Internal models of reaching and grasping *

Claudio Castellini[1] Francesco Orabona[1] Giorgio Metta[1,2] Giulio Sandini[2]

[1] *LIRA-Lab, University of Genoa, Italy*

[2] *Italian Institute of Technology, Genoa, Italy*

contact: Giorgio Metta, email: pasa@liralab.it

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 behaviours, and in general to plan actions well in advance. Neuroscience has started examining the neural basis of these skills with behavioural 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 behaviours 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 behaviours. 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 behaviours, 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.

*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 also Luciano Fadiga and Thierry Pozzo for the ongoing discussion on the topics of this paper.

These models are sometimes called "internal models", meaning that they are the internal rehearsal (or simulation) of the world enacted by the brain.

Interestingly, internal models are built not only because they are required to control movements, but also, as it has been determined more recently, to interpret the movements of others [1, 2, 3, 4]. There is now a large body of literature that links the observation of actions to action execution, like for example the study of the motor system conducted by Rizzolatti and colleagues in relatively recent years [5, 6, 7]. It seems then that building internal models of actions is a key feature of intelligent living systems.

Moreover, it has been shown in the context of object grasping, that the efficiency of an internal model can be improved by priors on the object to be grasped, 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.

In this paper we set forth to investigate whether a computer, equipped with enough sensory information about human movements, namely grasping, could acquire something like an internal model using machine learning methods. In particular we ask (a) 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 (b) whether the knowledge of the object to be grasped could improve the model efficiency, leading to a smaller error in prediction.

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 he 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, namely 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.

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 realised, optimised and implemented, such internal models could potentially be used in various ways including the control of semi-autonomous teleoperated / prosthetic robotic artifacts, the interpretation and possibly mimicry of human movements [8]. 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; lastly 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) [9].

In a study by Umiltà et al. [10] 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 action was occluded by an opaque screen with the monkey knowing of 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 internal 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, that is, 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 for example [11]). To complete the picture, the work of Graziano, Hu, and Gross [12] 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. [13] 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 [14].

Accordingly, Fadiga et al. [15, 16] have shown that motor imagery changes the excitability of the cortico-spinal connections specifically to the imagined action, that is, 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 [17], 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 [18], for example, goes to a great length 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 [19]. For a review refer to [20].

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 [21, 11]. 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 preprocessing 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 always incorporated in the global internal model.

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*, that is, 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

Devices

We collected data using a 22-sensors Immersion CyberGlove for the hand posture [22], an Ascension Flock-Of-Birds (FoB) for the hand position [23] and a Force Resistor Sensor (FSR) to detect the contact moment with the object. Figure 1 shows the devices, as worn by a subject.

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 between 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 degree [22], but the noise associated with the sensors has been experimentally determined to be 1.1 on average and 3 at the maximum [24]. 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

(a) (b)

Figure 1: The devices used for the experiment, as worn by a subject: (a) the CyberGlove, with the Flock-of-Birds just above the subject's wrist; (b) the Force Resistor Sensor attached to the subject's thumb.

being parallel to the palm plane in the resting position. The device returns 6 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 one meter away from the subject. The FoB's resolution is 0.1 inches and 0.5 degrees [23].

Lastly, 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, synchronised, and saved in real time at a frequency of 50Hz.

Subjects

Eleven subjects, four females and seven males aged 24 to 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.

