Multimodal space representation driven by self-evaluation of predictability

Mathieu Lefort, Thomas Kopinski, Alexander Gepperth

Multimodal space representation driven by self-evaluation of predictability

Mathieu Lefort*, Thomas Kopinski† and Alexander Gepperth*
*Computer science and system engineering department - ENSTA ParisTech
858 boulevard des Maréchaux, 91762 Palaiseau Cedex - France
Email: {mathieu.lefort,alexander.gepperth}@ensta-paristech.fr
Mathieu Lefort and Alexander Gepperth are also members of INRIA Flowers
†University of Applied Sciences Bottrop - Computer science institute
Postfach 100755, 45407 Mühlheim - Germany
Email: Thomas.Kopinski@hs-ruhrwest.de

***Abstract*—PROPRE is a generic and modular neural learning paradigm that autonomously extracts meaningful concepts of multimodal data flows driven by predictability across modalities in an unsupervised, incremental and online way. For that purpose, PROPRE consists of the combination of projection and prediction. Firstly, each data flow is topologically projected with a self-organizing map, largely inspired from the Kohonen model. Secondly, each projection is predicted by each other map activities, by mean of linear regressions. The main originality of PROPRE is the use of a simple and generic predictability measure that compares predicted and real activities for each modal stream. This measure drives the corresponding projection learning to favor the mapping of predictable stimuli across modalities at the system level (i.e. that their predictability measure overcomes some threshold). This predictability measure acts as a self-evaluation module that tends to bias the representations extracted by the system so that to improve their correlations across modalities. We already showed that this modulation mechanism is able to bootstrap representation extraction from previously learned representations with artificial multimodal data related to basic robotic behaviors [1] and improves performance of the system for classification of visual data within a supervised learning context [2]. In this article, we improve the self-evaluation module of PROPRE, by introducing a sliding threshold, and apply it to the unsupervised classification of gestures caught from two time-of-flight (ToF) cameras. In this context, we illustrate that the modulation mechanism is still useful although less efficient than purely supervised learning.**

I. Introduction

Biological agents are able to adapt efficiently to unknown situations and tasks using online, incremental and partially unsupervised learning. These capabilities are the result of millions of years of evolution. Thus, we think that taking inspiration from the architecture and processing of the brain may be one promising way to transfer these interesting learning properties to the developmental robotics field. Here, we study more precisely the merging of multiple data flows by extracting correlated stimuli with a cortically inspired paradigm having these learning properties. This paradigm processes each data flow in a generic way wherever it comes from. Thus, we define such a processing as multimodal as the data flows may come from different senses, even if in this article we consider multiple data flows coming from the same kind of sensor, as a first study. This work fits in the currently active research on autonomous and progressive construction of sensory-motor representations in the developmental robotics field [3], [4], [5].

Detection of correlated stimuli seems to play an important role in multimodal fusion as a single event can induce sensory changes in various channels. Multiple psychophysical experiments on humans reveal that consistent multimodal stimuli improve learning and detection of events compared to monomodal stimuli or inconsistent multimodal ones [6], [7], [8]. Moreover, such a detection of correlated signals across modalities is consistent with sensory-motor theories that claim that sensory-motor regularities are one key point for structuring the agent interaction with its environment [9].

From a macroscopic point of view, the cortex is composed of a set of multiple cortical areas defined by their functional processing, as for example visual or motor areas. Despite their functional specialization, cortical areas seem to have generic layered architecture [10] and data processing [11], [12]. Especially, self-organization (i.e. close neurons in one cortical area have close sensibilities) is a widely spread computational paradigm that is mainly observed in low level sensory areas [13], [14], [15].

Based on these considerations, we propose the PROPRE (projection-prediction) paradigm. Each modal data flow is projected on a low-dimensional manifold by a self-organizing map (SOM). From each modal projection, predictions of all other projections are computed. A correct prediction can only be obtained if the corresponding modal stimuli are correlated. A predictability measure quantifies this ability of a projection to predict the other ones and modulates this projection learning so that the representation of predictable multimodal stimuli is favored at the system level. One of our claims with PROPRE is that by biasing the representation of the input flow towards stimuli that are interesting for the current task, here the correlated stimuli across modalities, we can improve the global performance of the system. In this context, the predictability measure acts as a self-evaluation module of the extracted representations by the system that influences the learning in order to improve the correlations between these representations across modalities.

