Hindawi
Complexity
Volume 2018, Article ID 7609587, 15 pages
https://doi.org/10.1155/2018/7609587

*Research Article*

# Encoding Longer-Term Contextual Information with Predictive Coding and Ego-Motion

**Junpei Zhong**,[1,2] **Angelo Cangelosi**,[2,3] **Tetsuya Ogata**,[1,4] **and Xinzheng Zhang**[5]

[1]*National Institute of Advanced Industrial Science and Technology (AIST), Japan*
[2]*Plymouth University, UK*
[3]*University of Manchester, UK*
[4]*Waseda University, Japan*
[5]*Jinan University, China*

Correspondence should be addressed to Junpei Zhong; junpei.zhong@plymouth.ac.uk

Received 13 July 2018; Accepted 23 October 2018; Published 13 November 2018

Guest Editor: Yimin Zhou



Studies suggest that, within the hierarchical architecture, the topological higher level possibly represents the scenarios of the current sensory events with slower changing activities. They attempt to predict the neural activities on the lower level by relaying the predicted information after the scenario of the sensorimotor event has been determined. On the other hand, the incoming sensory information corrects such prediction of the events on the higher level by the fast-changing novel or surprising signal. From this point, we propose a predictive hierarchical artificial neural network model that examines this hypothesis on neurorobotic platforms. It integrates the perception and action in the predictive coding framework. Moreover, in this neural network model, there are different temporal scales of predictions existing on different levels of the hierarchical predictive coding architecture, which defines the temporal memories in recording the events occurring. Also, both the fast- and the slow-changing neural activities are modulated by the motor action. Therefore, the slow-changing neurons can be regarded as the representation of the recent scenario which the sensorimotor system has encountered. The neurorobotic experiments based on the architecture were also conducted.

## 1. Introduction

*1.1. Prediction in Systems.* Delays always exist in both biological and engineering systems. They are typically caused by the processing time in biological organs, electrical circuits, or computer programs. Besides, noises, nonlinearities, delays, uncertainties, and redundancies are also other factors that caused delays in the system. In order to react and move adaptively, the system needs to be compensated by two means: firstly the system extrapolates the upcoming percept within the delayed time taken for the signal to travel between the various components of system; secondly, it plans ahead the upcoming events based on the contextual background, including its own perception and action, until it reaches the end of the sensorimotor loop.

Basically, to solve such a delay problem, the critical point for the system is to control the independent variables and *plan ahead* the changes of the dependent variables. Here, from the perspective of a control system, the independent variables are those variables that can be adjusted by the regulatory controllers or by the cognitive processes, while the dependent variables are the processes that have relatively slower responses compared with the changes of the independent variables. For instance, in the muscular contraction case, the muscle power is the dependent variable because it runs relatively slowly compared with the changes of the sensory transition and the cognitive commands from the motor cortex, which makes the real-time feedback control of the motor control infeasible [1, 2].

Such a planning-ahead prediction may be not a trivially constant event as percept extrapolation [3] or may even be complicated events such as competitive sports, when they are related to the background knowledge or the contextual information. For instance, in the biological system, such a

delay in the dependent variables in processing time may be caused by the neural processing. But the elimination of such delays is crucial for safety and adaptivity in their behaviours. The observation of such predictive models in behavioural experiments showed that the prediction in the biological world helps the diving birds to retract their wings before entering water [4] and the houseflies as they land [5]. The predictive nature has widely been explored and demonstrated through studies in neural systems as well. For instance, experiments demonstrated that monkeys have the ability to conduct smooth pursuit movements with zero retinal slip [6].

To solve the delay problems in a more complicated world and to save the bandwidth in perception, it is widely accepted that the predictive models should be used in many ways of our brain system (see also, e.g., [7–10]). The predictive model in biology should be able to adaptively compensate the current delay and adjust the parameters of itself to adapt the environment, both of which are based on learning from the (noised) measurement of independent variables. When we regard the sensorimotor system as a control system, such a prediction is involved in the coordination in both perception and action and it is accomplished by at least two strategies:

(1) An observation to the independent variables and predicting the trend of them, which results in the predictive control to the dependent variables.

(2) An understanding and classification of the world which the system is interacting with, which provides a cue for changing different prediction strategies (i.e., meta-learning).

Specifically, these two perspectives of control correspond to the unconscious or conscious ways by which our cognitive functions work, and they depend on the complexity of the events in the sensorimotor loop. Note that none of them works alone in the complete predictive sensorimotor system of our own. Instead, the complete predictive system in our brain incorporates different degrees of combination of both strategies. For instance, catching a ball requires both prediction and reflexive (feedback) mechanisms [11, 12]. Such predictive responses can be even detected by the muscle signals of the arm muscles (biceps and triceps) prior to the ball reaching the hand, which indicates that such a prediction mechanism is triggered unconsciously. Following the contact, the reflexive feedback mechanisms are engaged by the receptors in the hand and the arm. In such processes, the cerebellar and the motor cortex output is a signal that can be combined with delayed sensory feedback as well as the predicted state of the limb from the history of motor commands, which is from our conscious brain dynamics.

*1.2. A Predictive Coding Answer.* A stronger theory about prediction is called predictive coding (PC). It is about our brain that is constantly acting in a predictive loop and such predictive perception not only works as a compensation of the delay, but also forms part of higher level of cognitive functions. It is realised by "predicting the error itself" in the hierarchical way. Considered to be a unified predictive framework to integrate the perception and action, generally, PC asserts that the sensorimotor loop works as a predictive machine which utilises both perception and action to minimize the prediction error. This minimization is actively processed by the integral model of the functional explanation to our sensorimotor experience and perceptual consciousness. This integrative process can be specified as follows: to integrate the predictive perception and action, the sensorimotor system needs to predict an incoming flow of sensory stimuli from the contextual factors and the internal model should be able to deal with a priori provided by its perceptual experience. In this context, the cognitive process runs as a predictive machine unconsciously that actively generates predictions of its sensory percept using the prelearnt internal model. In the meanwhile, the internal model also learns about the statistical structure of the world and infers the posterior of one scenario or event following another in order to generate a prediction of what the current state is likely to be and adjust the parameters or the strategy of itself, if necessary. This is based on the perceptual consciousness. Therefore, the PC theory provides a unified explanation about how our brain forms the conscious prediction from the unconscious anticipation.

Quantitatively, such an estimation of the uncertainty can be formulated by the (approximate) Bayesian inference [13], which calculates the posterior percept that has the highest posterior probability based on the prior:

$$P(E \mid A, I) \propto P(A \mid E)\, P(E \mid I) \tag{1}$$

where $E$ estimates the upcoming perception evidence given an executed action $A$ and other prior information the brain have already known ($I$). The term $P(A \mid E)$ suggests a prelearnt model representing the possibility of a motor action $A$ will be executed given a (possible) resulting sensory prediction ($E$) is perceived from the higher-level prediction (top-down computation). The equation assumes that the decision of the current action depends on the perception at the moment.

At the same time, it is also possible that we are allowed to select an appropriate movement given possible perceptual inputs as well as our goal.

$$P(A \mid E, G) \propto \frac{P(E \mid A)\, P(A \mid G)}{P(E \mid G)} \tag{2}$$

where $A$ represents a particular action selected given the (intended) sensory information $E$ and a goal $G$. Here we assume that one's action is only determined by the current sensory input and the goal may modulate the action selection through the whole hierarchy.

To summarize both perspectives, the brain is always attempting to build the active processes which minimize the estimation error between such posterior estimation and the truth, by changing its internal learning model ("perceptual inference") (see also [14, 15]) or by action execution ("active inference") (see also [16, 17]). As such, perceiving the world (perceptual inference) and acting on it (active inference) are two processes that aim at minimizing the prediction error in the hierarchical architecture. To learn the Bayesian models for perception and action, one needs to make continuous

learning throughout childhood, through interaction with the environment. For instance, the object-directed reaching [18] and grasping [19] during the early stages of infant development are considered to be the learning of movements and sensory effects with the forward models [20], with consideration of object affordances.

While the neural areas are organized in a hierarchical way [21], based on the PC framework, the neuronal representations on the higher level generate predictions of representation in lower levels. This kind of predicting neural populations suppresses or inhibits prediction-error populations. And on each level, in turn, the bottom-up activities only carry the residual errors which attempt to correct the predicting neural activities patterns that occur on each level. As [22] discovered from the neurobiology evidence: "Predictive coding theories posit that the perceptual system is structured as a hierarchically organized set of generative models with increasingly general models at high levels." In the meanwhile, on each level of this hierarchical generative model, the role of action can also be found to suppress the proprioceptive prediction errors at the level of the spinal cord and cranial nuclei. These prediction-error minimization rules have been formulated in [10] which suggested a mathematically elegant way, namely, the free-energy principle, to describe a unified story around perception, action, and their prediction.

