Towards a Theoretical Framework for Learning Multi-modal Patterns for Embodied Agents

Nicoletta Noceti[1], Barbara Caputo[2], Claudio Castellini[3], Luca Baldassarre[1,4], Annalisa Barla[1], Lorenzo Rosasco[1,5], Francesca Odone[1], and Giulio Sandini[3,6]

[1] DISI - University of Genova
[2] IDIAP - Martigny
[3] DIST - University of Genova
[4] DIFI - University of Genova
[5] MIT, Cambridge, MA
[6] IIT, Genova
{noceti,baldassarre,barla,odone}@disi.unige.it,
bcaputo@idiap.ch, claudio.castellini@unige.it,
lrosasco@mit.edu, giulio.sandini@iit.it

Abstract. Multi-modality is a fundamental feature that characterizes biological systems and lets them achieve high robustness in understanding skills while coping with uncertainty. Relatively recent studies showed that multi-modal learning is a potentially effective add-on to artificial systems, allowing the transfer of information from one modality to another. In this paper we propose a general architecture for jointly learning visual and motion patterns: by means of regression theory we model a mapping between the two sensorial modalities improving the performance of artificial perceptive systems. We present promising results on a case study of grasp classification in a controlled setting and discuss future developments.

Keywords: multi-modality, visual and sensor-motor patterns, regression theory, behavioural model, objects and actions recognition.

1 Introduction

Multi-modal learning, that is, learning from sensorial patterns associated with very different kinds of sensors, is paramount for biological systems. Coupled acoustic and visual information is essential, for instance, for animals to determine whether they are facing a predator or a prey, as well as in courtship rituals. From the point of view of artificial intelligence, multi-modal learning is a potentially excellent way of enriching the input space of pattern recognition problems which could be otherwise more difficult. Indeed, sensorial inputs are available to biological systems in an endless, inextricably mixed flow coming from various sensorial apparatuses. It is not completely clear, then, *how* this information can be used to improve pattern recognition. For example, one could argue that the sight of a certain kind of predator is generally associated with a particular (set of) sound(s) and smell(s), and that animals learn to associate these multi-modal

P. Foggia, C. Sansone, and M. Vento (Eds.): ICIAP 2009, LNCS 5716, pp. 239–248, 2009.

patterns during their infanthood; later on, this fact is employed in recognising the associated danger in a dramatically better way. It seems apparent, then, that there is a mapping among sensorial modalities; e.g., the auditory stimulus corresponding to a predator should be reconstructible from its visual appearance. Therefore, even though not all modalities are always available, it should be possible to recover one from another, to various degrees of precision.

In this work we focus upon *active* perception modalities vs. *passive* ones. By active modality we mean perception arising from the *action* an embodied agent performs in its environment; by passive modality, we mean perception of stimuli which are independent from the agent's will. Our paradigmatic case is grasping for an embodied agent: objects must be grasped in the right way in order to use them as desired. According to the so-called *learning by demonstrations*, that is learning a grasp by observing someone doing it, we build a mapping from the object appearance to the grasping action and assess its ability to accurately describe the grasp type. In a multimodal setting, the estimated mapping could be used to predict the motor data when the corresponding channel is inactive.

In order to reconstruct actions from perception we draw inspiration from the work on *mirror neurons* [13,1]. Mirror neurons are clusters of neural cells which will fire if, and only if, an agent grasps an object *or* sees the same object grasped by another agent; they encode the semantics of an action associated to an object, and form the basis of *internal models* of actions, by which animals reconstruct the grasping and can therefore plan the grasp with greater robustness and effectiveness. Following the path laid out, e.g., in [14,17], where perception-action maps have been built into artificial systems, we hereby propose a theoretical framework for multi-modal learning in which an active modality is reconstructed via statistical regression from a passive modality. In the worked example, visual patterns describing the sight of an object are used to reconstruct the related grasping postures of the hand, with the hope that the use of *two* modalities, one active and one passive, instead of the passive one only, will aid the recognition of the object itself. This framework can be theoretically extended to any such active-passive coupling. The paper is organized as follows: in Section 2 we present the framework, discussing motivations and implementation choices; vision issues are tackled in Section 3 where we deal with objects modelling; the regression techniques used to build the perception-action map are in Section 4. In Section 5 we describe preliminary experiments that motivate the pertinence of our approach, while the last Section discusses future work.

