Using Object Affordances to Improve Object Recognition

C. Castellini, T. Tommasi, N. Noceti, F. Odone, B. Caputo

***Abstract*—The problem of object recognition has not yet been solved in its general form. The most successful approach to it so far relies on object models obtained by training a statistical method on visual features obtained from camera images. The images must necessarily come from huge visual datasets, in order to circumvent all problems related to changing illumination, point of view, etc.**

We hereby propose to also consider, in an object model, a simple model of how a human being would grasp that object (its affordance). This knowledge is represented as a function mapping visual features of an object to the kinematic features of a hand while grasping it. The function is practically enforced via regression on a human grasping database.

After describing the database (which is publicly available) and the proposed method, we experimentally evaluate it, showing that a standard object classifier working on both sets of features (visual and motor) has a significantly better recognition rate than that of a visual-only classifier.

***Index Terms*—robot tactile systems, robot vision systems, learning systems, biologically inspired feature extraction**

I. Introduction

CONSIDER the objects in Figure 1. How do we know that they are all cups? The answer is intuitive: they can all be used to contain liquids and to drink, and have actually been designed to this end. Although very little in their visual appearance ties them together, a human being will immediately know what can be done with such objects since *she has done it* at some time in the past. As a matter of fact, the category of an object is often determined more by its *function* rather than by its visual appearance; this idea has led Gibson in the 70s [1], [2] to define objects in terms of their *affordances* — "what can be done with them". It is probably this key intuition that makes human object recognition so robust.

This idea, we believe, could be profitably used to solve the general problem of mechanical object recognition. Consider the above cups: traditionally, an object recognition system would be trained on a very large database of images of very diverse cups, shot in different conditions of illumination, from different points of view, etc. This is clearly incomplete and resource-consuming [3]. But what if the system had an idea of how to *grasp* something which looks like a cup? In that case a new, strong "semantic" element could be used to tie together the three objects above in the category of cups. The use of object affordances to improve the classic solution to object recognition, in which visual features only are exploited, has actually been circulating for a while but it is still unclear how to mechanically enforce it. We hereby propose the use of grasping motor data (i.e., kinematic data obtained from human hands while grasping) to encode the affordances of an object, and then to use the representation of object affordances to improve object recognition.

To test this idea, a total number of 5200 human grasping sequences have been recorded from 20 subjects. Each sequence consists of the video and kinematic recording of the act of grasping one of 5 objects with one of 7 grasping shapes, chosen from standard grasping taxonomies such as Cutkosky's [4]. (Recording of the hand kinematics is done via a sensorised glove and a magnetic tracker.) These sequences are collected in the CONTACT Visuo-Motor Grasping dataBase (VMGdB), presented and described in this very paper.[1] Using this database and a simple regression schema based upon artificial neural networks, we then build a function, called Visuo-Motor Map (VMM), mapping visual features of an object to an associated grasp. Since in general many different grasps are associated with the same objects, the VMM is here associating an "average" grasp posture to each object, which is not guaranteed to correspond to a physically feasible grasp but still is deemed to carry enough information on the affordances of that object.

At this point, to test the effectiveness of the idea, a standard classifier (namely a Support Vector Machine) is used to classify the objects in the database using either: (a) the visual features only, as is standard in object recognition, (b) the motor features only as recorded by the sensorised glove, (c) a combination of these features sets and, lastly, (d) a combination of the visual features and the motor features *as reconstructed by the VMM*. The latter scenario is of course more realistic since in most real-life applications (and in real life as well) the only available input is visual. The hope is that the augmented object classifiers perform dramatically better than the standard one when the real motor features are added; and significantly better when the reconstructed ones are used. Our experimental results confirm this hypothesis, even given the simplifying assumptions made in this work.

The paper is organised like this: after an overview of related work, in Section II we describe the VMGdB. Section III

This work is supported by the EU projects DIRAC (IST-027787, BC and TT) and Contact (NEST 5010, CC).

C. Castellini carried this work out at the LIRA-Lab, Università degli Studi di Genova, Italy; he is now with the DLR—German Aerospace Research Center, Oberpfaffenhofen, Germany. email claudio.castellini@dlr.de

B. Caputo and T. Tommasi are with the Idiap Research Institute, Martigny, Switzerland. email bcaputo@idiap.ch,ttommasi@idiap.ch

N. Noceti and F. Odone are with DISI - Università degli Studi di Genova, Italy. email odone@disi.unige.it,noceti@disi.unige.it

[1]The VMGdB is available at the following URI: http://slipguru.disi.unige.it/Research/VMGdB for download.

Fig. 1. Three very different cups: (left) the *Pick Up* mug by Höganäs (2009); (center) the *Black Flute Half Lace* coffee cup (1775) and (right) the *'Ole* mug (1997), both by Royal Copenhagen.

defines the framework; we then show the experimental results (Section IV) and draw conclusions in Section V.

A. Related work

The capability to recognise and categorise objects is a crucial ability for an autonomous agent; and in robotics, it is inextricably woven with the ability of grasping an object. In cognitive science, the theoretical link between vision and manipulation was provided by Gibson, according to whom an object is characterized by three properties: (1) it has a certain minimal and maximal size related to the body of an agent, (2) it shows temporal stability, and (3) it is manipulable by the agent. These properties imply that the object is defined in relation to an embodied agent able to manipulate the object. Therefore the set of possible manipulation actions are a crucial part of the object definition itself.

Interestingly, the theory of affordances has recently found neurological evidence, it is claimed, in the mirror neurons paradigm [5], [6]. According to it, structures exist in the high primates' brain which will fire if, and only if, an object is grasped (which mainly involves the sensorimotor system) or is seen grasped by an external agent (involving the visual system only, [7]). In addition to the original findings in monkeys, very recent evidence has been produced for the existence of such structures in humans [8]. If this is true, then the human object classification is so robust exactly because we *know what to do* with the objects we see — a capability which machines lack, so far.

