Discrimination of pedestrians visual data by combining projection and prediction learning

Mathieu Lefort[1] and Alexander Gepperth[1]

1- ENSTA ParisTech - UIIS division
858 Boulevard des Maréchaux, 91762 Palaiseau - France
Mathieu Lefort and Alexander Gepperth are members of INRIA FLOWERS

Abstract. PROPRE is a generic and semi-supervised neural learning paradigm that extracts meaningful concepts of multimodal data flows based on predictability across modalities. It consists on the combination of two computational paradigms. First, a topological projection of each data flow on a self-organizing map (SOM) to reduce input dimension. Second, each SOM activity is used to predict activities in all other SOMs. Predictability measure, that compares predicted and real activities, is used to modulate the SOM learning to favor mutually predictable stimuli. In this article, we study PROPRE applied to a classical visual pedestrian data classification task. The SOM learning modulation introduced in PROPRE improves significantly classification performance.

1 Introduction

An autonomous robot needs to detect and learn sensory-motor regularities that emerge from its interaction with the environment. This autonomous learning of representations is an active research field in developmental robotics [1, 2, 3]. In this article, to tackle this problem, we take inspiration from biological agents who are already able to interact with their environment in a complex way.

Multimodal correlation detection seems to be a key point for humans to perceive their environment. Indeed, multimodal stimuli improve learning and detection of events compared to monomodal stimuli [4]. From a computational point of view, the cortex is composed of cortical areas specialized in one modality as visual or motor areas. However, they seem to have generic architecture and data processing [5]. Especially, self-organization (i.e. close neurons having close sensibility) is a widespread computational paradigm in sensory areas [6, 7].

PROPRE is a neural paradigm for multiple data flow fusion by learning correlations across modalities, an idea already developped in [8, 9, 10]. PROPRE provides a neural implementation of continuous, semi-supervised learning consisting of the combination of projection and prediction (PROPRE means PROjection-PREdiction). Each data flow is projected on a self-organizing map (SOM). Learning of this SOM is modulated by a supervised predictability measure that quantify the ability of the projection to predict other data.

In our previous works [11, 12], we focused on the validation of PROPRE paradigm using artificial multimodal data related to some robotic behavior. In this article, we apply PROPRE on a the challenging task of visual pedestrian pose discrimination[13, 14]. In the next section, we introduce the dedicated

PROPRE architecture that we use for the pedestrian classification task. The task protocol and obtained results are presented in section 3.

2 PROPRE paradigm

2.1 General description

PROPRE is based on the combination of projection and prediction. The projection step (see section 2.2) aims to provide a low dimensional representation of the current input stimulus. Each projected representation is used to predict projected representation of all other data flows (see section 2.3). A predictive measure (see section 2.4) quantifies the quality of the prediction that reflects the correlation between the multimodal stimuli. Indeed, correlated stimuli are predictable and so are their projections. This predictive measure is used to modulate the projection learning to favor the mapping of correlated stimuli. For more details about the general PROPRE paradigm, please refer to [12].

In the case of the pedestrian discrimination task (see section 3.1), we used PROPRE with a visual data flow (representing the detected pedestrian) and a category data flow (representing the potential danger of the pedestrian). The aim of the task is to transfer the knowledge of the category data flow to the visual one, so that to be able to visually recognize potentially dangerous pedestrians. In this context, the category data flow is considered as an already processed stream (that may be the result of a learning of another part of the system) and is thus neither computed in the projection nor in the prediction step. In practice, PROPRE consists on the alternating of a computation and a learning stage (respectively equations 1.x and 2.x in figure 1 and in the next sections).

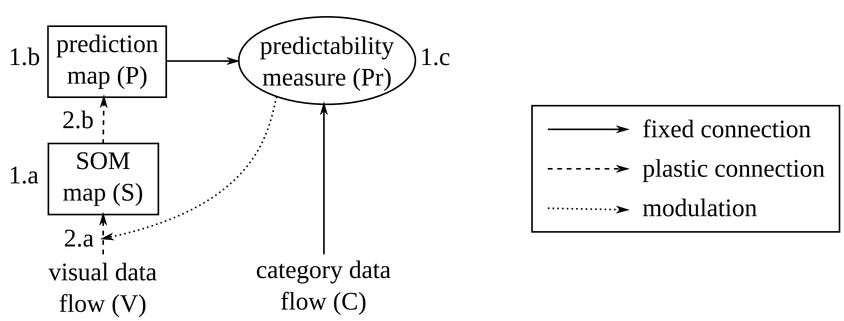

Fig. 1: PROPRE architecture used for the pedestrian visual data classification task. Please refer to text for details.