This use of predictability to influence representations in PROPRE is motivated by a conceptual work [16] arguing that symbolic quantities are defined by their power to predict other quantities. It is also conceptually very close to the predictive coding model of hierarchical visual processing [17], [18], [19]. This focus on predictability and on generic multimodal processing are the two main points that distinguish our work from other multimodal self-organizing maps models [20], [21], [22].

PROPRE was already successfully applied to the bootstrapping of selectivities based on previously learned representations with artificial multimodal data related to basic robotic behaviors [1]. We also show that the self-evaluation module of predictability leads to learning of representations that improve classification of visual pedestrian data in a supervised context. Moreover, these representations can be incrementally updated to take into account various changes in the data flows [2]. Following our target to use PROPRE for multimodal online and incremental learning on real developmental robotic platforms, in this article, we apply it to the unsupervised learning of hand gestures observed by two time-of-flight (ToF) cameras and also propose a new improved self-evaluation module. In the next section, we introduce the PROPRE paradigm and equations. The task protocol and obtained results with the various tested architectures are presented in section III.

II. PROPRE

A. Paradigm

Kohonen map [23] is a learning paradigm that represents a high dimensional input space by projecting it on a manifold with a fixed low dimension thus providing a low dimensional spatial coding of the input space. Kohonen map provides a vector quantization that is related to the mapping of the input space statistic [24]. However, achievement of a robotic task may need to have a granularity of the sensory-motor space representation that is different from the input space statistic. For that purpose, we propose to modulate the Kohonen learning rule so that to bias the obtained representations towards stimuli that are considered as relevant for the targeted task, here the correlated stimuli across modalities for a multimodal learning task.

PROPRE is a modular and generic neural paradigm for online, incremental and unsupervised learning of multimodal representations. It consists on the interaction between three modules (see figure 1):

- Projection: each modal flow is projected on a self-organizing map derived from the model of Kohonen. Each modal stimulus is represented by the spatial location of a Gaussian (centered on the best matching unit) that can be combined with any other modalities, e.g. in an incremental way.
- Prediction: each modal projected representation tries to predict the other ones. The spatial relationship between stimuli representation is learned at this stage.
- Predictability measure: the predictability measure quantifies the precision of the prediction with a simple and generic measure that does not do any assumption on the data flow. This measure is an indicator of the correlation between the stimuli and modulates the projection learning so that to favor stimuli correlated across modalities. This module can be considered as a simple autonomous self-evaluation module of the performance of the system to extract multimodal representations.

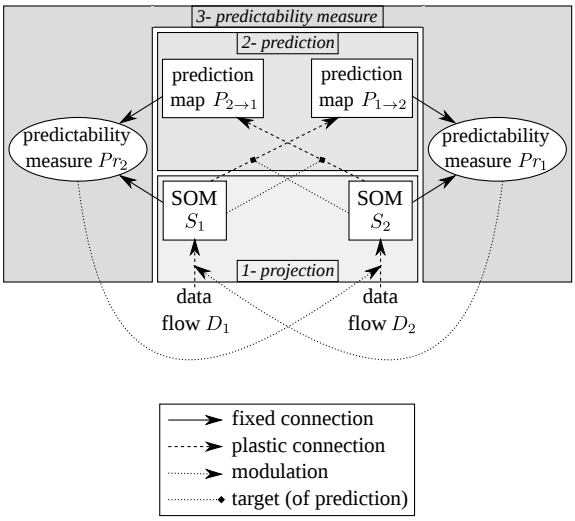

Fig. 1. PROPRE architecture is composed of three interacting modules. First, a projection module that provides a low dimensional representation of each modal stimulus. Second, a prediction of each modal representation by the other ones. Third, a predictability measure that quantifies the ability of a stimulus to predict the other ones and modulates the corresponding projection learning.

From a computational point of view, the reception of each multimodal stimulus in the model leads to one transmission and one learning steps 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. Then, the learning stage updates the weights of the plastic connections linking the modules. In the three next sections, we describe the equations used for each module of PROPRE. These equations are given for the processing and learning of the first data flow and are symmetrical for the second data flow.

