Joint Visual and Motor Learning for Object Modeling and Recognition

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

Abstract— **Object recognition is a key problem of artificial vision; in robotics, it is strongly connected to that of grasping. In fact, there is so far no general solution to either problem. Traditionally, visual features are evaluated from camera images and statistical methods are then trained on huge visual datasets in order to obtain a robust object classifier. The knowledge so obtained is then used to choose a model to perform a grasping action.**

Inspired, among others, by the neuroscientific framework of mirror neurons, we hereby propose to enhance the model of an object by adding to its visual features a probabilistic description of the grasps chosen by human subjects to grasp it. Since in a standard setting the grasps are not directly available to the system, they must be reconstructed from the visual features, and then used to augment the recognition system's input space. We achieve this by building a map from visual to motor features, which we call a Visuo-Motor Map (VMM), practically enforced via regression on a human grasping database.

We experimentally show that such a technique improves the recognition rate of a standard object classifier: in case the original motor features are used, the improvement is dramatic, whereas when we reconstruct them via the VMM we still obtain a statistically significant improvement.

I. Introduction

Consider the objects in Figure 1. How do we know that they are all cups? The answer is simple: all of them can be used to contain liquids and drink, and have actually been designed to this end. Although very little in their visual appearance ties them together, we all know what can be done with such objects since *we have done it* at some time in the past. The category of an object is often determined predominantly by its function rather than by its visual appearance alone; this led Gibson in the 70s [1], [2] to define objects in terms of their affordances — "what can be done with them".

This idea could be useful to improve the classic solution to object recognition, which uses visual features only. It is hard to figure out what visual features could lead an object recognition system to categorise as "cups" the three objects above. Traditionally, this problem is solved by training the system on a very large database of images of very diverse cups; but this is potentially incomplete and resource-consuming [3]. Now, what if our machines had an idea of how to *grasp* something which looks like a mug? In fact, if this paradigm is correct, this could be the reason behind the robustness of human object recognition. Such robustness could be obtained by automated systems if only they knew what to do with an object, as they see it. This implies that an object recognition system must be able to grasp objects, or it must know something about grasping. Once it does, it has a wholly new way to associate a category to an object it sees.

In order to test this hypothesis, we have first collected a number of human grasping sequences, recording at the same time a video sequence of the grasping act and the hand posture (using a sensorised glove). These sequences are collected in the CONTACT Visuo-Motor Grasping dataBase (VMGdB)[1].

Then, using the VMGdB, by means of a simple neural network we have built a map from the image of each object to the associated grasp(s) or, in other words, a Visuo-Motor Map (VMM) from visual to motor features. The VMM enables us to retrieve the "archetypal grasp" of an object when that object is seen. The VMM is then used to build a Visuo-Motor Classifier (VMC) which exploits the traditional visual information plus the associated motor information, either the "real" one, as it was recorded by the glove, or the VMM-reconstructed one. The latter scenario is of course more realistic, as in most real-life applications, and in real life as well, the only available input is visual. The hope is that this augmented object classifier performs 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 clearly confirm this hypothesis.

The paper is organised like this: after an overview of related work, in Section II we describe the Visuo-Motor Grasping dataBase (VMGdB). Section III defines the general multi-modal learning framework, and then it describes in detail the instance under examination: the visual and motor representation (III-A), the Visuo-Motor Map (VMM, Section III-B) and the Visuo-Motor Classifier (VMC, Section III-C). 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

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

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

C. Castellini is with the LIRA-Lab, University of Genova, Italy, and the DLR — German Aerospace Research Center, Wessling, Germany, claudio.castellini@unige.it

G. Sandini is with the Italian Institute of Technology, Genova, Italy, giulio.sandini@iit.it

N. Noceti and F. Odone are with the Computer Science Department of the University of Genova, Italy, odone@disi.unige.it, noceti@disi.unige.it

[1] Upon acceptance of the paper, the database will be made available online.

Preprint submitted to 2010 IEEE International Conference on
Robotics and Automation. Received September 15, 2009.

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.

