2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

PROPRE: PROjection and PREdiction for multimodal correlations learning. An application to pedestrians visual data discrimination.

Mathieu Lefort and Alexander Gepperth

***Abstract*— PROPRE is a generic and modular unsupervised neural learning paradigm that extracts meaningful concepts of multimodal data flows based on predictability across modalities. It consists on the combination of three modules. First, a topological projection of each data flow on a self-organizing map. Second, a decentralized prediction of each projection activity from each others map activities. Third, a predictability measure that compares predicted and real activities. This measure is used to modulate the projection learning so that to favor the mapping of predictable stimuli across modalities. In this article, we use Kohonen map for the projection module, linear regression for the prediction one and we propose multiple generic predictability measures. We illustrate the properties and performances of PROPRE paradigm on a challenging supervised classification task of visual pedestrian data. The modulation of the projection learning by the predictability measure improves significantly classification performances of the system independently of the measure used. Moreover, PROPRE provides a combination of interesting functional properties, such as a dynamical adaptation to input statistic variations, that is rarely available in other machine learning algorithms.**

I. Introduction

In order to interact with its environment, an autonomous agent needs to have a structured perception of its internal and external states. According to sensory-motor theories, sensory-motor regularities are one key point for structuring the agent's interaction [1]. Hence, autonomous and progressive construction of sensory-motor representations is currently an active research field in developmental robotics [2], [3], [4]. To tackle this complex problem, we propose to take inspiration from biological agents that are already able to interact with their environment in a structured way. Thus, our model is based on bio-inspired computational and learning paradigms.

Any biological agent observes its internal state and the one of its surrounding environment through multiple sensors. To obtain a consistent perception, this agent needs to merge and extract information from these multiple data flows. As a single event can induce sensory changes in multiple channels, multimodal co-occurrence detection is essential and is also consistent with sensory-motor theories. For example, in humans, consistent multimodal stimuli improve learning and detection of events compared to monomodal stimuli or inconsistent multimodal stimuli [5], [6], [7].

At a macroscopic point of view, the cortex can be described as a set of multiple cortical areas which are defined by their functional processing as for example visual areas (V1, V2, MT, ...) or motor areas. In spite of their functional specialization, cortical areas seem to have generic layered architecture [8] and data processing [9], [10]. Especially, self-organization (i.e. close neurons of a cortical area are sensible to close stimuli) is a widely spread computational paradigm that is mainly observed in low level sensory areas [11], [12], [13].

In this article, we propose a paradigm for multiple data flows fusion by multimodal correlation learning. This problem was already addressed in other studies with various approaches as for example maximum covariance analysis by principal component analysis [14], combination of sensory-motor anticipations [15] or mutually constrained modal self-organizations [16] among others. The originality of our work lies mainly in its biological inspiration to provide a generic neural implementation with an emphasis on preferential mapping of predictable stimuli across modalities with separated modal processing. This focus on predictability is motivated by a conceptual work postulating that even though perceptual representations can be diverse depending on their level of abstraction, their relevance depends on their ability to predict other quantities [17]. Another important aspect of our work is that our model was tested on a challenging applicative task using real world data and provides combined functional properties such as online and incremental learning that classical machine learning techniques usually miss.

PROPRE consists on the combination of projection and prediction (PROPRE stands for PROjection-PREdiction). Each modal data flow is projected onto a low-dimensional manifold by a self-organizing map. Based on each modal projection, predictions of all other projections are computed. A correct prediction can only be obtained if corresponding modal stimuli are correlated. A predictability measure, that defines the ability of a projection to predict the other ones, modulates this projection learning so that, at the system level, mapping of predictable multimodal stimuli is favored.

We previously studied the PROPRE paradigm with artificial multimodal data related to basic robotic behaviors [18], [19]. Targeting its use on a real developmental robotic platform, in this article we apply it to a challenging visual discrimination task of namely real-world pedestrian pose classification [20], [21] (see figure 1). Moreover, we illustrate the importance of the modulation mechanism by proposing various predictability measures and using simple bio-inspired projection and prediction algorithms.

In the next section, we introduce the generic PROPRE paradigm and its application to the pedestrian classification

Mathieu Lefort and Alexander Gepperth are with the UIIS division, ENSTA ParisTech, 858 boulevard des Maréchaux, 91762 Palaiseau, France. Mathieu Lefort and Alexander Gepperth are members of Inria FLOWERS, Inria Bordeaux Sud-Ouest, 200 avenue de la vieille tour, 33405 Talence, France (email: {mathieu.lefort, alexander.gepperth}@ensta-paristech.fr).

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

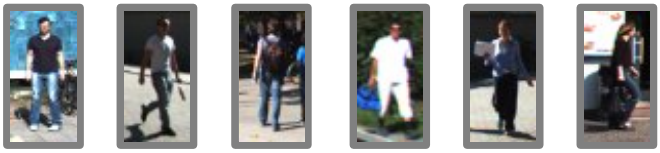

Fig. 1. Examples of pedestrian visual data used in the classification task.

task. The task protocol and obtained results illustrating the multiple PROPRE functional properties such as robustness and flexibility are presented in section III.