B. Projection

S_1 is a discrete bi-dimensional square grid of neurons. Let $\mathbf{w_{S_1D_1}}(\mathbf{x},\mathbf{t})$ be $(w_{S_1D_1}(x,y,t))_y$ with $w_{S_1D_1}(x,y,t)$ the weight between the y-st value in the current stimulus $D_1(t)$ and the unit at position x in S_1 at time t. The activity of S_1 at position x at time t is computed as

$$S_1(x,t) = (\mathbf{w_{S_1D_1}}(\mathbf{x}^*,\mathbf{t}).\mathbf{D_1}(\mathbf{t}))e^{\frac{-||x-x^*||_2^2}{\sigma^2}}$$

with x^* the best matching unit of the map defined as $x^* = \max_x \mathbf{w_{S_1D_1}}(\mathbf{x},\mathbf{t}).\mathbf{D_1}(\mathbf{t})$[1]. σ is the variance of the Gaussian neighborhood radius and $||\cdot||_2$ is the euclidean

distance.

The incoming weights of the unit at position x in S_1 at time t are updated as following:

$$\begin{aligned}\Delta \mathbf{w_{S_1D_1}}(\mathbf{x},\mathbf{t}) &= \eta\lambda_1(t)S_1(x,t)(\mathbf{D_1}(\mathbf{t}) - \mathbf{w_{S_1D_1}}(\mathbf{x},\mathbf{t})) \\ \lambda_1(t) &= \begin{cases} 1 \text{ if } Pr_1(t) \geq \theta_1(x^*,t) \\ 0 \text{ otherwise} \end{cases} \\ \theta_1(x,t) &= \begin{cases} 0 \text{ if } t = 0 \\ \theta_1(x,t-1) \text{ if } x \neq x^* \\ \tau Pr_1(t) + (1-\tau)\theta_1(x,t-1) \text{ if } x = x^* \end{cases}\end{aligned}$$

with η the constant learning rate, $Pr_1(t)$ the predictability measure (see section II-D) and θ_1 the sliding predictability threshold. Compared to our previous work [2], we introduce this sliding threshold $\theta_1(x,t)$, defined for each unit x as an iterative average on a sliding window of the predictability measure when x is the best matching unit, so that it is autonomously adapted to the data.

This learning equation of the SOM weights is the one of Kohonen map in which we introduce the modulation term $\lambda_1(t)$. Thus, only stimuli that are more predictable than the average, with respect to the best matching unit, are learned by the system. This influences the distribution statistic of stimuli and consequently the obtained quantization of the SOM towards predictable stimuli across modalities.

C. Prediction

The projection activity in S_1 is used to provide a prediction in $P_{1\to2}$ of the projection activity in S_2. Consequently, $P_{1\to2}$ and S_2 have the same size. The activity in $P_{1\to2}$ at position x at time t is:

$$P_{1\to2}(x,t) = \sum_y w_{P_{1\to2}S_1}(x,y,t)S_1(y,t)$$

with $w_{P_{1\to2}S_1}(x,y,t)$ the weight from the unit at position y in S_1 to the unit at position x in $P_{1\to2}$.

The weights of the connection between S_1 and $P_{1\to2}$ are learned with an online version of the classical linear regression algorithm [25] that minimizes the mean square error between the prediction $\mathbf{P_{1\to2}}(\mathbf{t})$ and the target activity $\mathbf{S_2}(\mathbf{t})$:

$$\Delta w_{P_{1\to2}S_1}(x,y,t) = \eta' S_1(y,t)(S_2(x,t) - P_{1\to2}(x,t))$$

with η' the constant learning rate.

D. Predictability measure

The predictability measure provides a quantification of the prediction quality. In one previous article [2], we studied multiple measures and showed that the precise choice of the measure does not significantly influence the model performances. Here, we slightly adapt one of these measures.