2.2 Projection

To project an input data flow, we use slightly modified Kohonen self-organizing maps [15] which provide some interesting properties, such as quantization [16]. In practice, S is a bi-dimensional Kohonen map that receives the visual data flow V (see figure 1). Let $\mathbf{w_{SV}(x,t)}$ 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

$$S(x,t) = e^{\frac{-||x-x^*||_2^2}{\sigma^2}} \tag{1.a}$$

with x^* so that $||\mathbf{w_{SV}(x^*,t)} - \mathbf{V(t)}||_2 = \min_x ||\mathbf{w_{SV}(x,t)} - \mathbf{V(t)}||_2$ with $\mathbf{V(t)}$ the current stimulus, σ the variance of the Gaussian neighboorhood radius[1] and $||\cdot||_2$ euclidean distances.

The incoming weights of S are updated as following:

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

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

with η the learning rate[1], $Pr(t)$ the predictability measure (see section 2.4) and θ the predictability threshold. Thus, only predictable stimuli (i.e. that have their predictability measure overcoming the threshold) are learned by the system. $\lambda(t)$ is fixed to 1 for some time steps at the beginning of the learning so that projection and prediction converge and predictability measure becomes relevant.

2.3 Prediction

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

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

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

The prediction is learned by linear regression [17] that minimizes the mean square error between the prediction and the current category stimulus $C(t)$:

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

2.4 Predictability measure

The predictability measure aims to quantify the quality of the category prediction P w.r.t. the real category C. Let define X_c as $\{x|C(x) \neq 0\}$ when C represents the c category, which is relevant as the category is represented as a spatial coding (see section 3.1). As predictability measure we use one of the three following measures with c^* the current real category represented by $C(t)$:

[1]To reduce convergence time of the SOM, the variance of the Gaussian and the learning rate decrease from high values to low constant values that keep the SOM able to adapt to changes in the data flow.

$$Pr(t) = \underbrace{\frac{\sum\limits_{x \in X_{c^*}} P(x,t)}{\max\limits_{c} \sum\limits_{x \in X_c} P(x,t)}}_{Pr_1(t)} \text{ or } \underbrace{\frac{\sum\limits_{x \in X_{c^*}} P(x,t)}{\sum\limits_{c} \sum\limits_{x \in X_c} P(x,t)}}_{Pr_2(t)} \text{ or } \underbrace{\frac{\left(\sum\limits_{x \in X_{c^*}} P(x,t)\right)^2}{\sum\limits_{c} \sum\limits_{x \in X_c} P(x,t)}}_{Pr_3(t)} \qquad (1.c)$$

Pr_1 represents if the prediction of the real category is maximal.
Pr_2 represents the proportion of the prediction of the real category compared to all predictions.
Pr_3 represents the strength of the prediction of the real category and its proportion compared to all predictions.

3 Results

3.1 Pedestrian classification task

We used data taken from the Daimler monocular pedestrian detection benchmark [14] to which we manually assigned one of four possible orientations (left, right, front and back) as in [13]. The left orientation is categorized as a potential danger whereas the other three categories are considered as not dangerous. The data set was split into a learning and an evaluation data set composed respectively of 11351 and 1333 pictures. In practice, each visual stimulus is a 18x42 vector corresponding to the HOG feature of a 32x64 image of a pedestrian. In the terms of [18], 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 to compute HOG features. The category stimulus is a 7x32 vector which represents the potential danger by the spatial position of a Gaussian (see figure 2).

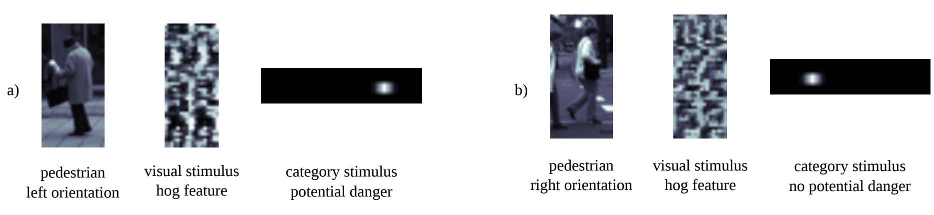

Fig. 2: Example of stimuli provided to PROPRE for a) (resp. b)) a dangerous (resp. non dangerous) pedestrian with left (resp. right) orientation.