This idea has so far been little exploited; among the positive cases there are [9], [10] who take an exquisitely robotic perspective, letting their systems acquire motor information about objects by having a humanoid robot manipulating them. Our work draws inspiration from [9] and it represents an extension and a further exploration of its topic. On the other hand, the vast majority of work on object recognition and categorization models objects starting from static images, without taking into account their 3D structure and their manipulability [11], [3]. An interesting exception is [12] where and-or trees and 3D features are used to categorise objects according to how well they fit a functional profile.

Few very recent attempts try to capture the Gibson's view. The approach proposed in [13] presents a Bayesian framework that unifies the inference processes involved in object categorization and localization, action understanding and perception of object reaction. The joint recognition of objects and actions is based on shape and motion, and the models take as input video data. In [14], the authors consider objects as contextual information for recognizing manipulation actions and vice versa. The action-object dependence is modelled with a factorial conditional random field with a hierarchical structure. In both approaches, objects and their affordances are first modelled separately, and combined together in a second step. This does not consider the embodiment of the agent manipulating the objects.

II. The Database

The CONTACT Visuo-Motor Grasping Database (VMGdB) is the result of recording the visual and kinematic content of grasping acts made by several human subjects, in changing conditions of illumination.

a) Experimental protocol: The subjects (all right-handed) would sit comfortably on a chair in front of a desk. Their right arm and hand would be resting on the arm of the chair. An object would be placed in a predefined position onto the desk. Then, the subject would be instructed to (a) reach for and grasp the object with his/her right hand (the grasping instant being signalled by a beep), (b) drop it somewhere else in the workspace (the releasing instant being signalled by another, different beep), (c) put the right arm and hand back in the resting position, (d) put the object back in the original position with the left arm and hand. The desk was uniformly dark green and non-reflective; the objects were chosen to be colourful; the illumination was provided by two windows looming over the desk. Intentionally we did not fix the illumination, which changed over time, since acquisition sessions spanned over a week, in the morning, afternoon and evening. Before each experiment we would fix the white balance of the cameras in order to avoid saturation. Figure 2 shows a typical reach-and-grasp sequence, as seen by the two cameras.

b) Data acquisition setup: The cameras are two Watec *WAT-202D* colour cameras, operating at 25Hz and connected to two Picolo PCI-bus frame grabbers. One camera is placed in front of the subject while the other was placed on the right-hand side of the subject, almost framing the object in a close-up and focussed upon it. The first camera has the view of what an external observer would be seeing of the grasp; the second would give an accurate representation of the act of grasping in full detail, including the last moments of the reaching sequence. Kinematics of the grasping act

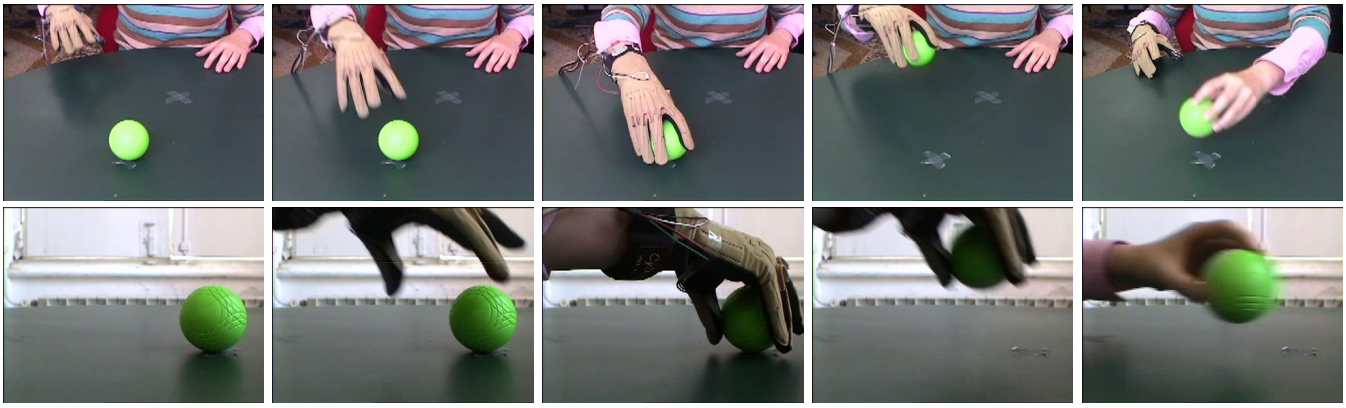

Fig. 2. Synchronised sample frames from the two video sequences, showing the the grasping act: hand and arm resting, reaching, grasping and dropping the object and, lastly, putting it back with the left arm and hand.

was captured using a 22-sensors Immersion *CyberGlove* [15] right-hand sided dataglove, which provides 22 8-bit numbers linearly related to the angles of the subject's hand joints. The resolution of the sensors is 0.5 degree. 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 database also includes data coming from an Ascension *Flock-Of-Birds* magnetic tracker [16] mounted on the subject's wrist, which would return six real numbers, the linear and angular coordinates of the wrist with respect to a base mounted on the far end of the desk. A force sensing resistor (FSR) glued to the subject's thumb is used to determine the instant of contact with the object. All data was collected on a fast and safe mass memory storage unit and synchronised using common timestamps.

c) Objects, subjects, grasps: The subjects pool includes 20 right-handed people, 6 females and 14 males, aged between 24 and 42 years (mean 31.5 years, median 31). They were engaged in grasping 7 different objects in 5 different ways. Figure 3 shows the objects and the grasp types. First we chose the grasps, using standard grasp taxonomies such as, e.g., Cutkosky's [4] as guidance, and trying to figure out which would be more useful in practical applications such as, e.g., dexterous teleoperation or hand prosthetics. Subsequently we chose the objects among everyday tools and toys, carefully selecting them in order for some of them to be graspable in several different ways, chosen among the grasps we had previoulsy selected. Table I sums up the total *(grasp,object)* pairs we have enforced (13). For instance, the pen would be grasped with either a pinch or a tripodal grip, the tape with the pinch, spherical or tripodal, the pig with the cylindrical power grasp, and so on.

