in a consistent way to obtain haptic information from objects. Six EPs are described by the authors for the acquisition of static properties. Depending on the task and the dimension involved (weight, shape, temperature and so on), different EPs will be performed as they do not provide the same kind of information. For example, unsupported holding is used to estimate the weight of an object while static contact is better for temperature. The enclosure is the action providing information for the broadest range of dimensions even if in a crude way. It is also fast and thus is usually the first one to be performed when exploring an object, before more specific gestures take over.

Thus, it is not surprising that grasping (enclosure in the seminal nomenclature) is the most common action used in robotics for object exploration [1], [2], [6], [8], [13], [14], [17]. Nevertheless, there are many examples of use of other EPs like pressure [2], static holding [2], regrasping [6], lateral motion [2], [3], [7], [9], contour following [11], or other actions which are not part of the original list from Lederman and Klatzky [1], [2], [17]. It has further been shown that combining different actions and modalities improve the perception capabilities and that the effectiveness of an actions depend on the modality used as well as the property you are interested in [2], [17]. Finally, it was observed in [7], [14] that, similarly to what happens for humans, performing the same EP multiple times on the same object significantly improves the recognition accuracy in both object and texture recognition settings.

When developing an artificial system for haptic recognition, a critical question is how to combine the multimodal information across multiple exploratory procedures. A straightforward approach is to concatenate all sensory inputs in a joint vector which is then used to train a learning machine to perform classification [1], [14]. This strategy however does not account for the differences (in magnitude and structure) of the difference sensory signals and can lead to sub-optimal results. A possible solution is to train an individual learning machine for each sensory modality and then group their results either using a majority voting strategy [7] or a weighted sum of predictions [2], [17]. In this work we propose novel strategies to combine the multi-sensory input in a haptic recognition setting and compare them with previous methods. Our approach builds on a hierarchical model organized in two layers: the first layer is composed by classifiers trained on a single sensory modality and can be interpreted as the process of learning individual objects' features. The second layer effectively learns how to combine these individual features in a joint haptic classifier. The experimental evaluation performed on the iCub demonstrates that our method allow recognition accuracy higher than other methods previously proposed in the literature

III. HAPTIC EXPLORATION OF OBJECTS

In this section we introduce our approach to haptic object recognition. In its general formulation, this task consists in training a robot to classify a set of objects by physically interacting with them and solely based on the information acquired through haptic/tactile sensors (i.e. without using visual cues). We assume that objects are sufficiently small to be held by the robot and that most of the interaction occurs while the robot is holding them. The robot will use various sensors, including proprioception and force and torque from F/T sensors mounted on the arms.

In this setting, one of the major challenges is to successfully combine multiple measurements gathered from different sensors of the robot. Furthermore, depending on the way the robot interacts with the object (e.g. grasping, pushing, etc.) such measurements could differ dramatically. We devised a set of exploratory procedures that the system can use to interact with hand-held objects and assessed different machine learning strategies to combine the multi-modal haptic data. Fig. 1 gives an overview of the overall pipeline.

A. Hardware

Our investigation was carried on the iCub robot. iCub is a full humanoid with a total of 56 Degrees of Freedom (DOF). In particular, the arms have 16 DOFs: 3 for the shoulder (pitch, roll, yaw), 1 for the elbow, 3 for the wrist (pronosupination, pitch, yaw) and 9 DOFs for the hand: 3 DOFs for the thumb, index and middle finger proximal phalanges, 3 for the corresponding distal phalanges (the two distal joints of each finger are coupled), 1 for the six joints of the ring and little fingers (they are all coupled together), 1 for the thumb opposition and 1 for fingers adduction/abduction (see [15] for further details). Coupling in the joints of the fingers is achieved using springs that allow the fingers to conform to the objects being grasped. Magnetic encoders in the phalanges allow retrieving the configuration of all joints of the hand. The robot arm is equipped with a 6-channels F/T sensor in each arm.