2 A Theoretical Framework for Multi-modal Learning

As outlined in the Introduction, we assume that there exists a mapping between (sets of) patterns belonging to different modalities — here we focus upon the relations which exist between a passive and an active modality. In the aforementioned example dealing with objects (as seen) and grasping them, something like what is shown in Figure 1 is sought for.

In general, active modalities are not available to a biological system during the prediction phase, but only during the training phase. A paradigmatic example is

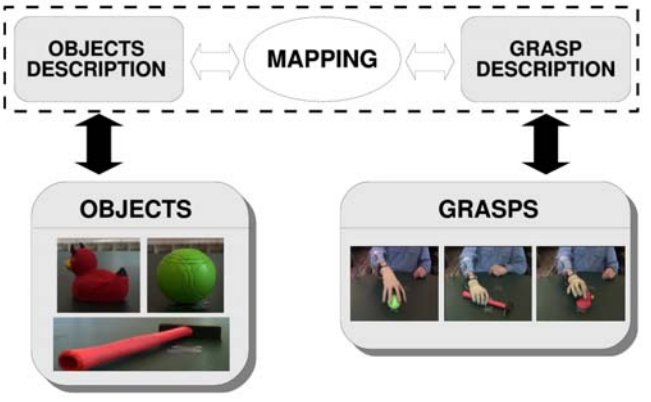

Fig. 1. An instance of the framework we propose: estimating a mapping between appropriate visual descriptions of objects and classes of grasp actions. For the time being, we assume that such relation is a one-to-one mapping.

that of a human infant learning how to grasp an object: by repeatedly trying to apply, e.g., a cylindric grasp to a bottle, he will learn not only to do it more and more efficiently, but also that a bottle is better be grasped cylindrically when moving it or bringing it close to the mouth. Later on, the sight of a bottle will remind the young human what one of the correct grasps is for that particular object. A *perception-to-action map* (PAM) is the equivalent of such training for a biological system: a model to reconstruct an active modality from a passive one. The PAM of our example is a mapping from visual features of an object to motor features of the grasping action used for that object. In general such a map is many-to-many: both a hammer and a bottle can be grasped cylindrically[1]), and as well a mug can be handled either cylindrically or by the handle. In this work we make the simplifying assumption that for a specific object there is just one acceptable grasping action — the PAM is one-to-one. A PAM is useful in passive pattern recognition (e.g., classifying an object just by seeing it) since it augments the input space with PAM-reconstructed active patterns (e.g., classifying the same object from its sight *and the associated grasp*). In this preliminary work we focus upon a simpler problem, namely that of checking whether, given the visual features of an object, the PAM-reconstructed grasp is (similar to) the one associated with that particular object. For example, we might train a PAM to reconstruct a pinch grip (hand posture) from the visual features of a pen; given then, in the prediciton phase, the visual features of another pen, will the PAM-reconstructed hand posture of a pinch grip look like a true pinch grip?

In particular, what is needed is: (*i*) a *vision unit* to extract visual features from an image or a series of images, and (*ii*) a *regression unit*, which will build the PAM.

[1] The nomenclature of grasp types loosely follows that of Cutkosky [16].