Therefore, the notion of hierarchies in the sensorimotor loop is essential for the PC system to make perceptual inference. Within the PC framework, many of the consciousness issues related to perceptual experience and the mental state may account for the intransient representations on the higher level in the hierarchy, where there is an amodal experience on all the modalities of transient inputs but without much effect from the top-down knowledge [23]. And such a high level understanding based on multiple modalities also provides a source for extrapolation of the perception on the lower level, through the top-down prediction in the hierarchies. Mathematically, such kind of hierarchical generative models enable the high level of neural representation to optimize our own prior beliefs, and therefore, based on such prior per se, that is the reason we can see (or predict) our interoceptive inputs.

## 2. Related Works

*2.1. The Amodal-Based Prediction.* To solve the engineering problem of controlling the dependent variables by predicting the independent variables, some approaches tried to model the trend of the independent variables. The amodal-based approaches for predictive control are designing models which can be used to predict the current values of the output variables based on the prebuilt model. A typical predictive model should consider the residuals, the differences between the actual and predicted outputs, to generate the feedback signal to this prediction model. The Model Predictive Control (MPC) method was firstly used to solve the control system with long delays and containing a set of dependent variables with long delays. It basically predicts the change in the dependent variables of the system that will be caused by changes in the independent variables. This is done by solving an optimization problem to predict the future outputs and control actions. It considers a finite sequence of control actions occurring in the past fixed length of time (aka. horizon). Since MPC controllers are mostly implemented in digital systems which have been constrained and discrete-time properties in an online manner, although MPC provides nonoptimal solution, this trade-off is still acceptable for control applications for which the optimal control problem could not be solved in real-time.

Because of its real-time properties, MPC control has been applied in various robotic applications, especially for trajectory planning and controls. For instance, [24–26] investigated the path-planning function of an autonomous mobile robot, and it was realised by the learning-based MPC which took a priori physical model of the world and a learned disturbance perception model into consideration. To guarantee the robustness of the performance, [27] used a linear MPC with bounds on its uncertainty to construct invariant in the path planning. Reference [28] applied a linearised tracking-error dynamics to predict future system behaviour using a quadratic cost function. On the other hand, besides the prediction of the robot itself, the predictive model can be used in the world model as well [29–31].

To summarize, studies were based on the amodal-based approaches focusing on the tracking of the change of the independent variables by only describing the properties of the variables, including the system and the environment. However, the difficulty of these predictive problems is that some of the independent variables of the system can be only described by the higher-order variables which are difficult to track. Such variables can be the goal positions and orientations, which are not reachable asymptotically by means of smooth and time invariant feedback control laws.

*2.2. Multimodal Prediction by Interacting.* In this article, the concept of the multimodal learning methods usually employs mechanisms to optimize both the perception and action as an integral part by the observation from the environment. Such embodied learning mechanism can be considered as a kind of the internal models for animals too, which can be considered foundational in cognitive science (e.g., [32, 33]). Since the multimodal aspect of the sensorimotor model, such kind of internal models usually emphasizes the embodied and the situated nature of the agents and learns from interacting with the world [34].

The predictive function of the internal model can range from short- and mid-term time-scale prediction/delay compensation to relatively long-term planning behaviours which emerge from the short-term simulations. The short-term predictive models are mostly related to sensorimotor control, especially to the maintenance of the consistency of visuomotor coordination (e.g., [35, 36]) or fast reaction (e.g., [37, 38]).

There are evidences that such kind of short-term neural prediction may result in some predictive behaviours as well. Papers [20, 39] studied how to apply internal models in controlling the actual motor actions. The research [40] also extends the model to learn imitation behaviours. All of the three models built a forward predictive model to control the robot and acquire certain behaviours. Similarly, a long-term

planning behaviour can also emerge by internal simulation if the prediction is well planned before (e.g., [41, 42]). Reference [43] reported that experiments with a mobile robot with a two-level recurrent architecture implemented were able to accomplish the linguistic and sensorimotor task. An extension model has also been examined in symbolic understanding tasks [44].

If we regard the unified learning scheme of prediction with different time-scales, the Multiple Time-scale Neural Network (MTRNN) was proposed by [45]. The model is able to represent different temporal scales of sensorimotor information in the hierarchical structure of the sensorimotor sequences, such as the spelling of words [46] and object features/movements [47]. Extended from the MTRNN model with multiple modalities, the multiple spatiotemporal scales RNN (MSTRNN) [48] integrates the MTRNN and the convolutional neural networks [49, 50]. It includes two modalities: the temporal properties and the receptive field, both of which differ in the spatial sizes and the time constants in different levels. The PredNet [51] also holds a similar concept of using the convolutional network to capture the local features of the visual streams, but the architecture only ensures the learning of deterministic property of the temporal sequences.

In this paper, we propose a computational-feasible model using deep neural network, based on the predictive coding (PC) model to learn the datasets in the real world. Similar to the PC framework, the model unifies two ways of prediction:

(1) the recognition of the scenarios based on the hierarchical sensorimotor interaction, which may result in the higher-level cognitive computation;

(2) the prediction in visual percept modulated both by the "understanding" of the world as well as the voluntary ego-motion (Note that muscle contraction should be considered as the dependent variables with delay and disturbances.).

Note that prediction of these two perspectives by changing of the dependent variables is not totally separate in the sensorimotor systems. Instead, they can be combined and mixed on different levels. This model also shows how this integration of conscious and unconscious prediction occurs determined only by a single parameter.

## 3. Model

*3.1. The Action Modulated Predictive Model.* In terms of the architecture, the Multiple Time-scale Action Feedback Augmented Predictive Network (MT-AFA-PredNet) (Figure 1) is similar to our previous work called AFA-Prednet [52]. Consistent with the general concept of perception-action integration, it integrates the motor action as an additional signal which modulates the top-down generative process via an attention mechanism. Moreover, the multiple time-scales of the different levels result in the difference in updating rate of such a perception-action integration.

As most of the deep learning architectures, the network consists of a series of repeating stacked modules in a hierarchical way which attempts to make local predictions of the visual inputs. In general, the MT-AFA-PredNet is functionally organized as an integration with two networks: the left part is equivalent to a generative recurrent network (*top-down*), while the right part is a standard convolutional network (*bottom-up*). Each layer of the network consists of three basic parts: a generative unit (*GU*, green) containing the recurrent convolutional networks, a discriminative unit (*DU*, blue) containing convolutional networks (CNN), and the error representation layer (*EL*, red). The generative unit, *GU*, is usually a generative model that is able to give a prediction of the next time-step given the current input. Here, the convolutional LSTM [53] is employed to generate the local prediction in the image region. We employ a number of independent recurrent units as one layer of the *GU* units to ensure they learn different possibilities of the prediction based on the modulation value from the motor action. During training with various action-perception paring occasions, each of these units implicitly memorizes different possibilities of the prediction (e.g., the moving direction) with respect to the motor action in a self-organized way.

The *DU* networks calculate the differences between convolutional output of the predicted signal from *GU* and the bottom-up signal as an error representation, *EL*, which is split into separate rectified positive and negative error populations. The error, *EL*, is then passed forward through a convolutional layer and becomes the input to the next layer. The recurrent prediction layer $R_l$ receives a copy of the error signal *EL*, along with top-down input from the representation layer of the next level of the network ($R_{l+1}$).

*3.2. Multiple Time-Scales.* The concept of the multiple time-scales in artificial recurrent neural networks was firstly proposed in [45]. In this hierarchical network, all neurons on the same layer have the same updating rate, but the updating rates of neurons on different layers differ. Specifically, neurons on the lower levels have faster updating rates, which are called fast context neurons (or layers), while neurons on the higher levels have slower updating rate, called slow context neurons (or layers). This difference between the fast and slow context layers is determined by the time constants $\tau$, which determine the speed of the adaptation given a time sequence with a specific length, when updating the neural activity. The larger the value of $\tau$, the slower the neuron adaptation. The difference of adaptation rate of the neurons further assembles features of the input sequences in various time-scales, which results in the representation of the long-term context: given the temporal states $S(0), S(1), \ldots, S(t)$, their spatiotemporal features will be self-organized on different levels of the network. Therefore, such oscillatory patterns in the RNN are formed by self-organizing as fixed points and the limited-cycle nonlinear dynamics are memorized.

In the context of the MT-AFA-PredNet, the time constants $\tau$ are set in the generative units, i.e., the convolutional LSTM units, in which the values are updated with influence from the previous state.

$$R_l^d(t) = \left(1 - \frac{1}{\tau}\right) R_l^d(t) + \frac{1}{\tau} R_l^d(t-1) \tag{3}$$

where $R(t)$ is the output of the $d$-th generative unit (*GU*, i.e., ConvLSTM here) at time $t$ on $l$-th layer.

