ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

Learning how to grasp objects

Annalisa Barla[1] and Luca Baldassarre[1,2] and Nicoletta Noceti[1] and Francesca Odone[1]

(1) DISI and (2) DIFI
Università degli Studi di Genova
via Dodecaneso 35, Genova - Italy

Abstract. This paper deals with the problem of estimating an appropriate hand posture to grasp an object, from 2D object's visual cues in a many-to-many *(objects,grasp)* configuration. A statistical learning protocol implementing vector-valued regression is adopted for both classifying the most likely grasp type and estimating the hand posture. An extensive experimental evaluation on a publicly available dataset of visuo-motor data reports very promising results and encourages further investigations.

1 Introduction: state of the art

This paper presents a machine learning approach to predict an appropriate hand posture to grasp an object from two-dimensional visual cues. First a simple description of the object appearance, tolerant to 3D object rotation in the 3D world, is extracted from an image of an unknown object. Second, without explicitly classifying the object, we associate it its most likely grasp type. Third, we estimate a measurements vector describing the hand position for the selected grasp most appropriate to the specific object. This final step implicitly embeds the notion of object affordance, defined as a quality of an object that allows an individual to perform an action. The whole procedure is data-driven and, from the algorithmic stand-point, it is based on a state-of-the-art method for vector-valued regression [1] that we use for both estimating the most probable grasp types and predicting an appropriate hand posture.

Grasping classification is a keystone of many robotics applications. Different methods have been proposed in the literature according to the amount of prior knowledge available and the input data at disposal. Focusing on methods that exploit visual information at some level, a rather common approach starts from the computation of a full 3D model of the object to be grasped, with a subsequent association of an appropriate grasp type. Machine learning methods have been applied to this setting — see for instance [2].A practical problem with this approach, otherwise effective, is that often a 3D model of the object is not available nor easy to compute. Methods based on 2D visual cues have been proposed [3, 4]. Grasp classification is a loose definition that may refer to associating a grasp type from a pre-defined taxonomy to an object, or may be based on the explicit estimation of measurements modeling the grasp (e.g., describing the relative angles between joints). Our data-driven approach is related to the latter choice: it allows us to learn a hand configuration appropriate to grasp a given object. A recent trend of robotic grasping relies on gathering some understanding on the models for human grasping. In this work we refer to

human grasp classification (i.e., our data are originated by grasping actions performed by humans) and discuss how a possible grasp appropriate for an object may be estimated before actual grasping occurs. If a mapping from human to robot grasping is available, the proposed work may be applied to human-oriented robotic grasp learning — see, for instance, [5].

The paper contributions are two-fold. First, a real world application of a recently proposed algorithm for vector-valued regression that we apply on two different levels of the abstraction process. Second, an effective grasp classification strategy, based on visual cues, that allows to predict an appropriate hand position before the actual grasping takes place. The perception-to-action-map we consider is many-to-many: different objects can be grasped in the same way and the same grasp can be applied to more than one object. This fact poses a serious problem when, given a visual instance of an object, we want to estimate the hand posture in order to grasp it, since we first have to determine which grasp type to apply. The approach adopted in a simplified one-to-one framework [6] to learn a vector valued model on all training examples will fail in this case. The learning model would average the hand postures associated to the same object, yielding a configuration that does not represent any actual grasp (although it could carry information on the object volume).

We deal with this problem with a two steps procedure: *first* we apply the vector-valued regression model to associate to a given image of an unknown object a set of possible grap types. Then the most likely grasp is used to activate a visuo-motor regression module trained on pairs (object, grasp) related to a specific grasp type. This module returns an estimate of the hand position for the given grasp type most appropriate for the (unknown) object under consideration. This final step is implicitly related to the object affordance, that is, the same