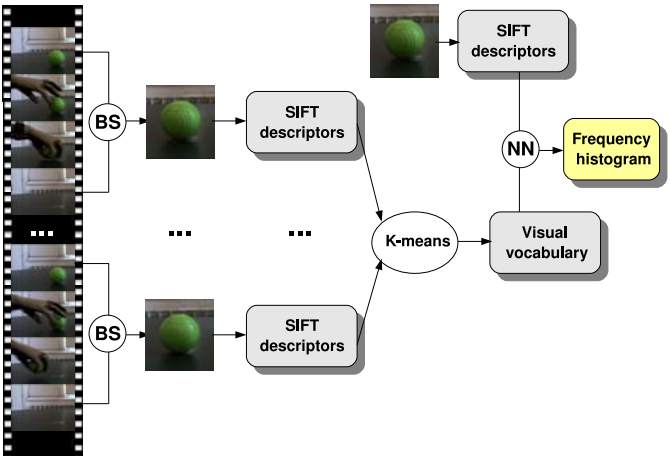

Fig. 2. A schema of the vision unit. First, suitable frames are extracted from the sequence and objects are located by means of background subtaction (BS). SIFT descriptors of a set of random points are input of a clustering step to get to the final visual vocabulary. Finally, each image is represented with respect to the vocabulary adopting a nearest neighbour (NN) strategy (see text for details).

3 Vision Unit

As we will discuss in Sec. 5, the system gathers, as one input, a video sequence acting as *spectator*, whose focus is on object appearance. The goal of the vision unit is to process the signal to obtain a global model of a set of given objects. Figure 2 shows the pipeline of the vision unit when considering only one object (the same procedure is applied to the whole set of objects). Among the sequence, we first select the frames showing only the object without any occlusion, then we locate more precisely its position by means of a simple background subtraction. Although in our application there is not an explicit object recognition step, it is clear from the architecture pipeline that a robust and specific object model is functional to subsequent analysis. It is worthwhile also to mention that with the terms *object recognition* we indicate the characterization of a specific object instance (againts the concept of categorizing classes of objects). We adopt an approach based on local features to describe image structures: because of their popularity a rich variety of local measurements have been proposed in the literature [2,3,4] and applied successfully to objects recognition and categorization problems (see [6,7] just to name a few). Local approaches tipically include two distinct steps: keypoints extraction and description. However, in our case, a keypoint based-representation often ends up into a poor description due to the limited size of the images. We thus built our representation by extracting enough random points guaranteeing a more homogenous sampling. We chose to adopt

SIFT descriptors [4,5] to model image patches around these points, obtaining a set of *words* for each image.

To avoid redundancy and include some global information in our model, we apply k-means [15], following the well-known bag-of-words approach [6]. We thus build a *global* vocabulary, containing SIFT descriptions of all known objects. Image representation is obtained by means of frequency histogram of visual words, selecting for each random point extracted from the image the most similar visual word as nearest neighbor. A normalization step may be advisable for the subsequent data processing.

4 Regression Model

The mapping between object description and grasp description (Fig. 1) corresponds to a vector-valued regression problem. Given a training set of input-output pairs $\{(\mathbf{x}_i, \mathbf{y}_i) : \mathbf{x}_i \in \mathbb{R}^p, \mathbf{y}_i \in \mathbb{R}^d\}_{i=1}^n$, the aim is to estimate a deterministic map from images of objects to sensor values able to generalize on new data. In other words, we want to estimate a function $\boldsymbol{f} : \mathbb{R}^p \rightarrow \mathbb{R}^d$, where p is the number of features representing the input images and d is the number of sensors.

This requires an estension of supervised learning methods to the vector valued setting. 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. On the other hand, the minimization of the *empirical risk*: $\mathcal{E}_n(\boldsymbol{f}) = \frac{1}{n}\sum_{i=1}^n V(\mathbf{y}_i, \boldsymbol{f}(\mathbf{x}_i))$ leads to solving an ill-posed problem, since the solution is not stable and achieves poor generalization. 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 data. The use of kernel functions allows to work with non-linearity in a simple and principled way. In [10] 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 first term in (1) is the *empirical risk* evaluated with the square loss and the second term is the norm of a candidate function $\boldsymbol{f}$ in the RKHS defined by the kernel K. The latter represents the *complexity* of the function $\boldsymbol{f}$, while 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 [11,10] 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 $nd \times nd$.

