Advanced Robotics, Vol. 21, No. 13, pp. 1545–1564 (2007)

Also available online - www.brill.nl/ar

Full paper

Internal models of reaching and grasping

CLAUDIO CASTELLINI [1], FRANCESCO ORABONA [1], GIORGIO METTA [1,2,*]
and GIULIO SANDINI [2]

[1] *LIRA-Lab, DISJ, University of Genoa, Viale Cansa 13, 16145 Genoa, Italy*
[2] *Italian Institute of Technology, 16163 Genoa, Italy*

Received 22 December 2006; revised 2 May 2007; accepted 16 May 2007

Abstract—One of the most distinguishing features of cognitive systems is the ability to predict the future course of actions and the results of ongoing behaviors, and in general to plan actions well in advance. Neuroscience has started examining the neural basis of these skills with behavioral or animal studies and it is now relatively well understood that the brain builds models of the physical world through learning. These models are sometimes called 'internal models', meaning that they are the internal rehearsal (or simulation) of the world enacted by the brain. In this paper we investigate the possibility of building internal models of human behaviors with a learning machine that has access to information in principle similar to that used by the brain when learning similar tasks. In particular, we concentrate on models of reaching and grasping, and we report on an experiment in which biometric data collected from human users during grasping was used to train a support vector machine. We then assess to what degree the models built by the machine are faithful representations of the actual human behaviors. The results indicate that the machine is able to predict reasonably well human reaching and grasping, and that prior knowledge of the object to be grasped improves the performance of the machine, while keeping the same computational cost.

Keywords: Manipulation and grasping; computational intelligence; teleoperation; cognitive robotics.

1. INTRODUCTION

One of the most distinguishing features of cognitive systems is the ability to learn to predict the future course of actions and the results of ongoing behaviors, and in general to plan actions well in advance. It is now relatively well understood that the brain builds models of the physical world through learning. These models are sometimes called 'internal models', meaning that they are the internal rehearsal (or simulation) of the world enacted by the brain [1].

Interestingly, these models are built not only because they are required to control movements, but also, as has been determined more recently, to interpret the

[*] To whom correspondence should be addressed. E-mail: pasa@liralab.it

movements of others [2–4]. There is now a large body of literature that links the observation of actions to action execution, e.g., the study of the motor system conducted by Rizzolatti *et al.* in relatively recent years [5–7]. It seems then that building predictions of the future course of action is a key feature of intelligent living systems.

Moreover, it has been shown in the context of object grasping that the efficiency of a model of grasping can be improved by using knowledge on the object to be grasped as priors [8, 9], i.e., the presence of a target object and its geometrical properties strongly constrain the type of grasp and the approach to the object, and, as a consequence, the brain might need to include this information when planning an appropriate course of action [10].

In this paper we set forth to investigate whether a computer, equipped with enough sensory information about human movements, i.e., grasping, could acquire a specialized model using machine learning methods. In particular we ask (i) whether the final configuration of the hand, i.e., at the very moment an object is grasped, could be predicted from the initial part of the movement and (ii) whether the knowledge of the object to be grasped could improve the model efficiency, leading to a smaller error in prediction. It is worth noting that we do not want to necessarily mimic the structure of the brain, but rather more simply analyze the human movement data with the best possible algorithm available.

To shed light on these questions, we have set up an experiment in which several able-bodied subjects have performed a highly repetitive grasping task on various daily life objects, and we have collected data about their hand position, orientation and posture. Then we have tried to put a computer in the same situation a human observer would be if they were to see only the initial part of a grasping action, the final part being occluded by a screen: a sub-sequence of each grasping sequence, i.e. the initial segment a human observer would be able to see, was used to train an efficient machine learning system based on Support Vector Machines (SVMs).

We have then analyzed the error in predicting the final hand configuration and we have analyzed whether the *a priori* knowledge of the grasped object makes a difference in performance as it should intuitively do. The results we present here, albeit still in a preliminary form, indicate that the machine is able to predict reasonably well human reaching and grasping, and that prior knowledge of the object to be grasped improves the performance of the machine, while keeping the same computational cost.

Once actually realized, optimized and implemented, such models could potentially be used in various ways including the control of semi-autonomous teleoperated/prosthetic robotic artifacts, and the interpretation and possibly mimicry of human movements [11]. For example, in controlling or teleoperating an anthropomorphic robotic platform, such models would be able to guess the user intention and ask the robot to complete the action autonomously. Predicting the user intention finds its natural role in building man–machine interfaces and possibly into the control of prosthetic devices.

The paper is structured as follows: after a brief review of related work, we describe the methods and the experimental setup in Section 2 and the results obtained in Section 3; finally we discuss them and comment on future development in Section 4.

1.1. Related work

In the monkey, premotor area F5 has been particularly well studied and it is in fact the location where 'mirror neurons' were first identified. In this respect, mirror neurons are the quintessential correlate of internal models since they are activated both when executing a specific grasping action and when observing a congruent action being executed by another individual (or the experimenter) [12].

In a study by Umiltà *et al.* [13] the response of mirror neurons to the observation of actions that terminate behind a screen has been investigated. In this case, the authors analyzed mirror neurons in situations where the final part of the trajectory of the hand was occluded by an opaque screen with the monkey knowing the presence/absence of an object to be grasped. As long as the object was shown to the monkey, the brain could easily supply the missing visual information by rehearsing the model of the action. The control experiment, in this case, was that of an identical hand kinematics, an identical screen, but the absence of the target object, i.e., identical visual stimulation apart from the knowledge of the presence of the object. Elsewhere it has been also shown that the presence of an object is required to elicit the mirror neurons response in the monkey [6].

A posteriori, given these results, it is easy to see how the presence of a target object and its geometrical properties strongly constrain the type of grasp and the approach to the object, and that, as a consequence, the brain might need to include this information when planning an appropriate course of action. In the monkey these constraints are so strong that mirror neurons do not fire unless the goal of the action is clearly perceivable. The brain codes for the object–motor identity in part via another class of F5 visuomotor neurons called 'canonical neurons' (for a discussion, see, e.g., Ref. [9]). To complete the picture, the work of Graziano *et al.* [14] has shown that the presence of objects is coded in the ventral premotor cortex and maintained even when the object is no longer visible as long as there is evidence for its presence at a particular location.

Relevant to this discussion, the work of Fogassi *et al.* [15] contributed to the identification of mirror neurons in the parietal cortex (inferior parietal lobule), which are thought to be related to the decoding of the intentions of others. Contextual information which links the enacted action to its final goal seems to be implicated in this type of neural response. The presence of objects is a clear contextual cue. In humans, it has been demonstrated that the activation of brain areas correlated to action observation is not simply a perceptual effect, but rather the activation of a precise sensorimotor model which includes, for example, the hand kinematics [16] and a precise muscular pattern of activations [17].

Accordingly, Fadiga *et al.* [18, 19] have shown that motor imagery changes the excitability of the cortico-spinal connections specifically to the imagined action, i.e., imagining a motor task causes the under-threshold activation of the same neural pathways required to execute the task. This under-threshold activation was revealed by transcranial magnetic stimulation. In a conceptually similar experiment [20], the excitability of cortico-spinal pathways was also examined as a consequence of the actual sensory input. In summary, the motor system is similarly activated when acting in first person, when imagining an action or when watching somebody else's action. Jeannerod [21], for example, goes to great lengths in showing how plausible is the fact that mental imagery uses the same internal models used by actual action generation. It is known in this respect that the time required to simulate an action is the same that is required to execute that action [22]. For a review, refer to Ref. [23].

