Object Recognition using Visuo-Affordance Maps

Arjan Gijsberts, Tatiana Tommasi, Giorgio Metta and Barbara Caputo

Abstract— One of the major challenges in developing autonomous systems is to make them able to recognize and categorize objects robustly. However, the appearance-based algorithms that are widely employed for robot perception do not explore the functionality of objects, described in terms of their affordances. These affordances (e.g., manipulation, grasping) are discriminative for object categories and are important cues for reliable robot performance in everyday environments.

In this paper, we propose a strategy for object recognition that integrates both visual appearance and grasp affordance features. Following previous work, we hypothesize that additional grasp information improves object recognition, even if we reconstruct the grasp modality from visual features using a mapping function. We considered two different representations for the grasp modality: (1) motor information of the hand posture while grasping and (2) a more general grasp affordance descriptor. Using a multi-modal classifier we show that having real grasp information significantly boost object recognition. This improvement is preserved, although to a lesser extent, if the grasp modality is reconstructed using the mapping function.

I. INTRODUCTION

The capability to recognize and categorize objects is one of the crucial competencies for autonomous agents operating in human-made environments. According to Gibson [1], 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 (i.e., the *affordances*) are a crucial part of the object definition itself. Results from neurophysiology seem to indicate that the human brain incorporates a similar strategy of associating objects with actions, as "canonical" visuomotor neurons presumably reconstruct motor information for grasping from visual information [2]. The vast majority of work in computer vision on object recognition and categorization, however, models objects starting from static images, training a model (in the case of generative approaches) or a classifier (in the case of discriminative approaches) on a very large image

A. Gijsberts and G. Metta are with the Department of Robotics, Brain and Cognitive Sciences, Italian Institute of Technology, Via Morego 30, 16163 Genoa, Italy {arjan.gijsberts,giorgio.metta}@iit.it

G. Metta is also with the Department of Communication, Computer, and System Sciences, Faculty of Engineering, University of Genoa, Viale F. Causa 13, 16145 Genoa, Italy

T. Tommasi and B. Caputo are with the Idiap Research Institute, Centre Du Parc, Rue Marconi 19, CH-1920 Martigny, Switzerland {ttommasi,bcaputo}@idiap.ch

Fig. 1. Images of three different mobile phones. We see that their visual appearance varies considerably, while they share the same functionality and manipulability.

database containing various instances of the object of interest. This approach is potentially incomplete and resource-consuming [3], [4]. Furthermore, it completely neglects the specific functionality of objects. This functionality, described in terms of affordances, is often far more informative than just the visual appearance of (manipulable) objects. Think for instance of a mobile phone: what you do with it (typing numbers, talking) defines it far better than how it looks (see Fig. 1).

Most of the work within the robotics community related to affordances has focused on predicting opportunities for interaction with an object. For instance, Paletta et al. use reinforcement learning to learn a causal relationship between visual cues and the associated anticipated interactions [5]. Additionally, Sun et al. propose a probabilistic graphical model that leverages visual object categorization for learning affordances [6]. An initial prediction of the object's category is subsequently used to predict its affordances, similar to earlier work by Fitzpatrick et al. [7]. Only few recent attempts exploit the relevance of affordances for object recognition. In the approach presented by Metta et al. [8], an agent observes its hand grasping an object from a first-person perspective and subsequently learns a mapping from visual perception to motor information using a neural network. Although limited in scope, their experiments indicate that this motor information indeed improves object recognition. More recently, Noceti et al. have confirmed on a larger dataset that it is possible to build joint models of the visual appearance of a specific object and the grasp type associated with that object [9]. In work by Griffith et al., a robot learns to disambiguate container and non-container objects by means of interactive exploration of affordances [10]. Gupta et al. instead present a Bayesian framework that unifies the inference processes involved in object categorization and localization, action understanding, and perception of object reaction [11]. The joint recognition of objects and actions is based on shape and motion, and the models take as input video data. Also Montesano et al. propose a probabilistic

model, although they consider the joint distribution of objects and actions with effects, rather than with localization [12]. Kjellstrom et al. consider objects as contextual information for recognizing manipulation actions and vice versa [13]. The action-object dependence is modeled with a factorial conditional random field with a hierarchical structure. In both approaches, affordances are demonstrated to the system by means of a human agent performing an action on an object. Furthermore, objects and their affordances are first modeled separately, and are combined together in a second step.