II. PROPRE

A. Main paradigms

PROPRE is a modular and generic neural paradigm for multiple modal data merging that is based on the coupling of projection and prediction (see figure 2). The projection step uses the self-organizing map paradigm (SOM) to obtain a dedicated topological projection of each modal input space (see section II-B). It provides a low dimensional spatial representation of all current stimulus. Each modal projection is used to predict projected representations of all other data flows (see section II-C). Such a prediction can only be accurate if the corresponding modal stimuli - and a fortiori their low level representations - are correlated. A predictability measure quantifies the quality of the prediction that indirectly reflects the correlation between the multimodal stimuli (see section II-D). This predictive measure is used to modulate the projection learning so that to favor the mapping of stimuli correlated across modalities.

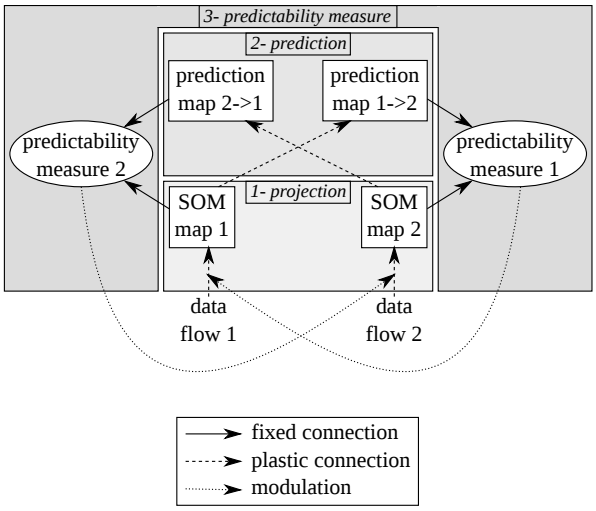

Fig. 2. PROPRE architecture is composed of three interacting modules. First, a projection module that provides a low dimensional representation of each modal stimulus. Second, a prediction of each modal representation by the other ones. Third, a predictability measure that quantifies the ability of a stimulus to predict the other ones and modulates the corresponding projection learning.

For the pedestrian discrimination task addressed in this article (see section III-A for details), we used PROPRE on a multimodal flow composed of visual data (representing the detected pedestrian) and category data (representing the potential danger of the pedestrian). The aim of the task is to transfer the knowledge contained in the category data flow to the visual one for the system to be able to visually recognize potentially dangerous pedestrians. In this context, the category is considered as an already processed stream for the model (that may be resulting of computation and learning of another part of a larger system). Thus, this data flow is neither processed in the projection nor in the prediction steps leading to the removal of the dedicated maps in the PROPRE architecture (see figure 3).

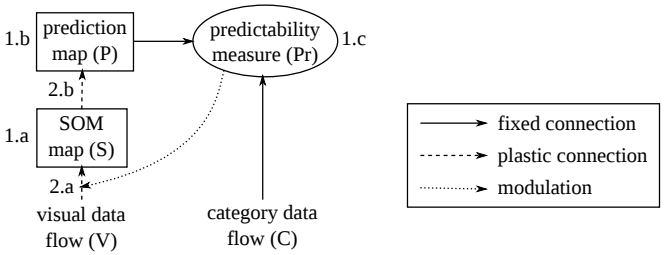

Fig. 3. Architecture of PROPRE applied to the pedestrian visual data classification task. For this task, the category flow is considered as a reference data flow so that it is not projected and the projection (S) of the visual data tries to predict (P) directly the category stimulus (C).

From a computational point of view, the reception of each multimodal stimulus in the model leads to one transmission and one learning step so that the model provides online learning (i.e. the stimulus is represented and learned at the same time). Technically speaking, the transmission stage consists on the evaluation of each module activity (equations 1.a-b-c in figure 3). Then the learning stage updates the weights of the plastic connections linking the modules (equations 2.a-b). The used equations are detailed in the three next sections describing respectively each of the three modules of PROPRE.

B. Projection

For the projection step, we use a slightly modified version of the classical Kohonen self-organizing map [22] in which we include the modulation by the predictability measure. Kohonen maps provide some interesting properties, such as quantization that is related to the mapping of the input space statistic [23]. Thus, by modulating learning, we influence the received input statistics so that the predictable stimuli space will be preferentially mapped and represented.