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 [4], [5]. 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, [6]). In addition to the original findings in monkeys, very recent evidence has been produced for the existence of such structures in humans [7]. 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 [8], [9] who take an exquisitely robotic perspective, letting their systems acquire motor information about objects by having a humanoid robot actually manipulating them. 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 [10], [3]. Few very recent attempts try to capture the Gibson's view. The approach proposed in [11] 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 [12], 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.

	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	

TABLE I

MAPPING GRASPS-OBJECTS. EACH ACTOR PERFORMS GRASPING ACTIONS ON 13 OBJECT-GRASP TYPE PAIRS.

II. THE VISUO-MOTOR GRASPING DATABASE

The VMGdB dataset is built considering 7 different objects ((see Fig. 2), top) and 5 grasps ((see Fig. 2), bottom). 20 different actors participated to the acquisition, during which each object was grasped in one or more ways, according to the many-to-many relationship reported in Table I. In total we consider 13 different pairs (grasp,object), and, for each triple *(object, grasp, actor)*, we acquired 20 replicates of the grasping experiment.

We obtained 5200 experiments *(object, grasp, actor, exp-num)*, and for each of them the VMGdB stores the following information:

- **Visual information.** 2 video sequences acquired by 2 color cameras with different focus – one is the object, the other one is the grasping action. The video sequences are associated to 2 data files reporting the video frames time-stamps, allowing for synchronization with the sensor data;
- **Hand posture sensor information.** 1 data file containing the hand posture sensor data acquired by a CyberGlove [13]. For each posture the glove returns 22 8-bit numbers linearly related to the angles of the actor's hand joints. 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. Again, the sensor data are associated to acquisition time-stamps for synchronization.

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

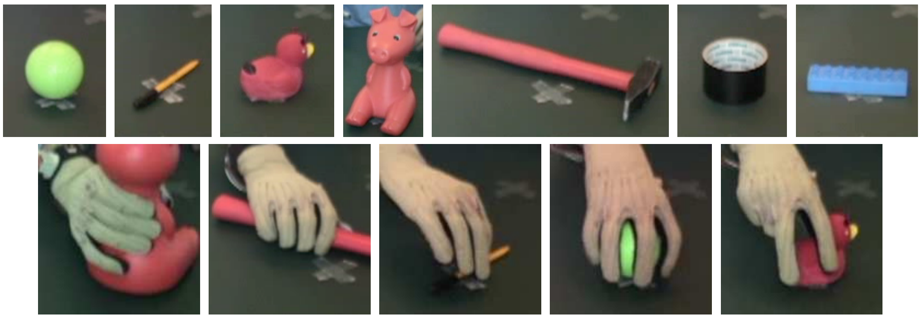


Fig. 2. Top row: the objects used in our experiments. Bottom, the grasp types we consider: *(left to right)* cylindric power grasp, flat grasp, pinch grip, spherical and tripodal grip.

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.

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. Following Gibson's idea, for example, the sight of an object would be inextricably associated by a human being to the ways it can be used. This association is primed by manipulation in the early development: at first randomly, then in a more and more refined way. According to this, human object recognition is so good also because we can reconstruct motor information associated to visual information, and use it too when dealing with a new object.

With the goal of applying this idea to automated object recognition, we propose to build an object classification system on a set of visual *and* motor features. Whenever the motor features are not perceived by the system (i.e. the agent is not grasping the object in the field of view), we infer them from the visual input through a mapping function, learned during training. This scheme mimics the early-development training mentioned above. We now proceed to illustrate in detail our framework, that is summarized schematically in Figure 3. We first describe the learning processes that occur during training, and then we describe the object classification procedure during testing.

Training Figure 3, left, illustrates the learning processes activated during training. The system receives as input visual and motor data. These data are used to learn

The Visuo-Motor Map (VMM) between the two modalities via regression.

The Visuo-Motor Classifier (VMC) that recognizes objects using visual and motor features.

Testing Figure 3, right, illustrates the two possible scenarios during testing:

The system receives as input vision and motor features. 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.