Here we further explore the approaches by Metta and Noceti et al. [8], [9], and we propose a framework for mapping from visual appearance to grasp affordances that (a) allows for a many-to-many correspondence between grasp types and objects and (b) can effectively be used to build object classification algorithms exploiting the graspability of each object to enhance recognition. Although a mapping from visual appearance to grasp affordances is useful in its own right [14], [15], [6], our particular interest is in pushing the current paradigm of object recognition from purely appearance based to one based on a mixture of visual appearance and affordances.

Specifically, we propose a multi-modal classifier that integrates both visual and grasp affordance modalities. As it is infeasible for an agent to attempt multiple grasps on each object of interest, a limited set of actual grasp actions is used to learn a mapping from visual appearance to possible grasps. The auxiliary grasp modality is therefore a reconstructed (or imaginary) variant of the actual grasps, as predicted by the mapping function from the visual features. We consider two separate variants to represent the grasp affordances, namely (1) motor information of the hand posture while in grasping position, and (2) an abstract representation of the set of possible grasp types. The latter variant has the advantage that it is able to capture the many-to-many relationship between grasps and objects. For both variants, experimental validation of our framework shows that having real information regarding the grasp affordances significantly boosts object recognition. This improvement is preserved, although to a lesser extent, if the grasp modality is reconstructed using the mapping function, possibly indicating that the degree of improvement is related to the quality of the reconstruction.

The rest of the paper is organized as follows: our proposed framework for object recognition using visual features and predicted affordances is described in Section II. The experimental setup is described in Section III, followed by the results in Section IV. Finally, we give concluding remarks in Section V.

II. FRAMEWORK

Our hypothesis is that object recognition can be improved by including information regarding affordances together with the visual appearance of the object. In general, this principle requires a multi-modal classifier that integrates information from a primary and an auxiliary modality. Here we assume that there is a correlation between percepts of both modalities and that the primary modality is always available, whereas

the auxiliary modality may only be available at selected instances. This general setting loosely resembles perception of most animals and advanced robots, which use multiple and possibly redundant sensory modalities (visual, auditory, tactile, etc.). Although we consider only a single auxiliary modality, extending the framework to multiple auxiliary modalities is straightforward.

In the following, we restrict ourselves to the visual perception and grasp affordances as primary and auxiliary modality, respectively. Grasp affordances are defined as the set of possible grasp actions that can reasonably be performed on an object. We consider two distinct representations for the grasp affordances. In the first, we describe a particular grasp in terms of the hand posture of the agent performing a grasping action. Representing a grasp directly in terms of motor information is straightforward and avoids the necessity for grasp categorization. A schematic overview of the framework in case of grasp motor information as auxiliary modality is shown in Fig. 2a.

There are several issues with representing grasp affordances in terms of motor information. Firstly, many objects can be grasped in multiple ways, depending for instance on the intended use of the agent for the object. Consequently, we would need to store a variable number of motor representations for each object, or describe the grasps in terms of a probability distribution function in motor space. We would need a large number of training samples, however, to obtain reliable statistics for the latter approach. Furthermore, the motor representation of a grasp is strongly dependent on the exact embodiment of the agent and may vary from trial to trial. It is therefore difficult, if not impossible, to transfer these representations onto other agents.

A second approach, which solves both these issues, consists in representing grasp affordances as the set of grasp types that can be used with an object. In effect, we thus model grasp affordances as an abstraction of multiple instances of motor data[1], as shown in Fig. 2b. A set of very similar hand postures are thus regarded as a grasp type, and we then record whether or not these grasp types can be used on an object. This representation is justified by the observation that humans use a limited set of (parameterized) grasp types [16]. However, by representing grasps in this manner we require the agent to be aware of the type of grasp it uses or to have other means of categorizing its grasps.

