A Theoretical Framework for Transfer of Knowledge Across Modalities in Artificial and Biological Systems

Francesco Orabona, Barbara Caputo
Idiap Research Institute
1920 Martigny, Switzerland
Email: {forabona, bcaputo}@idiap.ch

Antje Fillbrandt, Frank W. Ohl
Leibniz Institute for Neurobiology
Magdeburg, Germany
Email: {frank.ohl, antje.fillbrandt}@ifn-magdeburg.de

Abstract—Learning from sensory patterns associated with different kinds of sensors is paramount for biological systems, as it permits them to cope with complex environments where events rarely appear twice in the same way. In this paper[1] we want to investigate how perceptual categories formed in one sensory modality can be transferred to another modality in biological and artificial systems. We first present a study on Mongolian gerbils that show clear evidence of transfer of knowledge for a perceptual category from the auditory modality to the visual modality. We then introduce an algorithm that mimics the behavior of the rodents within an online learning framework. Experiments on simulated data produced promising results, showing the pertinence of our approach.

I. Introduction

One of the magnificent abilities of higher animals and humans is to cope with a highly complex environment in which events rarely appear twice in the same way. The ability to extract abstract categories from specific sensory experiences can be considered as basic for dealing with the variety of things. It is a hallmark of cognition that, once formed, categories can also be transferred to newly encountered, rare stimuli, leading to their quick identification.

Research on artificial cognitive systems has the ambitious goal to build autonomous agents exhibiting similar capabilities. Significant progress has been achieved in the last years in the fields of visual categorization [1], [2] and detection of incongruent events [3]. Still, artificial systems are far from achieving performance even remotely comparable to those of biological systems, particularly in terms of dealing with unexpected stimuli. A crucial aspect is that biological systems build categorical models using multi sensory information. A widely acknowledged advantage of multimodal category models for artificial systems is their robustness. A less explored, but equally fundamental property, is that information on one single modality can be exploited to boost learning in another modality. Specifically, in this paper we want to investigate how perceptual categories formed in one sensory modality (audition) can be transferred to another modality (vision). We

first looked at biological systems: our previous work in the auditory system has shown that already rodents (Mongolian gerbils) are capable of abstracting from learned experiences by forming categories [4]. Furthermore, using electrophysiological analysis, we demonstrated that the formation of auditory categories is accompanied by the emergence of specific spatiotemporal patterns of neuronal activity in auditory cortex. During training the electrocorticogramm was recorded from multielectrode arrays implanted onto the primary auditory cortex of the Mongolian gerbils. All electrodes in the array shared the same waveform and the amplitude of the common waveform was modulated in the two spatial dimensions forming Amplitude Modulated patterns (AM-patterns). The moment the animals were able to sort new events into the correct categories coincided with the evolvement of category-specific AM-patterns. These AM-patterns only correlated with the stimulus category but not with the specific physical features of the stimulus [5]. Here we move forward, and we address the transfer of formed categories from one sensory modality to another sensory modality. We trained Mongolian gerbils to associate two different presentation rates ('slow' versus 'fast') of stimuli with Go- versus No-Go- responses using an active avoidance paradigm (section II). Animals were trained first in the auditory modality; when they produced correct responses reliably, the modality was switched to vision. The observed behavioral results indicate that the animals could transfer the rate-response association learned in the auditory modality to the visual modality, i.e. to information derived from a disjunct set of sensors.

Armed with this knowledge, we designed a theoretical framework for transfer of knowledge across modalities that mimics the behavioral response of Mongolian gerbils. Our algorithm uses the online learning framework, where the learner receives one input at a time, and takes action based on its current knowledge. Then, the environment gives feedback to the learner on the validity of the selected action. The goal of the learner is to minimize the number of mistakes on the input sequences. Within this framework, we assume that the two modalities are connected via a mapping function, i.e. that it is possible to predict one modality, given the other one.

This results in an algorithm for transfer of knowledge across modalities. Experiments show both the effectiveness of the presented framework and its consistency with the biological findings. To the best of our knowledge, our model represents the first attempt to implement a biologically motivated algorithm for transfer of knowledge across modalities.

