ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Slowness-based neural visuomotor control with an Intrinsically motivated Continuous Actor-Critic

Muhammad Burhan Hafez, Matthias Kerzel, Cornelius Weber, and Stefan Wermter [*]

University of Hamburg - Department of Informatics
Vogt-Koelln-Str. 30, 22527 Hamburg - Germany
http://www.knowledge-technology.info

Abstract. In this paper, we present a new visually guided exploration approach for autonomous learning of visuomotor skills. Our approach uses hierarchical Slow Feature Analysis for unsupervised learning of efficient state representation and an Intrinsically motivated Continuous Actor-Critic learner for neuro-optimal control. The system learns online an ensemble of local forward models and generates an intrinsic reward based on the learning progress of each learned forward model. Combined with the external reward, the intrinsic reward guides the system's exploration strategy. We evaluate the approach for the task of learning to reach an object using raw pixel data in a realistic robot simulator. The results show that the control policies learned with our approach are significantly better both in terms of length and average reward than those learned with any of the baseline algorithms.

1 Introduction

A core challenge for an agent that interacts with the world is to learn visuomotor abilities that map from raw high-dimensional sensory input to raw control output. This ability, however, cannot be directly learned supervised since there is no labeled training data already available, and the agent must learn online using a reward signal from its environment that is frequently delayed and sometimes sparse. Deep Reinforcement Learning (RL) has been widely used to learn this mapping by finding a control policy that maximizes the expected total reward, utilizing deep neural architectures as powerful nonlinear function approximators [1] [2]. One problem these approaches face is the sample complexity and the training time necessary to adjust the very large number of parameters (weights) such architectures typically have. State representation plays a key role here since low-dimensional, highly informative representations that are learned unsupervised often reduce the training parameters and time and are thus preferred to other representations. Slow Feature Analysis (SFA) is an unsupervised learning method that learns representations with these desirable properties.

SFA is primarily used to learn compact and temporally stable abstraction from a temporal input sequence [3]. Recently, slow features have been successfully applied as a low-dimensional input for precise prediction of view-invariant attributes of objects, such as identity, position and orientation in the context of supervised learning [4], for simultaneous learning of place and head-direction cells for potential robot

[*] This work was supported by the German Academic Exchange Service (DAAD) funding programme (No. 57214224) with partial support from the German Research Foundation DFG under project CML (TRR 169) and the Hamburg Landesforschungsforderungsprojekt.

navigation [5][7], and for learning a sequence of task-independent and increasingly difficult-to-learn abstractions from raw images that are used to discover reusable skills on a humanoid [6]. While Legenstein et al. [7] were the first to explicitly study the effect of using SFA output as state encoding in standard RL, experiments were limited to low-dimensional action space in a simple 2D environment with non-sparse reward in the variable-target task.

Beside unsupervised learning of efficient representations, truly autonomous robots often exist in environments where external rewards are absent or sparsely spread. Intrinsically motivated RL addresses this by equipping the agent with intrinsic drives, such as fear, hunger or curiosity which enable it to meaningfully explore its environment [8]. Several Intrinsically motivated RL methods were proposed after [8], but very few have addressed tasks with continuous action spaces which is essential for realistic autonomous systems [9] [10].

In this paper, we use state representations learned by a hierarchical SFA network and extend the Intrinsically motivated Continuous Actor-Critic (ICAC) algorithm [10] to the case of learning from raw image data. The proposed system comprises two distinct phases: unsupervised training of the SFA network on a visual input stream, and training the ICAC's actor and critic neural networks using the trained SFA network as a preprocessing step for the visual input to the ICAC networks. Section 2 describes the training process and the architecture of the hierarchical SFA. The ICAC algorithm on the SFA output is then presented in Section 3, followed by experiments and results in Section 4.

2 Hierarchical Slow Feature Analysis

The learning problem in SFA is an optimization problem of finding the most slowly varying features in an input signal. Given an I-dimensional input signal $x(t)$, the task is to find a set of J input-output functions $g(\boldsymbol{x}) = [g_1(\boldsymbol{x}), \dots, g_J(\boldsymbol{x})]^T$ such that the output signals $y_j(t) = g_j(\boldsymbol{x}(t))$ minimize

$$\Delta\left(y_j\right) \coloneqq \langle \dot{y}_j^2 \rangle \tag{1}$$

under the constraints

$$\langle y_j \rangle = 0 \qquad (zero\ mean), \tag{2}$$

$$\langle y_j^2 \rangle = 1 \qquad (unit\ variance), \tag{3}$$

$$\forall\, i < j : \langle y_i y_j \rangle = 0 \qquad (decorrelation) \tag{4}$$

with $\langle . \rangle$ and $\dot{y}$ indicating temporal averaging and the time derivative of y respectively. Constraints (2) and (3) avoid a trivial solution of constant output and constraint (4) ensures that different functions g_j contribute different features. This is a variational calculus optimization problem and is generally difficult to solve. But if the functions g_j are constrained to be linear combinations of nonlinear functions, then the problem is simplified. The solution is obtained by an eigenvector approach and can be found with the SFA algorithm which is guaranteed to find a global optimum [3].