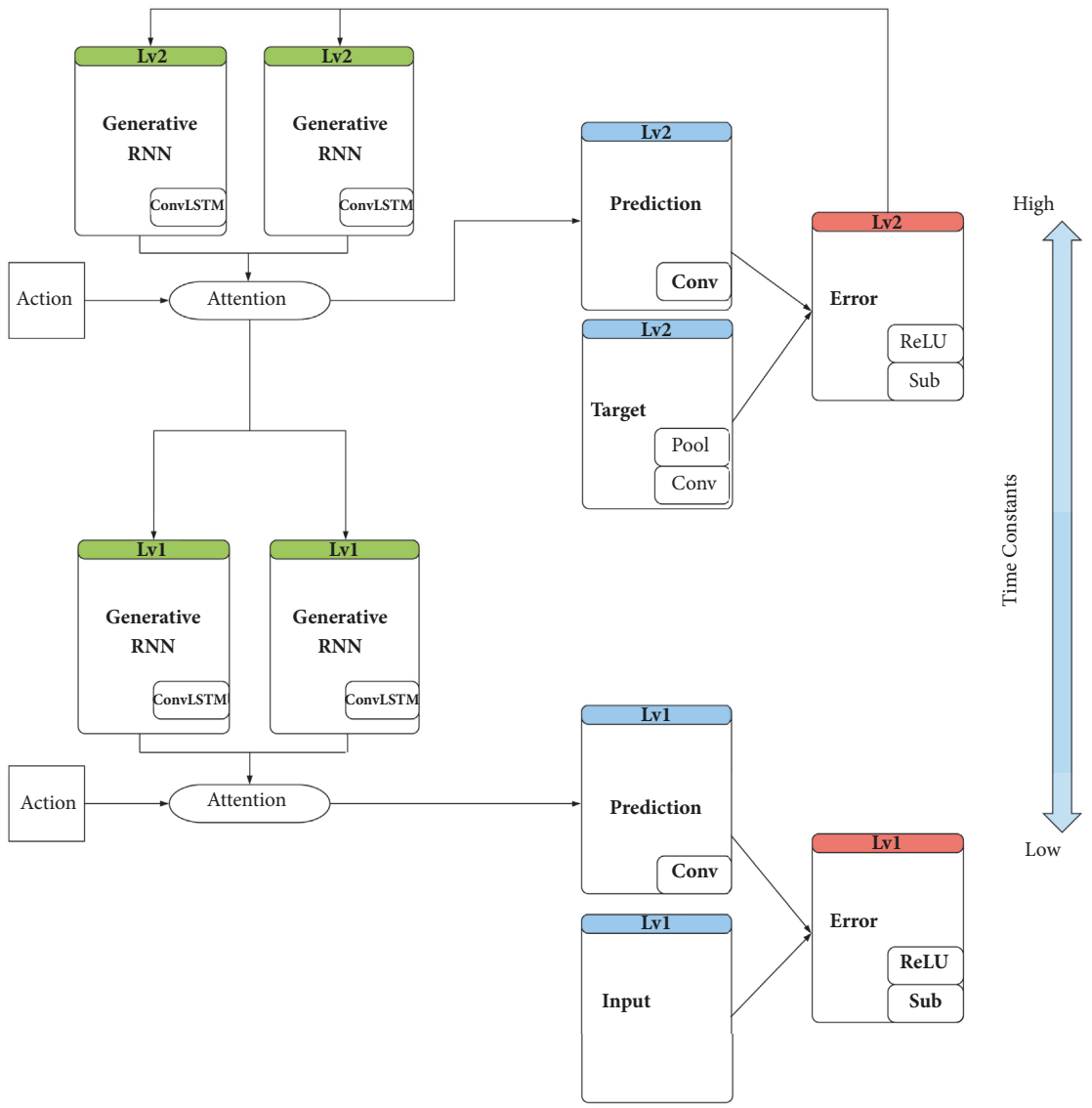



Figure 1: A 2-layer MT-AFA-PredNet. A 2-layer MT-AFA-PredNet Example: the green (generative) units calculate (Eq. (9)) for prediction and further combine proportionally by the attention unit (Eq. (10)). The red units calculate the error (Eq. (8)). The error is obtained from blue units (Eq. (7) (Prediction) and Eq. (6) (Target)). Moreover, the time constant changes on different levels.

*3.3. Action Modulation.* We employ number $N$ of independent recurrent units on one layer as a number of the *GU* units which independently memorize the possible changing possibility of the pixels. Since we have multiple generative units (*GU*), the role of the modulation of the motor action is to selectively integrate different prediction results from the *GU* units. This is loosely similar as the attention mechanisms of visual systems (see also [54]), because only part of the generative outputs contributed to the overall prediction. Such kind of attention mechanisms have recently been used in NLP [55], text translation [56, 57], and language models [58]. Inspired from the attention model, but considering the motor modulated role works as a Markov model, we use the motor action at the current time $t$ as the input of the attention model (5). This attention model is further used as the normalization term of the outputs from the multiple convolutional LSTM units (6). Such possibility is further integrated by the softmax functions in the attention unit, calculated by the motor actions.

$$R_l = \frac{attention\,(a) \times R_l^d\,(t)}{\sum_{d=1}^{N} attention\,(a) \times R_l^d\,(t)} \tag{4}$$

where

$$attention\,(a) = \frac{\exp\left(a^i \cdot w^{id}\right)}{\sum_{d=1}^{N} \exp\left(a^i \cdot w^{id}\right)} \tag{5}$$

in here, $N$ equals the number of convolutional LSTM units on each layer. With the soft-max function, we can have a normalized combination to all the outputs from the convolutional LSTM units.

*3.4. Algorithm.* We denote the indices of these perception input image as $i_t$, and the target of the network prediction at the lowest level is set to the actual percept at the next time-step $i_{t+1}$. We directly put the image as the input of the lowest layer, layer 1, so the input of the layer 1, $X_1$, equals the actual image data $X_1^t = i_t$.

The targets for higher layers at time-step $t$ is denoted as $X_l(t)$. Except layer 1, $X_l(t)t$ is obtained by the higher-level representation of the deep convolutional layer, which follows a usual calculation process of the convolutional network as shown in (6): the convolution kernel, the rectified linear unit (ReLU) calculation, and the max-pooling are sequentially used. This bottom-up process uses convolutional network to extract the local features of the error.

At the *GU* unit, the generative process is determined by the representation from the recurrent connection (i.e., from the previous time-step) $X$, the bottom-up error $E_l((t-1)$, and the top-down prediction $R_{l+1}(t)$. Such a prediction in a convolutional LSTM is calculated as (9): a deconvolution is used to reconstruct a larger size of the (predicted) representation $\widehat{A}$ after the rectified function (ReLU) (Eq. (7)).

To avoid the drawback of the ReLU which only captures the positive and negative error, the error representation $E_l(t)$ is calculated from the positive and negative errors (Eq. (8)), as the original PredNet does.

During training with various action-perception paring occasions, in the instance of mobile vehicle, with different turning directions at the cross, each of these units implicitly memorizes different possibilities of the prediction (e.g., the moving direction) with respect to the motor action in a self-organized way.

$$X_l(t) = \begin{cases} i(t), & \text{if } l = 1, \\ MAXPOOL\left(f\left(Conv\left(E_{l-1}(t)\right)\right)\right), & \text{if } l > 1 \end{cases} \quad (6)$$

$$\widehat{X}_l(t) = f\left(Conv\left(R_l(t)\right)\right) \quad (7)$$

$$E_l(t) = \left[f\left(X_l(t) - \widehat{X}_l(t)\right); f\left(\widehat{X}_l(t) - X_l(t)\right)\right] \quad (8)$$

$$R_l^d(t) = \left(1 - \frac{1}{\tau}\right) R_l^d(t) + \frac{1}{\tau} ConvLSTM\left(E_l(t-1), R_l(t-1), DevConv\left(R_{l+1}(t)\right)\right) \quad (9)$$

$$R_l(t) = attention\left(a(t)\right) \times R_l^d(t) \quad (10)$$

where $f(\cdot)$ is an activation function of the neurons, where we apply the ReLu function to ensure a faster learning in back-propagation, $X(\cdot)_l^t$ is the neural representation of the level $l$ at time $t$. The representation on the *EL* layer $l$ is $E(\cdot)_l$. The *MAXPOOL*, *Conv*, *ConvLSTM*, and *attention* are the corresponding neural algorithms. The overall algorithm for optimizing the network is shown in Algorithm 1.

Table 1: Parameters for Line Tracer Robot.

| Parameters | Value |
|---|---|
| $\tau_0$ | 1.0 |
| $\tau_1$ | 1.1 |
| $\tau_2$ | 1.3 |
| Kernel | $3 \times 3$ |
| Padding | 1 |
| Pooling | $2 \times 2$ |

## 4. Experimental Results

In this section, two experiments are conducted to investigate how the perception and action of the robots (or autonomous car) can be integrated in the predictive coding (PC) framework and how the hierarchical representation differ in different scenarios.