The rest of the paper is organized as follows: section II describes the experiments showing how Mongolian gerbils have the capacity to transfer categorical knowledge formed in the auditory modality to the visual one. Section III presents the algorithm that performs transfer of knowledge across modalities exhibiting a behavior similar to that of Mongolian gerbils. The paper concludes with a summary.

II. TRANSFER OF KNOWLEDGE ACROSS MODALITIES IN BIOLOGICAL SYSTEMS

In this section we describe our study on the transfer of formed categories from one sensor modality to another. We trained Mongolian Gerbils (*Meriones unguiculatus*) to associate two different presentation rates ("slow" versus "fast") of stimuli with Go- versus No-Go-responses using an active avoidance paradigm. Animals were trained in a shuttle box: in response to one of the stimulus rates (the Go-stimulus) the animals had to jump across a hurdle to avoid electrical foot shock (the Go-response). In response to the other stimulus rate (the No-Go-stimulus) animals should stay in their current shuttle box compartment (the No-Go-response), if they jumped nevertheless (false alarm) they would receive a mild foot shock. Animals were trained first in the auditory modality: here they had to discriminate the rates of tones. When they produced correct responses reliably the modality was switched. Now the animals had to discriminate the same stimulus rates in visual flash stimuli. For one group of animals (the congruent group, $n = 20$) the stimulus response association stayed the same from auditory to visual training (i.e. both during the auditory and the visual training the animals had to respond to the "fast" stimuli with the Go-Response and to the "slow" Hz stimuli with the No-Go-Response). For a second group of animals (the incongruent group, $n = 13$) the association between the stimulus rate and the response was reversed from auditory training to visual training (i.e. if they had to associate the Go-response with the "fast" stimuli during auditory training, then during visual training they had to associate the "slow" stimuli with the Go-response). Discrimination performance was quantified using the discriminability index d which is derived from signal detection theory [6]. Differences between the discrimination performance of the congruent and the incongruent group was tested using the Mann-Whitney-U-Test. The rate of spontaneous hurdle crossings determined from the 3-minute break before the start of the first trial was generally low (congruent group: 1.2 ± 1.3 1/min, incongruent group: 1.66 ± 1.53, mean $\pm$ standard deviation).

Expectedly, the auditory discrimination was easily learned by all animals: within a few training sessions they showed the appropriate discriminative behavior. After switching the modality, animals of the congruent group learned the visual

discrimination faster than animals of the incongruent group. Moreover, they acquired a higher overall rate of correct responses ($z = -3.71$; $p < 0.01$). The behavioral results indicate that the animals could use the rate-response association learned during auditory training for their visual discrimination. Figure 1 (a) shows as a representative example the performance of one animal in terms of the rate of correct responses to the Go-stimulus (hits) and the rate of incorrect responses to the Nogo- stimulus (false alarms). We see a high rate of correct responses to the Go-stimulus already in the first session after the modality switch (15th training session), which indicates a knowledge transfer.