In our work, we apply SFA hierarchically as in [4]. This is both computationally efficient and biologically realistic since hierarchical ordering requires only local communications similarly to how extensive connectivity is avoided in the neural circuits of the visual cortex [11]. Our hierarchical architecture consists of three layers of SFA nodes (see Fig. 1). Each node in the first layer has a receptive field of 8×8 pixels in the original 64×64 pixel input with 4-pixel overlap, resulting in 15×15 nodes with partially overlapping receptive fields. The second layer's nodes have receptive fields of 6×6 nodes in the lower first layer with 3-pixel overlap, resulting in 4×4 nodes. The top layer has a single SFA node that covers the full image frame. We compute 32, 32 and 16 SFA components for nodes in the lower, middle and top layers, respectively. All connections are topologically organized such that each SFA node receives inputs from neighboring nodes in the preceding layer.

Each individual SFA node implements four subsequent sub-processes, as shown in the inset in Fig. 1. First, a linear SFA is performed which reduces the effective dimension of the node input to 32; then the linear SFA output is quadratically expanded (original data with all quadratic combinations) to introduce non-linearities. White noise is then added to avoid unwanted singularities in the following SFA step. Finally, a second linear SFA is performed on the noise-added output. For training the network, we used 50K time points that correspond to a temporal sequence of input images (collected over 3.2 hours on the simulator) generated by a random walk described in Section 4. The training is done sequentially from bottom to top, each layer at a time using the same training sequence. Once trained, the network is used to generate an efficient state representation for the subsequent RL phase, as discussed in the next section.

3 Intrinsic Neuro-optimal Control

In [10] we proposed ICAC, an actor-critic RL algorithm that learns an optimal action policy by learning an ensemble of local predictive models of environmental dynamics online and generates an intrinsic reward based on the learning progress of each model. Using the learning progress of predictive models for selecting explorative actions is in accordance with the Goldilocks effect in infant cognition that attributes optimal learning to stimuli of an intermediate degree of difficulty [12]. This is evident in how infants seek increasingly complex learning samples by self-organizing their interaction with the world, moving from well-explored regions to others where they expect to learn new effects of motor activity. ICAC, however, has only been applied to low-dimensional inputs in non-visual control tasks. Here, we extend and evaluate ICAC for visually guided motor tasks using state representations learned unsupervised with hierarchical SFA.

ICAC, as shown in Fig. 1, has two components: a multi-predictor network and a control module. The first incrementally partitions the sensory space into local regions with local predictive models using the Instantaneous Topological Map (ITM) [13], a self-organizing network developed for strongly correlated stimuli as present in most robotics applications where data is generated by exploring along continuous trajectories and has been successfully used in RL [14]. The network gives an intrinsic reward based on the learning progress of the current region's predictor. The control module guides the action selection using the combined external and intrinsic reward as follows: We compute the change between two consecutive average prediction errors of a predictor associated with the best-matching ITM node n for the current world state (the SFA output). This change represents the learning progress the robot has made or expects to make and is combined with the perception error, which is the distance between the state encoding and the weight vector of n, to give the intrinsic reward. This self-generated reward encourages the robot to try actions that maximize its learning