The system receives as input vision features only. This corresponds to the case when the agent sees the object but does not grasp it. In this situation, the system first generates an archetypal grasp from the perceived visual features, using the VMM. Then, it uses the two features (one perceived, one inferred) to classify the object.

We now proceed to describe in detail each component of the system. We begin from the vision and motor features (section III-A). We then describe the algorithmic implementation of the VMM (section III-B) and we conclude the section with the algorithm behind the VMC (section III-C).

A. Perceptual representations

This section describes the visual and motor features used in our framework. We begin with the visual features (Section III-A.1) and then continue with the motor features (Section III-A.2).

1) Visual Features: The visual appearance of objects is captured by a dedicated item of the framework, which can be sketched as follows. We first select from the video sequence a set of interesting frames where the object is clearly visible. To avoid contamination due to background elements, we apply change detection by comparing the selected frames against a background model, and then restrict our attention on the region of interest (ROI) defined by the object bounding box. We apply to the ROI a bag-of-keypoints object description [14] designed as a two steps procedure [15]:

- We build the global vocabulary, by putting together keypoints extracted from images of all the objects into the dataset. As keypoints, we consider a set of randomly sampled points whose patch is modeled with a vector

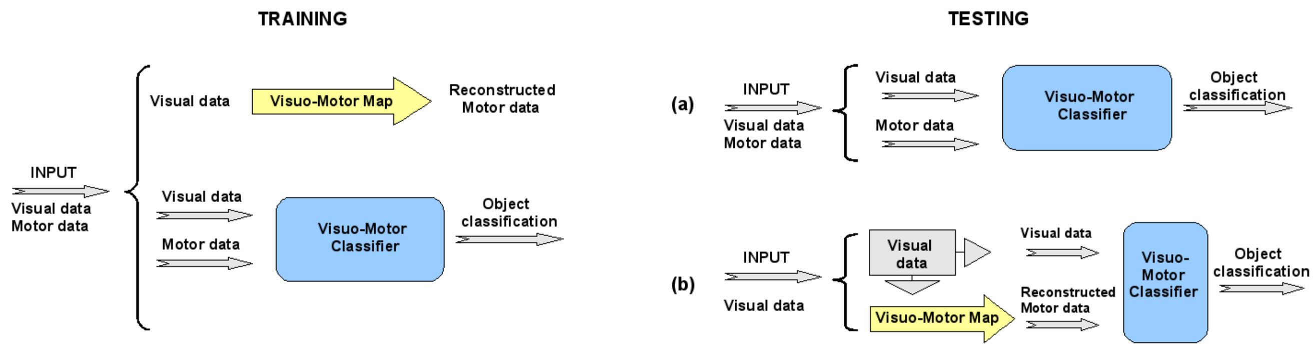


Fig. 3. A schematic representation of the theoretical framework. During training (left), the system receives in input visual and motor data, and it learns simultaneously a Visuo-Motor Map (VMM) and Visuo-Motor Classifier (VMC). During testing (right), whenever the agent sees but does not grasp the object, the VMM generates an archetypal grasp from the visual input, that is then used as input to the multi modal classifier jointly with the perceived visual features.

valued descriptor which can be seen as a fixed scale SIFT [16]. K-means is adopted to cluster the descriptors: the centroids (or virtual features) become the words of the visual vocabulary.

- Both training and test images are thus represented with respect to the vocabulary, with a simple nearest neighbor approach. At the end, visual appearance of objects is summarized with a frequency histogram, whose peaks should indicate which virtual features are the most important in modelling a specific object.

Notice that the vocabulary size is a system parameter which should be tuned with respect to the complexity of objects, to find a trade-off between sparsity of the descriptions and capability of characterizing the objects.

Finally a remark is in order. From the point of view of appearance-based object recognition, the experimental scenario is not challenging. We opted for such a setting in order to keep the focus of the work on the joint modelling of visual and motor inputs.

2) Motor Features: The MPR are simply the 22 angles returned by the dataglove, considered at the time of contact of the subject's hand with the object[2]. The MPR is therefore a "snapshot" of the subject's hand in the instant of grasping the object.