Let define $X_2(t)$ as $\{x | S_2(x,t) > \epsilon\}$ with ϵ low and strictly positive. $X_2(t)$ is the set of indices corresponding to the location of the Gaussian in S_2 at time t (see section II-B). The predictability measure is computed as

$$Pr_1(t) = \frac{\sum_{x\in X_2(t)} P_{1\to2}(x,t)}{\sum_x P_{1\to2}(x,t)}$$

This measure, whose possible values are between 0 and 1, represents the proportion of the prediction corresponding to the correct location of the Gaussian and consequently the ability of one stimulus to predict the other one.

[1]In practice, we normalize the weights $\mathbf{w_{S_1D_1}}(\mathbf{x},\mathbf{t})$ and the input $\mathbf{D_1}(\mathbf{t})$ 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.

III. Experiments

A. Protocol

We recorded a set of ten left hand poses: point, fist, grip, L, stop and counting from 1 to 5 (see figure 2). The data were obtained using two ToF cameras which provide depth images of resolution 165x120 pixels at 90 frames per second. Since the ToF principle works by measuring the time the emitted light needs to travel from the sensor to an object and back pixel-wise the light is modulated by a frequency of 30MHz in order to be able to distinguish it from interferences. In a multi-sensor setup, as the one we use, this may lead to a distortion of measurements since both sensors have the same modulation frequency. To avoid such measurement errors, the data were recorded by taking alternating snapshots from each sensor.

Fig. 2. Gestures recorded by the two ToF cameras.

The two cameras were positioned with a 30° shifted angle and about 20 cm away from each other. In order to have some variability in the data, each pose was recorded with a variation of the hand posture in terms of translation (range of 20-50 cm from each camera) and rotation of the hand and fingers ($\pm$ 15°). Moreover, all ten gestures were recorded for eight different persons independently. As for each pose and each person, a set of 2000 point clouds was recorded for each camera, this yields a multimodal dataset of 160.000 samples.

Each data was preprocessed using a descriptor built upon the PFH (Point Feature Histogram) descriptor described in [26]. The PFH descriptor tries to describe the relationship for two points in a point cloud by the calculation of each of the point's normals as well as the distance between the points. The main difference between the descriptor used and PFH is that the PFH considers all couples of points in some neighborhood of each point of the cloud to calculate a quadruplet of features