As far as gesture/hand configuration recognition is concerned, in a previous experiment we have analyzed the problem of recognizing hand gestures visually by incorporating a generative approach that used motor information explicitly [8, 9]. In that case we showed that an action recognition system that uses motor information in a preprocessing step can perform better (97% recognition rate *versus* 80% on the test set) than a traditional classifier built directly in terms of visual information. This justifies the fact that as a pre-processing step we can consider a visuo to motor mapping that transforms the available visual information into motor data. This procedure is consistent in that it can be trained through self-observation. We can imagine that the brain can exercise its control, and simultaneously acquire both the motor commands and the corresponding visual information and learn such a mapping. In the following, we will only consider motor information since we can safely assume that the visuo-motor map can be always incorporated in the global model.

Lastly, machine learning has already been used to classify the types of grasps (e.g., Refs [24–26]); but, to the best of our knowledge, nobody has applied it in order to perform regression on the position and configuration of the hand.

2. MATERIALS AND METHODS

In this section we detail the process of gathering data from human subjects and the processing that makes them suitable for analysis by a machine learning system. In particular, we address the problem of building a training set, i.e., a set of data effectively representing, for each user and object considered, the grasping process, that could be used to train the system.

2.1. Experimental Setup

2.1.1. Devices. We collected data using a 22-sensors Immersion CyberGlove for the hand posture [27], an Ascension Flock-of-Birds (FoB) for the hand position [28] and a force resistor sensor (FSR) to detect the contact moment with the object. Figure 1 shows the devices, as worn by a subject.

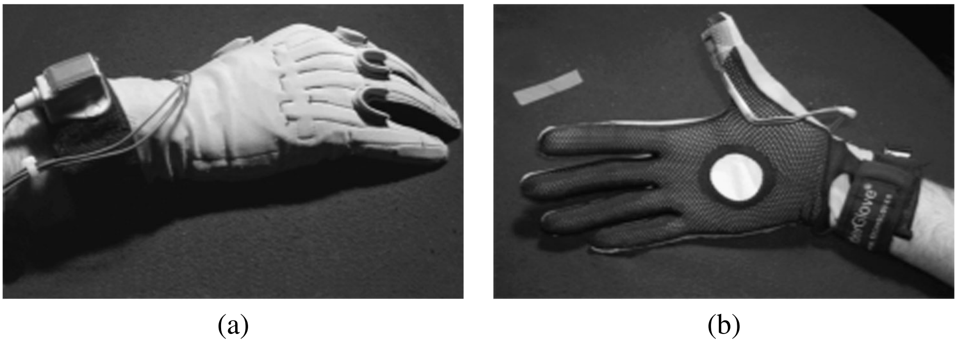


Figure 1. The devices used for the experiment, as worn by a subject. (a) The CyberGlove, with the FoB just above the subject's wrist. (b) The FRS attached to the subject's thumb.

The CyberGlove was worn by the subject on the right hand. The device returns 22 8-bit numbers linearly related to the angles between the ends of the sensors and roughly indicating the angles of the subject's hand joints; the sensors are embedded in the glove in order for them to be adherent to the subject's skin. The resolution of the sensors is on average about 0.5° [27], but the noise associated with the sensors has been experimentally determined to be 1.1 on average and 3 at the maximum [29]. The sensors describe the position of the three phalanxes of each finger (for the thumb, rotation and two phalanxes), the four finger-to-finger abductions, the palm arch, the wrist pitch and the wrist yaw.

The FoB was firmly mounted on the CyberGlove, just above the subject's wrist, with the X/Y plane being parallel to the palm plane in the resting position. The device returns six double-precision numbers describing the position (x, y and z in inches) and rotation (azimuth, elevation and roll in degrees) of the sensor with respect to a magnetic basis mounted about 1 m away from the subject. The FoB's resolution is 0.1 inch and 0.5° [28].

Finally, the FSR was mounted on the subject's thumb. It returns a 32-bit number approximately inversely proportional to the pressure applied to the surface of the sensor. We only used the FSR as an on–off indicator of when the subject made contact with the object.

All data were collected, synchronized and saved in real time at a frequency of 50 Hz.

2.1.2. Subjects. Eleven subjects, four females and seven males aged 24–34 of different nationalities, joined the experiment. They were all right-handed and fully able-bodied, and were given initially some knowledge of the aim of the experiment. They expressed their informed consent prior to their inclusion in the study.

2.1.3. Method. The subjects were asked to sit comfortably in front of a clean workspace of about 1 m^2, at the center of which an object was placed in a predefined position. The subjects were then asked to wear the devices and choose a resting position for their right hand and arm. They were then instructed to grasp the object with their right hand as they felt appropriate, not necessarily the same way each time, keeping a ‘natural’ attitude. After grasping the object, they had to drop it somewhere else in the workspace, and then return their right hand and arm in the initial resting position. Subsequently, they had to use their left hand to reposition the object roughly in the same place it was before.

We first had the subjects do a trial run of the experiment in order for them to gain confidence in the setup. A beeping sound was heard each time the subject made contact with the object (i.e., each time the FSR signaled a significant change), and they were asked to try and hear the beep each time they grasped the object. Although this ruled out grasps which made no use of the thumb, it enabled us to better determine the contact points.

After the trial run, subjects were asked to repeat the grasp/drop/reposition procedure 120 times for each object. We will call both this procedure and the data time sequence gathered during the procedure a session. We employed, in turn, three objects: a beer can, a duct tape roll and a mug (Fig. 2). The objects were chosen so that each of them could be grasped in several different ways, but with a certain degree of overlapping, e.g., both the beer can and the mug could be grasped cylindrically, but only the mug could be grasped using the handle.

Each experiment employed one subject and consisted of six sessions: first the can, then the roll and then the mug, all of them twice, for an approximate total of 720 grasps per subject, 240 per object. The numbers are not precise since now and then the subjects would grasp without properly activating the FSR. This problem has been corrected in the batch analysis of the data.

Each experiment lasted 35–56 min, depending on the subject’s confidence and speed; although almost no subjects reported tiredness, we allowed them to rest between sessions. It was reported by almost every subject that the experiment became rapidly boring, which lets us claim that almost all grasps were done in

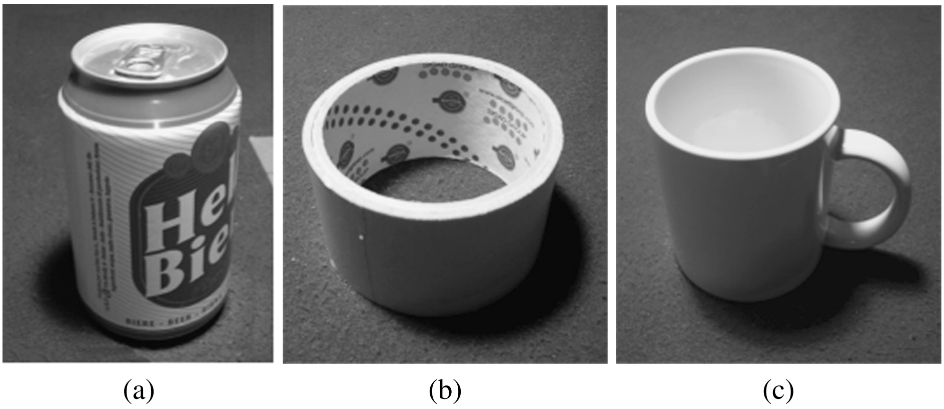

(a) (b) (c)

Figure 2. The objects used in the experiment: a beer can (a), a duct tape roll (b) and a mug (c).