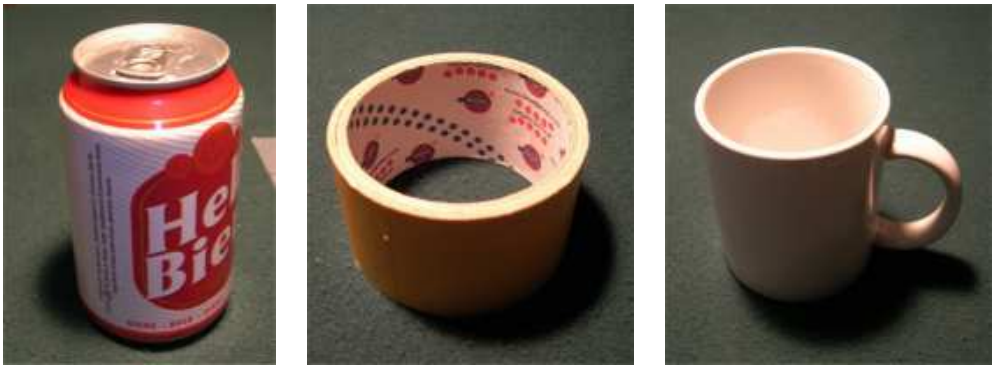
Method

The subjects were asked to sit comfortably in front of a clean workspace of about one square meter, 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 (that is, each

time the FSR signalled 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 (Figure 2 shows the objects). 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.



(a) (b) (c)

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

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 to 56 minutes 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 natural, almost unconsciously. Figure 3 shows the main phases of the experiment.

2.2 Building the training set

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, signalling, 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

(a) (b) (c)

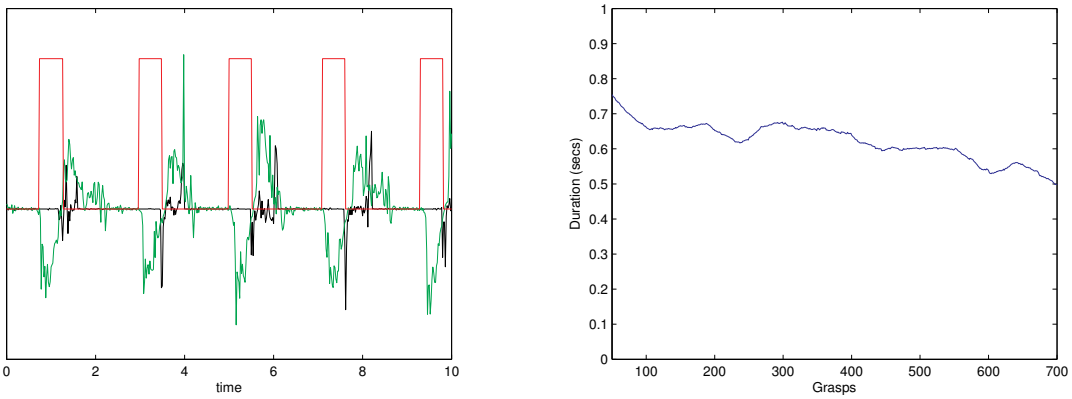
Figure 3: The experiment. The subject sits confortably in front of a clean workspace, at the center of which an object is placed (a), with his right hand in a resting position. He then grasps the object and drops it somewhere else in the workspace (b), bringing then his arm and hand in the resting position.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

value dropped by more than 10% of its overall minimum value. Moreover, after each spike, we ignored one second 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 seconds, 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 seconds just before the ending point; moreover, we ignored grasps which resulted shorter than 280 milliseconds. All these values were determined experimentally to be near optimal in order to catch as many grasps as possible while avoiding spurious ones.

Figure 4 (a) shows an example set of detected grasps. As one can see, the hand speed (green curve) shows the well known bell-shaped profile of a planar reaching movement [25]: the hand acceleration diminishes, changes sign and then goes back to zero at the end of the trajectory.

Overall, the procedure could recognise 716 ± 12 grasps for each subject, which matches the desired result of 720, that is 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.



(a) (b)

Figure 4: (a) Detecting the grasps. The Figure shows 10 seconds of a subject grasping an object. The red bands indicate the start and end of a grasp; the black line is the FSR response; and the green 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. (b) Grasps duration. The Figure shows the duraton of the grasps (moving average over 50 grasps) averaged for all subjects. As the experiments advance, the duration becomes shorter.