B. Sensory Modalities

In this work we focused on two main sources of haptic information, which correspond to the Kinesthetic sensory modalities in humans: *Proprioception* via the joint encoders readings and *Force/Torque* sensing, via the F/T sensor measurements. By adding minimal amount of prior information about the robot's model, we identified the following 5 types of sensory information:

- **Joints positions**: We recorded the joint's angle position of the main three fingers of the robot's hand for a total of 7 measurement at each time instant: thumb 3 DOFs, index 2 DOFs and middle 2 DOFs.
- **Fingers' spring model**: The actual position of each finger is measured using magnetic encoders. In absence of external forces the springs in the phalanges do not deform and the position of all the joints can be predicted knowing the motor encoders and the mechanical coupling between the phalanges. When grasping an object, forces exerted on the hand misalign the predicted and observed finger position. This misalignment is a cue that encodes the amount of force exerted by the finger on the object and, as such, it is a powerful cue that helps to measure physical properties of the object. Following this intuition, in this work we include the difference between predicted and measured joint position of each finger as a possible cue to discriminate objects.
- **Raw Force/torque**: The F/T sensor in the arm shoulder is a six-axis sensor measuring 3 forces and 3 torques at a $100Hz$ rate.
- **Wrist Force/torque**: if we assume that external forces are acting exclusively on the hand of the robot, we can exploit knowledge of the robot's model to estimate the forces and torques exerted by an object on the wrist. For this purpose we used the *iDyn* library available for the iCub developers community [4].
- **Root Force/torque**: The force and torque exerted by the object on the hand can be mapped to the root reference frame (i.e. the robot's waist), providing information about the weight of the object that is independent of the actual position of the hand.

C. Exploratory procedures

The measurements recorded from the sensors described in Sec III-B depend critically on the specific way the robot is interacting with the object. As a consequence, it is reasonable to expect that different exploration strategies would allow accessing information related to different properties of the object. Following this intuition, in this work we devised the following strategies (i.e. *Exploratory Procedures (EP)*), that the robot can use to gather haptic information about objects it holds in the hand (all represented in Fig. 2):

- **Grasp**: This procedure corresponds to the *enclosure* EP originally described in [10]. It consists in wrapping the fingers around the object and seems to be the most basic EP for humans, giving information about a broad range of physical properties in a short time. In our setup, a three-fingers grasp (thumb, index and middle finger) is used. To simplify our acquisitions process, at each trial we placed the object in the hand of the robot.
  In order to detect contact and avoid the robot's finger to close indefinitely, we used a threshold on the springs model misalignment (described in Sec III-B).
- **Weigh**: weighing the object can provide finer information about the inertial properties of an object. This procedure corresponds to the *unsupported holding* EP in humans ([10]). In our implementation the robot was programmed to move the object rapidly upwards and downwards three times.
- **Rotate**: We implemented an additional EP during which the robot rotates its wrist while keeping the object firmly within its grasp (this is similar to the rolling action in [1], except that the whole forearm is moving). This strategy provides further information about inertial properties of the object and turns out to be indeed particularly useful for object recognition. The Rotate EP does not have an equivalent in terms of human EPs originally identified in [10].

D. Features

Using carefully designed features is not unusual in haptic object recognition and can really improve the performances. However, in this work, we choose a simpler approach using raw sensor values or applying only simple transformations to them (i.e. Fast Fourier Transform). Both of the considered features can be encoded in a single vector and therefore used by machine learning tools, as described in Sec. IV (see also Fig. 1), to perform recognition:

- **Snapshot** These feature consists in the recording of all measurements (joint angle, F/T measure, etc.) in a specific instant in time. We acquired this kind of features after the Grasp EP was completed in order to record the state of the hand enclosing the object of interest and capture properties related to shape and surface hardness. Clearly this type of feature could be used also for novel EPs in the future, but for the Weigh and Rotate EPs adopted in this work, we did not find any significant point in time to be recorded.
- **Fourier** In order to capture the dynamic evolution of the sensory signal during a given EP, we also performed a frequency analysis on it. For our experiments, we chose the Fast Fourier Transform (FFT) and condensed each sensory signal in a vector of fixed length (concatenating the coefficients of the real and complex imaginary parts). One advantage of this technique is that it allows to encode sequences of different length into vectors of same size. Arguably, this approach is extremely coarse, and much careful analysis of the signal during an EP could lead to significant improvement at recognition time. However it is a very cost effective solution that can be applied to any EP.

E. Dataset

Using the exploratory procedures described in the previous section we collected a dataset in which the robot interacted with a set of 11 objects (Fig. 2). These objects were carefully selected in order to span a wide range of physical properties such as weight, shape and surface hardness. As a consequence, one of the major challenges of haptic recognition in this setting is that, for any sensory modality considered, there are always 2 or more objects that cannot be distinguished using such modality alone (e.g. the empty and full bottles have exactly the same shape but different weight).