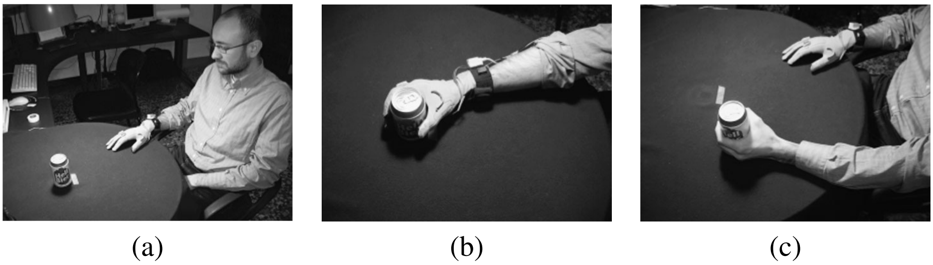


Figure 3. The experiment. The subject sits comfortably in front of a clean workspace, at the center of which an object is placed (a), with his right hand in a resting position. He then grasps the object and drops it somewhere else in the workspace (b), then bringing his arm and hand in the resting position. Lastly, he repositions the object in the initial position using his left arm and hand (c).

a natural way, almost unconsciously. Figure 3 shows the main phases of the experiment.

2.2. Building the training set

2.2.1. Detecting grasps. In order to figure out when each single grasp starts and ends in a session, we first observed the values of the FSR mounted on the subject's thumb. We manually verified that the FSR correctly reacted in almost all cases with a spike, signaling, whenever the subject made contact with the object, a significantly different value from that recorded elsewhere shortly before the contact. The spike instants were taken as the ending points of each grasp and were gathered by checking when the first derivative of the FSR value dropped by more than 10% of its overall minimum value. Moreover, after each spike, we ignored 1 s of the session to avoid detecting possible spurious spikes which happened immediately after the grasp due to object slippage and/or blurred values coming from the FSR.

Subsequently, in order to detect the starting point of each action, for each ending point we observed the hand speed and acceleration, averaged over 0.2 s, from the ending point backwards. Since we had instructed the subjects to always return to the resting position before initiating a new grasp, when the grasp starts, the speed must be close to zero and the acceleration must be negative (the subject's arm is moving toward the FoB's reference point). Therefore, we set the grasp starting point at the nearest moment in time before the ending point in which the hand speed was close to zero and the hand acceleration was negative. In order to avoid detecting spurious speed/acceleration glitches when the hand made contact with the object, we ignored 0.1 s just before the ending point; moreover, we ignored grasps which were shorter than 280 ms. All these values were determined experimentally to be near optimal in order to catch as many grasps as possible while avoiding spurious ones.

Figure 4a shows an example set of detected grasps. As one can see, the hand speed (dotted shallow curve) shows the well-known bell-shaped profile of a planar reaching movement [30]: the hand acceleration diminishes, changes sign and then goes back to zero at the end of the trajectory.

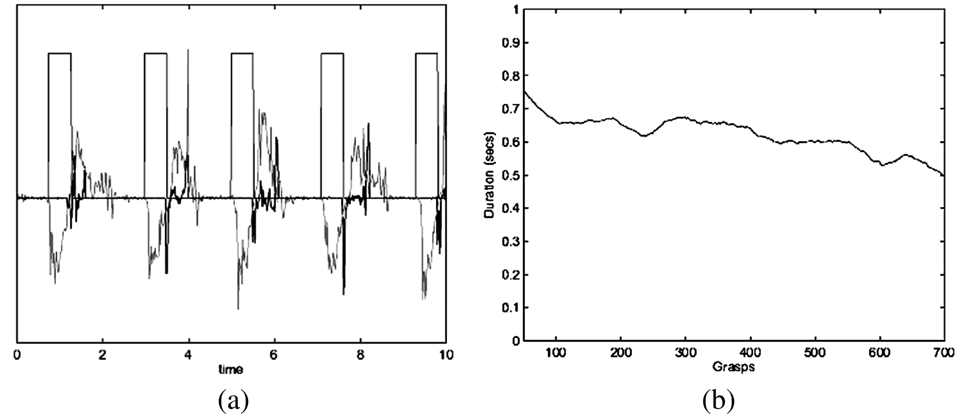


Figure 4. (a) Detecting the grasps. The figure shows 10 s of a subject grasping an object. The vertical bands indicate the start and end of a grasp; the thick, continuous line is the FSR response; and the dotted, shallow line is the hand speed. As one can see, the ending points are found near the FSR spike, indicating contact; moreover, the hand speed shows the well-known bell-shaped profile of planar reaching [30]. (b) Grasps duration. The figure shows the duration of the grasps (moving average over 50 grasps) averaged for all subjects. As the experiments advance, the duration becomes shorter.

Overall, the procedure could recognise 716 ± 12 grasps for each subject, which matches the desired result of 720, i.e., 120 per session, each user running six sessions (during two experiments, the FSR sensor broke down, resulting in the recognition of only 550 and 649 grasps). All data were also parsed by hand in order to verify that spurious detected grasps would be an insignificant fraction of the total grasps.

2.2.2. The blind window. In order to test the power of prediction of our machine, we needed to somehow hide to the system the final part of the grasping action. We therefore defined a blind window B, with $0 \leqslant B \leqslant 1$, representing what fraction of the grasp, from the contact point backwards, was hidden. Figure 5 shows a typical situation. It was intuitively expected that larger values of B would smoothly lead to larger errors.

Moreover, in general, in order to make time sequences suitable for regression, they all must have the same length so that they can be represented as vectors in a fixed-dimension input space. An alternative possibility appears, e.g., in Ref. [31]; this issue is the subject of future research. In order to accomplish this, since in general not all grasps have the same length, for each grasp we decided to stretch its visible window (i.e., a fraction of the grasp corresponding to $1 - B$) to a predefined length and it seemed reasonable to choose this length according to the average speed of the grasps in order not to lose any information.