Each subject replicated each grasp for 20 times, giving a total of $13 \times 20 \times 20 = 5200$ sequences, each sequence a *(grasp,object,subject,n)* tuple, where $n = 1, \ldots, 20$. (The correct number of grasping sequences was enforced by letting the subject hear the beeping sounds each time.)

TABLE I
THE 13 (OBJECT,GRASP) PAIRS ENFORCED IN THE VMGDB.

	ball	pen	duck	pig	hammer	tape	lego brick
cylindr. pow.				X			
flat					X		X
pinch		X	X			X	X
spherical	X					X	
tripodal	X	X	X			X	

III. THEORETICAL FRAMEWORK

We deal here with the problem of augmenting visual information about an object with motor information about it, that is the way the object can be grasped by a human being. This can be seen as an instance of a more general framework for multi-modal learning. Although a formal, abstract definition of this framework is out of scope here, we outline it in order to clearly frame the point of view from which we hope to improve classical object modelling and recognition. We first give a theoretical overview of the idea, and then go in deeper detail describing the visual and motor features used, the method for training the VMM and lastly the object classifier.

A. Affordances, and their role in object recognition

In everyday life, living beings use *distal* sensory modalities as their only means of "on-line" gathering information about the world (by distal here we mean, senses which operate at long distance such as, e.g., vision, hearing, smell, etc.). This is coherent with the basic needs of avoiding predators, finding food, mating and so on. Of course, (distal) sensorial information is multi-modal in nature, as, e.g., the smell, sight and noise characteristic of a predator come together in experience. But to our end, a more subtle form of multi-modal learning is considered, that is, associating distal and *proximal* modalities in the infanthood, where by proximal we mean sensorimotor and proprioceptive: those modalities which appeal to manipulation.

According to Gibson's concept of affordances, which is an instance of this general framework, the sight of an object is inextricably associated by a human being to the ways it can be used; this association is primed by manipulation in the

Fig. 3. Top row: the objects used in our experiments. Bottom, the grasp types we consider: *(left to right)* cylindric power grasp, flat grasp, pinch grip, spherical and tripodal grip.

early development: at first randomly, then in a more and more refined way. According to this, human object recognition is so good because we immediately associate to the sight of an object its affordances, and this generalises to the case of new objects.

So, object classification should be improved by motor information, be it the real motor information or reconstructed starting from the distal modality (in this case, sight). Checking whether this idea works would in principle involve riproducing the developmental phase of manipulation in infants. Of course this is so far impossible, so we resort to building a relation between the object seen and how it can be grasped. Notice that this is in general a *(non-functional) relation*, a many-to-many relationship, since many objects can be grasped with the same grasp, and one object might afford many different grasps. In our case, to simplify the problem, we build the relation from available data gathered from adult humans, and assume that the relation is functional, meaning that it will give us only one grasp for each object it sees.

The system is then built as follows: in the *training* phase (Figure 4, left) the system is input visual and motor data, which are used to train both the Visuo-Motor Map (VMM) and the Visuo-Motor Classifier (VMC), an object classifier which uses both visual and motor features. In the *testing* phase (Figure 4, right) the system is input either

- *visual and motor data (a)* This corresponds to the case when the agent sees and grasps the object. Here the classifier receives both modalities, and it classifies the object using these informations; or
- *visual data only (b)* This corresponds to the case when the agent sees the object but does not grasp it. In this situation, the system first reconstructs a grasp from the perceived visual features, using the VMM; then, it uses the two sets features (one perceived, one reconstructed) to classify the object.

B. Implementation

1) Visual features: From each of the 5200 sequences, a set of relevant frames in which the object is clearly visible is extracted from the object-close-up camera stream. This is easily accomplished since the sequences are quite similar to each other in length. Background subtraction and then change detection are applied, by comparing the selected frames against a background model, in order to select a region of interest (ROI) in which the object is found. Subsequently, a bag-of-keypoints object description [17] is applied to the ROI, in order to extract from it salient visual features which can be safely associated with the object itself. Building the bag-of-keypoints description of an object is a two-phases procedure (the same idea is applied in [18], where more details can be found).

In the first phase a vocabulary of 200 visual features is built: inside each ROI a random number of points is chosen and a fixed-scale and -orientation variant of the SIFT descriptors [19] is used to characterize them. The global set of descriptors is then clustered using k-means (see, e.g., [20]) with $k = 200$. This value was set after an initial round of experiments as the best found given the number of objects, sequences and characterstics of the dB. (Notice that the optimal value of k could be found automatically, e.g., using x-means [21].) The obtained 200 centroids (virtual features) are the words of the vocabulary.

In the second phase, each object is associated to a bag of words of the vocabulary, via a simple nearest-neighbour approach. The visual appearance of the object is therefore represented by a frequency histogram of 200 bins, the ith bin of the histogram indicating how many times the ith word of the vocabulary is seen belonging to that object, with $i = 1, \cdots, 200$.

2) Motor features: The motor features are the 22 numbers returned by the CyberGlove, considered at the time of contact of the subject's hand with the object. The value of the force-sensing resistor was used to determine the instant of contact. The motor features give a faithful snapshot of the subject's hand posture at the time of grasping the object.