*4.1. Line Tracer Robot.* The first experiment was conducted in a simulation scenario that multiple motor action possibilities were executed with the same image input. From this we analysed the performance of the network as well as the representation of the units, especially the *GU* units. The dataset from a robot simulation was recorded in the scenario where the line tracer robot car is moving along the line from the VRep simulator [59]. In this simulation (Figure 2), the robot is equipped with three vision sensors as well as three Line Finder sensors (Figure 2), so that the image sequences can be captured while the robot car is tracing the line autonomously. Using VRep, we were also able to record the wheel velocity data and the camera data to train the network. To gather the data, we captured the binary images with size of $8 \times 12$ pixels from camera in the middle every 0.02s.

A 3-layer MT-AFA-PredNet was used for training the sequence of both motor action vectors (i.e., the velocities of the wheels) and images, with the Adam optimizer [60]. Three different values of $\tau$ were applied in three different layers. With a larger *tau* on the upper levels, it indicates slower neural activities would be expected. Compared with the $\tau$ values selected in MTRNN works (e.g., [45, 47]), a much smaller $\tau$ values are chosen, because the LSTM networks perform longer-term memories by themselves. The parameters are shown in Table 1.

Figures 3 and 4 show the comparison between the samples of the original and the predicted images. Although Figure 3 is the black-and-white binary inputs for training while the gray-scale figures were used for generative outputs in Figure 4, we can still observe the similarities.

We further visualise the neural activities on different layers to examine how time parameters $\tau$ affect the representation. Corresponding to the prediction samples, the internal representations of the prediction on the 1st *GU* of each layer are shown (Figures 5, 6, and 7), from which we can observe that the predicted image on the higher level (Figure 7) remains steady during almost the whole movement of the robot compared with other two layers. A demo of the experiment can be found on Youtube (https://youtu.be/4w7RqeU42XY).

**Data:** $i(t)\&a(t) \in data$
**while** $error > threshold$ or $iteration > maximum\_iteration$ **do**
  **for** $t \longleftarrow 0$ **to** $T$ **do**
    **for** $l \longleftarrow 0$ **to** $L$ **do**
      **if** $l == L$ **then**
        $R_l^d(t) =$
          $(1 - 1/\tau)R_l^d(t-1) + 1/\tau \cdot ConvLSTM(E_l(t-1), R_l(t-1);$
      **else**
        $R_l^d(t) = (1 - 1/\tau)R_l^d(t-1) + 1/\tau \cdot ConvLSTM(E_l(t-$
        $1), R_l(t-1), DevConv(R_{l+1}(t)));$
      **end**
      $R_l(t) = attention(a(t)) \times R_l^d(t);$
    **end**
    /* Generative (top-down) Process */
    **for** $l \longleftarrow L$ **to** 0 **do**
      $\widehat{X}_l(t) = f(Conv(R_l(t)));$
        $E_l(t) = [f(X_l(t) - \widehat{X}_l(t)); f(\widehat{X}_l(t) - X_l(t));$
        /* Discriminative (bottom-up) Process */
    **end**
  **end**
**end**

Algorithm 1: MT-AFA-PredNet Computation.

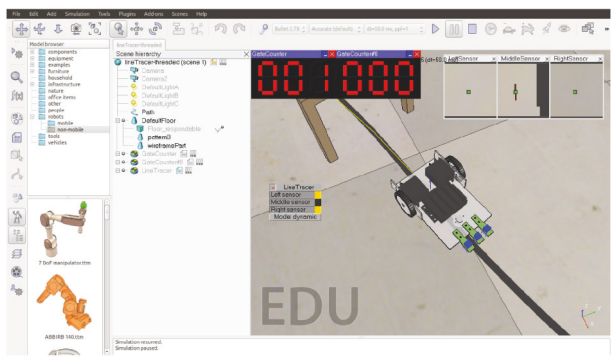
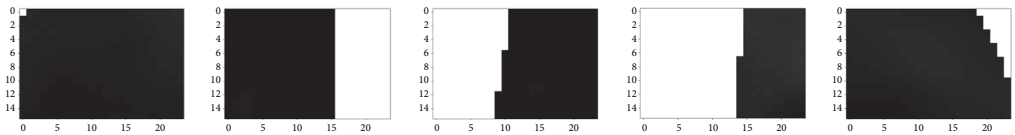
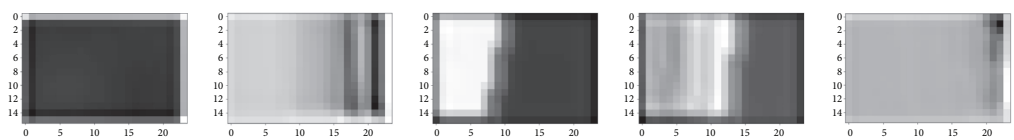



Figure 2: Data Collected from VRep Simulation.



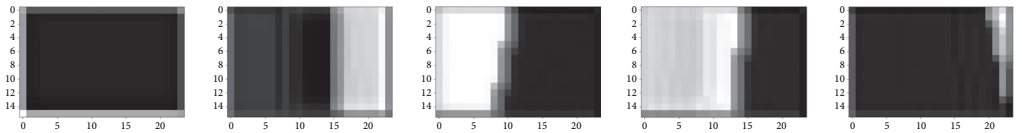
(a) Frame 1 (b) Frame 10 (c) Frame 50 (d) Frame 100 (e) Frame 130

Figure 3: Image Samples from the Middle Vision Sensor.



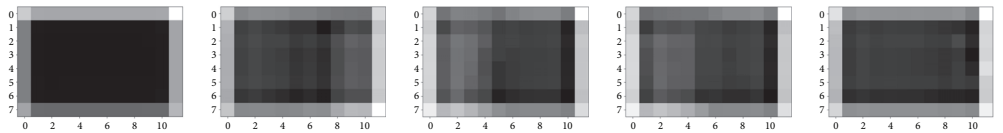
(a) Frame 1 (b) Frame 10 (c) Frame 50 (d) Frame 100 (e) Frame 130

Figure 4: Predicted Images after training.



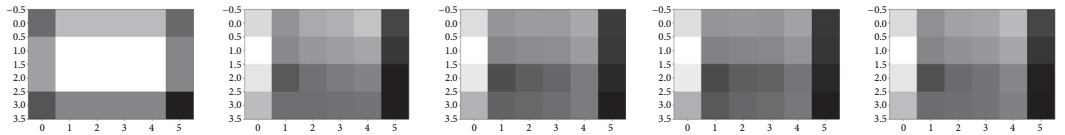
(a) Frame 1 (b) Frame 10 (c) Frame 50 (d) Frame 100 (e) Frame 130

Figure 5: Image generated from the 1st *GU* output (Layer 1), $\tau = 1.0$.



(a) Frame 1 (b) Frame 10 (c) Frame 50 (d) Frame 100 (e) Frame 130

Figure 6: Image generated from the 1st *GU* output (Layer 2), $\tau = 1.3$.



(a) Frame 1 (b) Frame 10 (c) Frame 50 (d) Frame 100 (e) Frame 130

Figure 7: Image generated from the 1st *GU* output (Layer 3), $\tau = 2.0$.

*4.2. Prediction of Different Probabilities.* To investigate how the *GU* representation differentiates various scenarios, we manipulated a bit different lines in three different virtual simulations. These differences are not very significant but the line around the area $x \in [-2, -0.5]$, $y \in [1, 1.5]$ differs. We trained the model with these three trajectories to inspect whether the differences in the sensorimotor reaction in these three scenarios can be encoded in different *GU* units. These trajectories of three scenarios are shown in Figure 8. Then we use 2-layer network with 3 *GU* units on each layer. Similar to the previous subsection, we visualise the representation of the *GU* outputs to see how they are different. We pick the time when the robot reaches the coordinates $(0, 0.5)$ and visualise them every $0.2s$ for $0.8s$.

We found the most differences occur on the $1st$ level, which visualise in Figures 9–12. Although it is not direct mapping of the image prediction, they emerged in different training scenarios. Relatively subtle differences can be found on the $2nd$ layer compared with the $1st$ layer, probably because all the generation can be done on one layer in the hierarchical architecture.

*4.3. Learning Scenarios in Driving.* We conducted the experiments to examine the performances of the multiple time-scale properties of the proposed network. The targets of the experiments are twofold:

(1) The prediction of the incoming images compared with other predictive models.

(2) The multitime scales properties of the proposed network.