Figure 4b shows the average grasp durations for all subjects over each experiment (moving average over 50 grasps); as one would expect, in general the subjects get rapidly used to the grasp/drop/reposition task and the grasps become faster and

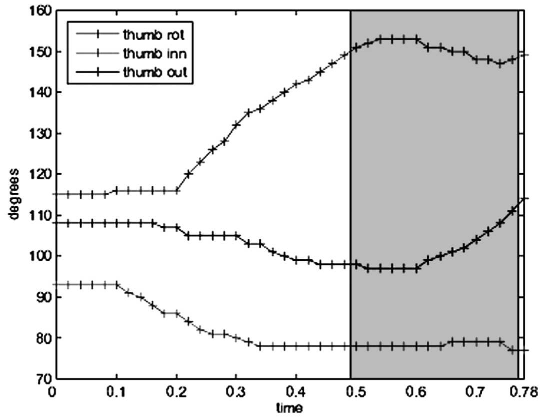


Figure 5. The blind window (the grey zone) indicates what fraction of each grasp, from the contact point backwards, is hidden to the learning machine. The data shown is a typical trajectory of the thumb (rotation, inner phalanx, outer phalanx) during a grasp. In this case the grasp lasts 0.78 s and $B = 0.375$. The last sample (for $t = 0.78$) is the target value.

faster. It must be remarked, though, that this is not the case for all subjects when considered individually. On average, the grasp duration was 0.62 ± 0.20 s. We decided then to stretch every visible window to 1 s by linear interpolation, obtaining fixed-length time sequences of 50 samples for each sensor and grasp; these time sequences were then represented as vectors in a 50-dimensional space.

2.3. SVMs

Our machine learning system is based upon SVMs in the particular variant for regression. Introduced in the early 1990s by Boser *et al.* [32], SVMs are a class of kernel-based learning algorithms deeply rooted in statistical learning theory [33], now extensively used in speech recognition, object classification and function approximation, for example, with good results [34]. We now give a very quick account of SVMs; for an extensive introduction to the subject, see, e.g., Ref. [35].

We are interested here in the problem of SVM regression, i.e. given a function whose value is known only for a finite number of points in its input domain, find its best approximation f drawn from a suitable functional space $\mathcal{F}$. In practice, let $S = \{\mathbf{x}_i, y_i\}_{i=1}^{l}$, with $\mathbf{x}_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$ be a set of l points and output values of the unknown function (the training set); then the resulting $f(\mathbf{x})$ is a sum of l elementary functions $K(\mathbf{x}, \mathbf{y})$, each one centered on a point in S, and weighted by real coefficients α_i:

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b, \tag{1}$$

where $b \in \mathbb{R}$. The choice of K, the so-called kernel, is done *a priori* and defines $\mathcal{F}$ once and for all; it is, therefore, crucial. According to a standard practice (see, e.g., Ref. [34]) we have chosen a Gaussian kernel, so that:

$$K(\mathbf{x}, \mathbf{y}) = \mathrm{e}^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\sigma^2}}, \tag{2}$$

where $\sigma \in \mathbb{R}^+$ is the standard deviation of the Gaussian function K.

Now, let $C \in \mathbb{R}$ be a positive number; then the α_is and b in (1) are found by solving the following minimisation problem (training phase):

$$\min\left(R(S, K, \boldsymbol{\alpha}) + C \sum_{i=1}^{l} L^{\epsilon}(\mathbf{x}_i, y_i, f) \right), \tag{3}$$

where R is a regularization term and L^ϵ is a loss functional. Minimizing the sum of R and L^ϵ together ensures that the solution will approximate well the values in the training set (thanks to L^ϵ), at the same time avoiding overfitting, i.e., exhibiting poor accuracy on points outside S (thanks to R). The regularization term controls the 'complexity' of $f(\mathbf{x})$. As is apparent from (3), C balances the relative importance of L^ϵ and R. In SVM regression, it is usually the case that $L^\epsilon(\mathbf{x}_i, y_i, f) = \max(0, |y_i - f(\mathbf{x}_i)| - \epsilon)$, where $\epsilon > 0$ controls the width of an 'insensitive band' around the output values, e.g., errors on the training set within this band are not considered.

In the end, there are three numbers to be tuned in our setting, called hyperparameters: C, σ and ϵ. In all our regression tests, we found the optimal values of C and σ by grid search with 5-fold cross-validation, whereas ϵ was chosen accordingly to the resolution of the sensors being examined (see next Section 3 for a more detailed discussion).

It is also usually the case that, after the training phase, some of the α_is are found to be zero; the $\mathbf{x}_i$s associated with non-zero α_is are called support vectors (SVs). Both the training time (i.e. the time required by the training phase) and the testing time (i.e., the time required to find the value of a point not in S) crucially depend on the total number of SVs; therefore, the total number of SVs is an indicator of how hard the problem is. An even better indicator is the fraction of SVs with respect to $|S|$, since in the standard SVM setting the number of SVs grows proportionally to the total number of samples in S [36].

Notice, finally, that the quantity to be minimized in (3) is convex; due to this, as well as to the use of a kernel, SVMs have the advantages that their training is guaranteed to end up in a global solution and that they can easily work in highly dimensional, non-linear feature spaces, as opposed to analogous algorithms such as, for example, artificial neural networks. Our system employs LIBSVM v2.82 [37], a standard, efficient implementation of SVMs.

According to the procedure described in the previous parts of this section, we defined $\mathbb{R}^{50}$ as the input space of our machines. We have then set up 28 such SVMs, each one approximating the value of a sensor at the time of contact.

3. RESULTS

We were mainly interested in answering two questions:

(i) How far in the future can our system predict well?

(ii) How does the knowledge of the grasped object affect the error?

In order to answer the first question, we have checked how the error on regression changes as B varies from 0.1 to 0.5. This procedure was repeated independently for each single sensor.

To obtain statistically meaningful results, we recorded each mean error obtained on a single fold of the cross-validation procedure; for each sensor and value of B, this means we have obtained five numbers. The errors for each sensor were then grouped accordingly to their measurement units and meaning: the position of the hand (three sensors, the x, y, z from the FoB), the hand orientation (three sensors, the azimuth, elevation and roll from the FoB) and the posture of the hand (22 sensors, the joint positions from the CyberGlove). According to the device resolutions (see Section 2), we set ϵ to 0.1 in. for the hand position, 0.5° for the hand orientation and 1° for the hand posture.

Finally, for each group of sensors, we averaged the errors per single cross-validaton fold, and evaluated the mean and standard deviation of the resulting five values. This gave us an indication of how well our machine performed on the hand position, orientation and posture. In all graphs, the points on the curves represent the mean values, whereas the error bars are placed at ±1 standard deviations which is common practice in machine learning.

In order to answer the second question, we first evaluated the error obtained as described above using all sessions for each single object, so to obtain an estimate of how complex it is to approximate the grasp for the can, roll and mug unbiased by the differences among the subjects. Subsequently we averaged these three errors and compared the averages with the overall error, obtained by joining all sessions together in a single training set.

3.1. Prediction power

With reference to Fig. 6, left column: first of all, as B increases, the error does, as it was intuitively expected: the more data is hidden, the harder the prediction becomes. Then, as one can see, as far as the hand position and orientation are concerned, the three objects show comparable errors. On the other hand, there is a precise ranking in the hand posture regression: the mug is more difficult than the roll, which is in turn harder than the beer can. This also is intuitively sensible, since it is possible to grasp the roll in more ways than the can and it is possible to grasp the mug in even more ways (especially using the handle).