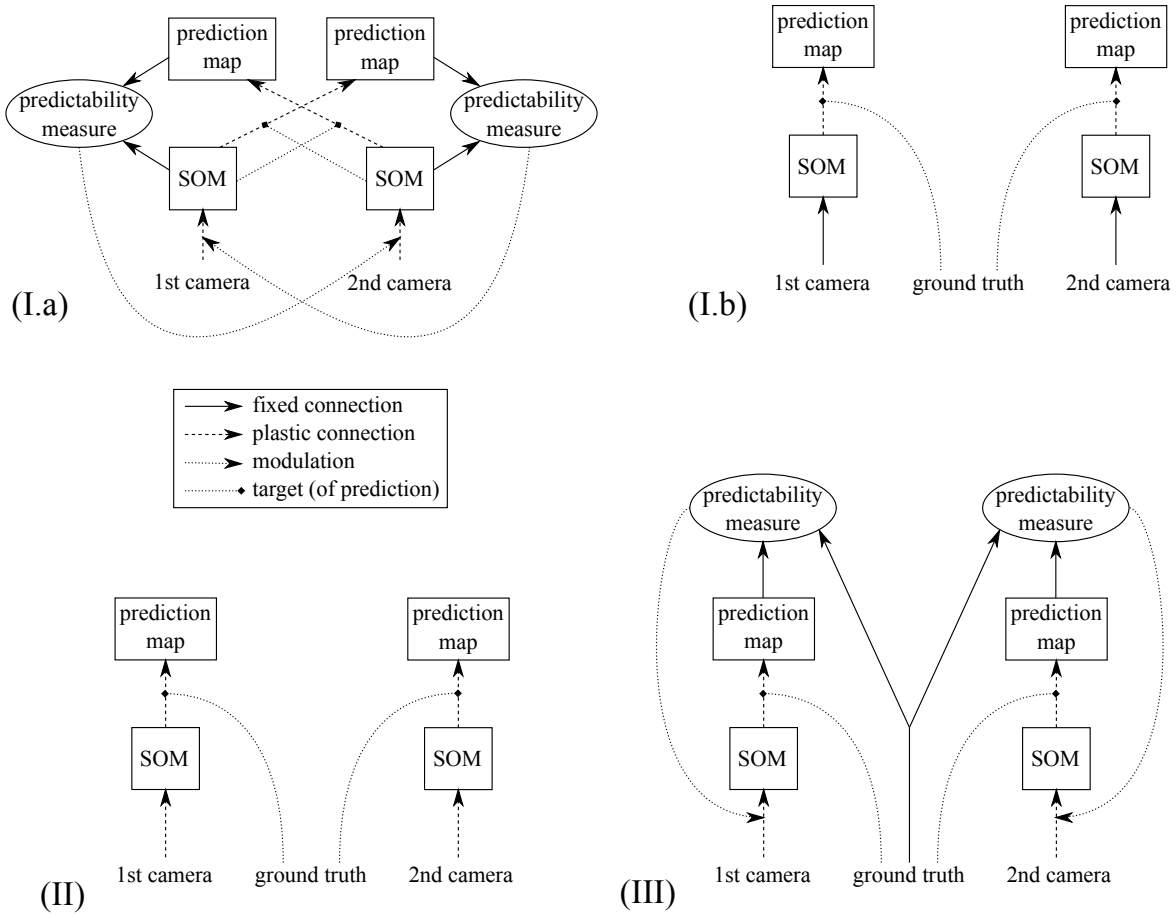

Fig. 3. Architectures compared in this article: unsupervised learning with PROPRE (I.a and I.b), supervised learning without the self-evaluation module (II) and supervised learning with PROPRE (III). Please refer to the text for more details.

(distance, pan, tilt and yaw) whereas we randomly subsample 10000 couples of points from the cloud and compute the quadruplet of features only on this sub-sampled set in order to discard variabilities in number of point detected, as already proposed in [27]. The resulting 10000 quadruplet values are normalized and grouped in 5 bins per feature leading to a 625 dimension vector histogram representing the hand pose for each camera. The gesture category is represented by a $(160, 7)$ discrete spatial coding vector.

B. Architectures

We tested our PROPRE architecture with the setup depicted in the previous section where preprocessed data provided by each camera defines one input flow for the system (see figure 3 I.a). This paradigm learns unsupervised representations of each data flow in the corresponding SOM with a focus on representations that are efficient to predict the ones of the other modality (see section II). After convergence of this unsupervised learning, we labeled these representations provided by the SOM. For this, we learn to predict the real category (ground truth) of the gesture presented to the cameras from each representation given by a SOM with previously learned and fixed connections (see figure 3 I.b). This supervised learning phase only targets to label data in order to quantify the classification performance of the model but is not mandatory for the use of PROPRE.

We compare the performance of this unsupervised learning with two baselines.

First, to quantify the influence of the modulation mechanism proposed in PROPRE, we tested the corresponding architecture without any modulation of the projection learning (see figure 3 II) which is equivalent to force the modulation term $\lambda_1(t) = 1,\ \forall t$ (see section II-C).

Second, we tested the performance of the predictability modulation corresponding to a supervised learning provided by a similar architecture. For that purpose, we use simultaneously two PROPRE architectures. Each architecture receives one flow provided by a camera and another one corresponding to the ground truth (see figure 3 III). In this case, the architecture is slightly adapted as the ground truth data flow is not projected as already proposed in [2].

C. Results

We randomly split our 160000 examples data set (see section III-A) in a learning and a test data sets respectively composed of 90% and 10% of the data. For each tested data,

the location of the induced highest peak in each prediction map encodes the gesture recognized from the corresponding camera. The multimodal classification of the system is obtained by the location of the highest peak in the sum of the prediction of both cameras. All presented results were obtained using 10×10 SOMs and averaged over 10 simulations with random initial weights for each setup.

1) Sliding threshold: In order to evaluate the influence of the new sliding threshold used in PROPRE, we compare the performance obtained by the model when using this sliding threshold or a fixed one. In the last case, we force both predictability thresholds to be equal to some constant going from 0.1 to 0.9 with a 0.1 incremental step (as the predictability measure is in $[0, 1]$). One of the advantage of the sliding threshold is that it is easy to parametrize (as independent of the input). But we can also observe that the performance of the system with this sliding threshold is equivalent or even better to the best performance that we can obtain with a fixed threshold (see figure 4).