3) Training the Visuo-Motor Map: The VMM has at its core a Multi-Layer Perceptron (MLP). One single MLP in our setting has 200 input units, one hidden layer with 20 units and 22 output units; the net is trained via the Scaled Conjugate Gradient Descent method [22] and the activation is a logistic

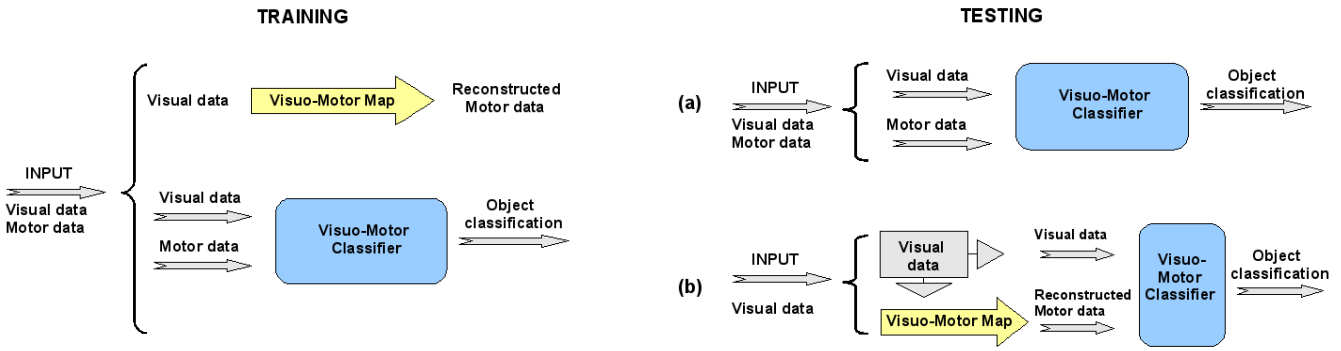

Fig. 4. A schematic representation of the theoretical framework. During training (left), the system receives in input visual and motor data, and it learns simultaneously a Visuo-Motor Map (VMM) and Visuo-Motor Classifier (VMC). During testing (right), whenever the agent can see the object but cannot grasp it (b), the VMM reconstruct a grasp from the visual input, which is then used as input to the multi-modal classifier jointly with the visual features.

sigmoidal function. Training is done via early stopping on the appropriate validation set. (These settings are inspired by the work of Richmond and others [23], [24] on audio-to-motor mapping, probably the most similar approach to what we are presenting here.)

After an initial phase of preliminary experiments, we noticed that the VMM could be largely improved by employing one MLP per each known object, and then selecting which MLP to use based upon the output of the visual classifier. If an object can be grasped in only one way (as is the case, e.g., of the hammer and pig, see Table I again), the reconstructed motor data will correspond to an estimate of this grasp; otherwise, it will represent a weigthed mixture of the available grasps.

4) Training the Visuo-Motor Classifier: The VMC should accept visual, motor or combined features. Algorithmically, this implies building a classifier over multiple cues. In the computer vision and pattern recognition literature some authors have suggested different methods to combine multiple cues. They can be all reconducted to one of the following three approaches: low-level, mid-level and high-level integration [25], [26]. In the low-level case the features are concatenated to define a single vector. In the mid-level approach the different features descriptor are kept separated but they are integrated in a single classifier generating the final hypothesis. The high-level method starts from the output of different classifiers each dealing with one feature: the hypotheses produced are then combined together to achieve a consensus decision.

To train the VMC here we implement these three strategies in a Support Vector Machine-based framework (SVM, see [27]). We use the Discriminative Accumulation Scheme (DAS, [28]) for the high-level, and the Multi-Cue Kernel (MCK, [29]) for the mid-level integration. As already mentioned, the low-level integration just consists in the feature concatenation, with the new vector fed to a standard SVM. A short description of the DAS and MCK schemas follows:

DAS (high-level). DAS is based on a weak coupling method called accumulation. Its main idea is that information from different cues can be summed together. Suppose we are given M object classes and for each class, a set of N_j training data $\{I_i^j\}_{i=1}^{N_j}$, $j = 1,\dots M$. For each, we have a set of P different features so that for an object j we have P training sets. We train an SVM on every set. Kernel functions may differ from cue to cue and model parameters can be estimated during the training step via cross validation. Given a test image $\hat{I}$ and assuming $M \geq 2$, for each single-cue SVM we compute the distance from the separating hyperplane $D_j(p)$, correspondent to the value of the margin obtained using the model j_{th} *class vs all* for cue p. After collecting all the distances $\{D_j(p)\}_{p=1}^{P}$ for all the M objects and the P cues, we classify the image $\hat{I}$ using the linear combination:

$$j^* = \underset{j=1}{\overset{M}{\operatorname{argmax}}} \left\{ \sum_{p=1}^{P} a_p D_j(p) \right\}, \quad \sum_{p=1}^{P} a_p = 1. \tag{1}$$

The coefficients $\{a_p\}_{p=1}^{P} \in \Re^+$ are determined via cross validation during the training step.

MCK (mid-level). The Multi Cue Kernel is a positively weighted linear combination of Mercer kernels, thus a Mercer kernel itself:

$$K_{MC}(\{T_p(I_i)\}_p, \{T_p(I)\}_p) = \sum_{p=1}^{P} a_p K_p(T_p(I_i), T_p(I)), \sum_{p=1}^{P} a_p = 1. \tag{2}$$

In this way it is possible to perform only one classification step, identifying the best weighting factors $a_p \in \Re^+$ through cross-validation while determining the optimal separating hyperplane. This means that the coefficients a_p are guaranteed to be optimal. Notice that cross-validation for MCK is quite time-consuming and could in principle make the obtained advantage untenable; but it is an off-line pre-processing step (as well as standard cross-validation, used to find the best hyperparameters) and, once the parameters are defined, training and testing have the same computational complexity of a standard SVM. For problems in which large sample databases are available or the input space is highly-dimensional, MCK can be substituted with a Multi-Kernel Learning algorithm (see, for instance, [30]).