Grasping speed

In general, in order to make time sequences suitable for a machine learning system, they all must have the same length;[1] in order to do this, since in general not all grasps have the same length, we stretched each sequence to a predefined length. The predefined length was chosen according to the average speed of the grasps. Figure 4 (b) 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 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 seconds. We decided then to stretch every grasp to 1 second by linear interpolation, obtaining fixed-length time sequences of 50 samples for each sensor and grasp.

2.3 Support Vector Machines

Our machine learning system is based upon Support Vector Machines (SVMs). Introduced in the early 90s by Boser, Guyon and Vapnik [27], SVMs are a class of kernel-based learning algorithms deeply rooted in Statistical Learning Theory [28], now extensively used in, e.g., speech recognition, object classification and function approximation with good results [29]. We now give a very quick account of SVMs; for an extensive introduction to the subject, see, e.g., [30].

[1] An alternative possibility appears, e.g., in [26]; this issue is the subject of future research.

We are interested here in the problem of SVM regression, that is: 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 (actually, 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., [29]) we have chosen a *Gaussian* kernel, which has one positive parameter $\sigma \in \mathbb{R}$ which is the standard deviation of the Gaussian functions used to build (1).

Let $C \in \mathbb{R}$ be a positive parameter; then the α_is and b are found by solving the following minimisation problem (*training phase*):

$$\min \left(R(S, K, \boldsymbol{\alpha}) + C \sum_{i=1}^{l} L^{\epsilon}(\mathbf{x}_i, y_i, f) \right) \tag{2}$$

where R is a *regularisation term* and L^{ϵ} is a *loss functional*. In practice, after the training phase, some of the α_is will be zero; the $\mathbf{x}_i$s associated with non-zero α_is are called *support vectors*. 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 support vectors; therefore, this number is an indicator of how hard the problem is.

In (2), minimising the sum of R and L^{ϵ} together ensures that the solution will approximate well the values in the training set, at the same time avoiding overfitting, i.e., exhibiting poor accuracy on points outside S. Smaller values of the parameter C give more importance to the regularisation term and vice-versa.

Moreover, in SVM regression, $L^{\epsilon}(\mathbf{x}_i, y_i, f) = \max(0, |y_i - f(\mathbf{x}_i)| - \epsilon)$, 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.

There are, therefore, three parameters to be tuned in our setting: C, σ and ϵ. In all our regression tests, we found the optimal values of C and σ by grid search with 2-fold cross-validation, whereas ϵ was chosen accordingly to the resolution of the sensors being examined (see next Section for a more detailed discussion).

Notice, lastly, that the quantity to be minimised in Equation (2) 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, e.g., artificial neural networks. Our system employs LIBSVM v2.82 [31], a standard, efficient implementation of SVMs.

According to the procedure described in the previous parts of this Section, we decided to define $\mathbb{R}^{50}$ as the input space of our machines, and to use, for each sensor in the setup, the stretched grasping sequences as points; the target values would then be the values of the same sensor at the time of contact with the object. This way we have obtained 28 SVMs, each one approximating the value of a sensor at the time of contact.

3 RESULTS

We were mainly interested in answering two questions:

1. how far in the future can our system predict well?
2. 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 the *blind fraction* $0 \leq B \leq 1$ of the grasp increases from 0.1 to 0.5. The blind fraction indicates what percentage of the grasp, from the contact point backwards, is hidden to the system (Figure 5 shows a typical situation). It was intuitively expected that larger values of B would smoothly lead to larger errors.