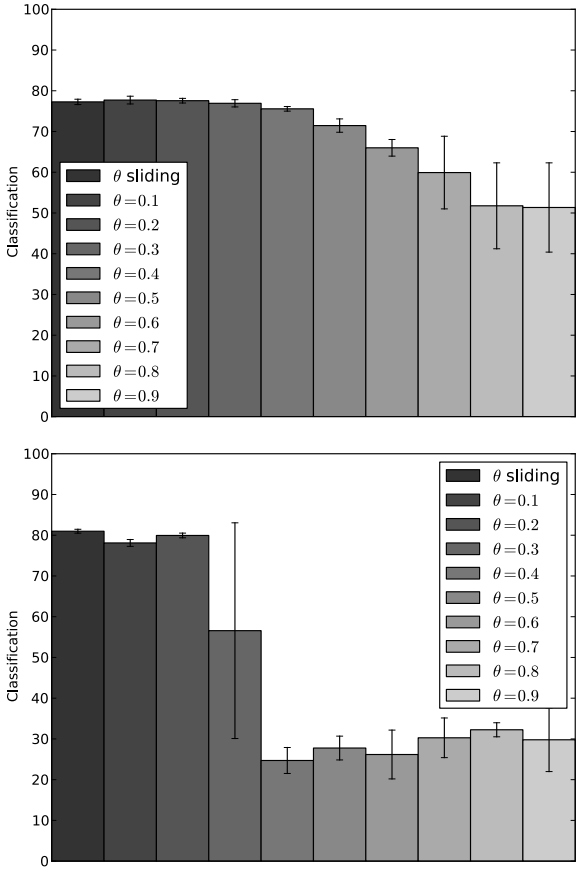

Fig. 4. Top (resp. bottom): Average multimodal classification performance obtained with unsupervised (resp. supervised) learning, i.e. architecture I (resp. III), depending on the threshold used (sliding or fixed).

2) Classification performance: Classification performance obtained with unsupervised learning (architecture I) are presented in figure 5 and compared to the one obtained without modulation (architecture II) and with supervised modulation (architecture III).

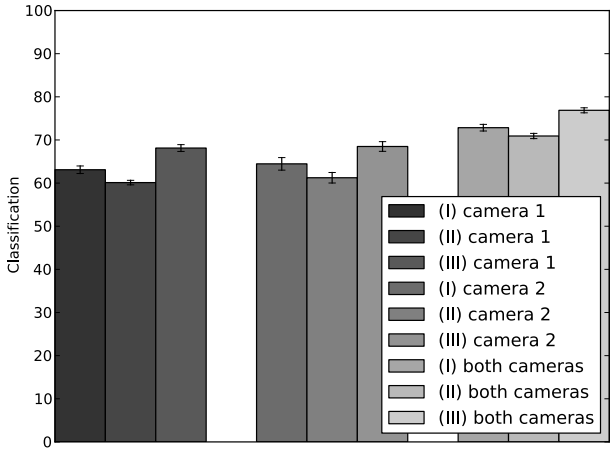

Fig. 5. Average classification performance over all ten gestures depending of the camera flow used for the prediction (one of two or both) and architectures.

We can observe that the predictability measure used with unsupervised learning improves slightly classification performance, compared to the baseline without modulation, in all cases (one camera or both). This seems to suggest that representations of stimuli correlated across modalities are meaningful as they improve the obtained classification performance. However, classification performance with unsupervised learning is not as good as the one achieved with supervised learning in PROPRE. Such a difference can be easily explained as the supervised learning task is easier as all data are labeled during learning. By the way, this supervised performance confirms the efficiency of the modulation introduced in PROPRE in this case as already shown in [2].