A. Reconstructing Auxiliary Modalities

In our framework, we do not assume that the auxiliary modality is always available. It is evident why we do not want to impose such a restriction in the context of grasp affordances, as it would require an agent to actually attempt to grasp each object prior to classification. Instead, we train a mapping function that reconstructs (or predicts) the auxiliary modality from the primary modality. In terms of visual and affordance modalities, we can imagine the agent visually

---

[1]We do not explore the relationship between abstract grasp affordances and motor information.
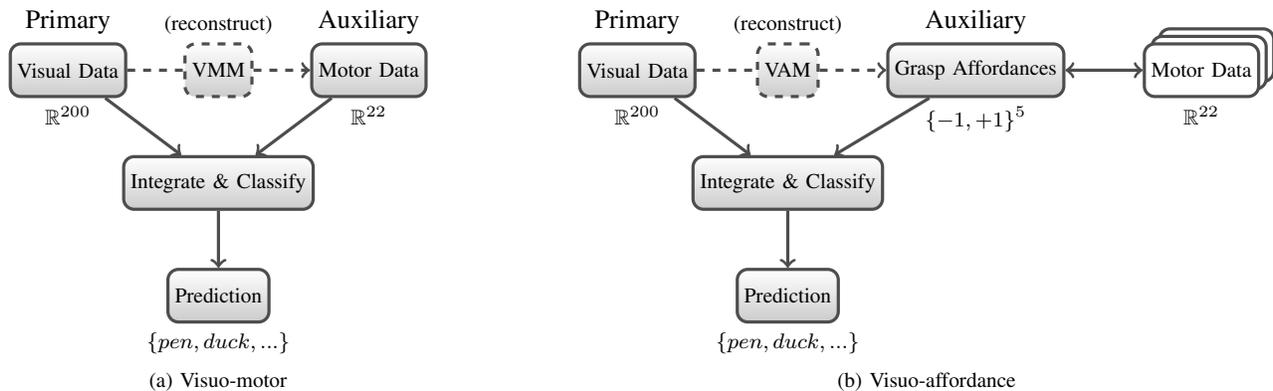
Fig. 2. Schematic overview of the multi-modal classification framework. The auxiliary modality is represented directly as (a) motor information of the hand posture of a grasp, or alternatively, by representing (b) grasp affordances as a set of possible grasp types. The mapping functions for each variant, respectively the Visuo-Motor Map (VMM) and Visuo-Affordance Map (VAM), are used to reconstruct the auxiliary modality from the primary modality. The labels denoting the input and output spaces are related to the experimental setup as described in Section III.

perceiving known objects (i.e., the object identity is known to the agent) without grasping them, or alternatively the agent both seeing and grasping objects for which it does not know the identity[2]. In our framework, the multi-modal classifier integrates information from both these scenarios using a reconstruction of the auxiliary modality (cf. Fig. 2).

III. EXPERIMENTAL SETUP

An identical set of experiments has been performed to validate each of the two variants of our theoretical framework. For the first variant, information of the visual appearance of an object is combined with the hand posture of a human subject while grasping the object. We investigate to which extent this additional motor information improves object recognition. A possible difficulty in this experiment is that a single object may allow for multiple grasps and, moreover, different subjects may perform the same type of grasp differently. In the second set of experiments, we combine the visual appearance of an object with a vector representing its affordances. Contrary to some earlier work (e.g., Metta et al. [8]), the images only contain the object itself, without showing the grasping action. The affordance modality is thus estimated purely from the visual appearance of the object, rather than the visual appearance of the grasping hand. The affordances considered in this study are limited to the possible types of grasp that can reasonably be used with the particular object. We expect to see that these grasp affordances are highly discriminative for the object class and therefore improve object recognition significantly.

All experiments were repeated on 10 stratified random splits of the data and with a varying number of training samples for both the object classifier and the mapping function. Furthermore, we consider the case that real motor and affordance information is available, and compare it to the case that this information is reconstructed from the visual features using separately trained visuo-motor and visuo-affordance maps. Regularized Least Squares has been used

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

TABLE I

THE 13 POSSIBLE OBJECT-GRASP COMBINATIONS.

for the object classifiers and for both types of reconstructing maps. In the following, we describe the database used and the extracted visual and motor features. Moreover, we briefly review the Regularized Least Squares algorithm for regression and classification and explain in detail how it has been used in the experiments.

A. The Visuo-Motor Grasping Database

The Visuo-Motor Grasping database (VMGdb[3]) consists of visual and motor data collected on 7 objects and 5 grasp types (cf. Fig. 3). On the basis of their affordances, each object allows different types of grasps, defining the many-to-many relationship as reported in Table I. The database contains 20 human subjects performing 20 trials for each of the 13 possible object-grasp types, yielding a total of 5200 samples. Visual information consists of video sequences of the performed grasping actions, acquired laterally with focus on the object. Of these video sequences, we selected frames showing the object in different poses and without any occlusions. Motor information was collected using a CyberGlove [17], which has sensors measuring the hand posture and a force-detector resistor on the fingertips that is used to determine the instant of contact with the object.