IV. Experimental results

This section reports the experimental validation of our model. We begin by using real motor data (section IV-A), showing that by joint modeling visual and motor information it is possible to achieve a significant boost in recognition, compared to using visual information only. We proceed by evaluating the quality of the reconstructed grasp via regression (section IV-B). We then show that, whenever the motor information is not perceived by the agent, it is still possible to get a better performance by using the VMM (section IV-C).

All classification experiments are performed as follows: a training set of 130 samples and a testing set of 2600 samples (disjoint from the training set) are randomly extracted from the 5200 samples found in the VMGdB; one such training/testing set pair is called *split*. This procedure is repeated 10 times, leading to 10 splits. Every classifier is evaluated on all splits, and then the average and standard deviations of the error rate over the 10 splits are reported. The error rate we use is the standard error rate for classification, i.e., the ratio of correctly predicted labels and the number of labels. This instance of cross-validation is used to choose the best classifier hyperparameters.

The classifier is a SVM, one-versus-all multiclass extension. In this extension, N-multiclass classification is achieved by solving N two-classes classification problems, where each classifier distinguishes between one of the labels and all others; the classifier with the highest confidence determines the winning label in a winner-takes-all strategy. We use the Gaussian Kernel for the visual and motor modalities, both when considered separately and in the integration approach (two Gaussian Kernels combined in the mid-level integration schema).

A. Classification with real motor features

The first set of experiments is conducted using the real motor features, namely those recorded by the users when grasping the objects, and the corresponding visual features.

Figure 5-a shows the overall recognition results obtained by using only visual information (V), only motor information (M), or the two combined together, with the three proposed approaches (low-, mid- and high-level). Using the visual features a better average performance is obtained ($86.37\% \pm 1.91\%$) than using the motor ones ($75.53\% \pm 1.22\%$); and their integration is clearly beneficial: the mid-level integration produces the best result ($93.94\% \pm 0.77\%$). The gain in accuracy between mid-level and visual only is 7.57% (difference in accuracy evaluated per split and then averaged on the 10 splits). The second best result is obtained by using the high-level integration ($92.65\% \pm 1.22\%$); the difference in performance between high- and mid-level is negligible.

Figures 5-b, -f show the confusion matrices obtained by using, in turn, the visual features (b), the motor features (c) and the low-, mid- and high-level integrations (d,e,f). Clearly the combination of the two modalities leads to considerable qualitative advantages in the recognition of each object, for all methods. Consider for instance the objects "ball" and "pig": the mean accuracy is respectively 88.6% and 75.1% using visual features and 77.2% and 96.6% using motor features. The ball was grasped in two different ways (tripodal and spherical grasp) while the pig was manipulated only with the cylindric grasp. Thus, grasp information is object-specific for the pig, and this leads to an impressive increase in performance when using mid-level integration (100% classification accuracy). Using integrated features is beneficial also for the ball, for which the accuracy is 96.5%. Analogous considerations hold for the two other approaches. We conclude that (a) feature integration leads to a dramatic improvement in performance and (b) the mid-level features integration is the most proficient.

B. Evaluation of the VMM

To evaluate performance of the VMM, the whole dataset was divided in a training set and a testing set, each one consisting of 2600 samples. Then:

(a) the 7 MLPs were trained and used to predict the motor features of the testing set;
(b) a SVM was trained on the real motor features to classify the *grasps*, and then tested on *reconstructed grasps* obtained at the previous step. A predicted grasp not being one of the possible grasps associated with the related object would count as an error.

This experiment was run on 10 such random (training/testing) splits, obtaining an average error rate of 10.7%, largely smaller than chance level (63%). This indicates that the grasp reconstruction is significantly faithful to the grasps associated to the objects during the training of the VMM.

C. Classification with reconstructed motor features

The experiments described in Section IV-A are here repeated using, instead of the real motor features, those reconstructed by the VMM. An appropriate MLP in the VMM is chosen accordingly to the prediction of the visual classifier.

Results are reported in Figure 6. Figure 6-a shows the recognition rates obtained by using only visual information (V – the same shown in the previous section), only motor information (M), and the two combined together (LOW, MID, HIGH). The performance of the motor only classifier decreases slightly in this case, if compared to the real features case ($71.90\% \pm 2.06\%$ versus $75.53\% \pm 1.22\%$). Still, the performance of the multi-modal classifiers show an increase in the overall performance, compared to the vision only approach. Once again, the best performance is achieved by the mid-level integration ($88.77\% \pm 1.29\%$), closely followed by the high-level ($88.38\% \pm 1.31\%$).

Figure 6-b, -f show the confusion matrices obtained by all classifiers, as reported in Section IV-A. The results for the reconstructed motor data are in general lower than that obtained with the real ones (Figure 5-c). To explain this behaviour there are two things to keep in mind: (1) the lower is the number of possible grasps associated with an object, the fewer are the data on which the corresponding neural network is trained; (2) if the first step of hypothesis generation fails, the error propagates on the motor data reconstruction.