The analysis of the obtained models (see Fig. 7, left column) shows that, accordingly, the percentage of SVs with respect to the total number of samples increases steadily as B grows, indicating that the problem becomes harder and

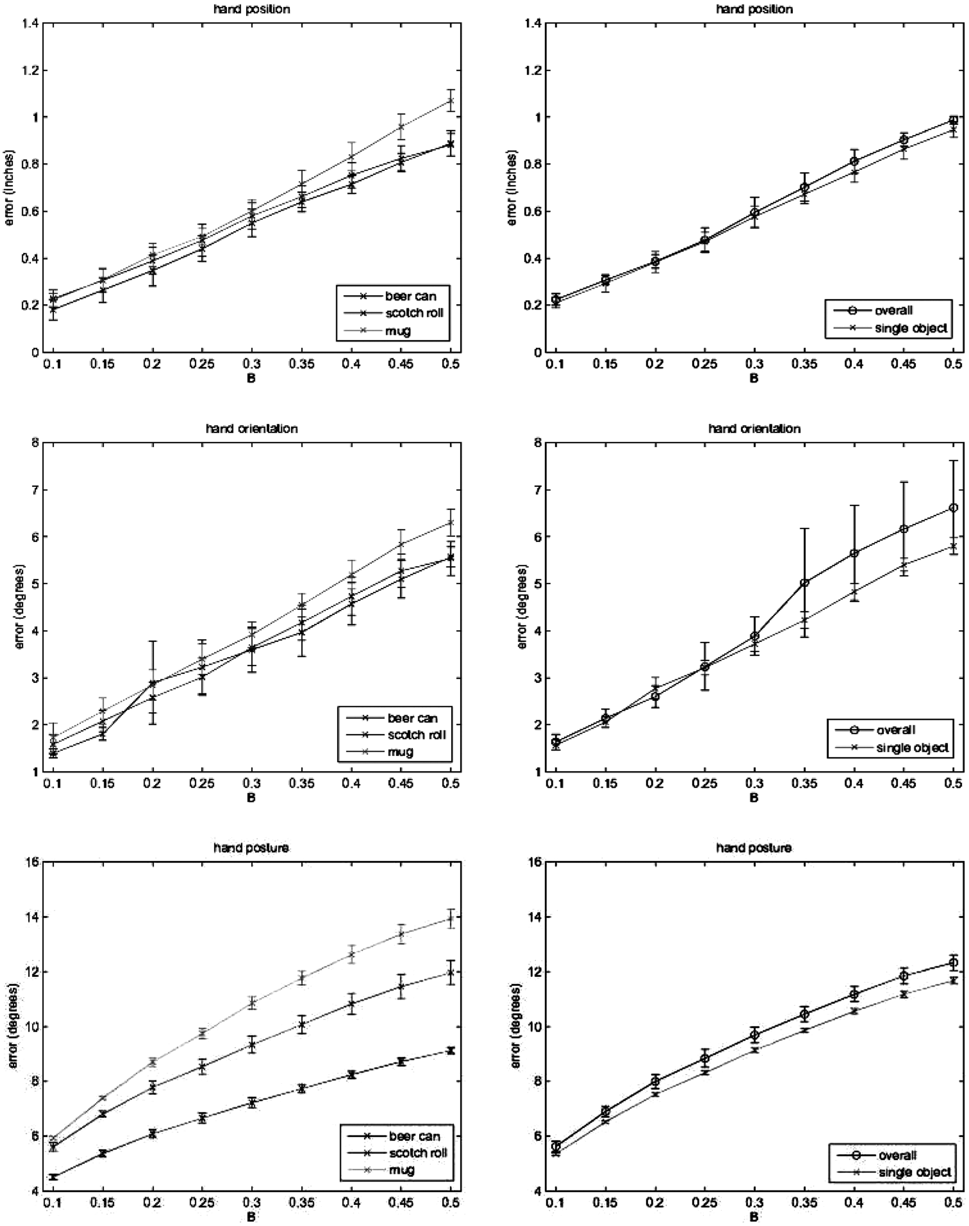


Figure 6. Regression results as the blind fraction *B* increases from 0.1 to 0.5. In each row the left-hand panels compare the errors on different objects, while the right-hand panels compare the average error on single objects and the overall error.

harder. The three curves also confirm that regression on the mug is the most difficult, followed in turn by the roll and the beer can.

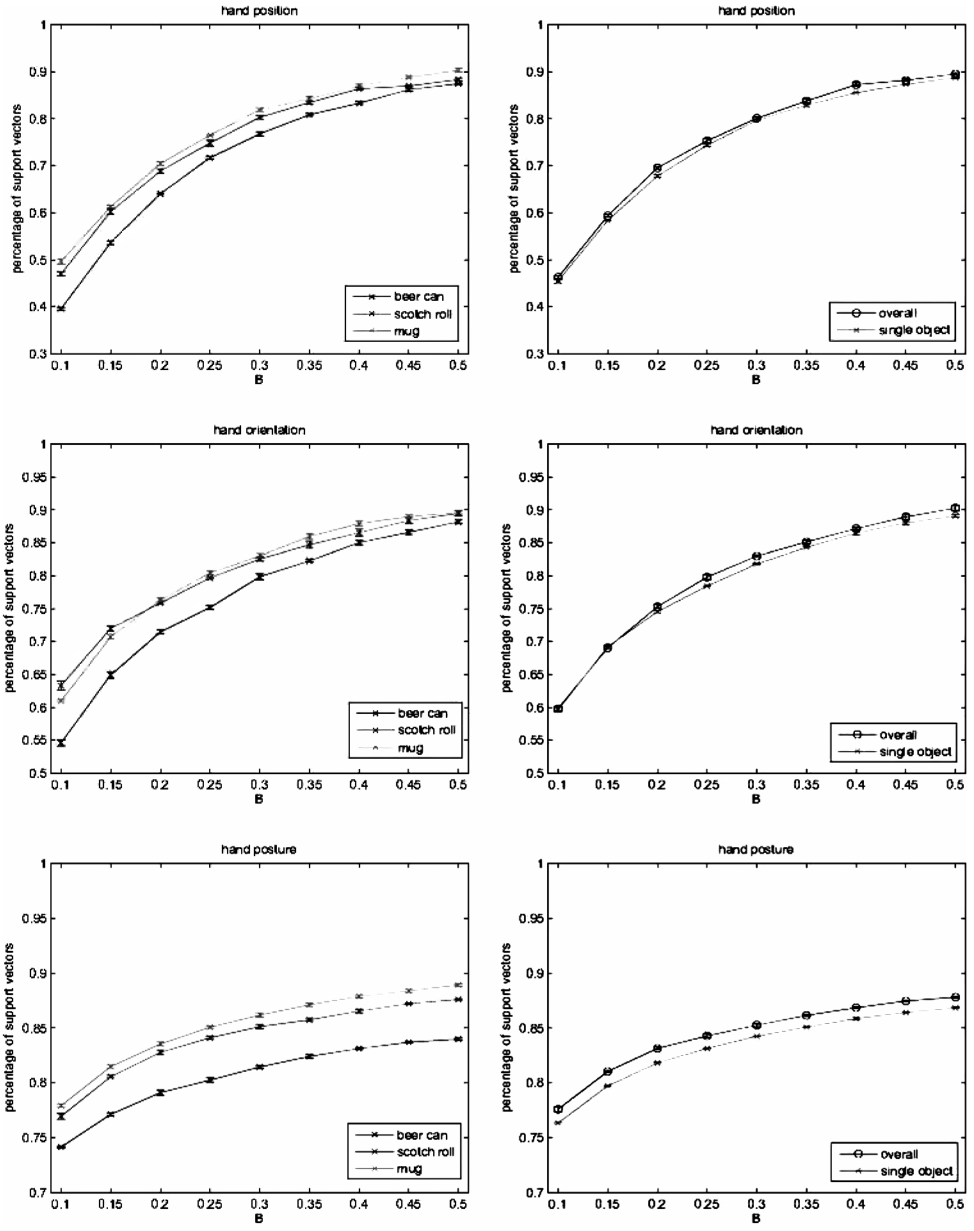


Figure 7. Percentage of SVs with respect to the total number of samples.