In practice, S is a discrete bi-dimensional square grid of neurons that receives the visual data flow V (see figure 2). Let $\mathbf{w_{SV}}(\mathbf{x},\mathbf{t})$ be $(w_{SV}(x,y,t))_y$ with $w_{SV}(x,y,t)$ the weight from the unit at position y in V to the unit at position x in S at time t. With these notations, the activity of S at position x at time t is computed as

$$S(x,t) = (\mathbf{w_{SV}}(\mathbf{x}^*,\mathbf{t}).\mathbf{V}(\mathbf{t}))e^{\frac{-||x-x^*||_2^2}{\sigma^2}} \quad (1.a)$$

with x^* the winning unit defined as the unit whose matching between its weights and the input stimulus, computed as $\mathbf{w_{SV}}(\mathbf{x}^*,\mathbf{t}).\mathbf{V}(\mathbf{t})^1$, is the highest (i.e. $\mathbf{w_{SV}}(\mathbf{x}^*,\mathbf{t}).\mathbf{V}(\mathbf{t}) =$

$\max_x \ \mathbf{w_{SV}}(\mathbf{x},\mathbf{t}).\mathbf{V}(\mathbf{t})$) with $\mathbf{V}(\mathbf{t})$ the current stimulus. σ is the variance of the Gaussian neighborhood radius and $||\cdot||_2$ is an euclidean distance.

The incoming weights of the unit at position x in S at time t are updated as following:

$$\Delta\mathbf{w_{SV}}(\mathbf{x},\mathbf{t}) = \eta\lambda(t)S(x,t)(\mathbf{V}(\mathbf{t}) - \mathbf{w_{SV}}(\mathbf{x},\mathbf{t})) \quad (2.a)$$

$$\lambda(t) = \begin{cases} 1 \text{ if } Pr(t) \geq \theta \\ 0 \text{ otherwise} \end{cases}$$

with η the constant learning rate, $Pr(t)$ the predictability measure (see section II-D) and θ the predictability threshold. This equation is the one of Kohonen map in which we introduce the modulation by $\lambda(t)$. Thus, only predictable stimuli (i.e. that have their predictability measure overcoming the threshold) are learned by the system.

C. Prediction

The projection activity of S is used to provide a prediction in P of the current category stimulus of the data flow C. Thus, size of P map is defined by the one of C. The activity in P at position x at time t is computed as a weighted sum of the S activity:

$$P(x,t) = \sum_y w_{PS}(x,y,t)S(y,t) \quad (1.b)$$

with $w_{PS}(x,y,t)$ the weight from the unit at position y in S to the unit at position x in P.

The weights of the connection between S and P are learned with a classical linear regression algorithm [24] that minimizes the mean square error between the prediction $\mathbf{P}(\mathbf{t})$ and the current category stimulus $\mathbf{C}(\mathbf{t})$. The update equation is:

$$\Delta w_{PS}(x,y,t) = \eta' S(y,t)(C(x,t) - P(x,t)) \quad (2.b)$$

with η' the constant learning rate.

D. Predictability measure

The predictability measure aims to quantify the quality of a prediction with respect to a projection so that to modulate the projection learning step to consider preferentially predictable stimuli (see sections II-A and II-B). In this article, we want to illustrate that the introduction of this modulation mechanism increases the performance of the system even associated with non tuned classical projection and prediction algorithms. For that purpose, we propose multiple generic measures that do not assume anything about the structure of processed multimodal data flow.

In the case of our architecture applied to the pedestrian pose classification task, the predictability measure quantifies the quality of the category prediction $\mathbf{P}(\mathbf{t})$ with respect to the real category $\mathbf{C}(\mathbf{t})$ (see figure 3). $\mathbf{C}(\mathbf{t})$ encodes the category as a spatial coding (see section III-A) which is consistent with the coding provided by the projection step, so that measures proposed in this article can be easily adapted to the generic PROPRE architecture.

In practice, let define X_c as $\{x|C(x) \neq 0\}$ when $\mathbf{C}$ represents the c category. The four proposed measures are the following with c^* the current category represented by $\mathbf{C}(\mathbf{t})$:

- $Pr_1(t) = \dfrac{\sum_{x\in X_{c^*}} P(x,t)}{\max_c \sum_{x\in X_c} P(x,t)}$ that represents if the prediction of the real category is maximal,
- $Pr_2(t) = \dfrac{\sum_{x\in X_{c^*}} P(x,t)}{\sum_c \sum_{x\in X_c} P(x,t)}$ that corresponds to the proportion of the predictive activity representing the true category comparing to the total predictive activity,
- $Pr_3(t) = \dfrac{\left(\sum_{x\in X_{c^*}} P(x,t)\right)^2}{\sum_c \sum_{x\in X_c} P(x,t)}$ that combines the strength of the prediction representing the true category with its proportion comparing to all predictions,
- $Pr_4(t) = -||\mathbf{P}(\mathbf{t}) - \mathbf{C}(\mathbf{t})||_2$ that is the opposite of the euclidean distance between the prediction and the stimulus[2].