Haptic measurements in our dataset were acquired in sessions during which the three EPs (Grasp, Weigh and Rotate) were executed in sequence. More precisely, each acquisition session would start with the robot presenting its open hand with the palm facing upwards, waiting for an object. Then, the experimenter would put the object to be learned/recognized in the hand. Using the variation of the F/T measurement as a contact detection cue, the robot would proceed with the Grasp EP, enclosing the object in its hand, then subsequently with the Weigh EP and finally the Rotate EP. We performed the above procedure between $29 - 32$ times per object. We split the data collected in a test set containing 9 sessions for each object and a training set containing the remaining sessions. We will make this dataset available online for the community with the plan of increasing it further the current 11 classes.

Unstructured/uncontrolled object pose: We care to point out that a further challenge of our dataset is that, during grasping, objects were not always positioned with the same pose in the hand. This simulates a realistic scenario in which the robot grasps the objects autonomously and, therefore, introduce a certain degree of variability in the acquired data.

IV. DATA ANALYSIS TOOLS

In Sec. V we will detail our experimental analysis on the 11-objects dataset collected in this work. Our goal is to evaluate the relevance of the sensory modalities available on the iCub, with respect to the task of haptic object recognition. To this end, in this section we review the machine learning techniques we adopted to perform object classification from the haptic features described in Sec. III-D, following the pipeline represented on Fig. 1.

A. Kernel methods

A wide range of powerful classification methods are available from the machine learning literature. In this work we focused on the family of non-parametric *kernel methods* [16], which do not assume any specific model for the classification function to be learned, but are completely data-driven. Typically, a machine learning problem can be formulated as the problem of finding the function $f : \mathcal{X} \to \mathcal{Y}$ from an input space $\mathcal{X}$ to an output space $\mathcal{Y}$ that best approximates the unknown functional relation between a finite set of observations or *training* pairs $\{(x_i, y_i)\}_{i=1}^n$. In our setting, the input examples x_i are the feature vectors (snapshot or fourier) described in Sec. III-D, while the corresponding output would be the label $y_i \in \{1, \ldots, 11\}$ associated to one of the 11 objects in our dataset.

Fig. 2. The 11 objects selected for our dataset. From left to right and going down: a cylindrical wooden toy, a sponge, a plastic cup, an empty water bottle, a turtle-shaped soft toy, a green bottle made of hard plastic, a tennis ball, a rectangular tea box, a blue ball made of soft foam, a bottle of water filled with paper and a round food box.

The learning problem is then formulated as the optimization

$$\underset{f \in \mathcal{H}}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^{n} V(y_i, f(x_i)) + \lambda R(f) \qquad (1)$$

where $V : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is the so-called *loss* function, penalizing prediction errors according to the task, $\mathcal{H}$ is a space of candidate functions where we aim to find our classifier and $R : \mathcal{H} \to \mathbb{R}$ is a *regularizer*, controlling the "complexity" of f and avoiding overfitting of the data points. We refer the reader interested to an in-depth introduction to machine learning to the excellent book on the subject [16].

In classification settings, a loss function often adopted is the least-squares loss $V(y, f(x)) = \|y - f(x)\|^2$, and class labels y are coded as binary vectors of length equal to the number of classes T (in our case $T = 11$), containing all -1 in their entries except for a $+1$ on the entry corresponding to the true class. Therefore, in these settings, the learned predictor is a vector-valued function $f : \mathcal{X} \to \mathbb{R}^T$ where each output can be interpreted as the confidence score that a point x belongs to a specific class. Indeed, the actual predicted class $\hat{c}(x)$ of an example x is typically obtained by taking

$$\hat{c}(x) = \text{argmax}_{c=1,\ldots,T} f(x)^{(c)} \qquad (2)$$

with $f(x)^{(c)}$ denoting the c-th element of the vector $f(x)$.

For our experiments we used the GURLS library [18], which implements different algorithms within the field of kernel methods implemented both in MATLAB and C++, allowing on one hand for flexible prototyping and on the other for real-time efficiency. Moreover, the GURLS library offers automatic parameter tuning, namely the choice of space $\mathcal{H}$ and regularization parameter λ from Eq. (1), making it a general "black-box" tool for data analysis.

B. Combining Sensory Modalities

One of the main interests of this work is to understand how multiple sensory modalities can be combined to discriminate grasped objects. Therefore, a natural question is how to adapt the machine learning tools introduced above to make the best use of the multiple channels available during both learning and prediction stages. In the following we describe different approaches considered in this work to perform sensory data combination:

Concatenation ([1], [14]). Sensory information can be combined by concatenating different feature vectors as described in Sec. III-D. One of the main limitations of this approach is that the scale and range of the values within different feature vectors cannot be compared, often leading to inconsistent results. A typical approach is to standardize the feature vectors independently so that they all have 0 mean and 1 standard deviation. However, standardization could remove relevant information from the signal.