A further analysis of the hyperparameters (see Fig. 8) confirms that, as B increases, more and more information is missing from the training set: both C and σ show a decreasing trend on all three groups of sensors (more pronounced in the case of C), meaning that the regularization term in (3) becomes more and more

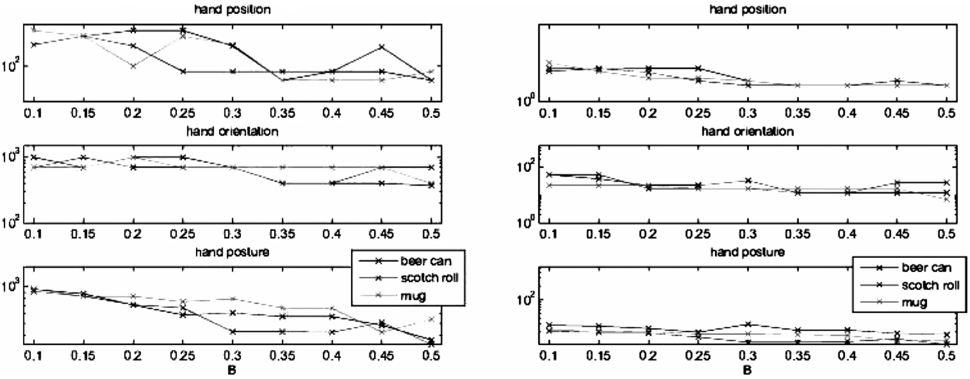


Figure 8. The trend of the hyperparameters C (left) and σ (right), as far as the hand position, orientation and posture is concerned. As B is increased, both parameters show a decreasing trend, more pronounced in the case of C.

important, and that the Gaussians used to build the solution become narrower and narrower.

To determine how far in the future SVMs can predict well, we need to decide what an acceptable error is. In general, this is application dependent. In this case we decided to accept an error as large as 5 times a minimum threshold, determined by taking into account the resolutions of the sensors as declared in the devices manuals and related publications (see Section 2). This led us to 0.5 in. for the hand position, 2.5° for the hand orientation and 7.5° for the hand posture. As far as the hand posture is concerned, it must be remarked that, in this paper, we have only considered the average of errors on all the 22 sensors, whereas in a more detailed analysis one should take into account that, for example, an error on the wrist pitch would lead to a worse displacement of the hand than an error on a phalanx would. This is a subject of future research.

As one can see from the graphs, the acceptable error is attained for the hand position at $B = 0.3$, for the hand orientation at $B = 0.2$, and for the hand posture at $B = 0.15$ (mug), 0.2 (roll) and $B = 0.35$ (beer can). Since the average grasp lasts on average 0.62 s, we can say that the system can predict reasonably well

- Something less than 200 ms in advance of the hand position.
- About 120 ms in advance of the hand orientation.
- About 100–200 ms in advance of the hand posture, the mug being the hardest and the beer can the easiest object.

This answers the first question.

3.2. Knowledge of the objects

As far as the second question is concerned, consider Fig. 6, right column: the curve representing the error on the single objects is basically always smaller than the other one, indicating that a specific SVM trained on a single object will on average be more precise than a SVM trained on all objects altogether: the *a priori* knowledge of the object improves the performance.

Notice that this effect is more pronounced in the case of the hand posture than for the hand position and orientation, where a substantial overlap of the error bars can be seen (the ANOVA test on the error indicates no significance for the hand position and orientation, but high significance for the posture, $P < 0.01$). This is sensible, since knowing what object one is going to grasp will tell a great deal about how to shape one's hand in order to grasp it, but it will definitely be less influential in order to determine where and how to reach unless there is a strong connection between the grasp type and the wrist orientation.

Let us now focus upon the fraction of SVs found by the SVMs: consider Fig. 7, right column. It turns out that SVMs trained on single objects have a definitely smaller fraction of SVs than the one trained on the overall sequence. This means that the machines trained on single objects are smaller and simpler than the overall one, while being more precise.

Summing up, we can say that if the problem is split into subproblems, each one regarding a single object, performances are better and the computational complexity is smaller. This phenomenon is particularly evident as far as the hand posture is concerned, as one can intuitively expect.

This answers the second question.

3.3. Single grasp types

A further interesting question is:

- How well does our system perform on specific types of grasps?

In order to shed light on this point, we have considered the most complex and interesting object under this point of view, i.e., the mug. All sessions in which the subjects were grasping the mug have been collected and all final grasping hand postures have been considered as representing the ways the mug was grasped. This resulted in slightly more than 2550 samples, according to the fact that each of the 11 subjects performed about 240 mug grasps.

The final positions were then clustered using a K-means clustering algorithm (see, e.g., Ref. [38]). We used the freely available Fuzzy Clustering and Data Analysis Matlab toolbox [39, 40]; due to the intrinsically stochastic nature of the algorithm, we ran the algorithm 100 times for $1, \ldots, 10$ clusters and employed Dunn's Index [41] to determine the optimal number of clusters. It was determined that the optimal number of clusters was 4.

We then split the mug grasps into four sets according to the clustering and ran separate experiments on each set for $B = 0.1, \ldots, 0.5$; the four clusters represent about one quarter each of the total mug grasps. Finally, we compared the four error and SV percentage curves to one another, and to the mug curve of Figs 6 and 7, left column, bottom graph. We hoped to get more insight on how the knowledge of what grasp is being examined changed the situation. Figure 9 shows the experimental results.