III. Experiments

A. Pedestrian classification task

We used data taken from the Daimler monocular pedestrian detection benchmark [20] to which we manually assigned one of four possible orientations (left, right, front and back) as in [21]. For each experiment, we associate to each orientation one of two categories, one that represents the potential danger of this pedestrian orientation, the other the absence of danger. In absence of specific mention, by default, in our experiments the pedestrian left orientation was associated to a potential danger whereas the other orientations are considered as not dangerous. This category is represented by the spatial position of a Gaussian in a 7x32 vector (see figure 4) that feeds the category data flow C (see figure 3). For the visual data flow, we use high dimensional data that corresponds either to the 32x64 pixel image of the pedestrian or to a preprocessed 18x42 vector corresponding to the HOG features of the image (see figure **??**). To compute HOG features, we use a cell size of 8x8 pixels, a block size of 16x16 pixels, a border of 0 pixels, and a window size of 32x64 pixels in the terms of [25].

For each experiment, the data set composed of 12684 samples was randomly split in a learning and an evaluation

[1]In practice, we normalize the weights $\mathbf{w_{SV}}(\mathbf{x},\mathbf{t})$ and the input $\mathbf{V}(\mathbf{t})$ so that the opposite of the dot product is directly related to the euclidean distance between the two values that is classically used as matching function in Kohonen map.

[2]According to our generic definition of the modulation signal for the projection learning (see section II-B), predictability measure has to be an increasing function of the quality of the prediction, so that we take the opposite of the distance.

Fig. 4. Examples of stimuli used in the experiments with a) (respectively b)) a pedestrian with a left (respectively right) orientation. In this example, left oriented pedestrians are considered as potentially dangerous contrary to the ones with right orientation.

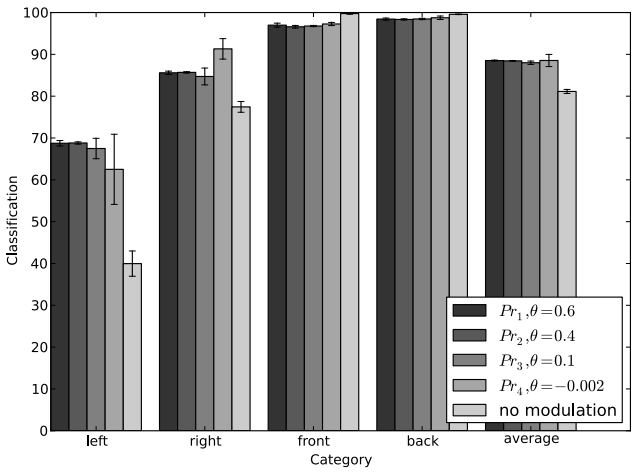

Fig. 5. Average and standard deviation of classification performance for each pedestrian orientation of the system receiving HOG features visual input in a 10×10 projection map depending on the predictability measure used. No modulation means that $\forall t, \lambda(t) = 1$ in equation 2.a.

data set composed respectively of 90% and 10% of the data. This split aims to clearly illustrate the ability of our paradigm to generalize its knowledge to unknown stimuli but is not mandatory for PROPRE as it provides online learning. At each time step, during typically 300,000 time steps, a multi-modal stimulus randomly picked up in the learning data set is presented to the model. Then, classification performance of the system is evaluated by presenting all visual stimuli of the evaluation data set and comparing its true category with the predicted one defined by the maximal prediction in P.

By the way, in order to reduce convergence time of the system in simulation, we made some adjustments to the projection equations presented in section II-B that do not qualitatively modify obtained results. First, the variance of the Gaussian and the learning rate are initially set to high values (so that the projection quickly maps the input space) and decrease to low non zero constant values that guarantee continuous learning which are respectively set to 1.0 and 0.01 in all our experiments. Second, $\lambda(t)$ is fixed to 1 for some time steps (20,000 in our experiments) at the beginning of the simulation so that projection and prediction quickly learn and predictability measure becomes relevant.

B. Classification performances

1) Influence of the predictability measure: In order to test the influence of the modulation mechanism introduced in our PROPRE paradigm, we tested our model over ten experiments without modulation (i.e. $\forall t, \lambda(t) = 1$ in equation 2.a) or with one of the four predictability measures proposed in section II-C. We present in figure 5 (respectively on figure 6) the results obtained for each measure with a 10×10 (respectively 30×30) projection map and HOG features (respectively pixel images) as visual input.

In the case of HOG features visual input (figure 5), we observe that the average classification performance with the use of a predictability measure is significantly higher than the one obtained without modulation. More importantly, this improvement is quite similar whatever predictability measure used (around +7%). This reinforces our main idea of introducing a modulation mechanism to guide the projection towards the mapping of predictable stimuli in order to improve system performance.

Moreover, the improvement of performance is particularly important for the left and right orientations which are the hardest ones to classify. Indeed, left and right orientations are visually similar but belong to different categories, contrary to front and back orientations that are visually similar and have the same category.