Averaging ([2], [17]). An alternative approach is to combine the confidence scores produced by classifiers trained on each modality independently. Indeed, as observed in Eq. (2), a predictor trained independently on a single modality would produce as output a vector of fixed length $T = 11$ whose elements represents the confidence score that the observed feature vector belongs to the corresponding classes. When each modality provides complementary information about the physical properties of the object, the combination of the confidence scores could lead to a more robust and reliable classifier. The *averaging* strategy therefore consists in taking the *argmax* of the sum $F(x) = \sum_{m=1}^{M} f_m(x)$ of the predictors f_m trained individually on each modality m. A related approach was considered in [17], where instead of the T-dimensional multiclass classfiers f_m, the authors useed $T(T-1)/2$ binary classifiers for each modality, each trained to discriminate between a single pair of objects. The scores of these binary classifiers were first recombined through pairwise coupling [5] to obtain modality and EP specific predictions, before averaging was performed on the probabilities obtained this way.

Hierarchical. Here we propose a novel method to combine sensory data, by learning how to combine the modality-specific classifiers f_m rather than simply averaging them. We do so by concatenating the scores predicted by each f_m, each consisting of a T-dimensional vector reporting the likelihood of a signal to be associated to one of the T classes and then use these MT dimensional vector as the new input to train a "unifying" classifier F. More precisely, when provided with a new observation $x \in \mathcal{X}$, such a classifier will take as input the scores produced by the individual predictors, namely $F(x) = F(f_1(x), \ldots, f_M(x))$. $F(x)$ will be again a T dimensional vector containing the likelihood of x belonging to a specific object class. We can interpret this strategy as "hierarchical" since we have the first layer of predictors f_m which are classifiers tuned to a specific modality, while the second layer is in charge of combining modality-specific response into the final prediction $F(x)$. Following [17], we also consider the possibility two use as first layer the outputs of binary classifiers trained to distinguish between only two classes out of the T available for the problem.

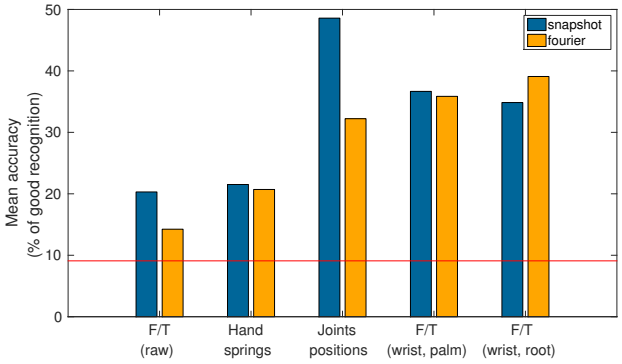

Fig. 3. Accuracy (percentage of good recognition) achieved by the condition-specific classifiers, using the Grasp EP. For each modality, two features (snapshot and fourier) are compared. The accuracy with random guesses is represented by the red line.

V. EXPERIMENTAL ANALYSIS

In this section we report the experimental analysis on haptic object recognition. Results are reported on the dataset described in Sec. III-E. We begin by investigating the role of each sensory modality independently and then proceed to analyze the effect of combining them together according to the strategies discussed in Sec. IV-B. We also take into account the effect of performing multiple, subsequent explorations of the objects in order to reduce uncertainty and achieve higher recognition accuracy.

A. Haptic Recognition by Grasping EP

In this section we focus on the single Grasping EP. Indeed, in the application scenario considered in this work, the robot starts to interact with other objects by first holding them in its hand. As a consequence, the Grasping EP is preliminary to both Rotate and Weigh EPs, and a natural question is therefore to ask what classification accuracy can be achieved by using this information alone.

1) Individual Sensory Modalities

We first assess the quality of each sensory modality independently. On one hand, this allows getting a better understanding of the role of each modality within the task of object recognition and on the other hand, it offers useful insights on possible ways to combine them in order to improve the classification accuracy.