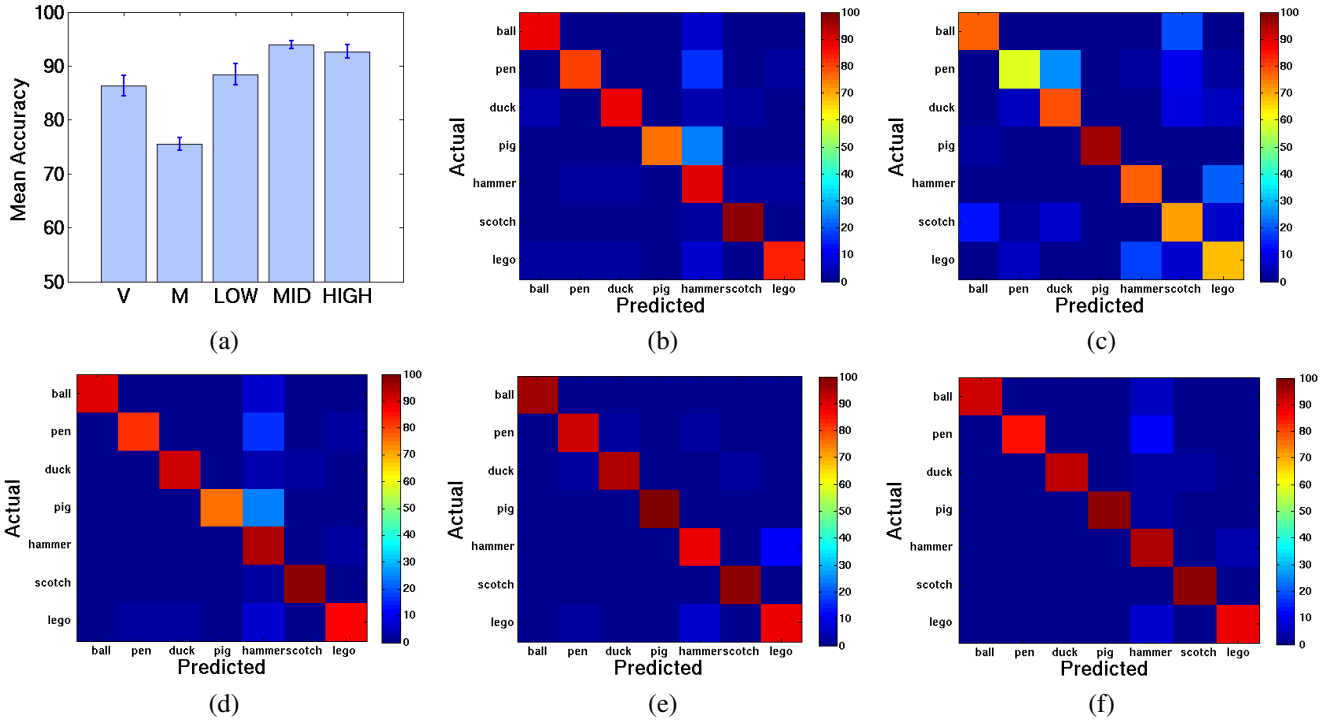

Fig. 5. (a) Classification mean accuracy (plus/minus one standard deviation) obtained while using visual features (V), real motor features (M), and their low-, mid- and high-level (LOW, MID, HIGH) integration. (b-f) confusion matrices using, in turn, V, M, LOW, MID and HIGH features.

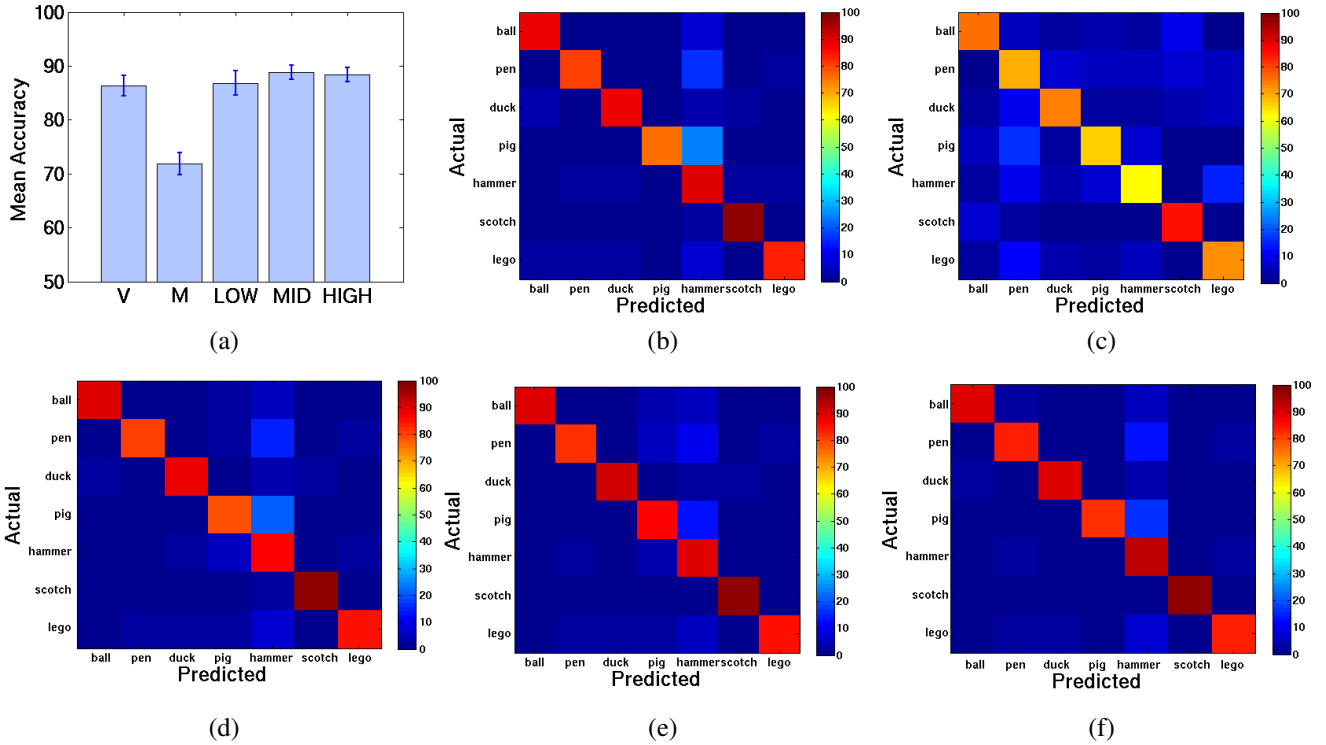

Fig. 6. (a) Classification mean accuracy (plus/minus one standard deviation) obtained while using visual features (V), reconstructed motor features (M), and their low-, mid- and high-level (LOW, MID, HIGH) integration. (b-f) confusion matrices using, in turn, V, M, LOW, MID and HIGH features.