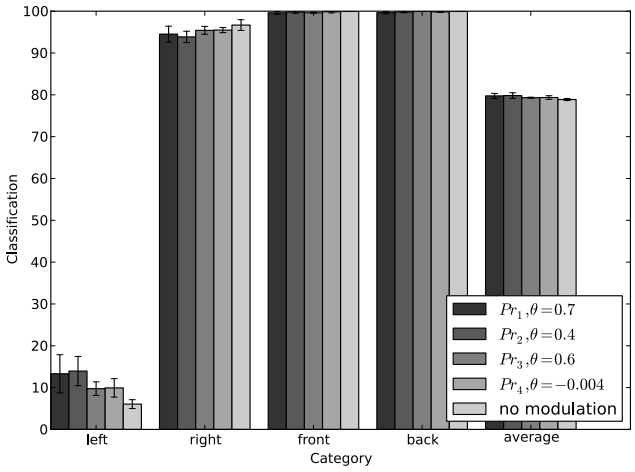

Fig. 6. Average and standard deviation of classification performance for each pedestrian orientation of the system receiving pixel images visual input in a 30×30 projection map depending on the predictability measure used. No modulation means that $\forall t, \lambda(t) = 1$ in equation 2.a.

With the use of pixel images as input, the problem is much more difficult compared to the preprocessed HOG features as the input space dimension is higher and the dimensions are less relevant (most of the pixels do not provide any

information about the pedestrian orientation as the ones of the background for example). This difficulty is illustrated by the drop of average classification performance provided by the reference SVM (support vector machine) algorithm [24] from 95.95% with HOG features to 76% with pixel images.

In this case, the modulation mechanism still provides an increase of average performance but only by around 1% (figure 6). However, it has to be noticed the most significant improvement of left orientation classification which reflects the functional consequence of the modulation mechanism. Indeed, thanks to the modulation mechanism, PROPRE algorithm maps the mostly predictive stimuli, independently of their category, whereas the original Kohonen algorithm maps the mostly presented stimuli, in this case the non dangerous ones. Thus, achieved PROPRE performances tend to be more diverse.

2) Influence of the predictability threshold: In order to study the dependency of system performance on the predictability threshold, we tested PROPRE with hog features as visual input, a 10×10 projection map, Pr_1 as predictability measure and ten different thresholds. Obtained results over ten experiments for each setup are presented on figure 7. Results provided by the use of the other proposed predictability measures are qualitatively similar.

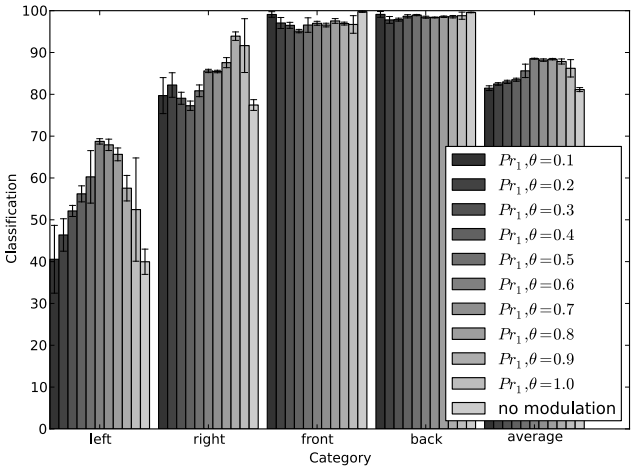

Fig. 7. Average and standard deviation of classification performance for each pedestrian orientation of the system receiving HOG features visual inputs in a 10×10 projection map depending on the threshold used with the Pr_1 predictability measure. No modulation means that $\forall t, \lambda(t) = 1$ in equation 2.a.

At one extreme, $\theta = 0$ means that every stimulus will be learned by the system. This configuration is equivalent to the no modulation case. At the other extreme, if $\theta = 1$, the system will only map stimuli that are correctly classified every time i.e. that the maximal prediction always corresponds to the true category. We can do two mains observations from figure 7 about the influence of the predictability threshold on the system performance.

First, significant improvement of the average performance is obtained for a large range of thresholds (for θ between 0.5 and 1 here). Thus, the model do not need a precise tuning of this parameter for the modulation to be efficient.

Second, even if stimuli are completely predictable in our setup, the increase of the predictability threshold (that imposes the system to be more selective) leads to a decrease of the average performance at some point (the shift occurs around $\theta = 0.8$ in this case). Thus, the threshold has to be chosen so that the system favors the mapping of clearly predictable stimuli but in the same time accepts some classification errors for the learned stimuli. This last point should be particularly relevant with noisy inputs as for example in real robotic tasks.

3) Influence of the SOM size: PROPRE's capacity of representation and consequently of prediction is limited by the size of the SOM. We illustrate in figure 8 the influence of this parameter on the system performance. Results presented are the average over ten experiments for each size, using HOG features as input and Pr_1 for the predictability measure associated to a 0.7 threshold. Once again, similar results are obtained with other predictability measures proposed.