Fig. 3 reports the average accuracy per class obtained by RLS classifiers trained individually on the different modalities available during grasp. 99 examples (9 per object) where put aside for testing while training was performed on the remaining 233 examples (~ 20 per object). A first observation is that results, although on average significantly higher than chance, are relatively low for a 11-class classification problem, with the highest accuracy slightly less than 50%. This is to be expected, since objects in our dataset where

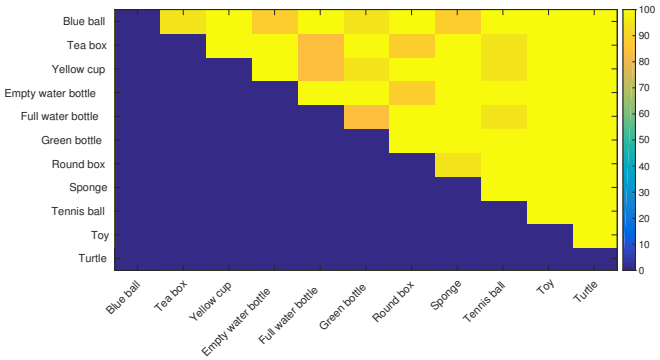

Fig. 4. Highest classification accuracy achieved by a classifier (across all modalities) for each pair of objects.

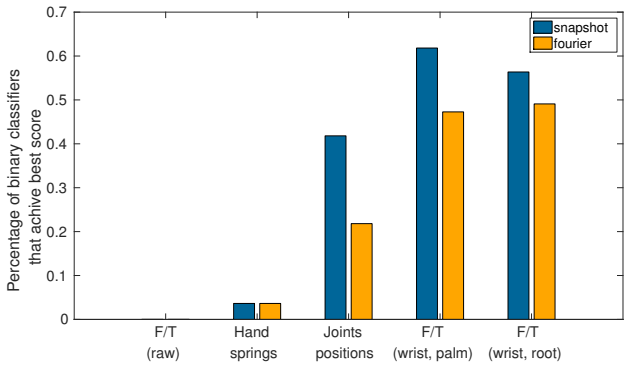

Fig. 5. Percentage of binary classifiers achieving highest accuracy across all modalities. The ratio is taken across all possible object pairs.

intentionally chosen to be "easily" confused and the grasping strategy results in high variability in the measurements.

From this experiment we further observe that adding prior information about the physical model of the robot can be extremely informative. While raw measurements from the F/T sensor exhibit the lowest performance among all modalities, it is possible to increase performance by projecting the F/T sensor measurements onto appropriate reference frames (i.e. the wrist and the root reference frame). This strategy leads to a remarkable boost of $\sim 20\%$ accuracy. This observation is particularly encouraging since it suggests that, while in this work we are focusing on (almost only) the raw signal acquired through the different sensors available, by using further information from the robot it could be possible to further improve the recognition capabilities of the system.

Classification of Objects Pairs. To better characterize the efficacy of each sensor with respect to the recognition task, we also performed a set of binary classification experiments where we evaluated the information provided by individual modalities in discriminating solely between two objects. This was done to confirm that the low performances observed in Fig. 3 are due to the fact that, for each modality, there exists some objects that cannot be distinguished one from the other. This can be observed in Fig. 4, where we have reported, for each object pair, the highest classification accuracy achieved with respect to any individual modality.

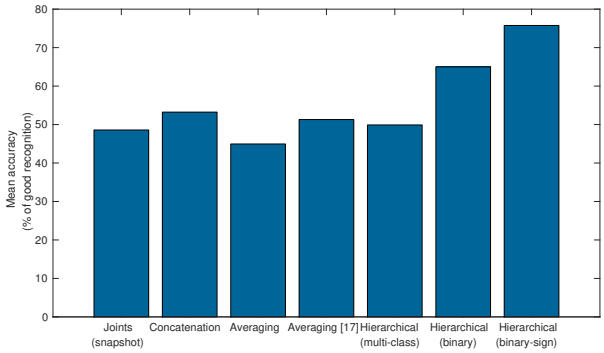

Fig. 6. Comparison of the accuracy (percentage of correct classification) achieved by combining the four modalities and the two features using different approaches. The first result is the result of the best condition-specific classifier and serves as a baseline.

For these experiments we tested the binary classifiers on 18 examples (9 per class) after training them on the remaining examples (~ 40 examples overall or ~ 20 per class).

As can be noticed, for most objects pairs a single modality is sufficient to achieve 100% accuracy, suggesting that carefully combining them in a single classifier should lead to a remarkable overall improvement for the 11-class classification scenario. To obtain a better understanding of which modalities are more significant for the classification task, in Fig. 5 we report, for each sensory modality, the percentage of object pairs for which such modality allows for the best possible accuracy. Note that in most cases, more than one modality achieves best accuracy and that, as previously observed in the multiclass setting, joint positions and F/T measurements seem to be the most relevant features.