In accordance with the targets, we chose the driving dataset (https://ccv.wordpress.fos.auckland.ac.nz/eisats/set-1/). This dataset, provided by Daimler AG, contains five different driving scenarios, each of which contains 250 or 300 images. Additionally, the driving information is also included corresponding to the very time-stamp of each image taken. The 5 units long of vector indicate the ego-motion information:

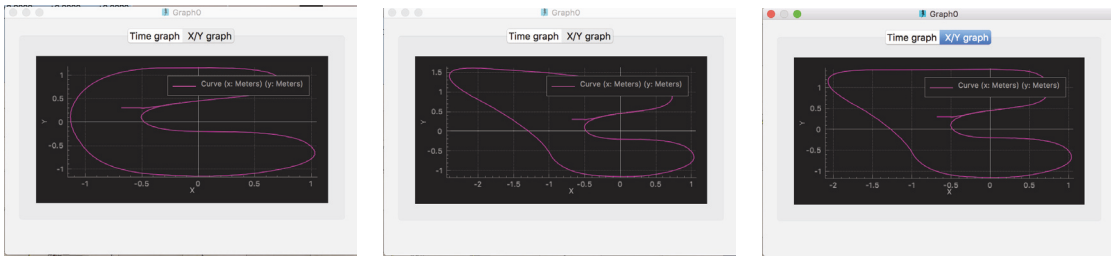
(1) $\theta \in [-\pi, \pi]$: the angle of the steering wheel.

(2) $\{v1, v2, v3, v4\} \in [0, 300]$: the velocities of each wheel.

A 3-layer MT-AFA-PredNet was used for training the sequence of both motor action vectors (i.e., the velocities of the wheels) and images, with the Adam optimizer [60]. Three different values of $\tau$ were applied in three different layers. With a larger *tau* on the upper levels, it indicates slower neural activities would be expected. Compared with the $\tau$ values selected in MTRNN works (e.g., [45, 47]), a much smaller $\tau$ values are chosen, because the LSTM networks perform longer-term memories by themselves. The parameters are shown in Table 2. The epoch equals 300, each of which includes 500 iterations for each sequence.

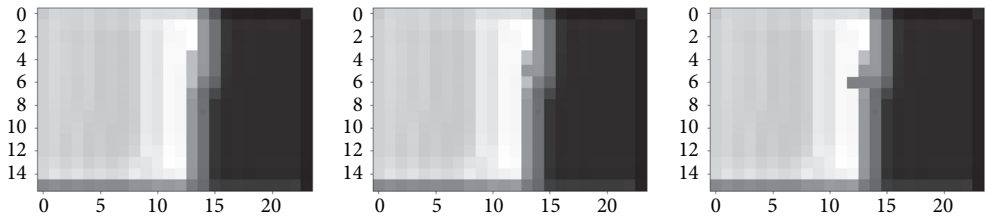
*4.4. Synthesis of Image.* Using the driving dataset, we calculated the error while the generative models predict the image sequences. Similar as before, the epoch is set to be 300, each of which includes 500 iterations for each sequence. After the training, the RMS of each image are calculated as

$$RMS = \frac{1}{T}\sqrt{\sum_{t\in(0,T]}\sum_{i\in pixels}\left(i(t) - \hat{i}(t)\right)^2} \tag{11}$$



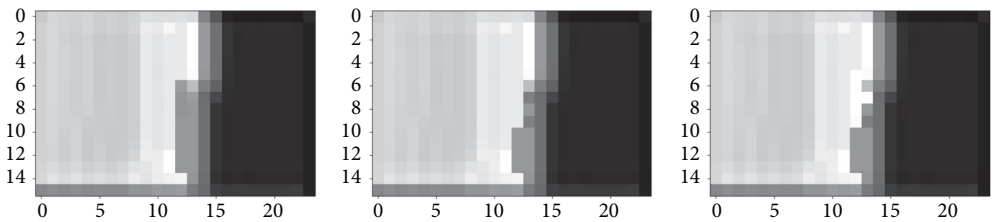

(a) Trajectory 1 (b) Trajectory 2 (c) Trajectory 3

Figure 8: Different trajectories of the line tracing robot.



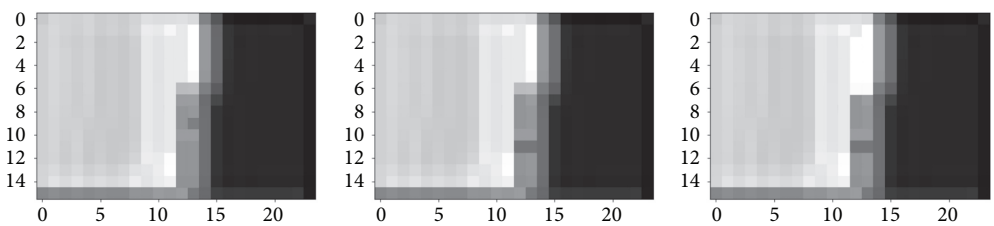
(a) *GU*1 (b) *GU*2 (c) *GU*3

Figure 9: The *GU* representation in different trajectory 1 of the line tracing robot.



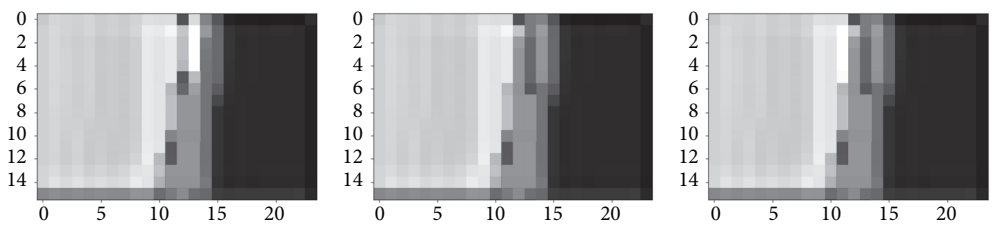
(a) *GU*1 (b) *GU*2 (c) *GU*3

Figure 10: The *GU* representation in different trajectory 2 of the line tracing robot.



(a) Trajectory 1 (b) Trajectory 2 (c) Trajectory 3

Figure 11: The *GU* representation in different trajectory 3 of the line tracing robot.



(a) Trajectory 1 (b) Trajectory 2 (c) Trajectory 3

Figure 12: The *GU* representation in different trajectory 4 of the line tracing robot.

(a) Frame 10 (b) Frame 20 (c) Frame 30 (d) Frame 40 (e) Frame 50

Figure 13: Image samples from the left camera (Crazy turn).

(a) Frame 10 (b) Frame 20 (c) Frame 30 (d) Frame 40 (e) Frame 50

Figure 14: Image samples from the left camera (Construction site).

Table 2: Parameters for EISATS Data-set.

| Parameters | Value |
|---|---|
| $\tau_0$ | 1.0 |
| $\tau_1$ | 1.3 |
| $\tau_2$ | 1.8 |
| Kernel | $3 \times 3$ |
| Padding | 1 |
| Pooling | $2 \times 2$ |

We compare the RMS error between the MT-AFA-PredNet and the LSTM in Table 3. As a baseline, a single layer LSTM is used to predict the image sequences without the motor action as inputs. As we can see, the MT-AFA-PredNet, which uses a multiple layer convolutional LSTM, performs better in prediction than the single layer LSTM, probably because the convolution calculation on each layer is beneficial to detect the image features.

The qualitative comparison between the ground-true and the synthesized images is shown below. Specifically, Figures 13 and 15 show the comparison between some samples of the original and the predicted images in the scenario "crazy turn," and Figures 14 and 16 show the scenario of "construction site." Since we did the normalization after the prediction, the synthesized images show inverted colour.

*4.5. Scenario Classification.* We further visualised the neural activities on different layers to examine how time parameters $\tau$ affect the representation. Due to the page limit, in this subsection, only the quantitative results are shown: we first visualise the representation on the layers 1 and 2 in the first two scenarios ("crazy turn" and "construction site"). The quantitative comparison will be conducted to see whether the update on each layer has been differentiated. Then we will observe if it has been categorized based on the representation on layer 2.

Corresponding to the prediction samples, the internal representations of 1st and 2nd layers of the GU units are shown, respectively, in Figures 17, 18, 19, and 20. We can observe that the higher-level representation (Layer 2) remains steadier than the lower levels. And the representation seems encoded in a sparse way. From this result, we can basically categorize different driving scenarios as shown in Figure 21, where we can see there are different representations with different training sequences.

## 5. Discussion

*5.1. From Predictive Perception to Planning.* The feedback affecting sensory input can be regarded as a kind of predictive information retrieved from the internal memory [61]. Based on the predictive coding theory, the hierarchical architecture uses the feedback signals (especially the top-down signals) to predict the forthcoming sensory input, while the sensory-driven bottom-up signals only deliver the error of the estimation. Our model further puts a hypothesis that, during event-based prediction, such feedback may be based on the categorization about the on-going sensorimotor events. They are performed with the similar Bayesian inference and are always updating the prior knowledge on the cognitive processes level.