Moreover, we can notice that a simple multimodal merging, based on the sum of the different predictions, achieves a significant increase of the classification performance of around 8% for all three architectures. This emphasis the importance of combining multiple sensors for object recognition. Even if this improvement can be easily explained as considering both camera increases the amount of data available, it is quite surprising that its range is similar for all architectures, including the one using unsupervised learning. Indeed, in this case, SOM representations learned are stimuli correlated across both cameras. A deeper study will be necessary to understand this phenomenon and especially the precise importance of the unsupervised and supervised steps in the obtained performance.

IV. Conclusion and perspectives

PROPRE is a generic cortico-inspired paradigm that autonomously extracts meaningful representations of multiple input flows with an online and incremental predictability driven learning. It combines the projection of each modal flow on a dedicated self-organizing map with the prediction of each projection by the other ones. The main originality of PROPRE consists in the use of a predictability measure, that quantifies the quality of the predictions obtained from one representation, to influence the corresponding projection

learning. This predictability measure acts as a simple self-evaluation module of the system performance, i.e. its ability to learn multimodal correlations, that will bias the learned representations towards stimuli correlated across modalities.

In this article, we introduce a new sliding predictability threshold and apply PROPRE to the classification of preprocessed data corresponding to ten gestures recorded by two ToF cameras for eight different persons. In addition to being easier to parametrize, we show that this sliding threshold provides equivalent or even better classification performance than a fixed one.

Moreover, the use of a predictability measure reflecting the ability of one camera to predict the content of the other one, proposed in PROPRE, slightly improves the classification performance compared to a baseline without predictability module. This suggests that these representations of stimuli correlated across modalities, obtained with unsupervised learning, are meaningful for this classification task. However, obtained performances are not as good as the one obtained with a supervised predictability module, which confirm the efficient of the modulation mechanism used in PROPRE, but can be explained as the task is more difficult. By the way, we observe that in all cases, multimodal classification of gestures, obtained by summing the prediction driven by each camera, is significantly higher than the classification provided by considering only one of the two cameras.

These first results on learning representations from multiple data flows are promising and a deeper study will be necessary to confirm them with the combination of data coming from different kind of sensors. It will be also interesting to compare the performance of the representations extracted with unsupervised or supervised predictability influence when the system has to learn multiple tasks. The relative importance of the unsupervised learning stage and the post supervised labeling stage in the classification performance has also to be precisely quantified. Moreover, targeting the use of PROPRE for a sensori-motor learning, we also want to study the influence of temporal evolution of performance in the self-evaluation module computation, as already proposed in [28] for robotic curiosity, in order to explore and represent simultaneously the multimodal input space.

ACKNOWLEDGMENT

The authors want to thank Louis-Charles Caron (UIIS department, ENSTA, France) for his help on the descriptor used in this article. We also want to thank all persons of our laboratory that participate to the gesture database recording.

REFERENCES

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

[2] M. Lefort and A. Gepperth, "Propre: Projection and prediction for multimodal correlations learning. an application to pedestrians visual data discrimination." in *International Joint Conference on Neural Networks*. IEEE, 2014.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

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

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

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] M. Spratling, "Predictive coding as a model of biased competition in visual attention," *Vision Research*, vol. 48, no. 12, pp. 1391–1408, 2008.

[18] K. Friston, "Learning and inference in the brain," *Neural Networks*, vol. 16, no. 9, pp. 1325–1352, 2003.

[19] R. P. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature neuroscience*, vol. 2, no. 1, pp. 79–87, 1999.

[20] T. Jantvik, L. Gustafsson, and A. P. Papliński, "A self-organized artificial neural network architecture for sensory integration with applications to letter-phoneme integration," *Neural computation*, vol. 23, no. 8, pp. 2101–2139, 2011.

[21] M. Johnsson, C. Balkenius, and G. Hesslow, "Associative self-organizing map," in *International Joint Conference on Computational Intelligence (IJCCI*. Citeseer, 2009, pp. 363–370.

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

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

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

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

[26] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz, "Aligning point cloud views using persistent feature histograms," in *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*. IEEE, 2008, pp. 3384–3391.

[27] L. Caron, Y. Song, D. Filliat, and A. Gepperth, "Neural network based 2d/3d fusion for robotic object recognition." in *European Symposium on Artificial Neural Networks*, 2014.

[28] P.-Y. Oudeyer, "Intelligent adaptive curiosity: a source of self-development," 2004.