From this experiment it is clear that the multiple signals provided by the sensors available on the robot indeed provide a very powerful cue to discriminate between objects. Therefore, in the following we investigate how to combine such multiple modalities to improve the overall recognition capabilities of our system.

2) Combining Sensory Modalities

We proceed by investigating the impact of combining multisensory information according to the methods introduced in Sec. IV. In particular, we considered two variants of the "hierarchical classifier" introduced in Sec. IV: one using as first "layer" the predictions of the multi-class classifiers trained independently on each modality while the other uses the binary predictors trained for each modality and each pair of objects in our dataset separately. We refer to these two models as respectively *hierarchical multi-class* and *hierarchical binary*. For the binary case, we also make the distinction between taking the predictors at the first layer as producing a score (*binary*) or just a ± 1 signal (*binary-sign*).

For comparison purposes, we also replicated the methodology used by Sinapov et al. in [17], which use a setting that is really close to ours. In that work, condition-specific pairwise classifiers (similar to our binary classifiers) are first grouped for each modality and EP separately through pairwise coupling [5]. Predictions from different modalities

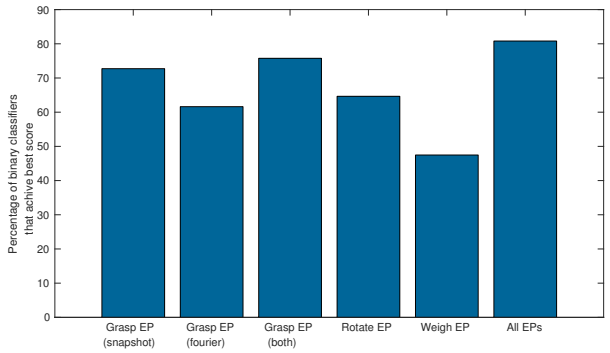

Fig. 7. Accuracy achieved with the *hierarchical binary-sign* approach by combining the information from a specific EP and all modalities (5 first columns) or from all data together (last column). Three classifiers have been trained for the grasping, one with snapshot feature, one with fourier and one with both.

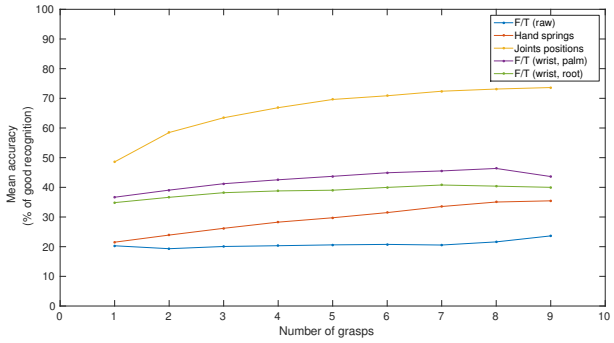

Fig. 8. Evolution of the mean accuracy achieved by combining predictions from 1 to 9 session for each modality separately, with the Grasp EP and the snapshot feature.

and/or EPs are then combined using *averaging*.

Fig. 6 reports the classification accuracy achieved by the *averaging*, *concatenation*, *hierarchical* (multi-class, binary or binary-sign), as well as the methodology from [17]. As a reference with respect to classifying using the sensory modalities individually, we also report the performance of the predictor trained exclusively on the *joint position* (snapshot), which is the best performing one in the previous experiments (Fig. 3). Experiments were performed by our systems on the same training set of 233 examples and tested on the remaining 99 as described in Sec. V-A.

It is interesting to notice that combining multiple sensory channels can lead to an overall improvement of $\sim 30\%$ in accuracy. However, it is quite surprising that the sole approach that seems to make correct use of the different modalities is the *hierarchical binary* (signed or not). In particular, we observe that averaging the responses of the individual predictors can be even more detrimental than considering each modality separately (the *averaging* approach exhibits lower performance than the one using the joints state alone).

The observation above suggests that in order to discriminate between the objects in our dataset it is fundamental to learn how the measurements provided by the sensors are correlated one to the other. In principle, this could be done by training a classifier on the concatenation of all feature vectors obtained from multiple modalities. Indeed, as shown in Fig. 6, the *concatenation* method appears to be promising, with higher accuracy than the *hierarchical multiclass* method. Arguably, as already mentioned in Sec. IV, the main challenge when training a classifier directly on the concatenation of all features, is that the measurements from different channels need to be preprocessed in order to be comparable one to the other. The way this is done affects dramatically the overall classification performance. While in this work we only standardized each feature before concatenating them, a much better approach could be available.