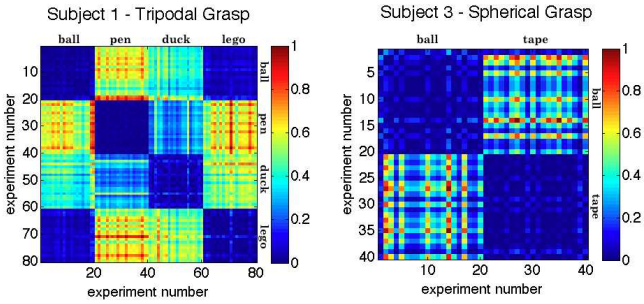

Fig. 1: Distance matrices between hand postures specific to a volunteer/subject. For a given (subject, grasp type) pair the images show the color-coded normalized distances between hand postures for different objects and repetitions. The block structure of the matrices indicates the distinctive affordances of the objects.

grasp type (say, tripodal) applied to different objects will originate different hand positions, closely related to the object's size and consistency, see Figure 1.

ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

The experimental analysis, based on a recently published multi-modal database of grasping actions, the VMGdB [7], shows that *(i)* the results obtained in the grasp type classification phase are very satisfactory, even with rather poor visual representations; *(ii)* preliminary results on the final regression phase speak in favour of the proposed strategy, but a more principled error evaluation is needed. Indeed, adopting conventional distance measures (e.g., the Euclidean distance) to estimate the similarity between real and estimated grasps does not take into consideration both the fact that, given a grasp, one or more fingers could be at rest (or, else, could be in any position without affecting the effectiveness of the grasp), as well as the presence of possible correlations among fingers. Ongoing research is addressing this issue.

2 Experimental Set up and data preparation

The set-up we refer to considers grasping actions performed by humans, and exploits multi-modal information: the object to be grasped is observed by a video-camera that registers the object appearance, while the grasping action is measured by a sensorised glove worn by the performing actor. The input data we use are about grasping actions performed by 20 human volunteers, grasping 7 different objects in various ways from a set of 5 possible grasp types. Table 1 reports the 13 *(object, grasp)* actions included in the dataset. Each volunteer repeats a given *(object, grasp)* action 20 times. Thus, the whole dataset contains 5200 grasping actions, where each pair is associated to 400 examples[1].

	Tripodal	Spherical	Pinch	Cylindrical	Flat
BALL	400	400	-	-	-
PEN	400	-	400	-	-
DUCK	400	-	400	-	-
PIG	-	-	-	400	-
HAMMER	-	-	-	-	400
TAPE	400	400	400	-	-
LEGO	-	-	400	-	400

Table 1: The (object,grasp) pairs included in the VMGdB [7] (see text).

The representation of visual information follows the solution proposed in [6]: in each image a set of keypoints is randomly sampled, every keypoint is represented with a SIFT descriptor. The keypoints are then clustered and a visual vocabulary is built. All images are divided in 4 quadrants and each quadrant represented with respect to the vocabulary, with a nearest neighbour approach. A frequency histogram of the visual features of each quadrant is built and, finally, the 4 histograms are concatenated, obtaining the final representation of the image. We set the size of the vocabulary equal to 20 words. For what con-

[1] A detailed account of the data used is available [7]. Here we consider image frames extracted from the lateral view, cropped on the object region, and the 22-dimensional sensor measurements acquired by the CyberGlove measuring the angles of the hand joints.

cerns motor information we adopt a simple normalization of the measurements acquired by the CyberGlove, dividing each element by $22 * 255$.

Since the proposed approach is based on learning from examples, the available 5200 data are organized in training and test sets as follows. Each group of 400 grasp actions is divided in two, where 200 actions are used for testing. From the remaining 200 actions we draw with no repetitions 10 actions for 10 times. We thus obtain 10 different training sets of 130 examples, allowing us to check the dependence of the solution with respect to the specific data choice, and a test set of 2600 examples.

3 Statistical Learning Protocol

The learning system consists of a cascade of two different learning modules. Both parts rely on a vector-valued regularization approach [1], which showed to be very flexible as a regression tool as well as a classifier.