For the analysis of the neurophysiological basis of audio-visual category transfer, the electrocorticogram was recorded from two 16-multielectrode arrays, chronically implanted onto primary auditory cortex and primary visual cortex in 10 animals (congruent group: $n = 6$; incongruent group: $n = 4$). The dynamics of crossmodal interactions were investigated by analysis of spatiotemporal activity patterns in the ongoing electrocorticogram (ECoG) using a multivariate pattern classification procedure [7], [8]. Cortical activity patterns associated with the Go- and the No-Go stimuli were determined in the spatial distribution of signal power: in the β- and γ-band (15 Hz to 80 Hz) of the ECoG the root mean squared amplitude was calculated in single time windows of 180 ms for every recording channel of the auditory and visual cortex. Data windows were stepped in 20 ms steps through each trial. The temporal development of the amplitude patterns across electrodes can be described by the trajectory of a vector in a space with one dimension for each channel. To compare trajectories from Go- and No-Go-trials the trajectories were averaged across subsets of trials to get the corresponding centroids. Trials not used for centroid computation were classified trial by trial as closer to a Go-centroid or closer to a No-Go-centroid on the basis of a Euclidean distance metric. The probability of finding at least the observed number correctly classified as $CS+$ and $CS-$ trials was determined for all 20 ms time steps of a trial under the assumption of the null hypothesis of randomly sorting trials into classes ($p = 0.5$) using the cumulative binomial distribution. A Bonferroni corrected α-level of $p = 0.001$ was used for significance determination [9]. During auditory training, patterns could be classified correctly only in the electrocorticogramm taken from training sessions where the animals started to discriminate the rates of tones. Interestingly, patterns could be observed both in the auditory and the visual cortex, but not each time a pattern occurred in the auditory cortex there was also a pattern in the visual cortex. At earlier time points during the course of a trial patterns occurred predominantly in the auditory cortex, especially during earlier phases of training there were no significant patterns in the visual cortex at that time. Only at later time points after the start of the stimulus patterns could be observed in both cortices. In addition, we observed that during early phases of auditory training, when the animal was just about to learn the discrimination (showing already some correct responses but not yet producing them reliably), patterns
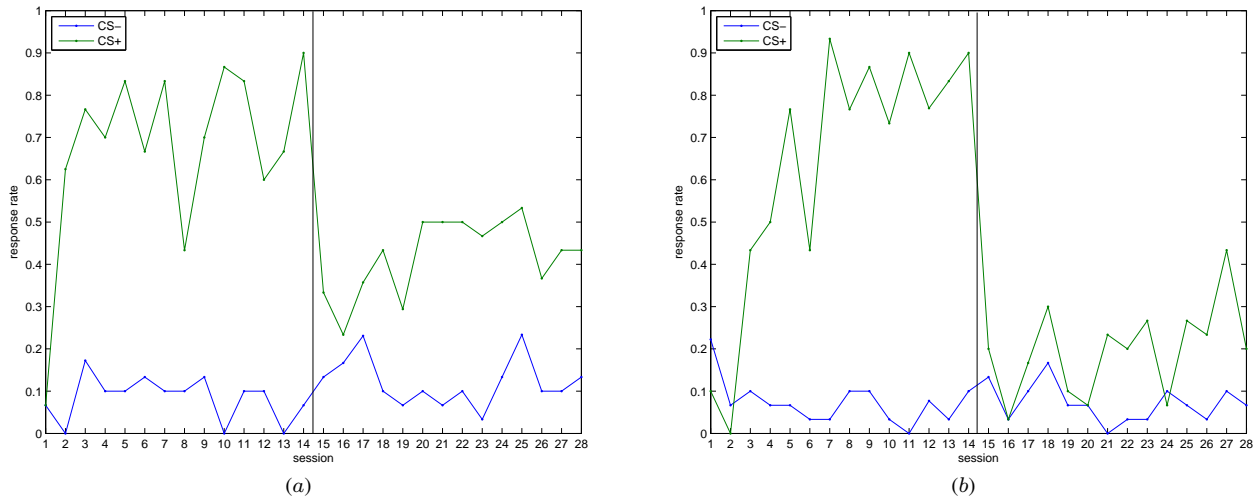
Fig. 1.   Performance of an animal with transferring of information (*a*) and one without transferring (*b*). Rate of correct responses to the Go-stimulus (CS+) and the rate of incorrect responses to the Nogo-stimulus (CS-) across the training sessions. First 14 sessions are auditory training, the following sessions visual training.