B. Visual, Motor and Affordance Features

The visual appearance of the objects is described in terms of SIFT features [18], which were extracted from a single

[2]In the second variant of our framework, the agent would always have to be aware of the object identity in order to update the possible grasp affordances for the object.

[3]We gratefully thank the LIRA-lab of the University of Genoa for having made their database available to us. For more information about the database, please contact giorgio.metta@iit.it.

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.

frame containing a lateral view of the object during each trial. We used a bag-of-words approach with a vocabulary of 200 elements, following the same strategy as used by Noceti et al. [9]. The resulting features are preprocessed independently by scaling them to a range of $[0, 1]$ by dividing by the maximum value present in the classifier training set. The CyberGlove returns 22 joint angles of the subject's hand posture, measured in a 8 bit resolution. These sensor measurements are used as motor information, after standardizing each of these input features to zero mean and unit standard deviation.

The grasp affordance information follows the relationships in Table I. It is important to note that these features do not simply report the single type of grasp that is performed in a given sample, but rather all of the possible grasps that were performed on the object under consideration. A bitwise encoding (i.e., $\{-1, +1\}^5$) is used to indicate whether or not each of the five grasps is associated with the object.

C. Regularized Least Squares

The classifier, the visuo-motor and visuo-affordance maps require an estimation function that predicts an output based on an input feature vector. Given a set $\mathcal{S}$ of m input-output pairs $\{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^n, y_i \in \mathcal{Y}\}_{i=1}^m$, we would like to construct a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ that accurately predicts an output for any given input $\mathbf{x}$. Regularized Least Squares (RLS) is an algorithm that can learn these functions both in case of binary classification problems (i.e., $\mathcal{Y} = \{-1, +1\}$, RLSC [19]), as well as regression problems (i.e., $\mathcal{Y} \subseteq \mathbb{R}$, ridge regression [20]). Furthermore, incorporating kernel functions allows the method to be used on non-linear problems, much like Support Vector Machines (SVM, [21]). In RLS, we optimize the functional

$$\frac{1}{m} \sum_{i=1}^m \|f(\mathbf{x}_i) - y_i\|^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad, \tag{1}$$

where $\mathcal{H}$ is the Reproducing Kernel Hilbert Space corresponding to the kernel k. The first term in (1) minimizes the squared errors on the training data, while the second term penalizes the complexity of the function by minimizing the norm of the function f in $\mathcal{H}$. The balance between minimizing these two terms is regularized by a constant hyperparameter $\lambda \geq 0$.

The representer theorem states that we can describe the solution of (1) in the form $f(\mathbf{x}) = \sum_{i=1}^m c_i k(\mathbf{x}, \mathbf{x}_i)$ [21], where we apply an additional sign function to the output in the case of binary classification problems. For RLS, the optimal set of coefficients $\mathbf{c}$ can be obtained by solving a system of linear equations

$$(\mathbf{K} + \lambda m \mathbf{I}) \mathbf{c} = \mathbf{y} \quad, \tag{2}$$

where $\mathbf{K}$ is an $m \times m$ matrix such that $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{I}$ is an $m \times m$ identity matrix.

The selection of a kernel function and the hyperparameters is crucial to the generalization performance of any kernel method. In our experiments, we consider the standard RBF kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$ [21]. The hyperparameters λ and γ are optimized using grid search, where $\lambda \in \{2^{-40}, 2^{-36}, \dots, 2^0, 2^4\}$ and $\gamma \in \{2^{-10}, 2^{-9}, \dots, 2^4, 2^5\}$.

D. Object Classifier

The output of the object classifier is one of the 7 object classes. We emulate multi-class classification using the one-vs-all classification scheme [22]. This scheme can be implemented efficiently using RLS, as we can solve the 7 systems of linear equations in (2) concurrently (for a given hyperparameter configuration) by writing an $m \times 7$ coefficient matrix $\mathbf{C}$ and output matrix $\mathbf{Y}$.

We perform experiments with a varying number of training samples for the classifier. For each of the 10 splits, we reserve 2600 samples for testing and use random subsets $\mathcal{S}_c$ of sizes $|\mathcal{S}_c| \in \{26, 52, 78, 104, 130\}$ of the remaining samples to train the classifier. These subsets are chosen such that they contain an equal amount of samples for each of the 13 object-grasp types.