Tikhonov Regularization is a specific instance of a larger class of regularized kernel methods studied by [8] in the scalar case and extended to the vector case in [preprint]. These algorithms, collectively know spectral regularization methods, provide a computational alternative to Tikhonov regularization and are often easier to tune. In particular we consider iterative regularization methods with early stopping, where the role of the regularization parameter is played by the number of iterations. Besides Tikhonov regularization, in the experiments we consider L2 boosting (Landweber iteration) [18,8] and the ν-method [8].

5 Experimental Setup

The experimental phase aims at testing the proposed framework in a highly controlled environment, where we focus on learning the mapping between image descriptors and motor-sensor data to predict the grasp associated to each object. In the following we present the experimental setup and the regression results.

5.1 Data Acquisition Setup

Data were collected using two Watec *WAT-202D* colour cameras for the images and a Immersion *CyberGlove* with 22-sensors for the hand posture. An Ascension *Flock-Of-Birds* magnetic tracker mounted on the subject's wrist, and an additional standard force sensing resistor glued to the subject's thumb were used to determine the hand position and speed, and the instant of contact with the object.

The cameras return two video sequences, one placed laterally with focus on the object (the *spectator*) and one placed in front of the subject (observing the *actor*).

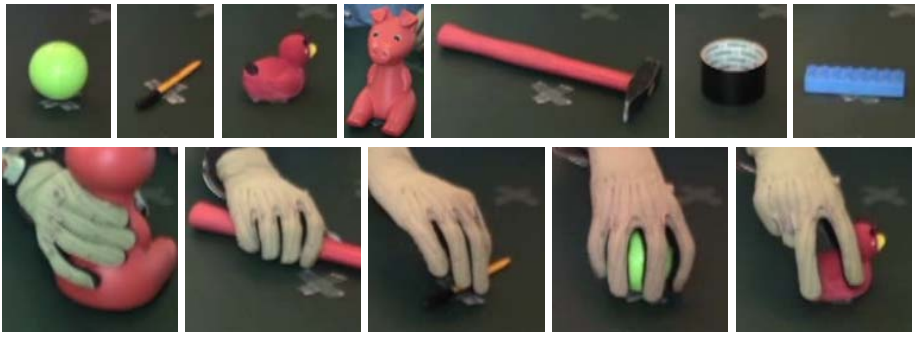

Fig. 3. Top row: the objects used in our experiments. Bottom, the grasp types we consider: *(left to right)* cylindric power grasp, flat grasp, pinch grip, spherical and tripodal grip.

We process only the *spectator* video sequence, because it supplies all the information required for preliminary testing. The video sequence is acquired at 25Hz by each camera, while the glove is sampled at 100Hz. Since the three devices are independent of one another a system of common time-stamps was used in order to synchronise the data.

The CyberGlove returns 22 8-bit numbers linearly related to the angles of the subject's hand joints. The resolution of the sensors is on average about 0.5 degrees. 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.

For these preliminary experiments we considered 7 objects and 5 grasping types identified by different hand postures (see Fig. 3); 2 subjects have joined the experiment: for each object, the actor was asked to perform the required grasping action 20 times.

5.2 Proof of Concept Experiments

Among the motor data, it is reasonable to consider only the 22 measures of hand joints as the most relevant for accurately describing the intrinsic properties of each grasping type. When a grasp occurs the pressure on the force sensing resistor increases, causing the signal to vary hence fixing the time-stamp of the event. Concurrently the values on each hand joint are stored as our output data.

By synchronising motor data and video sequence we select as input data the frames showing an object without clutter, going back along the sequence from the time-stamp in which the event occurs for a fixed amount of frames (see Fig. 2, left). Our data are thus generated as pairs of image descriptors and sensor-motor values, respectively input and output used to feed the regression model.

The regression methods discussed in Sec.4 are implemented in order to predict the expected sensor values of a grasp given the image of an object to be grasped. We compare four different image representations, based on bag-of-words descriptors where the histograms are computed for 20 and 50 words vocabularies on the entire image or on its four quadrants and then concatenated. We call the representations W20, W20conc, W50 and W50conc.