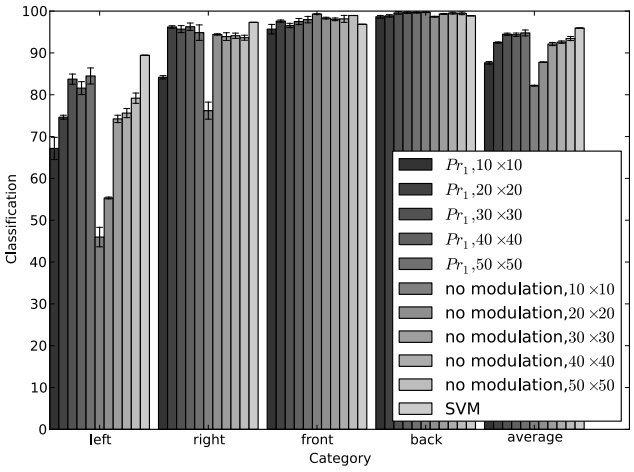

Fig. 8. Average and standard deviation of classification performance for each pedestrian orientation of the system receiving HOG features visual inputs with Pr_1 as predictability measure and a 0.7 threshold depending on the size of the projection map P. No modulation means that $\forall t, \lambda(t) = 1$ in equation 2.a. SVM performance are also represented for comparison.

As expected, classification performance increases with the size of the projection map. This increase is mainly obtained by better classification of left oriented pedestrians, which are one the hardest orientation to discriminate as previously mentioned. PROPRE performance reaches 94.78% in average with a 50×50 projection map which is very close to the 95.95% classification performance obtained with SVM that is the reference supervised linear classification algorithm. Moreover, preliminary results with higher projection map sizes (up to 80×80) seem to indicate that average PROPRE performance can slightly increase by 2% and then overcome SVM one. By the way, the increase of classification performance induced by the modulation is confirmed for each of the tested projection map size.

