Anticipating Noxious States from Visual Cues using Deep Predictive Models

Indranil Sur and Heni Ben Amor

Abstract—To ensure system integrity, robots need to proactively avoid any unwanted physical perturbation that may cause damage to the underlying hardware. In this paper, we investigate a machine learning approach that allows robots to anticipate impending physical perturbations from perceptual cues. In contrast to other approaches that require knowledge about sources of perturbation to be encoded before deployment, our method is based on experiential learning. Robots learn to associate visual cues with subsequent physical perturbations and contacts. In turn, these extracted visual cues are then used to predict potential future perturbations acting on the robot. To this end, we introduce a novel deep network architecture which combines multiple sub-networks for dealing with robot dynamics and perceptual input from the environment. We present a self-supervised approach for training the system that does not require any labeling of training data. Inference is performed using approximate Bayesian inference and moment matching. Extensive experiments in a human-robot interaction task show that a robot can learn to predict physical contact by a human interaction partner without any prior information or labeling. Furthermore, the network is able to successfully predict physical contact from either depth stream input or traditional video input.

I. INTRODUCTION

According to Isaac Asimov’s third law of robotics [1], “a robot must protect its own existence as long as such protection does not conflict with the First or Second Law”, i.e., as long as it does not harm a human. Aspects of safety and self-preservation are tightly coupled to autonomy and longevity of robotic systems. For robots to explore their environment and engage in physical contact with objects and humans, they need to ensure that any such interaction may not lead to tear, damage, or irreparable harm to the underlying hardware. Situations that jeopardize the integrity of the system need to be detected and actively avoided. This determination can be performed in either a reactive way, e.g., by using sensors to identify forces acting on the robot, or in a pro-active way, e.g., by detecting impending collisions. In recent years, a plethora of safety methods have been proposed that are based on reactive strategy. Approaches for compliant control and, in particular, impedance control techniques have been shown to enable safe human-robot interaction [2] in a variety of close-contact tasks. Such methods are typically referred to as post-contact approaches, since they react to forces exchanged between the system and its environment after they first occur.

In many application domains, however, robots need to proactively reason about impending damage before it occurs. Methods that tackle these scenarios, have mostly focused on proximity detection and collision avoidance. Motion tracking or other sensing devices are used to identify nearby humans and obstacles in order determine whether they intersect the robot’s path. To this end, a human expert has to reason about the expected obstacles before deployment and, in turn, hand-code methods for object recognition, human tracking, or collision detection. Such methods are largely based on the status quo of the environment and do not take in account the expected future states. For example, the behavior of a human interaction partner may already provide cues whether or not physical contact is to be expected. In addition, such methods suffer from limited adaptivity – the robotic system is not able to incorporate other or new sources of physical perturbations that have not been considered by the human expert. Changes in the application domain typically require the expert to specify the set of obstacles/perturbants and how they can be identified from sensor data. However, for robots to autonomously explore new goals, tasks, and environments they cannot be constantly relying on human intervention and involvement. In order to increase resilience of robotic systems, the processes responsible for ensuring safety should inherently be (a) adaptive to changes that occur during the cycle of operation and (b) anticipative in nature, so as to proactively avoid damage.

In biology such processes are common place: humans and animals experience pain as the guiding signal which ensures self-preservation and safe, “open ended” exploration of the.
world. Over time, biological creatures learn to anticipate impending pain sensations from environmental and proprioceptive cues, e.g., from the velocity and direction of an obstacle, or an imbalance in posture. The relationship between pain and learning is bi-directional in humans and vertebrate animals. Our perception of pain shapes the way we learn, but the perception itself is also shaped by the experiences we undergo during learning. In particular, repeated exposure to a sensory cue (e.g. a specific image or object) preceding a negative outcome (e.g. electric shock) will turn the cue into a conditioned stimulus that helps anticipate the shock in future trials. This ability to associate a certain stimuli with negative future sensations is considered to be essential to survival. Inspired by the relationship between pain and learning in biological creatures, we propose in this paper a new method for learning to anticipate physical perturbations in robotic systems. Robot safety is realized through an adaptive process in which prior experiences are used to predict impending harm to the system. These predictions are based on (1) environmental cues, e.g., camera input, (2) proprioceptive information, and (3) the intended actions of the robot. We introduce a deep predictive model (DPM) that leverages all three sources of information in order to anticipate contact and physical perturbations in a window of future time steps. An important characteristics of this deep predictive models, is that it differentiates between forces and perturbations that are the result of an executed robot behavior, e.g., fast arm movements, and external perturbations that are the result of outside influence. After learning, the DPM can be used to anticipate future states that result in forces being applied to the robot. In turn, this information can be used to provide visual feedback to the user, or actively avoid the predicted states, or change the impedance of the system.