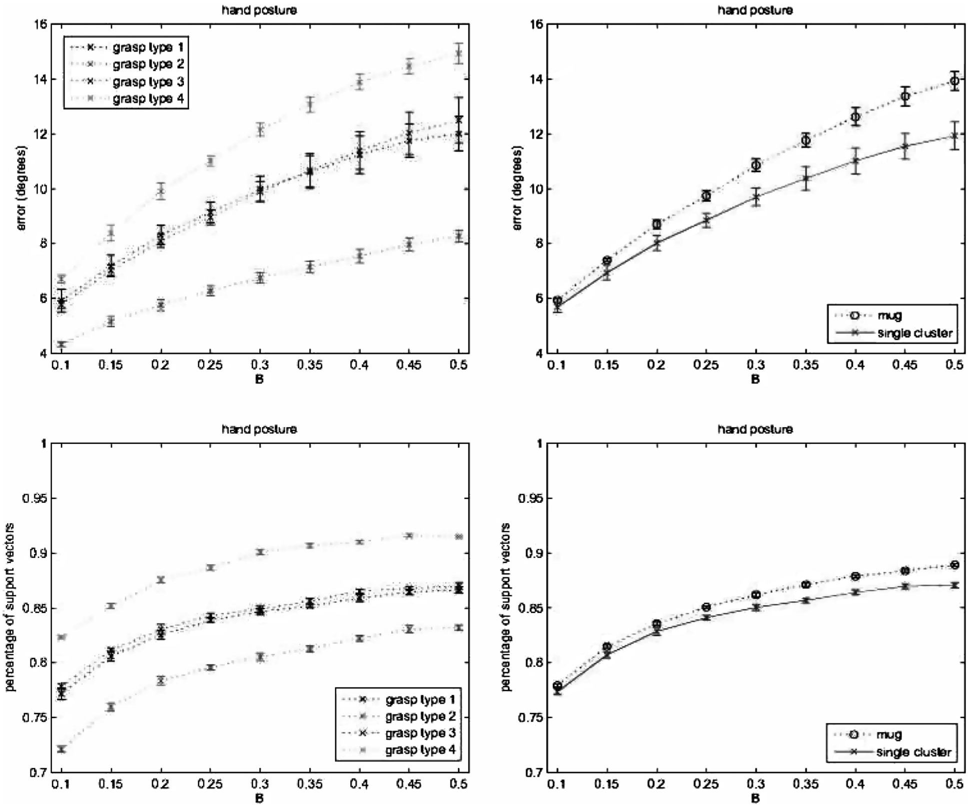


Figure 9. Comparison among regression on single grasp types (left) and between regression with/without knowledge of the grasp type (right).

First of all, it must be noticed (left column) that there is a considerably different error among the types of grasps, type 4 being the more complex, followed by 1 and 3 having the same complexity, and 2 being the easiest one. This ranking is confirmed by the percentage of SVs graph (bottom left). However, as well, the error on the single grasps is smaller than that on the mug in three cases out of four, and only slightly larger in the case of grasp type 4. This suggests that the machine is able to learn well a number of different grasp types without ‘specializing’ on one easy type.

Second, consider Fig. 9, right column, showing an analogous comparison as that shown in Fig. 6, right column: in that case we compared the errors obtained with and without prior knowledge of the object to be grasped; here we compare the errors on the mug, with and without prior knowledge of the type of grasp. The result is that knowing in advance the type of grasp makes the machine even more accurate and smaller than it used to be before.

4. DISCUSSION

With this initial experiment we really pose further questions and sketch future research rather than draw definite conclusions. The machine learning questions addressed in this paper do indeed have an answer, albeit partial; on the other hand, it remains difficult to say something other than speculations when comparing these results to neuroscience.

In short, the answer to the two questions posed in Section 2 is that we can predict well, given that we have access to motor information at least during learning and that knowing the objects to be grasped improves the ability to predict the outcome of an action. There are many caveats in this experiment, e.g., the question on whether a pre-processing of the data through clustering could improve performance further, i.e., given that objects afford certain grasping postures and they are executed with high probability. In humans the quality of the prediction of grasping is a function of the expectancies of the various possible grasp types which are in turn determined by the past experience of manipulation of the target object (Luciano Fadiga, personal communication).

The solution found by the SVMs detailed in the previous section is optimal, since the dependence from hyperparameters has been optimized out in our case by grid search and cross-validation that although expensive is known to provide good results. An analysis of the solution should thus provide an accurate characterization of the problem for the data set that has been collected.

In this sense (and only in this sense) we have shown that by partitioning the training set per object provides a general improvement of the quality of the solution and simultaneously of the training time (worst case $O(l^3)$ *versus* $O(3(l/3)^3)$ in our case with three objects and l the total number of samples). This can be an effective strategy when the world affords such an intuitive partitioning as for objects (seen as discrete entities).

This is also true from what is known about the brain structures that control grasping where the presence of a target object, its shape and affordance, and in general any contextual cue, are coded separately by different populations of neurons and influence simultaneously the response of the neurons that enact specific motor plans. After motor prediction is in place, the next step, that of recognizing the action of another individual, is conceptually simple since it amounts to building a classifier on highly predictable motor trajectories.

Another interesting question that is left to future research is whether we can investigate the complexity of the controllers of reaching and grasping (which are known to develop separately in humans) from the complexity of the learned models or as a consequence of the prediction error.

Clearly, the fact that we can train such models is prone to be applied in various contexts, as we mentioned, ranging from control of robots through interpretation and prediction of human behavior, in particular for man–machine communication.

Acknowledgments

The work presented in this paper has been supported by the European projects ROBOTCUB (IST-2004-004370), NEUROBOTICS (FP6-IST-001917) and CONTACT (NEST-005010). The authors would also like to thank Luciano Fadiga and Thierry Pozzo for the ongoing discussion on the topics of this paper, and the anonymous reviewers for many interesting and stimulating comments.

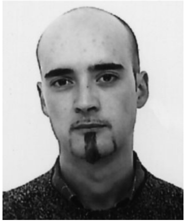
REFERENCES

1. M. Kawato, Internal models for motor control and trajectory planning, *Curr. Opin. Neurobiol.* **9**, 718–727 (1999).
2. D. M. Wolpert, K. Doya and M. Kawato, A unifying computational framework for motor control and social interaction, *Phil. Trans. R. Soc. B Biol. Sci.* **358**, 593–602 (2003).
3. F. A. Mussa-Ivaldi and E. Bizzi, Motor learning through the combination of primitives, *Phil. Trans. R. Soc. Biol. Sci.* **355**, 1755–1769 (2000).
4. J. R. Lackner and P. DiZio, Adaptation in a rotating artificial gravity environment, *Brain Res. Rev.* **28**, 194–202 (1998).
5. G. Rizzolatti and L. Craighero, The mirror-neuron system, *Ann. Rev. Neurosci.* **27**, 169–192 (2004).
6. V. Gallese, L. Fadiga, L. Fogassi and G. Rizzolatti, Action recognition in the premotor cortex, *Brain* **119**, 593–609 (1996).
7. G. Rizzolatti and G. Luppino, The cortical motor system, *Neuron* **31**, 889–901 (2001).
8. M. Lopes and J. Santos-Victor, Visual learning by imitation with motor representations, *IEEE Trans. Syst. Man Cybernet. B Cybernet.* **35**, 438–449 (2005).
9. G. Metta, G. Sandini, L. Natale, L. Craighero and L. Fadiga, Understanding mirror neurons: a bio-robotic approach, *Interact. Studies* **7**, 197–232 (2006).
10. H. Sakata, M. Taira, A. Murata and S. Mine, Neural mechanisms of visual guidance of hand actions in the parietal cortex of the monkey, *Cerebral Cortex* **5**, 429–438 (1995).
11. D. M. Wolpert, Z. Ghahramani and R. J. Flanagan, Perspectives and problems in motor learning, *Trends Cognitive Sci.* **5**, 487–494 (2001).
12. L. Fadiga, L. Fogassi, V. Gallese and G. Rizzolatti, Visuomotor neurons: ambiguity of the discharge or 'motor' perception?, *Int. J. Psychophysiol.* **35**, 165–177 (2000).
13. M. A. Umiltá, E. Kohler, V. Gallese, L. Fogassi, L. Fadiga, C. Keysers and G. Rizzolatti, I know what you are doing: a neurophysiological study, *Neuron* **31**, 1–20 (2001).
14. M. S. A. Graziano, X. Hu and C. G. Gross, Coding the location of objects in the dark, *Science* **277**, 239–241 (1997).
15. L. Fogassi, P. F. Ferrari, B. Gesierich, S. Rozzi, F. Chersi and G. Rizzolatti, Parietal lobe: from action organization to intention understanding, *Science* **308**, 662–667 (2005).
16. T. Pozzo, C. Papaxanthis, J. L. Petit, N. Schweighofer and N. Stucchi, Kinematic features of movement tunes perception and action coupling, *Behav. Brain Res.* **169**, 75–82 (2006).
17. P. Borroni, M. Montagna, G. Cerri and F. Baldissera, Cyclic time course of motor excitability modulation during the observation of a cyclic hand movement, *Brain Res.* **1065**, 115–124 (2005).
18. L. Fadiga, G. Buccino, L. Craighero, L. Fogassi, V. Gallese and G. Pavesi, Corticospinal excitability is specifically modulated by motor imagery: a magnetic stimulation study, *Neuropsychologia* **37**, 147–158 (1999).
19. C. D. Vargas, E. Olivier, L. Craighero, L. Fadiga, J. R. Duhamel and A. Sirigu, The influence of hand posture on corticospinal excitability during motor imagery: a transcranial magnetic stimulation study, *Cerebral Cortex* **14**, 1200–1206 (2004).