appeared only in the auditory cortex. During visual training in the congruent group there were both animals which learned the visual discrimination fast and achieved a reasonably rate of correct response, and animals which did not reach significant levels of discrimination performance. In animals not showing correct responses during the first sessions of visual training we can suspect that there was no transfer learning. Figure 1 (b) shows as an example the performance of an animal not showing signs of knowledge transfer. In these sessions also no significant patterns could be detected. Nevertheless, some animals started to produce some correct responses in later phases of the training. For these animals patterns could be identified in the visual cortex. When the animals became better in discriminating the stimuli patterns could also be found in the auditory cortex, but still they could be classified more reliably in the visual cortex. For animals showing a decent amount of correct response already during the first training sessions, we can suspect that they transferred the rate-response association learned during auditory training to the visual training. In these animals patterns could be detected in both the auditory and the visual cortex already during the first training sessions. Often patterns in the auditory cortex could be classified with higher reliability. These findings fit well to our expectation: in visual training phases we assume that there might be transfer learning across modalities. But we have to take into consideration that even during auditory training, when information transfer across the cortices was neither required nor helpful, patterns were found both in the auditory cortex and the visual cortex. Apparently, the default state of the system is to share the meaning of a stimulus among cortices even though there is no immediate need to do so implied by the experimental situation. This could give an explanation on why transfer learning might work fast and easily: when other cortices are already informed about the rate-response association, they could make use of this information when stimuli are presented in the modality they are specific for. But this also means that the emergence of patterns in multiple cortices is not the neural correlate which distinguishes between transfer learning and simple discrimination learning. Nevertheless some of our findings seem to be specific for transfer learning. We already noted that, during auditory training, patterns appeared in the auditory cortex first. Also during visual training, when the animals showed no indication of transfer learning, significant patterns could be detected first and predominately in the visual cortex. Obviously the cortex primarily involved in the feature extraction process also seems to have a priority in the production of patterns. It seems that the evolvement of patterns in only one single cortex indicates that the animal somehow is involved in the process of category learning, whereas patterns in both cortices indicate that the animals successfully sorts stimuli into categories which have already been formed. In contrast to this, in animals showing correct responses already in the first training sessions, after the modality switch patterns could be observed in both the auditory cortex and the visual cortex. Also these patterns appear early in the course of the trials. We suggest that in these animals both cortices are involved already during the process of category learning.

III. TRANSFER OF KNOWLEDGE ACROSS MODALITIES: A THEORETICAL FRAMEWORK FOR ARTIFICIAL SYSTEMS

We now turn to artificial systems, with the aim to develop an approach able to simulate the behavior of the knowledge transfer observed in the experiments with the Mongolian Gerbils. In particular our aim is to design a learning algorithm that is able to *replicate the finding that using knowledge from another source helps to improve the performance*. Here the performance is measured by the number of mistakes the system does on a sequence of inputs. We want to stress that we do not claim that the proposed method has a biological plausibility. All we aim for is that it can capture some critical aspects of the observed biological behavior.

In the rest of the section we present qualitatively our architecture (section III-A). In section III-B we give the details of the implementation and we provide its theoretical analysis within the online learning framework. We report our experimental findings, proving the soundness of our approach, in section III-C.

A. The Architecture

The proposed architecture is composed by a cascade of different classifiers, see figure 2. In the first layer we have two classifiers that work independently on the two modalities, i.e., audio and video. The second layer combines the outputs of the first layer with a weighted sum. The weights are continually and automatically adjusted to guarantee good performance. As shown in the next section, this layer grants to the system the ability to exploit the old knowledge gained when trained with a different modality. Still the system is robust, in the sense that if the old knowledge is useless or even adversarial, the performance will not be damaged that much. Moreover it is adaptive, that is it wisely uses the old and new knowledge to have always the best performance possible.

We also assume that there is a "link" between the two modalities (the "Mapping" box in the figure). That is, it is possible to predict one modality from the other one. Note that this mapping does not have to be perfect, a noisy reconstruction of one modality from the other one will suffice. In the biological experiments this mapping function is represented by the fact that the audio and video signal have the same frequency. In our experiments we learn the mapping offline (this can be done using any standard regression method). Note that, during this offline learning phase, the system is fed with inputs from the two modalities at the same time. However, during this stage the inputs are not associated with any task.

The learning takes place in two different phases. In the first phase the system is learning using only the first modality, figure 2.a. The second classifier is inactive, so the output of the weighted sum is just equal to the output of first classifier. In the second phase, figure 2.b, the environment switches to the second modality and the system starts learning with the second classifier. At the same time the input is transformed into the first modality by the mapping function and fed to the first classifier. As a result, the output is given by classifiers in *both* modalities, even if only one was received as input.