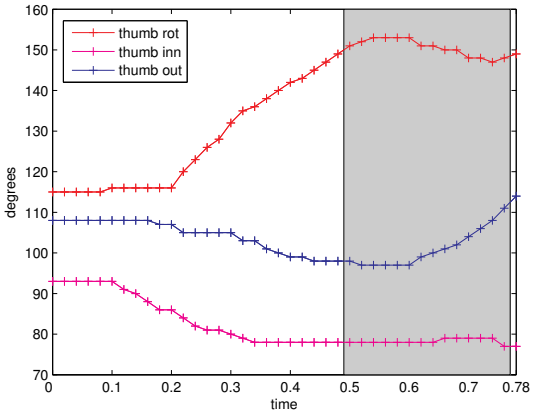


Figure 5: The blind window (the grey zone in the Figure) indicates what fraction of each grasp, from the contact point backwards, is hidden to the SVM. The data shown is a typical trajectory of the thumb (rotation, inner phalanx, outer phalanx) during a grasp. In this case the grasp lasts 0.78 seconds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

This procedure was repeated independently for each single sensor; the regression error obtained is the error attained by 2-fold cross validation with the optimal C and σ, determined using the procedure detailed in the previous Section. The errors for each sensor were then grouped and averaged accordingly to their measurement unit and meaning: the position of the hand (3 sensors, the x, y, z from the FoB), the hand orientation (3 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 the previous Section), we set ϵ to 0.1 inches for the hand position, 0.5 degrees for the hand orientation and 1 degree for the hand posture.

In order to answer the second question, we first compared 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 average with the overall error, obtained by joining *all* sessions together in a single training set. Figure 6 shows the experimental results.

Consider Figure 6, left column: as one can see, as far as the hand position and orientation are concerned, the three objects show a comparable error. On the other hand, there is a precise ranking in the hand posture regression: the mug is more difficult than the scotch roll, which is in turn harder than the beer can. This is intuitively sensible, since it is possible to grasp the scotch roll in more ways than the can, and it is possible to grasp the mug in even more ways (especially, using the handle).

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 the previous Section). This lead us to 0.5 inches for the hand position, 2.5 degrees for the hand orientation, and 7.5 degrees 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, e.g., an error on the wrist pitch would lead to a worse displacement of the hand, than an error on a phalanx would. This is 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 and scotch roll) and $B = 0.3$ (beer can). Since the average grasp lasts on average 0.62 seconds, we can say that the system can predict reasonably well

- something less than 200 milliseconds in advance the hand position,
- about 120 milliseconds in advance the hand orientation, and
- about 90 milliseconds in advance the hand posture while grasping the mug or the scotch roll, and about 200 milliseconds in advance the hand posture while grasping the beer can.

This answers the first question.

As far as the second question is concerned, consider Figure 6, right column: the curve representing the error on the single objects is always consistently 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. A further analysis of C (see Figure 7), indicates that the optimal values found for C show a decreasing trend. This correctly suggests