We consider two settings to evaluate the prediction performance of the proposed algorithms. In the first setting (V1-V2) we build training and test sets with the first and second volunteer's data respectively (140 examples each). In the second setting (MIXED) we mix the data of both volunteers and perform a 5 fold cross validation (5-CV). For both settings 5-CV on the training data only is used to select the regularizing paramenter for the RLS method and the stopping iteration for the Landweber [18,8] and ν-method [8]. The optimal regularization parameter is chosen among 25 values ranging from 10^{-6} to 10^{-2}, according to a geometric series. The maximum number of iterations for the iterative methods is set to 800. Tab.1 summarizes the prediction errors evaluated according to the square loss on all 22 components. The prediction errors are consistent among the three learning methods, homogenous with respect to the setting and there are no significant differences among the four representations. The values for the second

setting are markedly lower because mixing the data of both volunteers reduces the variance between training and test sets in each split of the 5-CV. Therefore if we aim at building a model generalizing on several people, it is crucial to collect data from a large variety of volunteers.

Table 1. Data analysis results. We considered two different settings, which differ on the data splitting between training and test sets. Four distinct visual data representations are compared by feeding three learning methods, namely regularized least square (RLS), Landweber (Land) and ν-method (see text for details). For each method we report the prediction accuracies expressed as mean square error and the average number of iterations for the iterative methods. In the MIXED setting the associated variance is reported as well. Results are consistent among the different learning techniques.

Setting	Representation	RLS	Land		ν-method	
		err $[10^3]$	err $[10^3]$	iterations	err $[10^3]$	iterations
V1-V2	W20conc	48	47	630	47	60
	W20	37	38	580	38	60
	W50conc	41	40	340	40	30
	W50	43	43	540	43	40
MIXED	W20conc	6.1(1.1)	6.4(1.2)	670	6.4(1.2)	80
	W20	7.9(1.3)	8.0(1.2)	630	8.0(1.3)	70
	W50conc	6.1(0.8)	6.1(0.9)	630	6.3(0.7)	70
	W50	7.4(2.0)	7.2(2.0)	620	7.3(2.0)	60

Finally, we aim at classifying the grasp type given the estimated sensor values. We restrict at the MIXED setting, using the best regression outcome case, W50conc/RLS. The input data are the sensor measures and the output data are the grasp classes. Again, a 5-CV is performed. For each split the training set is the actual set of measures from the sensors paired with the corresponding grasp type, while the test set is the set of estimated measures. We train a RLS classifier [20] in a One-vs-All configuration obtaining a prediction accuracy of 99.6 (0.8)%. This result indicates that the regression models perform well and guaranteeing the validity of the idea underlying the framework.

6 Discussion and Future Work

In this paper we proposed a general architecture for learning multi-modal patterns of data. The underlying assumption is that the system we want to model has several perceptual channels available, but among them some might be inactive. We adopted a regression-based approach to build a behavioral model of the system that can be exploited to amend such inactivity. As a validation attempt, we presented an application for grasp prediction by means of vector valued regression: the experimental phase produced very promising results that encourage us to further investigate this framework. Even though the regression problem is inherently vector-valued, we restricted our analysis to the simple scalar-valued

case. A preliminary analysis on the covariance matrix of the sensors measures shows some correlation among the sensors, both positive and negative, pointing at the usefulness of a full-fledged vector-valued approach. Recently, much work has been devoted on how to best exploit the similarity among the components and learn all of them simultaneously. The main idea behind most of the literature is to use prior knowledge on the components relatedness to design a particular penalization term or a proper matrix-valued kernel [19]. In absence of prior knowledge, one approach is to design an heuristic to evaluate the similarity among the components from the available data, e.g. by computing the sample covariance of the sensor measures. Our current research is focused on how to translate this information into a viable matrix-valued kernel. Alternatively one can learn the vector structure directly in the training phase [21,22].