B. Implementation and Analysis

To choose the algorithms for the different blocks of our architecture we first note that the classic framework of batch learning is not suited to model the learning task of the Mongolian Gerbils. We need instead to use the online learning framework [10]. In this framework the learner receives one input at the time and it acts based on its current knowledge. After its action, the environment gives feedback to the learner on the validity of the selected action. After each feedback, the learner can update its internal model. The sequence of inputs can be of any length and no assumption is done on the order of the data. The aim of the learner is to minimize

the number of mistakes on any sequence of inputs. Note that in this framework there is no distinction between a training and a testing phase, as in batch learning. On the contrary, the system is continuously used to predict (and decide which action to do) and updated after each action. This models exactly the situation of the experiment with the Mongolian Gerbils. Algorithms of this kind are for example the well-known Perceptron [11], ALMA [12], Passive-Aggressive [13]. For a more comprehensive introduction to the topic see [10].

As said above, there are not hypothesis on the order of arrival of the inputs nor any I.I.D. (Independent and Identically Distributed) assumption. Still the algorithms have theoretical guarantees on the worst case performance. More formally the performance is measured with a non-negative loss function, for example the number of shocks in the biological experiment. It is possible to show for the above algorithms an upper bound on the *cumulative regret*, that is the difference between its total loss over the input sequence and the loss of the *best* sequence of actions it could have done. In the following we show that the design of the system allows us to prove an upper bound on the cumulative regret.

Let us define formally the setting. The system receives at each time step t an example $(\boldsymbol{x}_t, y_t)$ where $\boldsymbol{x}_t \in \mathbb{R}^d$ is called an *instance* and $y_t \in \{-1, +1\}$ is called a *label*. We assume without loss of generality that $\|\boldsymbol{x}_t\| \leq 1, \forall t$. We also define the hinge loss with margin 1 of a hypothesis $\boldsymbol{u}$ on the example $(\boldsymbol{x}_t, y_t)$ as

$$\ell(\boldsymbol{u}, \boldsymbol{x}_t, y_t) = \max\{0, 1 - y_t \boldsymbol{u}^T \boldsymbol{x}_t\} , \qquad (1)$$

and we define the cumulative loss, D, of $\boldsymbol{u}$ on the first T examples as

$$D = \sum_{t=1}^{T} \ell(\boldsymbol{u}, \boldsymbol{x}_t, y_t) . \qquad (2)$$

The online algorithms of the first layer learn a mapping $f : \mathbb{R}^d \to \mathbb{R}$. In the following we consider the linear setting, but it can be easily generalized to the non-linear version through the use of Kernels [14]. The algorithms we have chosen are Passive-Aggressive (PA) classifiers, the simplest possible [13]. The prediction of the algorithm is $\mathrm{sign}(\boldsymbol{w}_t^T \boldsymbol{x})$, where $\boldsymbol{w}_t$ is an hyperplane and represents the hypothesis of the system at time t. Starting from the null hypothesis, $\boldsymbol{w}_0 = 0$, the classifier is updated with the following rule

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \alpha_t \boldsymbol{x}_t \qquad (3)$$

$$\alpha_t = \min\left(\frac{\ell(\boldsymbol{w}_t, \boldsymbol{x}_t, y_t)}{\|\boldsymbol{x}_t\|^2}, 1\right) . \qquad (4)$$

Hence the hypothesis is updated each time there is a prediction error or the prediction is correct but the magnitude of the prediction is too low, i.e. the algorithm is not confident enough. Note that the update of the first classifier is stopped in the second phase, that is we are just using its knowledge without updating it anymore.