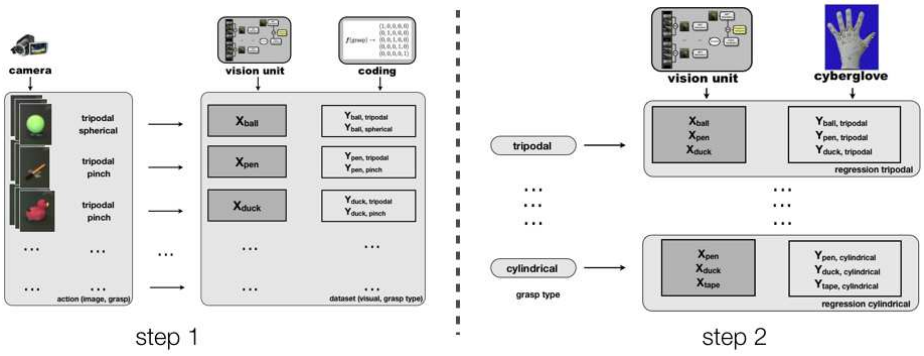

Fig. 2: A schema of the two statistical learning modules. On the left the multi-category classifier associating to a visual representation the most likely grasp types; on the right the 5 vector-valued regressors (one per grasp type) associating measurements of hand positions to visual representations.

Let us first describe some key points of regularization methods in the vector-valued case. Following the classical schema of statistical learning, we assume to be provided with a training set of input-output pairs $\{(\mathbf{x}_i, \mathbf{y}_i) : \mathbf{x}_i \in \mathbb{R}^p, \mathbf{y}_i \in \mathbb{R}^d\}_{i=1}^n$. Our aim is to estimate a function $\boldsymbol{f} : \mathbb{R}^p \rightarrow \mathbb{R}^d$, where p is the number of features representing the input images $\mathbf{x}_i$ and d is the dimension of the corresponding output label $\mathbf{y}_i$. Assuming that the data is sampled *i.i.d.* on $\mathbb{R}^p \times \mathbb{R}^d$ according to an unknown probability distribution $P(\mathbf{x}, \mathbf{y})$, ideally the best estimator minimizes the prediction error, measured by a loss function $V(\mathbf{y}, \boldsymbol{f}(\mathbf{x}))$, on all possible examples. Since P is unknown we can exploit the training data only. Regularized methods tackle the learning problem by finding the estimator that minimizes a functional composed of a data fit term and a penalty term, which is introduced to favour smoother solutions that do not overfit the training

ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

data. In [8] the vector-valued extension of the scalar Regularized Least Squares method was proposed, based on matrix-valued kernels that encode the similarities among the components f^ℓ of the vector-valued function $\boldsymbol{f}$. In particular we consider the minimization of the functional:

$$\frac{1}{n}\sum_{i=1}^{n}||\mathbf{y}_i - \boldsymbol{f}(\mathbf{x}_i)||_d^2 + \lambda||\boldsymbol{f}||_K^2 \tag{1}$$

in a Reproducing Kernel Hilbert Space (RKHS) of vector valued functions, defined by a kernel function K. The second term in (1) represents the *complexity* of the function $\boldsymbol{f}$ and the regularizing parameter λ balances the amount of error we allow on the training data and the smoothness of the desired estimator. The representer theorem [9, 8] guarantees that the solution of (1) can always be written as: $\boldsymbol{f}(\mathbf{x}) = \sum_{i=1}^{n} K(\mathbf{x}, \mathbf{x}_i)\mathbf{c}_i$, where the coefficients $\mathbf{c}_i$ depend on the data, on the kernel choice and on the regularization parameter λ. The minimization of (1) is known as Regularized Least Squares (RLS) and consists in inverting a matrix of size $nd \times nd$. RLS is a specific instance of a larger class of regularized kernel methods [10] extended to the vector case in [1]. More specifically — Fig. 3 — the first module consists of a multi-category classifier. The multiclass problem is transformed into a vector-valued regression problem by assigning to the examples of each class a vector-valued coding. For instance, examples associated with grasp 1 (tripodal) are given the coding $(1, 0, 0, 0, 0)$. Given a new example, the classifier will estimate the probabilities associated to each grasp type [1] and return the most probable grasp type or the grasp types whose probability is greater than a fixed threshold. The second module consists of 5 vector-valued regressors, one for each grasp type. Each regressor is trained only on examples that correspond to its specific grasp type. In the testing phase, a new visual representation of a given object is associated to a grasp type by module 1, which in turn activates the corresponding regressor in module 2. The final outcome is a vector estimating the hand posture specific for that *(object,grasp)* pair.