20. L. Fadiga, L. Craighero and E. Olivier, Human motor cortex excitability during the perception of others' action, *Curr. Biol.* **14**, 331–333 (2005).
21. M. Jeannerod, *The Neural and Behavioural Organization of Goal-Directed Movements* (*Vol. 15*). Clarendon Press, Oxford (1988).
22. A. Sirigu, J.-R. Duhamel, L. Cohen, B. Pillon, N. Dubois and Y. Agid, The mental representation of hand movements after parietal cortex damage, *Science* **273**, 1564–1156 (1996).
23. M. Jeannerod and V. Frak, Mental imaging of motor activity in humans, *Curr. Opin. Neurobiol.* **9**, 735–739 (1999).
24. C. de Granville, J. Southerland and A. H. Fagg, Learning grasp affordances through human demonstration, in: *Proc. Int. Conf. on Development and Learning*, Bloomington, IN, (CD-ROM) (2006).
25. G. Heumer, H. Ben Amor, M. Weber and B. Jung, Grasp recognition with uncalibrated data gloves—a comparison of classification methods, in: *Proc. IEEE Virtual Reality Conf.*, Charlotte, NC, pp. 19–26 (2007).
26. S. Ekvall and D. Kragić, Grasp recognition for programming by demonstration, in: *Proc. IEEE/RSJ Int. Conf. on Robotics and Automation*, Barcelona, pp. 748–753 (2005).
27. Virtual Technologies, *CyberGlove Reference Manual*. Virtual Technologies, Palo Alto, CA (1998).
28. Ascension Technology, *The Flock of Birds—Installation and Operation Guide*, Ascension Technology, Burlington, VT (1999).
29. G. D. Kessler, L. F. Hodges and N. Walker, Evaluation of the cyberglove as a whole-hand input device, *ACM Trans. Comput.–Hum. Interact.* **2**, 263–283 (1995).
30. P. Morasso, Spatial control of arm movements, *Exp. Brain Res.* **42**, 223–227 (1981).
31. H. Shimodaira, K. Noma, M. Nakai and S. Sagayama, Dynamic time-alignment kernel in support vector machine, in: *Proceedings of Neural Information Processing Systems 14* (T. G. Dietterich, S. Becker and Z. Ghahramani, Eds), pp. 921–928. MIT Press, Cambridge, MA (2002).
32. B. E. Boser, I. M. Guyon and V. N. Vapnik, A training algorithm for optimal margin classifiers, in: *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory* (D. Haussler, Ed.), pp. 144–152. ACM Press, Washington, DC (1992).
33. V. N. Vapnik, *Statistical Learning Theory*. Wiley, New York, NY (1998).
34. N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods)*, Cambridge University Press, Cambridge (2000).
35. A. J. Smola and B. Schölkopf, A tutorial on support vector regression, *Statistics Comput.* **14**, 199–222 (2004).
36. I. Steinwart, Sparseness of support vector machines, *J. Machine Learn. Res.* **4**, 1071–1105 (2003).
37. C.-C. Chang and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines* (2001), Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
38. J. B. MacQueen, Some methods for classification and analysis of multivariate observations, in: *Proc. 5th Symp. on Mathematical Statistics and Probability*, Berkeley, CA, vol. 1, pp. 281–297 (1967).
39. J. Abonyi and F. Szeifert, Supervised fuzzy clustering for the identification of fuzzy classifiers, *Pattern Recognit. Lett.* **24**, 2195–2207 (2003).
40. B. Balasko, J. Abonyi and B. Feil, *Fuzzy Clustering and Data Analysis Toolbox*, Freely available Matlab package. See http://www.fmt.vein.hu/softcomp/fclusttoolbox
41. J. C. Dunn, A fuzzy relative of the isodata process and its use in detecting compact well separated clusters, *J. Cybernet.* **3**, 32–57 (1973).

ABOUT THE AUTHORS

Claudio Castellini is a Postdoctoral Researcher at LIRA-Lab, University of Genova, Italy. After graduating in Electronic Engineering, University of Genova, 1998, he obtained a PhD in Artificial Intelligence in 2005 at the University of Edinburgh, UK, supervised by Dr Alan Smaill. He currently works on machine learning, neuroscience and developmental robotics. He is involved in two European FET projects. His PhD thesis has received the AIIA prize for best Italian dissertation in 2005.



Francesco Orabona received the MS degree in Electronic Engineering at the University of Naples 'Federico II', in 2003, and the PhD degree at the University of Genoa, in 2007. He is currently a Visiting Researcher at the IDIAP Research Institute. His research interests include active vision in particular visual attention modeling, machine learning and information theory.

Giorgio Metta is Assistant Professor at the University of Genoa where he teaches the courses of 'Anthropomorphic Robotics' and 'Operating Systems' and Research Scientist at the Italian Institute of Technology (IIT) in Genoa. He received his PhD in Electronic Engineering in 2000. He was a Postdoctoral Associate at MIT, AI-Lab from 2001 to 2002. Since 1993 he has been at LIRA-Lab where he has developed various robotic platforms with the aim of implementing bioinspired models of sensorimotor control.

Giulio Sandini is Director of Research at the Italian Institute of Technology (IIT) and Full Professor of Bioengineering at the University of Genoa. His main research interests are in the fields of computational and cognitive neuroscience and robotics with the objective of understanding the neural mechanisms of human sensorimotor coordination and cognitive development from a biological and an artificial perspective. He graduated in Electronic Engineeering (Bioengineering) at the University of Genova and was a Research Fellow and Assistant Professor at the Scuola Normale Superiore in Pisa until 1984 working at the Laboratorio di Neurofisiologia of the CNR. He has been a Visiting Research Associate at the Department of Neurology of the Harvard Medical School and Visiting Scientist at the Artificial Intelligence lab at MIT. Since 2006 he has been Director of Research at the IIT where he leads the Department of Robotics, Brain and Cognitive Science.