This multifaceted framework can be further extended in different directions. Regarding the experimental setup, we plan to enrich the dataset with a higher number of subjects, and multiple grasps for each object. Indeed, this will let us relax the one-to-one assumption we adopted in this paper and investigate a more realistic many-to-many mapping between objects and grasp classes. As anticipated in the introduction, the modeled mapping will be used in the context of multimodal learning to investigate whether, by reconstructing a missing modality, the object recognition rate improves. From the statistical learning viewpoint, we plan to explore new solutions drawing inspiration from the mentioned works on multitask learning.

Acknowledgments

This work was supported by the EMMA project sponsored by the Hasler Foundation (B. C.)

References

1. Rizzolatti, G., Craighero, L.: The Mirror-Neuron System. Annual Review of Neuroscience 27, 169–192 (2004)
2. Harris, C., Stephens, M.: A Combined Corner and Edge Detector. In: Proceedings of The Fourth Alvey Vision Conference, pp. 147–151 (1988)
3. Mikolajczyk, K., Schmid, C.: Scale and Affine Invariant Interest Point Detectors. IJCV 60(1), 63–86 (2004)
4. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
5. Mikolajczyk, K., Schmid, C.: A Performance Evaluation of Local Descriptors. Trans on PAMI 27(10) (2005)
6. Csurka, G., Dance, C.R., Fan, L., Bray, C.: Visual Categorization with Bag of Keypoints. In: ECCV (2004)
7. Ferrari, V., Tuytelaars, T., Van Gool, L.: Simultaneous Object Recognition and Segmentation from Single or Multiple Model Views. IJVC 67(2) (2006)
8. Lo Gerfo, L., Rosasco, L., Odone, F., De Vito, E., Verri, A.: Spectral Algorithms for Supervised Learning. Neural Computation 20(7) (2008)

9. Yao, Y., Rosasco, L., Caponnetto, A.: On Early Stopping in Gradient Descent Learning. Constructive Approximation 26(2) (2007)
10. Micchelli, C.A., Pontil, M.: On learning vector-valued functions. Neural Computation 17 (2005)
11. De Vito, E., Rosasco, L., Caponnetto, A., Piana, M., Verri, A.: Some Properties of Regularized Kernel Methods. Journal of Machine Learning Research 5 (2004)
12. Baldassarre, L., Barla, A., Rosasco, L., Verri, A.: Learning vector valued functions with spectral regularization (preprint)
13. Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G.: Action Recognition in the Premotor Cortex. Brain 119, 593–609 (1996)
14. Metta, G., Sandini, G., Natale, L., Craighero, L., Fadiga, L.: Understanding Mirror Neurons: A Bio-Robotic Approach. Interaction Studies 7, 197–232 (2006)
15. Hartigan, J.A., Wong, M.A.: A K-Means Clustering Algorithm. Applied Statistics 28(1) (1979)
16. Cutkosky, M.: On grasp choice, grasp models and the design of hands for manufacturing tasks. IEEE Transactions on Robotics and Automation (1989)
17. Castellini, C., Orabona, F., Metta, G., Sandini, G.: Internal Models of Reaching and Grasping. Advanced Robotics 21(13) (2007)
18. Buhlmann, P.: Boosting for High-Dimensional Linear Models. Annals of Statistics 34(2) (2006)
19. Micchelli, C.A., Pontil, M.: Kernels for Multi-task Learning. In: NIPS (2004)
20. Rifkin, R., Yeo, G., Poggio, T.: Regularized Least-Squares Classification. In: Advances in Learning Theory: Methods, Models and Applications (2003)
21. Argyriou, A., Maurer, A., Pontil, M.: An Algorithm for Transfer Learning in a Heterogeneous Environment. In: ECML/PKDD (1), pp. 71–85 (2008)
22. Jacob, L., Bach, F., Vert, J.P.: Clustered Multi-Task Learning: a Convex Formulation. In: NIPS (2008)