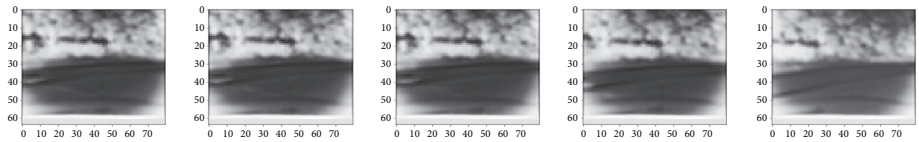
Similar, on the cognitive process level, if such kind of prediction lasts as a closed-loop and is long enough in a temporal domain, the higher level of representation might play as a mental simulation about the future events. Such an event can be represented in the higher level of the hierarchy, in which the neural activities are updated slower. Such a prediction is also about multimodality too. It captures the structural regularities in each of the modality, and in both spatial and temporal domains. As [62] suggested, the accomplishment of such tasks about decision making and planning is related to the Bayesian inference on the higher level of cognitive functions. As such, the difference between the sensorimotor prediction, the categorization of the events, and the planning behaviour can be unified. This model specifically focuses on the unification of the sensorimotor prediction and the categorization of the events.

As specified at [63], such a planning process inherited from the predictive process only exists when

(1) the specific goal is already determined at the very first beginning;

Table 3: RMS between LSTM and MT-AFA-PredNet.

| | LSTM | MT-AFA-PredNet |
|---|---|---|
| Construction site | 8.332 | 7.219 |
| Crazy turn | 10.315 | 8.287 |
| Dancing light | 9.834 | 7.314 |
| Intern on bike | 5.411 | 5.131 |
| Safe turn | 9.314 | 8.107 |
| Squirrel | 7.908 | 7.781 |



(a) Frame 10 (b) Frame 20 (c) Frame 30 (d) Frame 40 (e) Frame 50

Figure 15: Predicted images after training (Crazy turn).

(2) at a short- or mid-term planning problem.

When human is solving more complex planning problems, such as the multiobjective optimization problem (e.g., Traveller Salesman Problem, TSP), it needs a higher level of cognitive processes to make decision of computation power and time. Nevertheless, from the engineering perspective, the short- and mid-term planning are sufficient in some short- and mid-term planning applications, e.g., autonomous driving, where the PredNet model was already examined to predict the next frame of images from the vehicle camera.

*5.2. Model-Free Perception-Action Learning.* Traditional control problem usually requires an explicit description of the plant model. Therefore, the quality of control relies not only on the chosen control strategy, but also on the precision of assumed plants model. For instance, the agent should acquire the dynamics about the environment as well as the description of the physical model of itself. However, analytical descriptions are usually hard to define because of the systems nonlinearities and high order dynamics.

The development of recent reinforcement in robot control (e.g., locomotion [64], manipulation [65, 66]), and autonomous vehicle control [67]) usually requires explicit mathematical representation of the plant and environment model. For instance, a reinforcement learning algorithm with Monte-Carlo learning, whose value function is of a certain policy without a concrete model, learns the perception-action pairing by random sample placement.

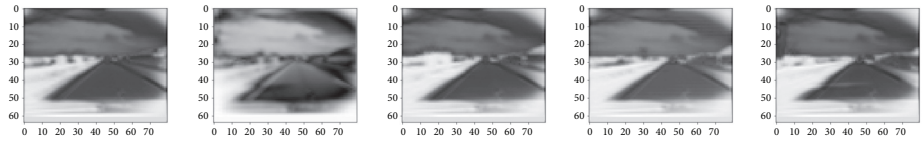
Although our proposed model does not include the reinforcement learning paradigm, it also maps such kind of the perception and action pairing without the prior knowledge of reward. It also makes our proposed model more practical since an appropriate representation for the policy or value function cannot be always chosen so that the temporal sequence of the perception-action pairing is difficult to achieve in the practical environment. On the other hand, as the target of practical real-world applications of reinforcement learner usually learns the reward after the episodic tasks. The internal memory of LSTM can accomplish similar job without explicit engineering job about handcrafting the reward function itself.

This also raises a problem in the bioinspired learning problem for agents: when the "state space" and "action space" are explicitly defined in reinforcement learning, the reward-driving learning usually learns the mapping between these two spaces with a trivial assumption that the action is purely driven by the perception, which is not an obvious case if we study the development of human cognition. Instead, most of the cognitive studies propose that the perception and action should hold an integrated relation. With the proposed architecture, the perception-action integration can be easily driven by the feedback loop in our future work. This kind of perception-action pairing might become another alternative to solve the sensorimotor problem of robot learning.

*5.3. Multimodal and Cross-Modal Prediction.* The first version of proposed network (AFA-PredNet) model integrates the perception and action under the predictive coding framework. It depicts the possibility that perception and action can be integrated in a hierarchical architecture. Furthermore, the bidirectional feedback connections in this architecture give rise to the cross-modal influence of both perception and action.

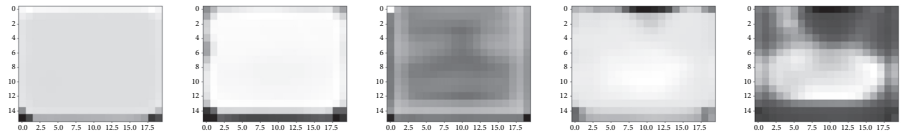
In the human brain, such kind of influence has been widely found in perception. For instance, the McGurk effect indicates the possibility of interaction between the auditory and visual modalities. Furthermore, under the framework of predictive coding, both perception and action attempt to minimize the prediction error.

Hierarchical architectures have been used for cross-modal understanding, in which some intriguing higher level of representation has been discovered (e.g., [68, 69]). Such kind of representation may give a hint that our higher level of cognition also exerts top-down signals to the lower level of sensory data, about how the incoming sensory data looks like. Such kind of higher-level cognition is also obtained when the agent is interacting with the environment, so that such kind of

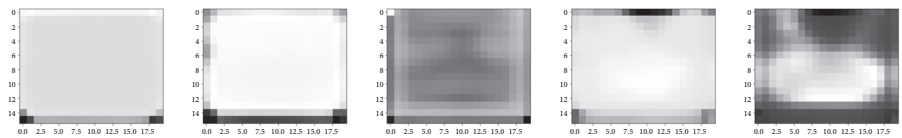


(a) Frame 1 (b) Frame 10 (c) Frame 50 (d) Frame 100 (e) Frame 130

Figure 16: Predicted images after training (Construction site).

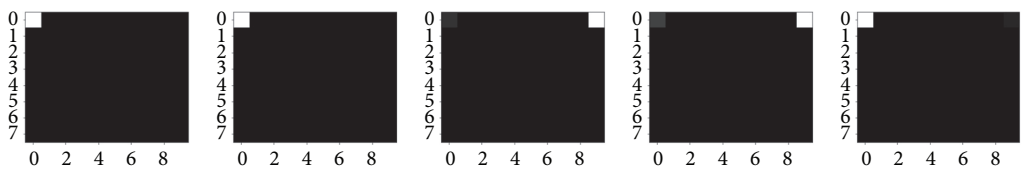


(a) Frame 10 (GU 1) (b) Frame 50 (GU 1) (c) Frame 100 (GU 1) (d) Frame 150 (GU 1) (e) Frame 200 (GU 1)

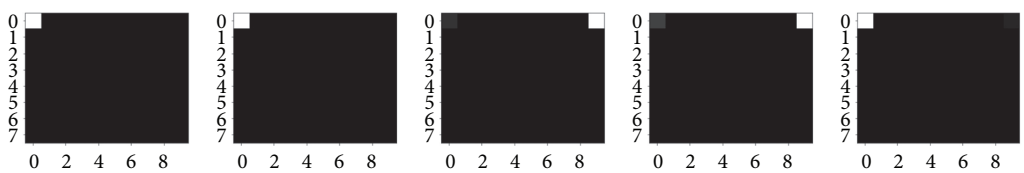


(f) Frame 10 (GU 2) (g) Frame 50 (GU 2) (h) Frame 100 (GU 2) (i) Frame 150 (GU 2) (j) Frame 200 (GU 2)

Figure 17: Representation in GU Units (Layer 2) (Crazy turn).

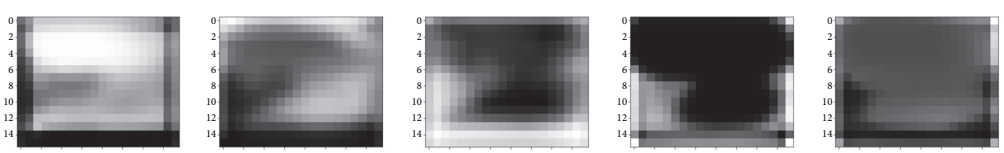


(a) Frame 10 (GU 1) (b) Frame 50 (GU 1) (c) Frame 100 (GU 1) (d) Frame 150 (GU 1) (e) Frame 200 (GU 1)

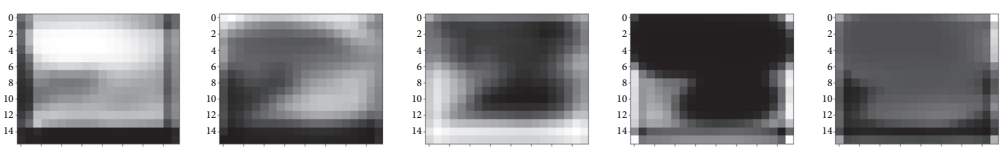


(f) Frame 10 (GU 2) (g) Frame 50 (GU 2) (h) Frame 100 (GU 2) (i) Frame 150 (GU 2) (j) Frame 200 (GU 2)

Figure 18: Representation in GU Units (Layer 3) (Crazy turn).