B. Learning the Visuo-Motor Map

The VMM is supposed to be a regression strategy from visual to motor features, as defined above. Since the output is multivariate (the motor features, consisting of 22 numbers) and the input is very highly dimensional (the visual features, consisting of 200 numbers), we decided to enforce the VMM using neural networks. Each network was kept as simple as possible: one hidden layer with 20 neurons, log-sigmoid transfer function and scaled conjugate gradient backpropagation. The training procedure used the early stopping strategy, i.e. the training set was divided in a new training and validation set. The network is evaluated on the validation set: when the performance stops improving, the algorithm halts.

Most of these settings are inspired by the work of Richmond and others [17], [18] on audio-to-motor mapping. In fact, since each object may correspond to several grasps as it happens in reality (recall the Section above), the relationship between the visual and the motor features is highly non-functional and it is in general hard, if not pointless, to model it using a single NN. Richmond's idea was to model a *probability distribution* rather than a functional map; here we follow a somewhat more naive approach: we define an "archetipal grasp" related to the specific object observed. In the case of an object that can be grasped in only one way (for instance "pig"), then the archetipal grasp will correspond to it. In case of an object graspable in different ways, then the archetipal grasp will correspond to an "average" grasp between those possible. We expect that this reconstructed grasp will have a positive effect on the overall performance of our object recognition system; at the same time we hope that such representation won't get messed up with other ones since the output space is also rather high-dimensional.

C. Learning the Visuo-Motor Classifier

Our goal in classification is to demonstrate that the motor information is useful in object learning and recognition. Specifically, we want to show that integrating it with the visual information can produce a better performance, namely higher classificaton rate and robusteness.

To this end we consider both the visual and the motor features labelled in terms of objects. The idea is that a classifier should predict which is the inspected object when the input is visual, motor or the combination of the two. 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 [19], [20]. 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

[2] A force-sensing resistor was used to determine the instant of contact.

of different classifiers each dealing with one feature: the hypotheses produced are then combined together to achieve a consensus decision.

To learn the Visuo-Motor Classifier here we decided to implement these three strategies in an SVM-based framework, and to evaluate experimentally their suitability for the task. Specifically, we used the Discriminative Accumulation Scheme (DAS, [21]) for the high-level, and the Multi-Cue Kernel (MCK, [22]) for the mid-level integration. As already mentioned, the low-level integration consisted only in the feature concatenation, with the new vector fed to a standard SVM.

DAS. It 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 new 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 $D_j(p)$, $p = 1 \dots P$. 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:

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

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

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

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

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.

IV. Experimental results

This section reports the experimental validation of our model. We begin by testing the model on real 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 evaluating the quality of the reconstructed archetypal 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 our VMM to generate an archetypal grasp as input for the classifier (section IV-C).

A. Results with real motor data

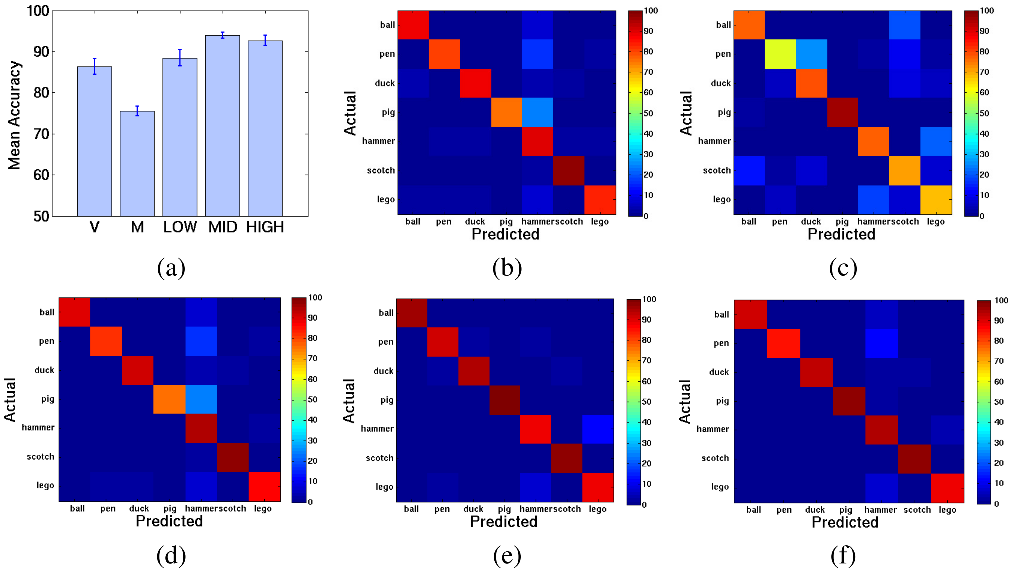
The first set of experiments was conducted on real data, namely motor data registered by the users when grasping the objects, and the corresponding images. Our goal here was to show the advantage in recognition achieved by modelling the object on both modalities, and at the same time compare the three possible joint modelling strategies presented in section III-C, to chose the best one.