Integration of visual features with either motor or affordance features is done using the multi-cue kernel approach [23]. In this approach, multiple feature types are integrated by means of a linear combination of kernels, such that each kernel operates only on a single feature type. In our setting, we thus obtain the composite kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = w^p k^p\left(\mathbf{x}_i^p, \mathbf{x}_j^p\right) + w^a k^a\left(\mathbf{x}_i^a, \mathbf{x}_j^a\right) \quad,$$

where $\mathbf{x}^p$ are the primary visual features and $\mathbf{x}^a$ are the auxiliary motor or affordance features. Further, we chose

both k^v and k^a to be the RBF kernel function, and set $w_v \in \{0.05, 0.15, \ldots, 0.85, 0.95\}$ and $w_a = 1 - w_v$. One of the advantages of the multi-cue kernel over simple concatenation of the features is that we can independently tune the kernel parameter γ for each set of features. Preliminary experiments on our problem indicate that this results in higher generalization performance, a finding that is supported both theoretically and empirically in related literature [23].

E. Reconstruction of Motor and Affordance Information

RLS is also used for the visuo-motor and visuo-affordance maps. In the former case, the learning problem is to reconstruct the hand posture of a grasp from the visual representation of an object. Learning this mapping is problematic, as the visual appearance of an object does not contain information about which of the possible grasps will be used in a particular sample. Furthermore, the exact hand posture for any given grasp may depend on the subject and trial. RLS will therefore learn a mapping from the visual appearance of an object to a prototypical grasp, which is approximately the average of all the possible grasp types for the object as performed by all the subjects. The visuo-affordance mapping is not affected by these variances, as it stores an abstract representation of the possible grasp types. Further, the ambiguity regarding the multiple possible grasp types is resolved by representing all possible grasp affordances.

The mappings are trained on a random subset $\mathcal{S}_m$ of the training samples, ensuring that the classifier training set $\mathcal{S}_c$ and $\mathcal{S}_m$ are disjoint. An equal number of training samples is used to train both the classifier and the mappings (i.e., $|\mathcal{S}_c| = |\mathcal{S}_m|$). For the visuo-affordance mapping, the samples in $\mathcal{S}_m$ are also used to find the set of possible grasp types for each object. The training samples that are not used for training either the classifier or the mapping are used as validation set for hyperparameter tuning of the mapping function. The hyperparameters are restricted to be equal for all outputs (22 for visuo-motor, 5 for visuo-affordance), such that these can be learned at once rather than independently. After finding the optimal hyperparameters, the mapping is used to reconstruct the motor and affordance features of both the classifier training set and the test set.

IV. RESULTS

The experimental results are separated according to the two variants of our theoretical framework, as described in Section II. For both variants, we compare the results obtained using real motor and affordance data with those obtained using reconstructed data. The goals of these experiments is to demonstrate (1) that including additional motor or affordance information improves object recognition, and (2) that an improvement persists even if this information is reconstructed from the visual features using a mapping function.

A. Results with Visuo-Motor Reconstruction

In the first set of experiments, we evaluate classifiers using visual features, motor features and an integration of visual and motor features (i.e., visuo-motor). For the cases involving motor features, we consider both the real motor information and the motor information as reconstructed using the visuo-motor mapping. Fig. 4a shows the balanced classification error of these classifiers with an increasing number of samples. These results show that integrating visual features with real motor information of a grasp outperforms all other methods. Even in case of reconstructed motor data, the integrated classifier still outperforms a classifier that only uses visual information, albeit the improvement is less substantial. A sign test indicates that this improvement is nonetheless statistically significant in the case of 52 or more training samples ($p \leq 0.0215$) [24]. Interestingly, the generalization performance of the classifier trained only with reconstructed motor features converges faster (with respect to the number of training samples) than the classifier trained with real motor features. The most likely explanation for this behavior is that the reconstruction actually resolves the ambiguities and acts as a noise filter, therefore aiding the final classifier in constructing a sensible prediction function. Unfortunately, this effect is not significantly noticeable for the integrated classifier that uses both visual and motor features.