In particular, both points give an intuition about why the objects "pig" and "hammer" (which were manipulated with only one grasp each) present the worst recognition results using motor information (66.65% and 61.45% respectively). Nevertheless, in the "pig" case, the reconstructed grasp data added to the visual features brings the mean accuracy for object recognition from 75.1% (only visual) to 87.0% (using mid-level integration). As a last remark, we see once again

that mid-level obtains the best performance (gain in accuracy of 2.40%) and therefore it appears to be the most suitable candidate for the VMC module.

V. Discussion

The grand goal of this research is to show that the reconstruction of proximal sensory modalities can help the interpretation of the distal modalities. This very abstract idea stems from recent findings in neuroscience which seems to indicate that this is the case for a number of living beings; its theoretical value extends in principle to any distal sensory modality that one wants to understand (e.g., sound, vision). Priming and coupling of distal and proximal modalities in the infanthood is supposed to be the phase when we learn to associate, e.g., the visual appearance of a cup with the possibility of using it for drinking and carrying a liquid around.

The work hereby presented is a modest step in this direction. We focus on the problem of visual object recognition and show that the use of kinematic hand features ("motor" features) representing the ways in which the object can be grasped can significantly improve the recognition performance of a standard classifier, with respect to the case in which visual features only are employed. Actually, the experimental results indicate that motor features recorded by a sensorised glove can improve the recognition rate by some 7.6%, and by 2.4% when reconstructed through the VMM. Although the latter result is not so impressive as the former, one should remember that all accuracies are here already around 90%, where it is hard to gain even a few points more. Moreover, since the real motor features are so useful, chances are that a better VMM could improve by far the current performance of the VMM-reconstructed motor features.

In fact, the VMM is here realised via a simple regression schema based upon a MLP trained in a completely standard way; moreover, the VMM is a function whereas the relation being modeled is non-functional, with the result that, in some cases, its output is no physically feasible grasp, but rather a weighted average of the possible grasps. Still, it turns out to be effective, and this matches what is reported by Richmond and others in a series of papers [31], [32], [24], [33] about *speech* recognition; this seems to indicate that (reconstructed) motor information is really of great help, even when the reconstruction is suboptimal. Richmond's move to counter this problem was to enforce a probabilsitic model of all possible grasps rather than a function, and that is also what we plan to do as immediate future work. In that case, the VMM would be able, when seeing, e.g., the pen, to say that it would be likely grasped with a pinch or tripodal grip (consider Table I again) rather than with a power, cylindric or flat grasp. Such a VMM would enforce quite closely Gibson's concept of the affordance of the pen, somehow fully accomplishing what we said in the introduction.

Another interesting point about the VMM is that it consists of 7 MLPs, hierarchically subordinated to the visual classifier; its effectiveness with respect to the simpler single MLP schema has been tested in an initial phase. How can it still be of help when the visual classifier is wrong? The answer gives more interesting insight into the problem. Let us take a step behind and consider "real" motor information. As it stands now in our framework, every object is naturally associated with the real grasp(s) it was grasped with during the experiments, and this could well be detrimental in some cases. As an example (consider again Table I and Figure 5-(b,e), visual features only versus mid-level integration), the visual classifier tends to confuse the pig and the hammer, due to analogies in the local descriptors, but the integration with motor features essentially eliminates the problem, since the pig is univocally associated with the cylindrical grasp and the hammer with the flat grasp. On the other hand, the motor integrated classifier shows a somehow higher confusion between the hammer and the lego brick since the lego brick too can be grasped with a flat grasp. From the motor point of view, the two objects are similar.

Let us now turn to the VMM-reconstructed features (Figure 6-(b,e)). In this case, too, the pig/hammer ambiguity is resolved thanks to the motor features, and this is not surprising — it just means that the wrong MLP is input ambiguous local descriptors which still result in something close to the correct grasp for the other objects, or at least that can correct the ambiguity. On the other hand, in motor-ambiguous cases such as the hammer/lego pair, this time the VMM can *correct* the error since in the case of the lego it returns a weighted-average-grasp composed of the flat and pinch grasps, rather than one *or* the other, as it was the case with the real motor features. This mechanism is likely to explain the improvement obtained by the VMM-reconstructed features, *even though* the visual classifier might be wrong in the first place: it turns out that a weakness of our system, given by a simplifying assumption, is beneficial. Notice, once again, that a probabilistic description of the grasps associated to each object would solve the problem and, very likely, boost the results.

The database we presented here, the CONTACT Visuo-Motor Grasping dataBase, (which has been using for all our experiments) is now available online for downloading (see the Note to Section I) and we hope it will represent an interesting dataset for those members of the scientific community who are willing to pursue the same path of research hereby described. The VMGdB consists of 5200 synchronised stereo video + kinematic sequences, each one representing an act of grasping. Twenty human subjects, 7 objects and 5 grasp types are involved; ground truth is available for all sequences; and the grasping/objects relationship enforced is a many-to-many relationship.