Experiments were performed considering the whole set of 5200 data and choosing randomly 130 samples for training and 2600 samples for testing. The random extraction was repeated defining 10 different splits and the classification was executed using SVM, one-vs-all multiclass extension. We used the Gaussian Kernel for the visual and motor modalities, both when considered separately and in the integration approach (two Gaussian Kernels combined in MCK). The best learning parameters were selected through cross validation.

Figure 4-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, HIGH). We see in general that using the visual information we obtain better average performance ($86.37\% \pm 1.91\%$) than using the motor one ($75.53\% \pm 1.22\%$), and that their integration is clearly beneficial. The mid-level integration produces the best result ($93.94\% \pm 0.77\%$): the gain in accuracy between MCK and only using visual features is 7.57% (difference in accuracy evaluated per split and then averaged on the 10 splits). The second best result is obtained by using DAS ($92.65\% \pm 1.22\%$); we see that the difference in performance between DAS and MCK is not statistically significant, and therefore both are suitable candidates for the VMC module.

Figure 4-b, -f shows the confusion matrices obtained by the vision only classifier (Figure 4-b), by the motor only classifier (Figure 4-c) and by the three integration methods: low-level (Figure 4-d), MCK (Figure 4-e) and DAS (Figure 4-f). It is clear that the combination of the two modalities leads to considerable 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 information. The ball was grasped in two different ways (with a 'tripodal" and a "spherical" grasp) while the pig was manipulated only with the "cylindric" grasp, which was used just for this object. Thus, the grasp information is object-specific for the pig. This led to an impressive increase in performance when using MCK, as we achieved a 100% classification rate. Using visuo-motor information is beneficial also for the ball, for which we obtained a multi modal recognition rate of 96.5%. Analogous considerations can be done for the two other approaches, and are omitted here for space reasons.

From these experiments we can conclude that: (a) using a joint visual and motor object model leads to a very concrete advantage in performance during recognition, and (b) the MCK algorithm seems the most suitable for the joint



(a) (b) (c)

(d) (e) (f)

Fig. 4. (a) Classification mean accuracy of the seven objects averaged on the ten splits; (b) confusion matrix using visual features; (c) confusion matrix using motor features; (d) confusion matrix using the low level feature integration; (e) confusion matrix using the mid level feature integration; (f) confusion matrix using the high level feature integration.

modelling of the two modalities.

B. Evaluation of reconstructed data

We now turn to the evaluation of the archetypal grasps generated by the VMM. We learn a neural network for each object seen during training; this results here in seven specific VMMs. If an object can be grasped in only one way, the reconstructed motor data correspond to an estimate of this grasp type. If the possible grasps are more than one, the reconstructed motor data represent an estimate of the "average" grasp for that object.

To evaluate the goodness of the VMM in producing "archetypal grasps", we performed the following experiment: we divided the whole dataset in two halves (2600 data each), using one for training and the other for testing. Specifically we used the samples to:

(a) Train the neural networks and predict the motor feature vectors of the testing set, for each VMM associated to its specific object.
(b) Train a "grasp classifier" on the real motor information. The testing phase consisted in predicting the grasp label of the reconstructed motor vectors obtained from (a). We counted as an error every time the predicted grasp was not one of the possible grasps associated with the relative object.