B. Results with Visuo-Affordance Reconstruction

It is evident from Table I that the grasp affordances are highly discriminative for the object class for our database. Moreover, the representation with possible affordances is not affected by ambiguities of having multiple grasp types associated with an object, or by the exact hand posture used by a subject to perform a grasp. We therefore expect the classifiers that use affordance information to outperform those using motor information. In case of real affordance features, this claim is supported by the results for the visuo-affordance classifier, as shown in Fig. 4b. The weighted classification error based on affordance data alone is very close to $1/7$, regardless of the number of samples. This is due to the classifier being unable to distinguish between a "pen" and a "duck", for which the possible grasp affordances are identical (cf. Table I). Integration with visual features helps to resolve this ambiguity, such that the recognition rate for the multi-modal classifier is nearly 100% for all of the tested training set sizes.

The results with reconstructed grasp affordances, however, are again less profound. This indicates that the mapping from visual features to grasp features is difficult. This is not surprising, as there is a close relationship between the grasp affordances and object identity, and it is thus likely that both problems are of similar complexity. The improvement of the reconstructed visuo-affordance classifier over the visual classifier, however, is statistically significant for 26, 52 or 78 training samples ($p \leq 0.0215$), and even more so for 104 or more samples ($p \leq 0.002$). This result confirms that there is a benefit of including a reconstructed auxiliary modality. When using 130 training samples, for instance, the visuo-affordance classifier achieves a weighted classification error of $7.6\% \pm 1.4\%$, which compares favourably to the
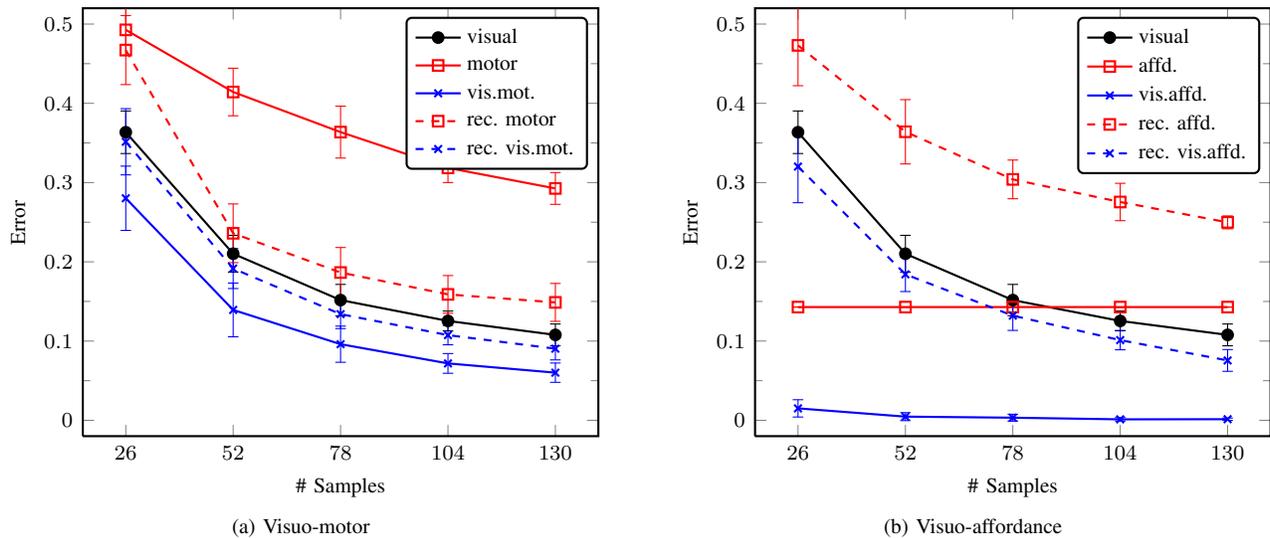
(a) Visuo-motor

(b) Visuo-affordance

Fig. 4. Balanced classification error (i.e., normalized for the number of samples in each class) for classification using (a) visual and motor data and (b) visual and affordance data, while varying the size of the training set. Both plots show performance of each of the two modalities individually, as well as performance of the multi-modal classifier using real (solid lines) and reconstructed features (dashed lines) for the auxiliary modality.
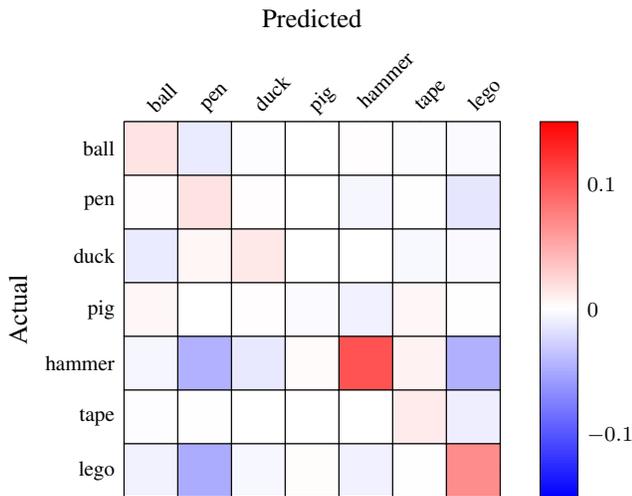