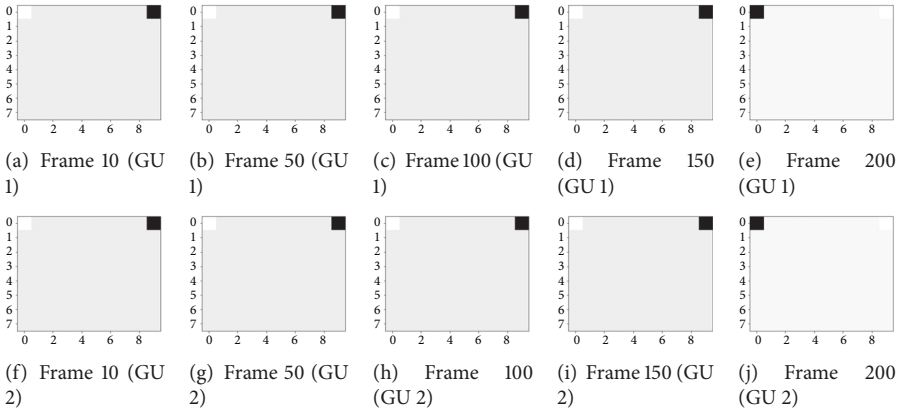


(a) Frame 10 (GU 1) (b) Frame 50 (GU 1) (c) Frame 100 (GU 1) (d) Frame 150 (GU 1) (e) Frame 200 (GU 1)



(f) Frame 10 (GU 2) (g) Frame 50 (GU 2) (h) Frame 100 (GU 2) (i) Frame 150 (GU 2) (j) Frame 200 (GU 2)

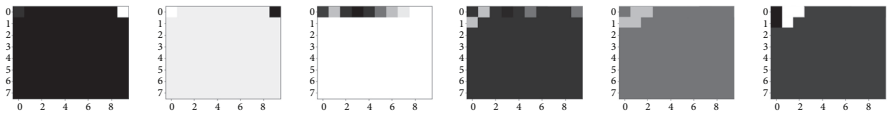
Figure 19: Representation in GU Units (Layer 2) (Construction site).



(a) Frame 10 (GU 1) (b) Frame 50 (GU 1) (c) Frame 100 (GU 1) (d) Frame 150 (GU 1) (e) Frame 200 (GU 1)

(f) Frame 10 (GU 2) (g) Frame 50 (GU 2) (h) Frame 100 (GU 2) (i) Frame 150 (GU 2) (j) Frame 200 (GU 2)

Figure 20: Representation in GU Units (Layer 3) (Crazy turn).



(a) Crazy turn (b) Construction site (c) Dancing (d) Intern on bike (e) Safe turn (f) Squirrel

Figure 21: Representation in GU Units (Layer 3, Frame 10) (6 Scenarios).

"world model", corresponding to the slow context layer in our MT-AFA-PredNet, cannot be explicitly modelled. Instead, the pairing of different modalities, as well as perception-action, is a model-free learning process.

*5.4. Conclusion.* The top-down prediction happens through the whole brain, predictive coding proposes a framework about such top-down prediction: the error runs through the bottom-up in the hierarchical brain and corrects the top-down prediction in the internal models. The higher level of cognition plays a role about explaining the current sensorimotor scenario in a mode of world models, which are dynamically changing, and makes a prediction based on such a model. During this process, the motor action also works in as one of the prior knowledge to proceed with the prediction. To build such an integration between perception and action in the predictive coding, the Multiple Time-scale Action Feedback Augmented Predictive Network (MT-AFA-PredNet) is proposed, which realises the following functionalities in the predictive coding framework in neurorobotic experiments:

(1) The top-down feedback pathways in perception are applied for sensory prediction. The prediction is realised by extracting the prior knowledge from perception and building the world models, embedded in the higher level of cognition, about the sensorimotor information of the agent itself.

(2) Such a representation of cognition about the long-term sensorimotor context is encoded in a slow context layer in our artificial neural network architecture. The fast and slow context neurons are determined by a time parameter.

(3) This "predictive coding" theory has been further integrated with ego-motor action. And it is realised by the attention mechanism.

Experiments were conducted in two simulations. The first line tracer robot experiment shows that the independent generative units encode the prediction of different movements, which are further modulated by the attention mechanism. The second experiment on the driving dataset demonstrates its ability to categorize different driving events, and its result of prediction outperforms the LSTM.

## Data Availability

The Pytorch implementation of the model and the driving dataset used to support the findings of this study have been deposited on GitHub: https://github.com/jonizhong/MT_AFAPrednet.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

The research was partially supported by New Energy and Industrial Technology Development Organization (NEDO).

## References

[1] J. H. R. Maunsell and J. R. Gibson, "Visual response latencies in striate cortex of the macaque monkey," *Journal of Neurophysiology*, vol. 68, no. 4, pp. 1332–1344, 1992.

[2] M. T. Schmolesky, Y. Wang, D. P. Hanes et al., "Signal timing access the macaque visual system," *Journal of Neurophysiology*, vol. 79, no. 6, pp. 3272–3278, 1998.

[3] M. Singh and J. M. Fulvio, "Visual extrapolation of contour geometry," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 102, no. 3, pp. 939–944, 2005.

[4] D. N. Lee and P. E. Reddish, "Plummeting gannets: A paradigm of ecological optics," *Nature*, vol. 293, no. 5830, pp. 293-294, 1981.

[5] H. Wagner, "Flight Performance and Visual Control of Flight of the Free-Flying Housefly (Musca Domestica L.) I. Organization of the Flight Motor," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 312, no. 1158, pp. 527–551, 1986.

[6] E. J. Morris and S. G. Lisberger, "Different responses to small visual errors during initiation and maintenance of smooth-pursuit eye movements in monkeys," *Journal of Neurophysiology*, vol. 58, no. 6, pp. 1351–1369, 1987.

[7] H. von Helmholtz et al., *Handbuch der Physiologischen Optik*, Voss, Hamburg, 1909.

[8] R. P. N. 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.

[9] A. Clark, "Whatever next? Predictive brains, situated agents, and the future of cognitive science," *Behavioral and Brain Sciences*, pp. 1–86, 2012.

[10] K. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[11] F. Lacquaniti and C. Maioli, "The role of preparation in tuning anticipatory and reflex responses during catching," *The Journal of Neuroscience*, vol. 9, no. 1, pp. 134–148, 1989.

[12] C. Ghez and J. Krakauer, "The organization of movement," *Principles of neural science*, vol. 4, pp. 653-73, 2000.

[13] J. Zhong, *Artificial neural models for feedback pathways for sensorimotor integration*, 2015.

[14] E. M. Segal and T. G. Halwes, "The influence of frequency of exposure on the learning of a phrase structural grammar," *Psychonomic Science*, vol. 4, no. 1, pp. 157-158, 1966.

[15] K. Friston and S. Kiebel, "Cortical circuits for perceptual inference," *Neural Networks*, vol. 22, no. 8, pp. 1093–1104, 2009.

[16] K. Friston, J. Mattout, and J. Kilner, "Action understanding and active inference," *Biological Cybernetics*, vol. 104, no. 1-2, pp. 137–160, 2011.

[17] G. Pezzulo, F. Rigoli, and K. Friston, "Active Inference, homeostatic regulation and adaptive behavioural control," *Progress in Neurobiology*, vol. 134, pp. 17–35, 2015.

[18] A. L. Woodward, "Infants selectively encode the goal object of an actor's reach," *Cognition*, vol. 69, no. 1, pp. 1–34, 1998.

[19] P. Rochat, "Mouthing and grasping in neonates: Evidence for the early detection of what hard or soft substances afford for action," *Infant Behavior & Development*, vol. 10, no. 4, pp. 435–449, 1987.

[20] D. M. Wolpert, Z. Ghahramani, and M. I. Jordan, "An internal model for sensorimotor integration," *Science*, vol. 269, no. 5232, pp. 1880–1882, 1995.

[21] S. Zeki and S. Shipp, "The functional logic of cortical connections," *Nature*, vol. 335, no. 6188, pp. 311–317, 1988.