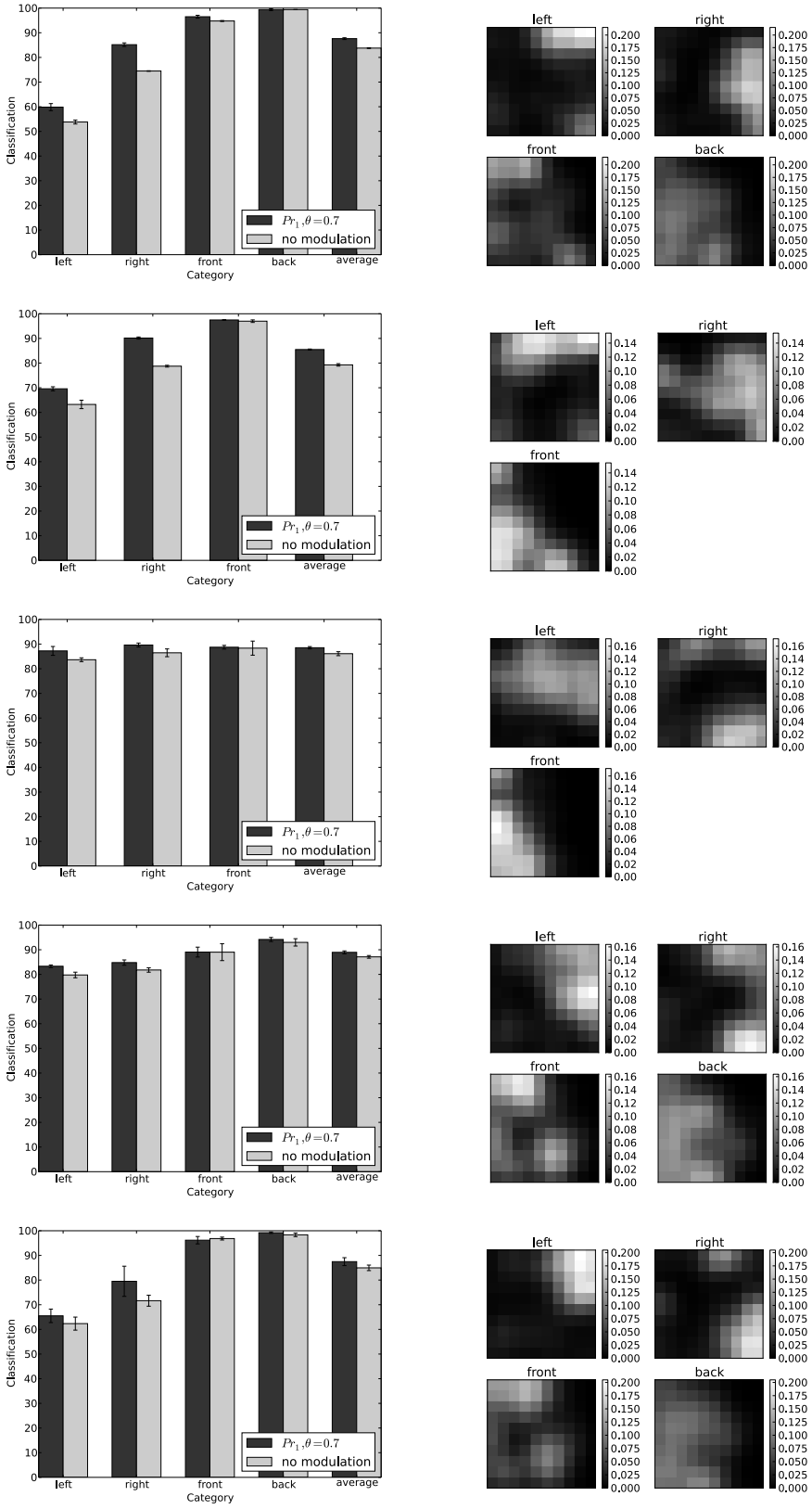

Fig. 9. Each line corresponds to the system results obtained for the corresponding stage in the input scenario (please refer to text for details). Left column: Average and standard deviation of classification performance for each pedestrian orientation of the system with a 10×10 projection map and Pr_1 as predictability measure with a 0.7 threshold. No modulation means that $\forall t, \lambda(t) = 1$ in equation 2.a. Right column: Average activity in the projection map for each pedestrian orientation at the end of the stage when using the modulation mechanism.

C. Plasticity

We showed in the previous section that PROPRE classification performance can be very close to the one provided by the reference SVM algorithm if we use a sufficiently width projection map. Moreover, contrary to SVM, PROPRE learning is incremental as based on the combination of Kohonen map and online prediction learning. In order to illustrate this plasticity property, we tested PROPRE with HOG features as input, a 10×10 projection map, Pr_1 for the predictability measure associated to a 0.7 threshold and changing input statistic over time (no external cue is provided to the model to signal the change in the inputs). Learning and evaluation data sets are evolving according to the following protocol that changes visual and category inputs (each stage last 200,000 time steps):

- in the first stage, the four pedestrians orientations are used and only the left orientation is considered as potentially dangerous (as in the experiments presented in previous sections),
- in the second stage, the back oriented pedestrians are removed from the datasets,
- in the third stage, the right oriented pedestrians are now categorized as potentially dangerous,
- in the fourth stage, back oriented pedestrians are reintroduced in the dataset and considered as previously as not dangerous,
- in the fifth stage, right orientation was again considered as not dangerous so that the fifth and first datasets are the same.

Classification performance over ten experiments and an example of obtained self-organization in P are presented in figure 9.

PROPRE achieves good classification performance with the five input statistics (close to 88% in average) confirming the plasticity of our system. Once again, performances are increased at each stage with the use of the modulation. By the way, we can notice that the performance for the first and fifth stage, that correspond to the same input statistic, are very similar in average, even if the distribution of performance over orientations has changed because of the learning history of the system.

Interestingly, we can observe that the orientation mapping provided by the projection map is quite stable over the different input statistics. It slightly spreads (respectively shrinks) when a visual pedestrian orientation is removed (respectively added) during stage 2 (respectively stage 4) or changes when the category of a pedestrian orientation is modified during stages 3 and 5. This behavior is very interesting as it reveals that the performance of the system for previously learned stimuli that are stable over time is not substantially modified by changes in other parts of the input space. Thus, PROPRE provides an interesting compromise between plasticity and stability.

IV. Conclusion and perspectives

PROPRE is a bio-inspired unsupervised learning paradigm for multimodal data merging that consists on the combination of projection and prediction. Each modal data flow is projected on a dedicated low dimensional self-organizing map that is used to predict all other modalities projections. The originality of PROPRE consists on the use of a predictability measure, that quantifies the ability of a projection to predict the other ones, to influence the corresponding projection learning. Thus, projections tend to map preferentially stimuli correlated across modalities.

In previous articles [18], [19], PROPRE was already applied on artificial data representative of a robotic task. In this article, targeting the use of PROPRE on real robotic platform, we illustrate multiple of its functional properties when applied to real visual pedestrian data on a challenging supervised classification task. This task consists on the visual classification of pedestrians with four possible orientations in a dangerous or not dangerous category.

PROPRE architecture is generic so that it can be applied on any multimodal flow. Here, we tested it with pixel images or preprocessed HOG features as visual inputs. In both cases, the modulation mechanism improves average classification performance especially for the left oriented pedestrians which are among the hardest ones of the protocol to classify. Moreover, this modulation tends to spread the classification performance over the different pedestrian orientations so that PROPRE abilities are more diverse.

PROPRE performance depends directly on the size of projection maps that determines the width of the input space that can be projected and then predicted. Thus, by increasing the size of the projection map, PROPRE performance can be easily improved and tends to be equivalent to the one of the reference SVM classification algorithm.

Thanks to its combination of a modified version of Kohonen map and continuous predictive learning, PROPRE provides an incremental learning of the input space. Thus, it is able to dynamically adapt to shrink or spread of the multimodal input space and to changes in the multimodal correlations between stimuli. Moreover, this plasticity does not significantly influence performance over non variable learned part of the input space. Such a compromise between stability and plasticity can be interesting for robotic applications.