Fig. 5. Difference between the confusion matrix of the multi-modal visuo-affordance classifier with reconstructed affordance features and the visual classifier for 130 training samples. Red elements on the diagonal mark an improvement in correct classifications, whereas blue elements off the diagonal indicate a reduction of misclassifications.

$10.8\% \pm 1.4\%$ error of the visual only classifier. The benefit becomes more evident in Fig. 5, which depicts the difference between the confusion matrices in both these cases. We can verify that the additional affordance data, even though reconstructed, helps to disambiguate visually similar objects (cf. "lego" and "pen", "hammer" and "pen", or "hammer" and "lego"). When comparing the reconstructed visuo-motor classifier with the reconstructed visuo-affordance classifier, we can observe that the latter performs slightly better, although this difference is not statistically significant.

C. Discussion

One may argue that, by reconstructing the auxiliary modality from the primary one, we do not actually add any new

information for the classifier. Therefore, reconstructed motor or affordance modalities should not have any beneficial effect on the classification performance, contradicting the results just presented. Though compelling, this argument only holds in case of an infinite number of training samples when a consistent classifier could potentially make perfect use of all available information in the visual data. With a limited number of samples, however, the additional set of extracted features allow the classifier to make efficient use of available training data. Another view is that the extra set of features – and hence the extra kernel – help to form an induced RKHS in which the data is more easily separable.

It is important to note that reconstruction is not guaranteed to improve classification; this will depend strongly on the type and representatation of the modalities. Trivially, both modalities have to contain relevant information for determining the output. If the auxiliary modality only contains noise, for instance, the reconstructed features may actually deteriorate classification performance. Furthermore, the modalities also need to be correlated, in order to be able to (partially) reconstruct the auxiliary modality from the primary modality. Similar requirements have been formed for reconstruction of missing features (i.e., imputation) in other fields, such as statistical matching [25] or sensor fusion [26].

Another notable observation is that the absolute improvement of the multi-modal classifier over the visual classifier remains relatively stable as the number of samples increases. Nonetheless, it can easily be confirmed that the performance of both classifiers will converge given infinite training samples. For our database, for example, the visual classifier obtains perfect classification with less than 2600 samples. The reason that convergence is not present in our results, is that we increase the number of training samples both for the reconstruction map and the classifier concurrently. We can observe in Fig. 4 that the quality of classification using only

the reconstruction features improves drastically. Therefore, even though the visual classifier gets better with an additional number of samples, the reconstructed features do so as well. The absolute improvement of the multi-modal classifier can consequently remain stable or even (slightly) improve. This behavior will diminish as soon as the reconstruction reaches it optimal performance, at which point the performance of the visual and multi-modal classifier will be guaranteed to converge.

V. CONCLUSIONS

In this paper, we hypothesize that object recognition can be improved by considering both appearance and functional aspects of objects. To this extent, we present a multi-modal framework that integrates information of both the visual appearance and the grasp affordances of objects. Two different representations of the grasp affordances were considered, namely the motor configuration of the hand posture while grasping and a more abstract representation of all possible grasp types that can be used with the object of interest. Our experiments show that this auxiliary modality is indeed discriminative for object recognition and significantly improves the recognition rate.

In realistic settings, however, the acting agent will not always have grasp information of objects available, as it would require the agent to grasp each object of interest. In order to relax this requirement, we propose to predict the grasp affordances from the visual appearance of an object using a mapping function. Besides the fact that such a mapping function is useful for acting agents, we expect it to extract functionality related features from the visual appearance that will be discriminative for the classification problem. Experiments show that object recognition improves when integrating visual appearance features with the reconstructed grasp affordances, although to a lesser extent than using real grasp information.

Future work will investigate possible ways to make the mapping between different modalities more robust by improving the quality of the reconstruction. The abstract affordance description seems very promising and we plan further verification of this approach using a larger database with more objects and grasp types, and possibly considering more affordances than just grasps. A possible difficulty in the current affordance representation is that the grasp type must be known, which could potentially be avoided by clustering grasps in motor space.