For PA it is possible to prove an upper bound on the maximum number of mistakes, M, the algorithm does on a sequence of examples, relatively to any hypothesis $\boldsymbol{u} \in \mathbb{R}^d$,
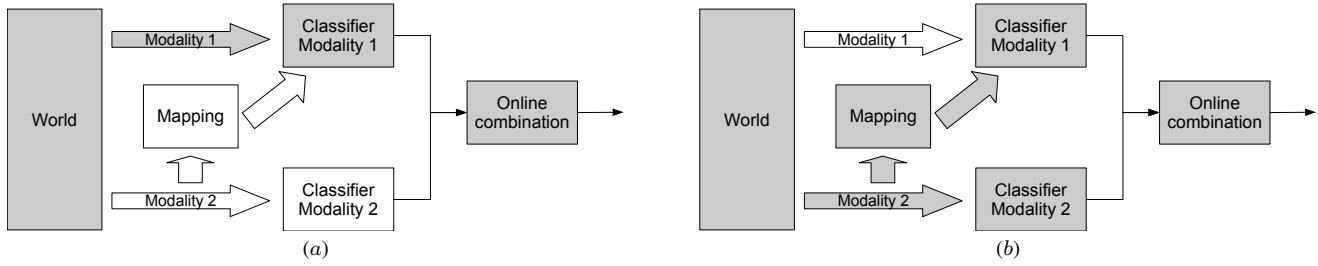
Fig. 2. Diagram of the system in the two phases. The gray boxes are active.

even chosen with the prior knowledge of the entire sequence of examples:

$$M \leq \|\boldsymbol{u}\|^2 + 2D \ . \tag{5}$$

The algorithm used to update the weights of the second layer is the Incrementally Adaptive Weighted Majority (IAWM) [15]. Denote with $o_t^{(1)}$ and $o_t^{(2)}$ the predictions of the first and second classifier at time t. The output of IAWM is

$$\beta_t^{(1)} o_t^{(1)} + \beta_t^{(2)} o_t^{(2)} \ , \tag{6}$$

where $\beta^{(1)}$ and $\beta^{(2)}$ are two positive weights, with sum equal to one, selected by the algorithm in order to minimize the upper bound on the cumulative regret. We refer the reader to [15] for the exact details of the update rule of IAWM. Chaining the bound of the PA with the bound of IAWM we have that

$$M_2 + 2\sqrt{2M_2 \log 2} + 3 \log \left(1 + M_2 \right) + 8 \ , \tag{7}$$

where M_2 is the number of mistakes done by the classifier 2, trained on the new data. This bound tells us that in the worst case scenario, that is when the mapping function is useless, we can expect the performance of the system to be close to the one of second classifier. In particular the number of mistakes of the overall system will be equal to the ones of the second classifier alone plus a term that is just $\mathcal{O}(\sqrt{M_2})$. In other words, the system is betting on the utility of the mapping function, but in the case the mapping is wrong or too noisy it will still recover the performance of the learning without any transfer of knowledge. However when the mapping is not noisy we expect to obtain better performance, as we show in the experimental results.

Moreover it is reasonable to think that old knowledge is useful only for a short period of time, that is when the second classifier did not acquire enough samples. After a certain amount of time the performance of the second classifier will reach the performance of the first one. When this happens, the IAWM will automatically online set its weights to give more importance to the new knowledge. In the next section we will show experimentally this intuition.

A very important case is when the task at hand is linearly separable, that is D is equal to zero. So from (5) we recover a convergence bound, i.e. after a finite number of mistakes the algorithm converges to a solution that will correctly classify any future instance. This means that also the bound in (7) will

become a finite number. In particular we have

$$\|\boldsymbol{u}\|^2 + 2\|\boldsymbol{u}\|\sqrt{2 \log 2} + 3 \log \left(1 + \|\boldsymbol{u}\|^2 \right) + 8 \ , \tag{8}$$

that is a constant, independent of the ordering of the inputs and their number.

C. Experimental Results

As preliminary tests, we have used the '3' vs '8' discrimination task of the MNIST database and a subset of the census-income (Adult) database[2], widely used to benchmark machine learning algorithms. For both the databases we have randomly split the features in two groups. Each group models one modality of training inputs. In this way we are sure that each modality carries more or less the same amount of information and that it should be possible to build the mapping function from one set of features to the other one.

We have used a standard Support Vector Regression method to build the mapping function, using 1000 examples. We use 4000 examples for each of the two phases of learning. The kernel used is a homogeneous polynomial of degree 5 for MNIST and a Gaussian Kernel for Adult.