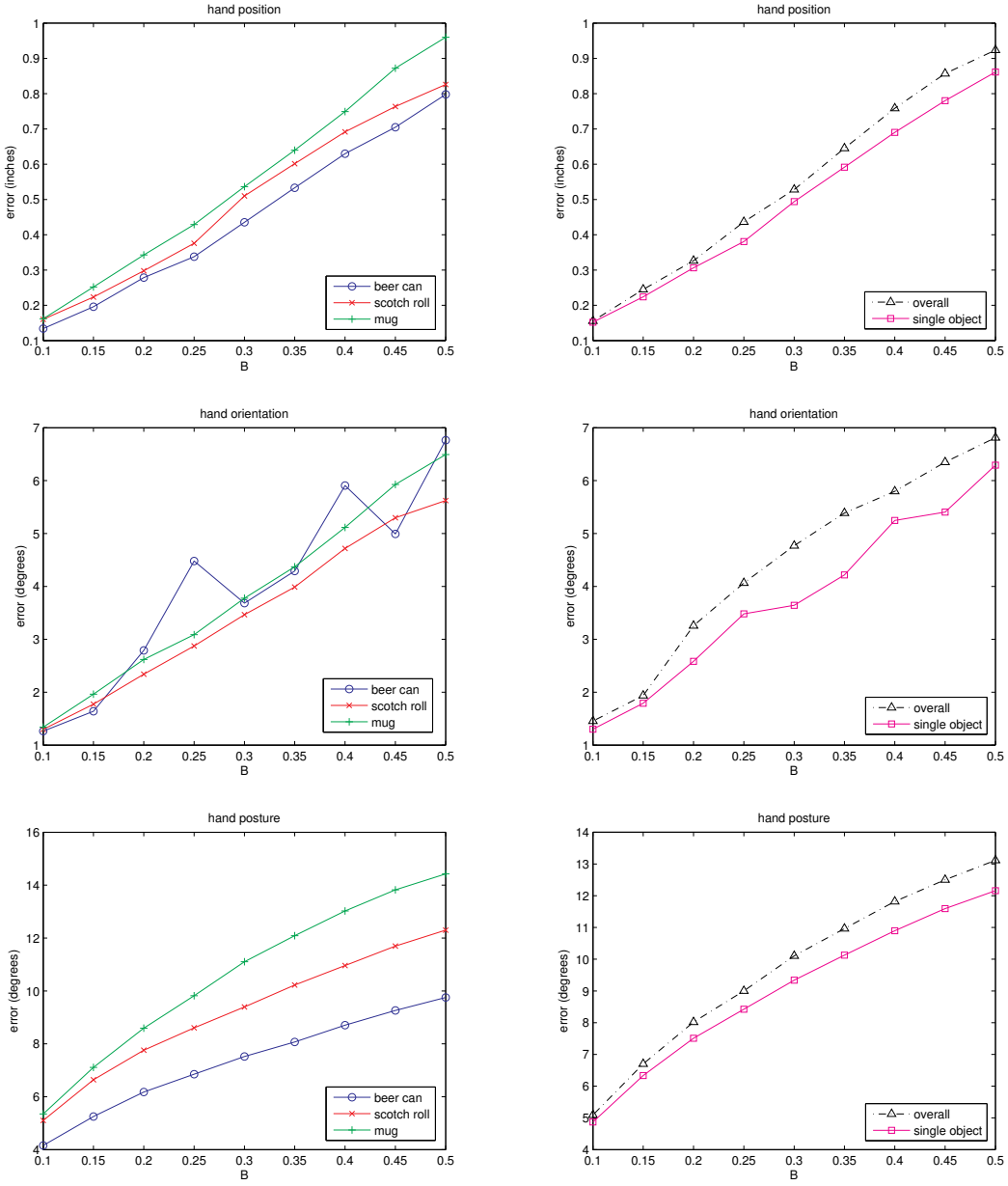


Figure 6: Regression results as the blind fraction B increases from 0.1 to 0.5. In each row, representing a different set of sensors (in turn, hand position, orientation and posture), the left-hand side pictures compare the errors on different objects, while the right-hand side pictures compare the average error on single objects and the overall error.

that, as B increases, more and more information is missing from the training set (the regularisation term in Equation (2) becomes increasingly important).

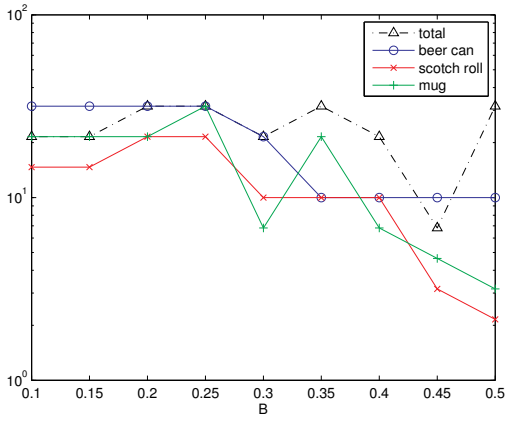


Figure 7: The trend of the hyperparameter C as B is increased. C has a definite decreasing trend.

Let us now focus upon the number of support vectors found by the SVMs. It turns out that SVMs trained on single objects have roughly the same number of support vectors as the one trained on the overall sequence (figures show about 1% more in the first case). This means that the single-object problems are computationally as hard as the overall one, but, as we have seen, they are 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 total computational complexity of the solution remains roughly the same. This answers the second question.