REFERENCES

[1] J. J. Gibson, *The Theory of Affordances*. Lawrence Erlbaum, 1977.
[2] L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti, "Visuomotor neurons: Ambiguity of the discharge or 'motor' perception?" *International Journal of Psychophysiology*, vol. 35, pp. 165–177, March 2000.
[3] G. Griffin and D. Perona, "Learning and using taxonomies for fast visual categorization," in *CVPR 2008: IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2008.
[4] 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.
[5] L. Paletta, G. Fritz, F. Kintzler, J. Irran, and G. Dorffner, "Perception and developmental learning of affordances in autonomous robots," in *KI '07: Proceedings of the 30th annual German conference on Advances in Artificial Intelligence*. Springer-Verlag, 2007, pp. 235–250.
[6] J. Sun, J. L. Moore, A. Bobick, and J. M. Rehg, "Learning visual object categories for robot affordance prediction," *The International Journal of Robotics Research*, vol. 29, no. 2-3, pp. 174–197, February 2010.
[7] P. M. Fitzpatrick, G. Metta, L. Natale, A. Rao, and G. Sandini, "Learning about objects through action -initial steps towards artificial cognition," in *ICRA 2003: Proceedings of the 2003 IEEE International Conference on Robotics and Automation*, September 2003, pp. 3140–3145.
[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] 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 *ICIAP '09: Proceedings of the 15th International Conference on Image Analysis and Processing*. Springer-Verlag, 2009, pp. 239–248.
[10] S. Griffith, J. Sinapov, M. Miller, and A. Stoytchev, "Toward interactive learning of object categories by a robot: A case study with container and non-container objects," in *DEVLRN '09: Proceedings of the 2009 IEEE 8th International Conference on Development and Learning*. IEEE Computer Society, 2009, pp. 1–6.
[11] A. Gupta and L. S. Davis, "Objects in action: An approach for combining action understanding and object perception," in *CVPR 2007: IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2007.
[12] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: From sensory-motor coordination to imitation," *IEEE Transactions on Robotics*, vol. 24, no. 1, pp. 15–26, 2008.
[13] H. Kjellström, J. Romero, D. Martínez, and D. Kragić, "Simultaneous visual recognition of manipulation actions and manipulated objects," in *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*. Springer-Verlag, 2008, pp. 336–349.
[14] M. Lopes and J. Santos-Victor, "Visual learning by imitation with motor representations," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 35, no. 3, pp. 438–449, 2005.
[15] D. Kraft, R. Detry, N. Pugeault, E. Baseski, J. H. Piater, and N. Krüger, "Learning objects and grasp affordances through autonomous exploration." in *ICVS 2009: 7th International Conference on Computer Vision Systems*, vol. 5815. Springer, October 2009, pp. 235–244.
[16] M. R. Cutkosky and R. D. Howe, "Human grasp choice and robotic grasp analysis," *Dextrous robot hands*, pp. 5–31, 1990.
[17] *CyberGlove Reference Manual*, Virtual Technologies, Inc., 2175 Park Blvd., Palo Alto (CA), USA, August 1998.
[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[19] R. Rifkin, G. Yeo, and T. Poggio, "Regularized least squares classification," in *Advances in Learning Theory: Methods, Model and Applications*, vol. 190. VIOS Press, 2003, pp. 131–154.
[20] C. Saunders, A. Gammerman, and V. Vovk, "Ridge regression learning algorithm in dual variables," in *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 1998, pp. 515–521.
[21] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.
[22] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *Journal of Machine Learning Research*, vol. 5, pp. 101–141, 2004.
[23] T. Tommasi, F. Orabona, and B. Caputo, "Discriminative cue integration for medical image annotation," *Pattern Recognition Letters*, vol. 29, no. 15, pp. 1996–2002, 2008.
[24] J. D. Gibbons, *Nonparametric Statistical Inference*. New York, USA: McGraw-Hill, 1970.
[25] M. D'Orazio, M. D. Zio, and M. Scanu, *Statistical Matching: Theory and Practice (Wiley Series in Survey Methodology)*. John Wiley & Sons, 2006.
[26] D. L. Hall and S. A. H. McMullen, *Mathematical Techniques in Multisensor Data Fusion (Artech House Information Warfare Library)*. Norwood, MA, USA: Artech House, Inc., 2004.