In figure 3 we show the average online number of mistakes during the second phase, that is the number of mistakes until time t divided by t, for MNIST. The performance of the system is compared to the performance of the classifier that is trained only using the second modality. This corresponds to using only the new knowledge available to the system. The other baseline is given by the performance of the first classifier used with the mapping function, that corresponds to using only the old knowledge. As said above, in the first steps the performance of first classifier, fed with the mapping function, is better than the performance of the second one. After about 1000 examples, the second classifier reaches a performance similar and after that point it gets better. The performance of the proposed system is always close to the best one between the two systems. This can be seen even better in figure 4, where we have plotted the weights of the last layer. In fact around 1000 examples the two weights are equal, while before the weight of first classifier is higher (more importance to old knowledge) and after it is lower (more importance to new knowledge).

Similar results are shown in figure 5 and figure 6, showing results obtained from the Adult database. Again the performance of the system, measured by the number of mistakes at
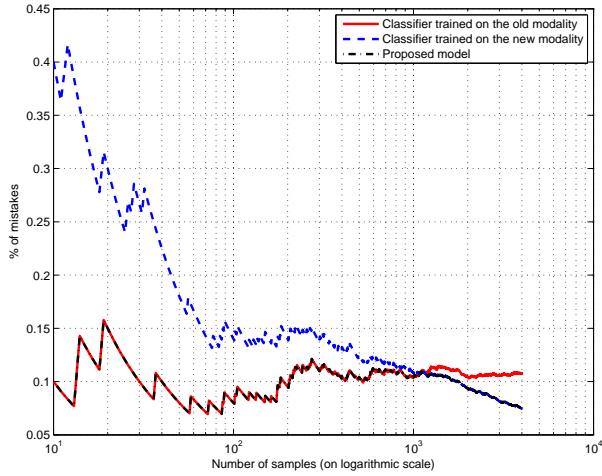
Fig. 3. Average online error on MNIST dataset as a function of the number of examples, during the second phase of learning.
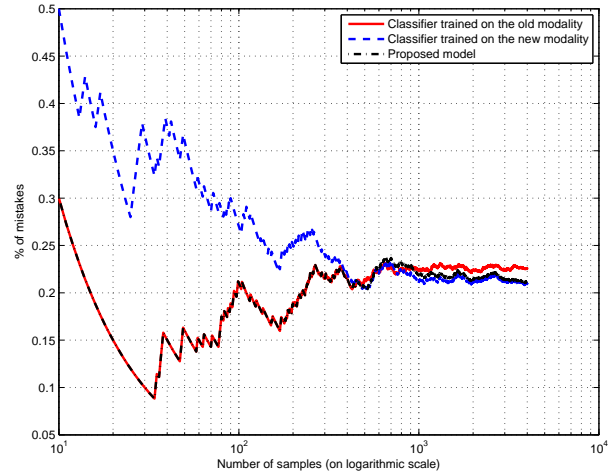

Fig. 5. Average online error on Adult dataset as a function of the number of examples, during the second phase of learning.
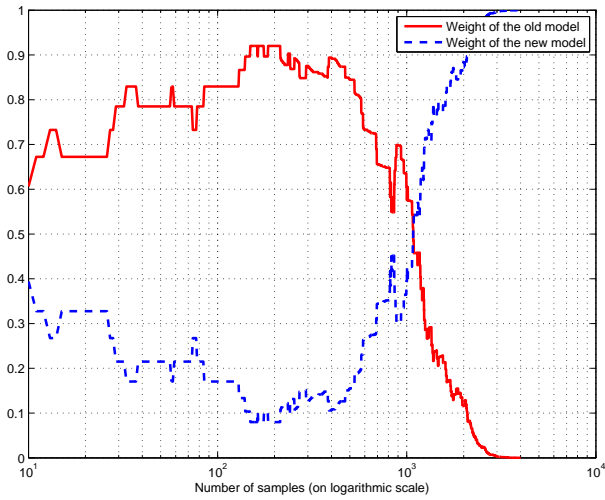