The dB is as yet not comparable with other similar efforts built for pure computer vision, as far as size is concerned; but its focus is, rather than on the number of samples/objects/grasps, on the association between grasping and vision, and the variability of the subjects involved aims at giving a broad spectrum of human reaching/grasping. From this point fo view, the dB has the potentiality to support much more research than is described here; in fact, here we have been neglecting the *orientation* of the hand at the time of grasping, the *dynamics* embedded in the reaching phase (containing a lot of information more, see, e.g. [34], [35]) and the possibility of exploiting the two points of view (we only

use one of the cameras); but these data are available in the dB. Such a research could finally lead to a significant advance also in robotic grasping, too, as the reconstructed grasp might be somehow mapped onto the robotic end-effector in a teloperated setup.

Lastly, the approach here described is supposed to scale up to many more objects, grasp types and subjects (and in the future to the generic, online case) *once the VMM is amended in the probabilistic way above described*. Moreover, so far the use of fixed-scale and -orientation SIFT descriptors does not allow us to claim that this system would work in changing viewpoint conditions, but this simplifying assumption can of course be lifted. The VMGdB is collected in changing conditions of illumination, and if the local image descriptors are fully fledged SIFT, then there is a reasonable hope of robustness to illumination changes, distortion, size and orientation changes. Future research will also be aimed at looking for better visual features.

REFERENCES

[1] J. J. Gibson, *The Theory of Affordances*. Erlbaum Associates, 1977.

[2] ——, *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, 1986.

[3] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 259–289, 2008.

[4] M. R. Cutkosky, "On grasp choice, grasp models, and the design of hands for manufacturing tasks," *IEEE Transactions on Robotics and Automation*, vol. 5, no. 3, pp. 269–279, 1989.

[5] V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti, "Action recognition in the premotor cortex," *Brain*, vol. 119, pp. 593–609, 1996.

[6] G. Rizzolatti and L. Craighero, "The mirror-neuron system," *Annual Review of Neuroscience*, vol. 27, pp. 169–192, 2004.

[7] M. Umiltà, E. Kohler, V. Gallese, L. Fogassi, L. Fadiga, C. Keysers, and G. Rizzolatti, "I know what you are doing: A neurophysiological study," *Neuron*, vol. 31, pp. 1–20, 2001.

[8] J. M. Kilner, A. Neal, N. Weiskopf, K. J. Friston, and C. D. Frith, "Evidence of mirror neurons in human inferior frontal gyrus," *J Neurosci.*, vol. 29, pp. 10 153–10 159, 2009.

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

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

[11] G. Griffin and P. Perona, "Learning and using taxonomies for fast visual categorization," in *Proc 21st IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[12] K. Woods, D. Cook, L. Hall, K. Bowyer, and L. Stark, "Learning membership functions in a function-based object recognition system," *Journal of Artificial Intelligence Research*, vol. 3, pp. 187–222, 1995.

[13] A. Gupta and L. Davis, "Objects in action: an approach for combining action understanding and object perception," in *Proc 21st IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[14] H. Kyellstrom, J. romero, D. Martinez, and D. Kragic, "Simultaneous visual recognition of manipulation actions and manipulated objects," in *Proc European Conference on Computer Vision (ECCV)*, 2008.

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

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

[17] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[18] N. Noceti, B.Caputo, C.Castellini, L.Baldassarre, A.Barla, L.Rosasco, F.Odone, and G.Sandini, "Towards a theoretical framework for learning multi-modal patterns for embodied agents," in *Proc 15th International Conference on Image Analysis and Processing (ICIAP)*, 2009.

[19] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[20] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21—27, Jan 1967.

[21] D. Pelleg and A. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," in *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann, 2000, pp. 727–734.

[22] M. F. Moller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, pp. 525–533, 1993.

[23] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zacks, and S. Levy, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on X-ray microbeam data," *J Acoust Soc Am.*, vol. 92, no. 2, 1992.

[24] K. Richmond, "Trajectory mixture density networks with multiple mixtures for acoustic-articulatory inversion," in *Advances in Nonlinear Speech Processing, International Conference on Non-Linear Speech Processing, NOLISP 2007*, ser. Lecture Notes in Computer Science, M. Chetouani, A. Hussain, B. Gas, M. Milgram, and J.-L. Zarader, Eds., vol. 4885. Springer-Verlag Berlin Heidelberg, Dec. 2007, pp. 263–272.

[25] R. Polikar, "Ensemble based systems in decision making." *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[26] C. Sanderson and K. K. Paliwal, "Identity verification using speech and face information." *Digital Signal Processing*, vol. 14, no. 5, pp. 449–480, 2004.

[27] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc 5th Annual ACM Workshop on Computational Learning Theory (COLT)*, D. Haussler, Ed. ACM press, 1992, pp. 144–152.

[28] M. Nilsback and B. Caputo, "Cue integration through discriminative accumulation," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 2, pp. 578–585, 2004.

[29] T. Tommasi, F. Orabona, and B. Caputo, "Discriminative cue integration for medical image annotation," *Pattern Recogn. Lett.*, vol. 29, no. 15, pp. 1996–2002, 2008.

[30] F. Orabona, L. Jie, and B. Caputo, "Online-batch strongly convex multi kernel learning," in *Proc 23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[31] K. Richmond, S. King, and P. Taylor, "Modelling the uncertainty in recovering articulation from acoustics," *Computer Speech and Language*, vol. 17(2), pp. 153—172, 2003.

[32] A. A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *Proceedings of the International Conference on Spoken Language Processing*, 2000.

[33] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 121(2), pp. 723—742, 2007.

[34] C. Bard, J. Troccaz, and G. Vercelli, "Shape analysis and hand preshaping for grasping," in *Intelligent Robots and Systems '91. 'Intelligence for Mechanical Systems, Proceedings IROS '91. IEEE/RSJ International Workshop on*, Nov 1991, pp. 64–69 vol.1.

[35] S. A. Winges, D. J. Weber, and M. Santello, "The role of vision on hand preshaping during reach to grasp," *Exp Brain Res*, vol. 153, pp. 489–498, 2003.