We run the experiment on 10 splits of the whole dataset and we obtained an average error rate of 10.7%. This is significantly low with respect to a random grasp labelling (error rate of 63%). We can conclude that the reconstructed grasp information is coherent with the real one, and therefore we expect that the archetypal grasp will turn out to be an informative feature when used for classification.

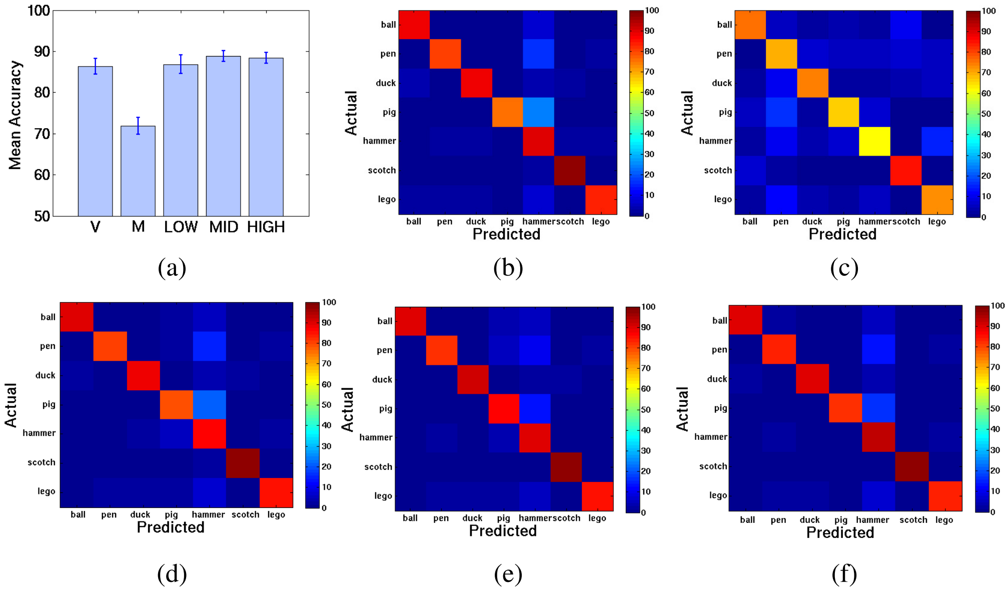
C. Results with reconstructed data

The most frequent case is of course that of an agent seeing an object without grasping it. In that case, our approach still permits to take advantage of the VMC, using as motor input the archetypal grasp generated by the VMM. More in detail, the system performs three steps (see Figure 5 for a schematic representation):

1) We extract the visual features from the object's view. Based upon it, we generate an hypothesis on the label of the object using only visual data.
2) The hypothesis is used to choose the appropriate VMM.
3) The VMM reconstructs the grasp associated with the object. This motor feature is then used, alone or jointly with the visual feature, to recognize the object.

We evaluated this strategy by repeating the experiments described in Section IV-A, using as input only visual data. For the implementation of the first step described above, we used the vision only classifier trained on real data (see Section IV-A).

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). We see that using the archetypal grasps as motor information, the performance of the motor only classifier slightly decreases compared to what we obtained on real motor features: $71.90\% \pm 2.06\%$ obtained with the archetypal grasps, as opposed to the $75.53\% \pm 1.22\%$ obtained using real motor features. 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 MCK (88.77%± 1.29%), closely followed by DAS (88.38%± 1.31%).



(a) (b) (c)

(d) (e) (f)

Fig. 6. (a) Classification mean accuracy of the seven objects averaged on the ten splits; (b) confusion matrix using visual features; (c) confusion matrix using motor features; (d) confusion matrix using the low-level feature integration; (e) confusion matrix using the mid-level feature integration; (f) confusion matrix using the high-level feature integration.