4 Results

We tested the classification performance of the first module, by considering whether the most probable grasp type returned by the multi-category classifier is at least one of the possible grasp types associated to the given object. The average classification error over the 10 samplings of the training sets is $6.4\% \pm 1.5\%$. A more detailed assessment of the grasp classifier is presented in Figure 3. The r.o.c. curve on the left side is computed counting as *false negative* every grasp whose probability is lower than the fixed threshold. Conversely, for computing the r.o.c. curve on the right side, we consider a false negative only when none of the probabilities corresponding to the actual grasp types are greater than the threshold. In order to give a preliminary assessment of the overall system performance, we used a simple Nearest-Neighbor (NN) procedure. For each example in the test set, we computed its NN in the training set according to the Euclidean distance between the estimated and the true hand postures. We then compared

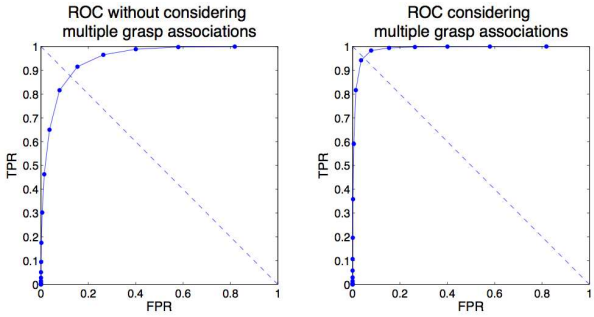

Fig. 3: R.O.C. curves evaluating the grasp classifier performance

the object and grasp classes of the NN with the true classes of the test example under consideration. We obtain an object *recall rate* of $52.6\% \pm 4.2\%$ and a grasp *recall rate* of $50.8\% \pm 1.8\%$. These results are better than a random guess (14% and 20%, respectively), but suggest that a more appropriate measure is needed to evaluate the regression module.

Acknowledgements

This work has been partially supported by the EU Integrated Project Health-e-Child IST-2004-027749. The authors would like to thank Barbara Caputo for useful discussions and insights, and Tatiana Tommasi for the vision unit.

References

[1] L. Baldassarre, L. Rosasco, A. Barla, and A. Verri. Multi-output learning via spectral filtering. *DISI - Technical Report*, pages 1–44, Dec 2009.

[2] R. Pelossof, A. Miller, P. Allen, and T. Jebara. A svm learning approach to robotic grasping. In *ICRA*, 2004.

[3] J. H. Piater. Learning visual features to predict hand orientations. In *ICML - Workshop in spatial knowledge*, 2000.

[4] A Saxena, J. Driemeyer, J. Kearns, and A. Y. Ng. Robotic grasping of novel objects. In *NIPS*, 2006.

[5] S. Ekvall and D. Kragic. Interactive grasp learning based on human demonstration. In *ICRA*, 2004.

[6] 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. *15th ICIAP*, Jun 2009.

[7] N. Noceti, C. Castellini, B. Caputo, and F. Odone. Vmgdb –the contact visuo motor grasping database, 2009. preprint - http://slipguru.disi.unige.it/Research/VMGdB.

[8] C. A. Micchelli and M. Pontil. On learning vector–valued functions. *Neural Computation*, 17:177–204, 2005.

[9] E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri. Some properties of regularized kernel methods. *Journal of Machine Learning Research*, 5, 2004.

[10] L. Lo Gerfo, L. Rosasco, F. Odone, E. De Vito, and A. Verri. Spectral algorithms for supervised learning. *Neural Computation*, Jan 2008.