It should also be noted that, even though it performs better than *averaging* and *hierarchical multiclass* methods, the methodology from [17] doesn't reach the performances of *hierarchical binary* technique, staying slightly behind the *concatenation* one.

B. Combining Multiple EPs

In this section we investigate how multiple EPs can be combined, leading to a further improvement in classification accuracy. While we have already observed that the Grasping EP achieves remarkable performance, such exploration strategy provides mainly information about static qualities of the object (e.g. weight, shape, hardness). In contrast, the Weigh and Rotate EPs offer the possibility to gather data about the inertial/dynamical properties of the object, which could be extremely useful for recognition.

We report the accuracy per individual EP in Fig. 7, together with the performance of a classifier trained on all measurements obtained during all three EPs together. To make these observations clear, we report only the results obtained with the *hierarchical binary-sign* approach, which exhibited significantly better results than its competitors. Interestingly, the Weigh and Rotate EPs achieve remarkably lower results than Grasping, suggesting that this modality allows capturing more accurately the properties of the object. In particular we care to point out that the Grasp action records how the finger closes around the object and this could provide information about the hardness of the material, which Weigh and Rotate EP cannot capture. However, it is also clear that such EPs are not redundant with respect to Grasping, since by combining them together (concatenating the output of binary classifiers from all modalities and EPs) we achieve a remarkable 81% of accuracy, around 5% improvement over using Grasp information alone.

C. Multiple Grasps (Weighs, Rotations)

In this section we investigate the impact of performing actions multiple times in order to gather a richer characterization of the object. One of the challenging aspects of our haptic dataset is that we provided the objects to the robot in a somewhat uncontrolled way. This was done to reproduce a realistic setting, in which the robot, in distinct trials, may grasp the objects in different ways. Clearly, very different grasp poses would cause a dramatic change in the haptic measurements, leading to lower recognition performance. However, a viable strategy to mitigate this effect is to have the robot perform each action multiple times, varying each time how the object is grasped. In this way it should

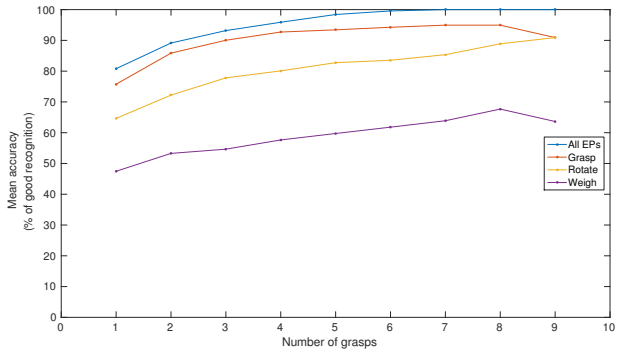

Fig. 9. Evolution of the mean accuracy achieved by combining predictions from 1 to 9 session for each EP separately or all together.

be possible to improve the confidence of the classification system that was previously trained on individual actions.

Here we quantify the improvement in performance when combining up to 9 subsequent actions. This analysis was done by averaging the confidence scores of the predictors trained on individual grasps according to the protocol described in Sec. V-A and Sec. V-B. In particular, Fig. 8 reports the improvement in average classification accuracy of multiclass classifiers trained on each modality independently (see Sec. V-A). We averaged across multiple runs (i.e. different combinations of individual grasps), in order to account for statistical variability. It can be noticed that for inertial measurement, multiple grasps do not appear to provide meaningful insight on the object identity. This is expected since the object weight does not change according to how it is held in the robot hand. On the opposite, for the fingers' joints sensory channel, multiple grasps lead to a dramatic improvement of $\sim 10\%$ for just 1 additional grasp to an overall improvement of $\sim 25\%$ of average classification accuracy when all 9 grasps were combined together.

Finally, we report in Fig. 9 the benefit of performing multiple actions, for binary classifiers combining all modalities during each EP. In particular, it can be noticed that, similarly to what is observed for the joints sensory channel in Fig. 8, even a single additional trial improves the accuracy by $\sim 5-10\%$ for each EP, and an overall improvement of $\sim 20-25\%$ is obtained when all grasps are considered. Furthermore, using this strategy together with the combination of all EPs, we managed to achieve perfect recognition (with 7 grasps or more). These results are really encouraging and clearly show the efficacy of the technique to compensate for the uncertainty in the pose imposed on our setup.