Fig. 4. Weights of the online combination as a function of the number of examples, on the MNIST dataset.
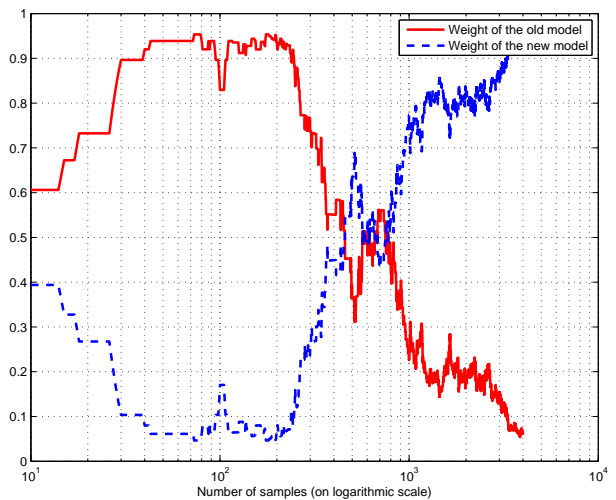

Fig. 6. Weights of the online combination as a function of the number of examples, on the Adult dataset.

tome t divided by t, is close to the best one between the classifier using only the old knowledge and the one using only the new knowledge. As predicted by the theory, when the performance of the system is worse than the one of the second classifier, for example around 1000 samples, the difference is small.

IV. CONCLUSIONS

In this paper we addressed the problem of knowledge transfer across modalities in biological and artificial systems. We specifically investigated how perceptual categories formed in one sensory modality could be transferred to another sensory modality in biological systems, and how to design learning algorithms able to mimic this behavior for artificial systems. We presented a study on Mongolian gerbils indicating quite clearly that the animals could use associations learned in the auditory modality when engaged in a visual discrimination task. In addition, we presented an algorithm, developed within

the online learning framework, that improves performance on one modality by transferring knowledge from another one, via a mapping between the two modalities learned off line. Experiments on two different databases show the effectiveness of our approach. In the future we will explore the connection between our approach and multi-task learning [16], [17]. We would also like to extend our algorithm so to be able to build more complex mappings across modalities. Specifically, we would like to have algorithms able to build one-to-many, or even better many-to-many multi sensory mapping, with the aim to cope with stimuli that can be associated with more than one category at a time.

REFERENCES

[1] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *IJCV*, 2008.
[2] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid, "Group of adjacent contour segments for objec detection," *IEEE Trans. PAMI*, 2008.

[3] D. Weinshall, H. Hermansky, A. Zweig, L. Jie, H. Jimison, and M. Pavel, "Beyond novelty detection: Incongruent events, when general and specific classifiers disagree," in *Proc NIPS*, 2008.

[4] W. Wetzel, T. Wagner, F. Ohl, and H. Scheich, "Categorical discrimination of direction in frequency-modulated tones by Mongolian gerbils," *Behav Brain Res*, 1998.

[5] F. Ohl, H. Scheich, and W. Freeman, "Change in pattern of ongoing activity with auditory category learning," *Nature*, 2001.

[6] C. C. MacMillan NA, *Detection theory: a user's guide*. Cambridge University Press, 1991.

[7] J. Barrie, W. Freeman, and M. Lenhardt, "Spatiotemporal analysis of the prepyriform, visual, auditory and somestethic surface EEGs in trained rabbits," *J. Neurophysiol.*, 1996.

[8] W. Freeman and J. Barrie, "Analysis of spatial patterns of phase in the neocortical gamma EEG of rabbits," *J. Neurophysiol.*, 2000.

[9] S. H. F. W. Ohl FW, Deliano M, "Early and late patterns of stimulus-related activity in auditory cortex of trained animals," *Biological Cybernetics*, pp. 374–379.

[10] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[11] F. Rosenblatt, *Psychological Review*.

[12] C. Gentile, *JMLR*.

[13] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *JMLR*, 2006.

[14] B. Schölkopf and A. Smola, *Learning with Kernels*. Cambridge, MA, USA: MIT press, 2002.

[15] P. Auer, N. Cesa-Bianchi, and C. Gentile, "Adaptive and self-confident on-line learning algorithms," *Journal of Computer and System Sciences*, 2002.

[16] Y. S. Abu-Mostafa, "A method for learning from hints," in *Proc NIPS*, 1993.

[17] Y. Jin and B. Sendhoff, "Knowledge incorporation into neural networks from fuzzy rules," *Neural Processing Letters*, 1999.