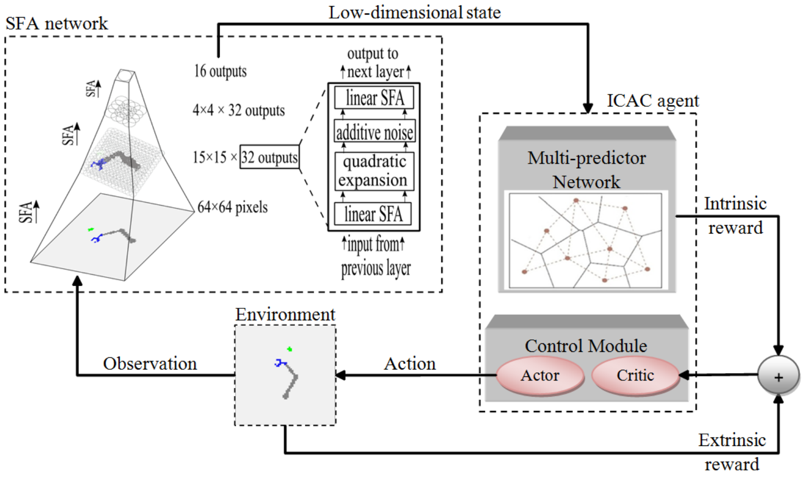


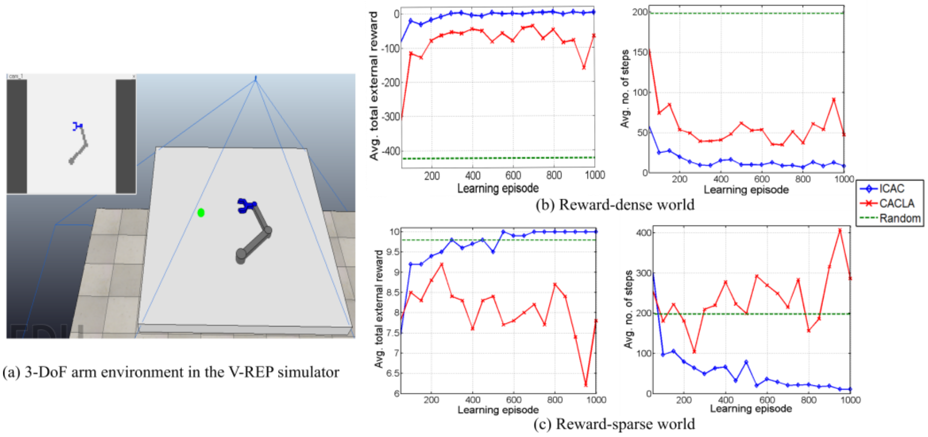
Fig. 1: SFA-based ICAC system. The ICAC agent takes an action chosen by the actor and the environment returns an observation. The SFA network then computes a state representation which the Multi-predictor Network uses to adapt its topology and, along with the current action, to update the learning progress of the predictor corresponding to the best-matching node from the previous timestep. The updated learning progress serves as an intrinsic reward that is combined with the external reward, if any, and fed to the critic to update its estimate of the utility of the current action. Finally, the actor is updated towards the current action if it is found to improve the critic's estimated utility.

progress and lead to perceptually novel states. For action selection, the Continuous-Actor-Critic-Learning Automaton (CACLA) [15], a state-of-the-art continuous action RL method, is used to learn an optimal control policy based on the combined intrinsic and external reward. We update the actor only when the critic's estimate increases, as opposed to gradient ascent on the value [2]. Both the critic and actor are represented by feed-forward neural networks and updated online from the sample transitions the robot experiences while interacting with the world (for details on ICAC refer to [10]).

4 Experiments and Results

We evaluate our algorithm on the random-target reaching task with a 3-DoF arm shown in Fig. 2 (a) in the V-REP robot simulator. The robot arm's joints can move within $[-\pi/2, \pi/2]^3$ representing the joint values. The robot gets a reward of 10.0 on reaching the goal region (green sphere in Fig. 2(a)) or, otherwise, a negative reward based on the Euclidean distance of the gripper tip to the goal. We also performed experiments on the more challenging sparse-reward setup (reward of 10.0 on reaching the goal region or 0.0 elsewhere). To train the SFA network, we performed random walk of 50K steps (3.2 hours on the simulator). In each step the robot takes a random action (max of 1 degree per joint) in the simulator and records an image of the world. The image sequence is then used as a single batch to train the network on three passes corresponding to the three layers of SFA nodes (see Sec. 2).

The actor and critic in ICAC and CACLA are represented by 2-layer, fully connected MLPs of 20 tanh hidden units. The output units are linear: three in the actor



(a) 3-DoF arm environment in the V-REP simulator

(b) Reward-dense world

(c) Reward-sparse world

Fig. 2: The environment with the 3D arm model and the output of the vision sensor in the upper left corner (a). Learning curves, averaged over 10 runs, for reaching with 3-DoF arm using the baselines and ICAC in different reward setups: reward-dense (b) and reward-sparse (c). The average over 50 episodes is shown for readability.

and one in the critic. The input to the actor and critic is the 16-dimensional SFA-output and the goal's Cartesian coordinates (the goal is not rendered in the visual input since only its current position is relevant for reaching). The learning rate was set to 0.01. We stored the most recent 1M state transitions in an experience replay buffer and performed mini-batch SGD with batch size 1000 for updating the critic during online training. The reward discount was 0.99. All actions (joint value changes) were drawn from a Gaussian distribution with a mean at the actor's present output and standard deviation of 0.3 and were clamped to a max of |10.0| degrees per joint. The predictors are 2-layer MLPs trained to predict future states (SFA-output) from previous state/action pairs. The above values were determined empirically. In each learning episode, the robot is given a maximum of 1000 steps to reach the current goal after which a new episode begins with a random goal position.