VI. CONCLUSIONS

In this paper we studied object recognition with haptic information and proposed a novel method to combine sensory information. Our approach departs from previous literature in that it learns how to combine different input sources by exploiting a hierarchical model. We assessed our approach on a novel dataset for haptic object recognition collected with the iCub robot. Experiments showed that our method achieves higher recognition rate than previous work. In particular, when applied to data gathered from different modalities, our approach outperformed previous methods by $> 25\%$ in classification accuracy.

We observed encouraging results when combining multiple grasps actions. These results suggest that a promising direction for research is the active selection of EPs. In future work, we would also like to investigate additional EPs and explore how each can prove useful to discriminate along different dimensions (shape, weight, temperature or softness among others).

REFERENCES

[1] Sachin Chitta, Jrgen Sturm, Matthew Piccoli, and Wolfram Burgard. Tactile Sensing for Mobile Manipulation. *IEEE Transactions on Robotics*, 27(3):558–568, June 2011.

[2] Vivian Chu, Ian McMahon, Lorenzo Riano, Craig G. McDonald, Qin He, Jorge Martinez Perez-Tejada, Michael Arrigo, Naomi Fitter, John C. Nappo, Trevor Darrell, and Katherine J. Kuchenbecker. Using robotic exploratory procedures to learn the meaning of haptic adjectives. In *2013 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3048–3055, May 2013.

[3] Patrick Dallaire, Philippe Gigure, Daniel mond, and Brahim Chaibdraa. Autonomous tactile perception: A combined improved sensing and Bayesian nonparametric approach. *Robotics and Autonomous Systems*, 62(4):422–435, April 2014.

[4] Matteo Fumagalli, Serena Ivaldi, Marco Randazzo, Lorenzo Natale, Giorgio Metta, Giulio Sandini, and Francesco Nori. Force feedback exploiting tactile and proximal force/torque sensing. *Autonomous Robots*, 33(4):381–398, April 2012.

[5] Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. pages 507–513. MIT Press, 1998.

[6] Koh Hosoda and Tomoki Iwase. Robust haptic recognition by anthropomorphic bionic hand through dynamic interaction. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 1236–1241. IEEE, 2010.

[7] Nawid Jamali and Claude Sammut. Majority Voting: Material Classification by Tactile Sensing Using Surface Texture. *IEEE Transactions on Robotics*, 27(3):508–521, June 2011.

[8] Magnus Johnsson and Christian Balkenius. Experiments with proprioception in a self-organizing system for haptic perception. *Proceedings of TAROS 2007*, pages 239–245, 2007.

[9] Oliver Kroemer, Christoph H. Lampert, and Jan Peters. Learning Dynamic Tactile Sensing With Robust Vision-Based Training. *IEEE Transactions on Robotics*, 27(3):545–557, June 2011.

[10] Susan J Lederman and Roberta L Klatzky. Hand movements: A window into haptic object recognition. *Cognitive Psychology*, 19(3):342–368, July 1987.

[11] Uriel Martinez-Hernandez, Tony J. Dodd, Lorenzo Natale, Giorgio Metta, Tony J. Prescott, and Nathan F. Lepora. Active contour following to explore object shape with robot touch. In *World Haptics Conference (WHC), 2013*, pages 341–346, April 2013.

[12] Lorenzo Natale, Giorgio Metta, and Giulio Sandini. Learning haptic representation of objects. 2004.

[13] Lorenzo Natale and Eduardo Torres-Jara. A sensitive approach to grasping. *Proceedings of the sixth international workshop on epigenetic robotics*, pages 87–94, 2006.

[14] Stefan Escaida Navarro Nicolas Gorges. Haptic Object Recognition using Passive Joints and Haptic Key Features. pages 2349–2355, 2010.

[15] Alexander Schmitz, Ugo Pattacini, Francesco Nori, Lorenzo Natale, Giorgio Metta, and Giulio Sandini. Design, realization and sensorization of the dexterous iCub hand. In *2010 10th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 186–191, December 2010.

[16] John Shawe-Taylor and Nello Cristianini. Kernel Methods for Pattern Analysis. July 2004.

[17] Jivko Sinapov, Connor Schenck, Kerrick Staley, Vladimir Sukhoy, and Alexander Stoytchev. Grounding semantic categories in behavioral interactions: Experiments with 100 objects. *Robotics and Autonomous Systems*, 62(5):632–645, May 2014.

[18] Andrea Tacchetti, Pavan K. Mallapragada, Matteo Santoro, and Lorenzo Rosasco. GURLS: A Least Squares Library for Supervised Learning. *Journal of Machine Learning Research*, 14:3201–3205, 2013.