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, as for example, 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[2].

[2]personal communication with Luciano Fadiga.

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 qua 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 $O(3 \cdot (l/3)^3)$ in our case with 3 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 internal models or as a consequence of the prediction error.

Clearly, the fact that we can train such internal 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.

REFERENCES

[1] M. Kawato. Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, 9:718–727, 1999.

[2] D.M. Wolpert, K. Doya, and M. Kawato. A unifying computational framework for motor control and social interaction. *Philosophical Transaction of the Royal Society: series B, Biological Sciences*, 358:593–602, 2003.

[3] F.A. Mussa-Ivaldi and E. Bizzi. Motor learning through the combination of primitives. *Philosophical Transaction of the Royal Society: Biological Sciences*, 355:1755–1769, 2000.

[4] J.R. Lackner and P. DiZio. Adaptation in a rotating artificial gravity environment. *Brain Research Reviews*, 28:194–202, 1998.

[5] G. Rizzolatti and L. Craighero. The mirror-neuron system. *Annual Review of Neuroscience*, 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] D.M. Wolpert, Z. Ghahramani, and R.J. Flanagan. Perspectives and problems in motor learning. *Trends in Cognitive Sciences*, 5:487–494, 2001.

[9] L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti. Visuomotor neurons: ambiguity of the discharge or 'motor' perception? *Internation Journal of Psychophysiology*, 35:165–177, 2000.

[10] 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.

[11] G. Metta, G. Sandini, L. Natale, L. Craighero, and L. Fadiga. Understanding mirror neurons: a bio-robotic approach. *Interaction Studies*, 7:197–232, 2006.

[12] M.S.A. Graziano, X. Hu, and C.G. Gross. Coding the location of objects in the dark. *Science*, 277:239–241, 1997.

[13] 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.

[14] T. Pozzo, C. Papaxanthis, J.L. Petit, N. Schweighofer, and N. Stucchi. Kinematic features of movement tunes perception and action coupling. *Behavioural Brain Research*, 169:75–82, 2006.

[15] 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.

[16] 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.

[17] L. Fadiga, L. Craighero, and E. Olivier. Human motor cortex excitability during the perception of others' action. *Current Biology*, 14:331–333, 2005.

[18] M. Jeannerod. *The Neural and Behavioural Organization of Goal-Directed Movements*, volume 15. Clarendon Press, Oxford, 1988.

[19] 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.

[20] M. Jeannerod and V. Frak. Mental imaging of motor activity in humans. *Current Opinion in Neurobiology*, 9:735–739, 1999.

[21] M. Lopes and J. Santos-Victor. Visual learning by imitation with motor representations. *IEEE Transactions on Systems, Man, and Cybernetics, Part B Cybernetics*, 35:438–449, 2005.

[22] Virtual Technologies, Inc., 2175 Park Blvd., Palo Alto (CA), USA. *CyberGlove Reference Manual*, August 1998.

[23] Ascension Technology Corporation, PO Box 527, Burlington (VT), USA. *The Flock of Birds — Installation and operation guide*, January 1999.

[24] G. Drew Kessler, Larry F. Hodges, and Neff Walker. Evaluation of the cyberglove as a whole-hand input device. *ACM Trans. Comput.-Hum. Interact.*, 2(4):263–283, 1995.

[25] P. Morasso. Spatial control of arm movements. *Experimental Brain Research*, 42:223–227, 1981.

[26] H. Shimodaira, K. Noma, M. Nakai, and S. Sagayama. Dynamic time-alignment kernel in support vector machine. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Proceedings of Neural Information Processing Systems 14*. MIT press, 2002.

[27] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM press, 1992.

[28] Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.

[29] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines (and Other Kernel-Based Learning Methods)*. CUP, 2000.

[30] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.

[31] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.