We run ICAC and two baselines of CACLA and uniform random policy for 1K episodes and averaged the results over 10 independent runs. As shown in Fig. 2, the ICAC reached much better policies in a much smaller number of episodes in both reward setups. While CACLA was able to converge to a relatively good policy in the reward-dense setup, it failed in the reward-sparse setup where ICAC was converging to a near optimal policy after 600 episodes of learning. In Table 1, we give a more concise comparison. Again, the table shows that ICAC outperformed the baseline approaches and that its performance was significantly higher in the difficult setting of the reward-sparse world, reaching a success rate of 100% over the last 200 episodes.

	Learning speed			Final performance		
	Random	CACLA	ICAC	Random	CACLA	ICAC
reward-dense world	-423.58	-85.93	**-7.81**	-434.70	-97.55	**0.87**
reward-sparse world	**9.80**	8.07	9.64	9.95	7.45	**10.00**
reward-dense world	198.15	58.17	**15.20**	201.08	63.26	**10.63**
reward-sparse world	198.15	236.58	**57.65**	201.08	250.50	**30.82**

Table 1: Mean reward (upper half) and mean no. of steps (lower half) over the entire learning period (learning speed) and for the last 200 episodes (final performance).

5 Conclusion

In this paper, we present SFA-based ICAC as an algorithm for reinforcement learning of visually guided reaching tasks. We demonstrate that deep slow features can be used to learn efficient and informative state representations critical for fast online learning of optimal control policies. SFA, however, is susceptible to learning task-irrelevant slow features like a change in the scene's background. Enhancing SFA to address such issues gives an interesting direction for future work. The results presented in this paper prove that the exploration strategy of ICAC driven by the intrinsic reward is effective for finding optimal policies in a reward-sparse world. This is in line with theories from developmental psychology that stress the importance of intrinsic reward in complex environments [10] [12].

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature,* vol. 518, no. 7540, pp. 529-533, 2015.

[2] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver and D. Wierstra, "Continuous control with deep reinforcement learning," in *International Conference on Learning Representations (ICLR)*, 2016.

[3] L. Wiskott and T. Sejnowski, "Slow feature analysis: unsupervised learning of invariances," *Neural Computation,* vol. 14, pp. 715-770, 2002.

[4] M. Franzius, N. Wilbert and L. Wiskott, "Invariant object recognition and pose estimation with slow feature analysis," *Neural Computation,* vol. 23, no. 9, pp. 2289-2323, 2011.

[5] X. Zhou, C. Weber and S. Wermter, "Robot localization and orientation detection based on place cells and head-direction cells," in *Proceedings of the 26th International Conference on Artificial Neural Networks (ICANN)*, Sardinia, Italy, 2017.

[6] V. R. Kompella and L. Wiskott, "Intrinsically motivated acquisition of modular slow features for humanoids in continuous and non-stationary environments," *arXiv preprint arXiv:1701.04663,* 2017.

[7] R. Legenstein, N. Wilbert and L. Wiskott, "Reinforcement learning on slow features of high-dimensional input streams," *PLoS Computational Biology,* vol. 6, no. 8, 2010.

[8] S. Singh, A. G. Barto and N. Chentanez, "Intrinsically motivated reinforcement learning," in *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, British Columbia, Canada, 2005.

[9] C. Florensa, Y. Duan and P. Abbeel, "Stochastic neural networks for hierarchical reinforcement learning," in *5th International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

[10] M. B. Hafez, C. Weber and S. Wermter, "Curiosity-driven exploration enhances motor skills of continuous actor-critic learner," in *Proceedings of the 7th Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, Lisbon, Portugal, 2017.

[11] A. A. Koulakov and D. B. Chklovskii, "Orientation preference patterns in mammalian visual cortex: a wire length minimization approach," *Neuron,* vol. 29, no. 2, pp. 519-527, 2001.

[12] K. E. Twomey and G. Westermann, "A neural network model of curiosity-driven infant categorization," in *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, Rhode Island, USA, 2015.

[13] J. Jockusch and H. Ritter, "An instantaneous topological mapping model for correlated stimuli," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, Washington, 1999.

[14] M. B. Hafez and C. K. Loo, "Topological Q-learning with internally guided exploration for mobile robot navigation," *Neural Computing and Applications,* vol. 26, no. 8, pp. 1939-1954, 2015.

[15] H. Van Hasselt, "Reinforcement learning in continuous state and action spaces," in *Reinforcement Learning,* Springer, Berlin, Heidelberg, 2012, pp. 207-251.