[22] I. Winkler and I. Czigler, "Evidence from auditory and visual event-related potential (ERP) studies of deviance detection (MMN and vMMN) linking predictive coding theories and perceptual object representations," *International Journal of Psychophysiology*, vol. 83, no. 2, pp. 132–143, 2012.

[23] K. Friston, "Consciousness and hierarchical inference," *Neuropsychoanalysis*, vol. 15, no. 1, pp. 38–42, 2013.

[24] C. J. Ostafew, A. P. Schoellig, and T. D. Barfoot, "Learning-based nonlinear model predictive control to improve vision-based mobile robot path-tracking in challenging outdoor environments," in *Proceedings of the 2014 IEEE International Conference on Robotics and Automation, ICRA 2014*, pp. 4029–4036, China, June 2014.

[25] H. Roggeman, J. Marzat, M. Sanfourche, and A. Plyer, "Embedded vision-based localization and model predictive control for autonomous exploration," in *Proceedings of the IROS Workshop on Visual Control of Mobile Robots (ViCoMoR 2014)*, pp. 13–20, 2014.

[26] T. M. Howard, C. J. Green, and A. Kelly, "Receding Horizon Model-Predictive Control for Mobile Robot Navigation of Intricate Paths," in *Field and Service Robotics*, vol. 62 of *Springer Tracts in Advanced Robotics*, pp. 69–78, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[27] A. Aswani, H. Gonzalez, S. S. Sastry, and C. Tomlin, "Provably safe and robust learning-based model predictive control," *Automatica*, vol. 49, no. 5, pp. 1216–1226, 2013.

[28] G. Klančar and I. Škrjanc, "Tracking-error model-based predictive control for mobile robots in real time," *Robotics and Autonomous Systems*, vol. 55, no. 6, pp. 460–469, 2007.

[29] D. Gu and H. Hu, "Receding horizon tracking control of wheeled mobile robots," *IEEE Transactions on Control Systems Technology*, vol. 14, no. 4, pp. 743–749, 2006.

[30] Y. Gao, C. G. Lee, and K. T. Chong, "Receding horizon tracking control for wheeled mobile robots with time-delay," *Journal of Mechanical Science and Technology*, vol. 22, no. 12, pp. 2403–2416, 2008.

[31] T. P. Nascimento, A. P. Moreira, and A. G. Scolari Conceição, "Multi-robot nonlinear model predictive formation control: Moving target and target absence," *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1502–1515, 2013.

[32] K. J. W. Craik, "The nature of explanation," *CUP Archive*, vol. 445, 1967.

[33] E. C. Tolman, "Cognitive maps in rats and men," *Psychological Review*, vol. 55, no. 4, pp. 189–208, 1948.

[34] R. A. Brooks, "How to build complete creatures rather than isolated cognitive simulators," *Architectures for intelligence*, pp. 225–239, 1991.

[35] E. von Holst and H. Mittelstaedt, *he reafference principle: Interaction between the central nervous system and the peripheral organs. selected papers of erich von holst: The behavioural physiology of animals and man*, 1950.

[36] R. C. Miall and D. M. Wolpert, "Forward models for physiological motor control," *Neural Networks*, vol. 9, no. 8, pp. 1265–1279, 1996.

[37] N. L. Cerminara, R. Apps, and D. E. Marple-horvat, "An internal model of a moving visual target in the lateral cerebellum," *The Journal of Physiology*, vol. 587, no. 2, pp. 429–442, 2009.

[38] J. Zhong, C. Weber, and S. Wermter, "A predictive network architecture for a robust and smooth robot docking behavior," *Paladyn. Journal of Behavioral Robotics*, vol. 3, no. 4, pp. 172–180, 2012.

[39] D. M. Wolpert and M. Kawato, "Multiple paired forward and inverse models for motor control," *Neural Networks*, vol. 11, no. 7-8, pp. 1317–1329, 1998.

[40] Y. Demiris and B. Khadhouri, "Hierarchical attentive multiple models for execution and recognition of actions," *Robotics and Autonomous Systems*, vol. 54, no. 5, pp. 361–369, 2006.

[41] H. Hoffmann, "Perception through visuomotor anticipation in a mobile robot," *Neural Networks*, vol. 20, no. 1, pp. 22–33, 2007.

[42] R. Möller and W. Schenck, "Bootstrapping cognition from behavior," *Cognitive Science*, vol. 32, no. 3, pp. 504–542, 2008.

[43] Y. Sugita and J. Tani, "Learning semantic combinatoriality from the interaction between linguistic and behavioral processes," *Adaptive Behavior*, vol. 13, no. 1, pp. 33–52, 2005.

[44] J. Zhong, A. Cangelosi, and S. Wermter, "Toward a self-organizing pre-symbolic neural model representing sensorimotor primitives," *Frontiers in Behavioral Neuroscience*, vol. 8, article no. 22, 2014.

[45] Y. Yamashita and J. Tani, "Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment," *PLoS Computational Biology*, vol. 4, no. 11, Article ID e1000220, 2008.

[46] T. Ogata and H. G. Okuno, "Integration of behaviors and languages with a hierarchal structure self-organized in a neuro-dynamical model," in *Proceedings of the 2013 IEEE Workshop on Robotic Intelligence in Informationally Structured Space (RiiSS)*, pp. 89–95, Singapore, Singapore, April 2013.

[47] J. Zhong, M. Peniak, J. Tani, T. Ogata, and A. Cangelosi, "Sensorimotor input as a language generalisation tool: a neuro-robotics model for generation and generalisation of noun-verb combinations with sensorimotor inputs," https://arxiv.org/abs/1605.03261.

[48] H. Lee, M. Jung, and J. Tani, "Recognition of Visually Perceived Compositional Human Actions by Multiple Spatio-Temporal Scales Recurrent Neural Networks," https://arxiv.org/abs/1602.01921.

[49] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998.

[50] J. Donahue, L. A. Hendricks, S. Guadarrama et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 2625–2634, USA, June 2015.

[51] W. Lotter, G. Kreiman, and D. Cox, "Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning," https://arxiv.org/abs/1605.08104.

[52] J. Zhong, A. Cangelosi, X. Zhang, and T. Ogata, "AFA-PredNet: The Action Modulation Within Predictive Coding," in *Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, Rio de Janeiro, Brazil, July 2018.

[53] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proceedings of the 29th Annual Conference on Neural Information Processing Systems, NIPS 2015*, pp. 802–810, Canada, December 2015.

[54] M. Denil, L. Bazzani, H. Larochelle, and N. de Freitas, "Learning where to attend with deep architectures for image tracking," *Neural Computation*, vol. 24, no. 8, pp. 2151–2184, 2012.

[55] Y. Wang, M. Huang, x. zhu, and L. Zhao, "Attention-based LSTM for Aspect-level Sentiment Classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 606–615, Austin, Texas, November 2016.

[56] T. Luong, H. Pham, and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal, September 2015.

[57] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[58] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," *Advances in Neural Information Processing Systems*, pp. 2773–2781, 2015.

[59] E. Rohmer, S. P. Singh, and M. Freese, "V-rep: A versatile and scalable robot simulation framework," in *Proceedings of the2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1321–1326, 2013.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," https://arxiv.org/abs/1412.6980.

[61] J. R. Anderson and L. J. Schooler, *The adaptive nature of memory*, 2000.

[62] M. Toussaint, "Probabilistic inference as a model of planned behavior," *KI*, vol. 23, no. 3, pp. 23–29, 2009.

[63] D. Basso, "Planning, prospective memory, and decision-making: Three challenges for hierarchical predictive processing models," *Frontiers in Psychology*, vol. 3, 2013.

[64] N. Kohl and P. Stone, "Policy gradient reinforcement learning for fast quadrupedal locomotion," in *Proceedings of the IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*, pp. 2619–2624 Vol.3, New Orleans, LA, USA, April 2004.

[65] E. Theodorou, J. Buchli, and S. Schaal, "Reinforcement learning of motor skills in high dimensions: A path integral approach," in *Proceedings of the 2010 IEEE International Conference on Robotics and Automation (ICRA 2010)*, pp. 2397–2403, Anchorage, AK, May 2010.

[66] J. Peters and S. Schaal, "Reinforcement learning of motor skills with policy gradients," *Neural Networks*, vol. 21, no. 4, pp. 682–697, 2008.

[67] P. Abbeel, A. Coates, M. Quigley, and A. Y. Ng, "An application of reinforcement learning to aerobatic helicopter flight," in *Proceedings of the 20th Annual Conference on Neural Information Processing Systems, NIPS 2006*, pp. 1–8, Canada, December 2006.

[68] Y. Aytar, C. Vondrick, and A. Torralba, "See, hear, and read: Deep aligned representations," https://arxiv.org/abs/1706.00932.

[69] L. Kaiser, A. N. Gomez, N. Shazeer et al., "One model to learn them all," https://arxiv.org/abs/1706.05137.