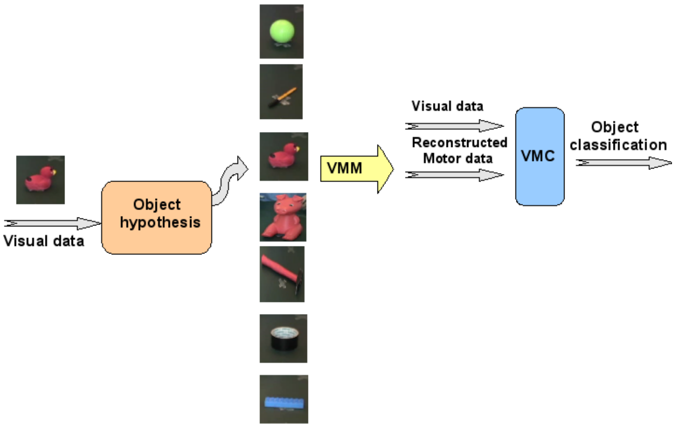


Fig. 5. A schematic representation of how the reconstructed motor features are used in the VMC.

Figure 6-b, -f shows the confusion matrices obtained by all classifiers, as reported in Section IV-A. We see that the results for the reconstructed motor data are in general lower than that obtained with the real ones (Figure 4-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. 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 MCK). As a last remark, we see once again that MCK 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. Conclusions and Future Work

This paper presented a theoretical framework for joint modeling of visual and motor data for multimodal object recognition. The key feature of our approach is the learning of a Visuo-Motor Map between the two modalities during training. The existence of this map makes it possible to benefit from the multimodal nature of the model even when the motor data is not perceived by the system. Experiments confirm the validity of our approach, showing a gain in performance of up to 7.6% and 2.4% when using both modalities, compared to results achieved using vision only.

The data upon which our experiments have been carried on are collected in the CONTACT Visuo-Motor Grasping dataBase (VMGdB), which we envision as a testbed and a benchmark for all researchers interested in investigating the nature of (human and robotic) grasping, and its ties to object recognition.

Future Work. The current implementation of the framework contains several simplifying assumptions, each corresponding to ongoing and future research directions:

1) *Dynamic of the data.* In this work we have neglected a lot of potentially useful information coming from the dynamics associated with the reaching phase, prior to grasping. This is well-known to carry substantial information about it [23], [24]. We plan to include the dynamic in the representation of motor and also visual features: indeed the dynamic changes in the object state associated with its manipulation are an importan cue on the object's identity [11], [12]. This research direction

will likely lead us to move from grasping postures to grasping actions, and therefore affordance-based object representations.

2) *Shape-based visual representations.* While here we used an appearance-based visual representation for the object, we are fully aware that this visual information is weakly correlated with the grasping and therefore makes the life of the mapping function much harder. We plan in the future to represent objects based on shape information. This will lead to visual information complementary to the grasp hand posture (the configuration of the hand at the moment of the grasp can be seen as a motor-based information regarding the shape of the object). We also expect that a shape-based visual representation will make it possible to build categorical object models based on their shape, i.e. graspability. This might lead to better defined VMMs, and it would greatly help in the case of large number objects.

3) *Learning of the Visuo-Motor Map.* Experimental results indicate that the "real" motor features, that is, the grasping hand postures as recorded by the data glove, contain much more information than the visual features alone. Therefore, if one were able to extract more (or better) motor information from the sight of an object, that is, to build a better VMM, the situation could improve further. One immediate direction for this line of research is that of abandoning the somewhat artificial notion of an archetypal grasp associated with an object, and start training a VMM in order for it to reconstruct a *probability distribution* over grasps. This would correspond to enhancing an object model with a possible set of grasps, rather than with one grasp only; a further, important step toward the re-definition of an object in terms of its affordances. We are currently studying an extended version of the VMM architecture based on vector-valued regression to explicitly take into account the many-to-many relationship between objects and grasps. In this case, instead of learning the map between an object and an average grasp, we would learn the map between an object and a vector of the possible grasps. At run time this would allow us to associate the most probable grasp to the object under consideration.

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

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

[6] M. Umiltá, 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.

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

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

[14] 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.

[15] 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. of ICIAP*, 2009.

[16] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60.

[17] 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.

[18] 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.

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

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.