All these properties were obtained with the four new predictability measures proposed in this article that are designed to be applicable to the computation of any multimodal data flow. This independence to precise predictability measure validates our main claim of modulating projection learning towards predictable stimuli across modalities to improve system performance. Moreover, the performance is quite stable over a large range of predictability thresholds, facilitating the parametrization of the model.

Based on these promising results, we plan to test the scalability of the proposed predictability measures by applying PROPRE on real multimodal data as for example visual and laser data for pedestrian classification. Moreover, in order

to reduce parametrization of the model to simplify its use in large scale robotic applications, we want to introduce a sliding predictability threshold. Preliminary results on this last point are promising.

REFERENCES

[1] M. Mossio and D. Taraborelli, "Action-dependent perceptual invariants: From ecological to sensorimotor approaches," *Consciousness and cognition*, vol. 17, no. 4, pp. 1324–1340, 2008.

[2] S. Kirstein, H. Wersing, and E. Körner, "Towards autonomous bootstrapping for life-long learning categorization tasks," in *International Joint Conference on Neural Networks*. IEEE, 2010, pp. 1–8.

[3] P.-Y. Oudeyer, "Developmental robotics," *Encyclopedia of the Sciences of Learning*, 2011.

[4] B. Ridge, D. Skocaj, and A. Leonardis, "Self-supervised cross-modal online learning of basic object affordances for developmental robotic systems," in *Robotics and Automation (ICRA)*. IEEE, 2010, pp. 5047–5054.

[5] I. Bernstein, M. Clark, and B. Edelstein, "Effects of an auditory signal on visual reaction time." *Journal of Experimental Psychology*, vol. 80, no. 3p1, p. 567, 1969.

[6] M. Doyle and R. Snowden, "Identification of visual stimuli is improved by accompanying auditory stimuli: The role of eye movements and sound location," *PERCEPTION-LONDON-*, vol. 30, no. 7, pp. 795–810, 2001.

[7] L. Shams and A. Seitz, "Benefits of multisensory learning," *Trends in cognitive sciences*, vol. 12, no. 11, pp. 411–417, 2008.

[8] E. Kandel, J. Schwartz, T. Jessell, S. Siegelbaum, and A. Hudspeth, *Principles of neural science*. Elsevier New York, 1991, vol. 3.

[9] K. Holthoff, E. Sagnak, and O. Witte, "Functional mapping of cortical areas with optical imaging," *NeuroImage*, vol. 37, no. 2, pp. 440–448, 2007.

[10] K. Miller, D. Pinto, and D. Simons, "Processing in layer 4 of the neocortical circuit: new insights from visual and somatosensory cortex," *Current opinion in neurobiology*, vol. 11, no. 4, pp. 488–497, 2001.

[11] W. Bosking, Y. Zhang, B. Schofield, and D. Fitzpatrick, "Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex," *The Journal of Neuroscience*, vol. 17, no. 6, p. 2112, 1997.

[12] C. Schreiner, "Order and disorder in auditory cortical maps," *Current Opinion in Neurobiology*, vol. 5, no. 4, pp. 489–496, 1995.

[13] C. Wessinger, M. Buonocore, C. Kussmaul, and G. Mangun, "Tonotopy in human auditory cortex examined with functional magnetic resonance imaging," *Human brain mapping*, vol. 5, no. 1, pp. 18–25, 1997.

[14] O. Kroemer, C. Lampert, and J. Peters, "Learning dynamic tactile sensing with robust vision-based training," *Robotics, IEEE Transactions on*, vol. 27, no. 3, pp. 545–557, 2011.

[15] J.-C. Quinton and J.-C. Buisson, "Multilevel anticipative interactions for goal oriented behaviors," *Proceedings of EpiRob*, pp. 103–110, 2008.

[16] M. Lefort, Y. Boniface, and B. Girau, "Somma: Cortically inspired paradigms for multimodal processing," in *International Joint Conference on Neural Networks*, 2013.

[17] P. König and N. Krüger, "Symbols as self-emergent entities in an optimization process of feature extraction and predictions," *Biological Cybernetics*, vol. 94, no. 4, pp. 325–334, 2006.

[18] A. Gepperth, "Efficient online bootstrapping of sensory representations," *Neural Networks*, 2012.

[19] A. Gepperth and L.-C. Caron, "Simultaneous concept formation driven by predictability," in *International conference on development and learning*, 2012.

[20] M. Enzweiler and D. Gavrila, "Integrated pedestrian classification and orientation estimation," in *CVPR*, 2010.

[21] A. Gepperth, M. Garcia Ortiz, and B. Heisele, "Real-time pedestrian detection and pose classification on a GPU," in *IEEE ITSC*, 2013.

[22] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[23] M. Cottrell, J. Fort, and G. Pagès, "Theoretical aspects of the som algorithm," *Neurocomputing*, vol. 21, no. 1-3, pp. 119–138, 1998.

[24] C. Bishop and N. Nasrabadi, *Pattern recognition and machine learning*. springer New York, 2006, vol. 1.

[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition (CVPR)*, vol. 1. IEEE, 2005, pp. 886–893.