3.2 Classification performances

In table 1, we present the performance of various algorithms on the pedestrian classification task described in the previous section:

- SOM+prediction[1] that means that we deactivated the modulation of the projection learning by the predictability measure (i.e. $\forall t,\ \lambda(t) = 1$),
- PROPRE[1] with the three different predictability measures (see section 2.4) and adaptated predictability thresholds (see section 2.2),
- classical support vector machine (SVM) algorithm, which is the reference supervised classification algorithm [17].

Algo / Orientation	SOM + prediction	PROPRE			SVM
		$Pr_1 > 0.7$	$Pr_2 > 0.5$	$Pr_3 > 0.5$	
left	81.85	86.57	85.34	85.08	89.44
right	21.92	59.49	57.14	62.73	97.31
front	82.70	87.89	87.37	87.61	96.84
back	90.96	94.64	93.84	91.85	98.88
average	71.74	83.53	82.35	82.92	96.62

Table 1: Average percentage of correct classification over 10 experiments.

We can observe that the modulation of the projection learning introduced by PROPRE significantly improves the classification results for all pedestrian orientations, especially for right one, whatever the predictability measure used. The improvement of right orientation pedestrian classification is quite remarkable as these pedestrians are visually extremely similar to the one with a left orientation, but correspond to different "dangerousness" categories.

PROPRE performance is not as good as the one provided by SVM but this is not an aim of our model. Indeed, PROPRE provides unsupervised and plastic learning [12] and input space mapping contrary to SVM. Moreover, PROPRE performance can be strongly improved by increasing the SOM size.

4 Conclusion and perspectives

PROPRE is a semi-supervised learning paradigm for multimodal data, that consists on the combination of projection and prediction. Predictability measure, that quantifies the ability of a projection to predict the other ones, influences the corresponding projection learning. Thus, stimuli correlated across modalities are mainly mapped by the projections.

In this article, we apply PROPRE to an important real-world object discrimination task. PROPRE receives a visual data flow (representing a pedestrian in one of four orientations) and a category data flow (representing the potential

[1] We use a 10x10 map for the projection map S and the predicted category of a visual stimulus is determined by the localization of the maximum of induced activities in P.

danger of the pedestrian). With the modulation of the projection learning introduced by PROPRE, classification performance is significantly improved whatever the predictability measure chosen between the three proposed in this article.

Based on these promising results, we plan to apply PROPRE to multimodal real data as for example visual and laser data for pedestrian detection. Moreover, in order to reduce parametrization of the model, we want to introduce a sliding predictability threshold rather than a fixed one. This may be possible as current experiments tend to show that PROPRE performance does not depend on the precise tuning of the fixed threshold.

References

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

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

[3] 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)*, pages 5047–5054. IEEE, 2010.

[4] L. Shams and A.R. Seitz. Benefits of multisensory learning. *Trends in cognitive sciences*, 12(11):411–417, 2008.

[5] K. Holthoff, E. Sagnak, and O.W. Witte. Functional mapping of cortical areas with optical imaging. *NeuroImage*, 37(2):440–448, 2007.

[6] W.H. 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*, 17(6):2112, 1997.

[7] C.E. Schreiner. Order and disorder in auditory cortical maps. *Current Opinion in Neurobiology*, 5(4):489–496, 1995.

[8] P. König and N. Krüger. Symbols as self-emergent entities in an optimization process of feature extraction and predictions. *Biological Cybernetics*, 94(4):325–334, 2006.

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

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

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

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

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

[14] M. Enzweiler and D.M. Gavrila. Integrated pedestrian classification and orientation estimation. In *CVPR*, 2010.

[15] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

[16] M. Cottrell, J.C. Fort, and G. Pagès. Theoretical aspects of the som algorithm. *Neurocomputing*, 21(1-3):119–138, 1998.

[17] C.M. Bishop and N.M. Nasrabadi. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[18] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893. IEEE, 2005.