II. RELATED WORK

Safety plays a critical role in the field of robotics. Recently, various methods have been put forward in order to protect a human interaction partner from harm. The work in [3], for example, uses proprioceptive sensing to identify collisions and, in turn, execute a reactive control strategy that enables the robot to retract from the location of impact. In a similar vein, the work in [4] describes methods for rapid collision detection and response through trajectory scaling. Many approaches to safe human-robot interaction are based on rapid sensing through force-torque sensors or tactile sensors [5]. Another approach is to generate estimates of external forces acting on a robot using disturbance observers [6]. These approaches, however, require a model of the underlying plant, which in the case of complex humanoid robots can become challenging to derive. In addition, nonlinearities underlying the current robot or task can often lead to instabilities in the system [7]. To circumvent such challenges, several approaches have been proposed for learning perturbation filters using a data-driven machine learning method [8], [9], [10]. An alternative, bio-inspired approach was proposed in [11]. In particular, an “Artificial Robot Nervous System” was introduced which emulates the human skin structure, as well as the spikes generated whenever an impact occurs.

All of the above techniques are reactive in nature. Only after contact with its environment, can a robot detect a physical impact or perturbation and react to it. However, many critical situations require a proactive avoidance strategy. To this end, various methods for human motion anticipation have been proposed [12], [13]. Given a partially observed human motion, a robot can anticipate the goal location and intermediate path and accordingly generate a collision free navigation strategy. However, such approaches are brittle in that they require some form of human tracking. If the source
of the perturbation is non-human, e.g., a moving object then no anticipation can occur. In this paper, we are interested in dynamic approaches to anticipation of perturbations. Robots learn to predict contacts or impact by associating them with visual cues. This is similar in spirit to [14] where dashcam videos were used to predict car collisions. However, an important difference is that our predictions are based on both visual cues, proprioceptive sensors, as well as next robot actions.

III. DEEP PREDICTIVE MODELS OF PHYSICAL PERTURBATIONS

Our goal is to enable an intelligent system to anticipate undesired external perturbations before they occur. Following the biological inspiration, repeated exposure to a sensory cue preceding a physical perturbation will turn the cue into a predictive variable that helps anticipate the perturbation. To this end, we propose the deep predictive model seen in Fig. 2. A key insight of our approach, is that a robot can learn to correlate perceptual data, e.g., camera image, to its future internal states of the robot. By establishing a mapping between perceptual information and internal states, the it can anticipate harm before it occurs.

The DPM is based on a modular architecture, in which two modules generate the necessary relationship in order to map perceptual data to future anticipated noxious signals. The first module of the DPM is a deep dynamics module that learns to discriminate between external perturbations caused by outside events and natural variations of sensor readings due to the currently executed behavior and sensor noise. The deep dynamics module is trained to generate a signal, whenever the recorded sensor values cannot be explained by the actions of the robot.

The second module of the DPM is the perceptual anticipation module—a deep network that takes visual percepts, proprioceptive states, and actions taken as inputs and generates predictions over future noxious signals. It learns to associate specific visual and proprioceptive cues to harmful states. The perceptual anticipation module contains convolutional recurrent layers [15], [16] that process the visual input in both space and time. Learning is performed using the paradigm of self-supervised learning. More specifically, the training signal produced by the deep dynamics module is used as a target signal. No human intervention is need in order to either provide new training data or label existing data.

Subsequently, we will explain the elements of the introduced network architecture and describe the underlying training and inference process in more detail.

A. Deep Dynamics Model

The task of the deep dynamics model is to identify exogenous perturbations affecting the robot. Detecting such perturbations in the sensor stream can be challenging when the robot is performing dynamic movements that by themselves cause fluctuation in the sensor readings. To discriminate between exogenous and endogenous perturbations, we will use a strategy inspired by the human motor system.

Before sending an action \( a_t \in \mathbb{R}^Q \) to the actuators, the robot creates a copy of \( a_t \), the so-called efference copy, and predicts the expected sensory stimuli after execution. This prediction is performed by the deep dynamics model, a neural network that maps the current state \( s_t \) and intended action \( a_t \) onto the expected next state \( s_{t+1}^* \).

After executing action \( a_t \), we can measure the discrepancy between the expected sensations and the measured sensor values, also called the reafference. The degree of discrepancy is an indicator for external physical perturbations.

Data Collection: Without loss of generality, we define for the remainder of the paper the system state to be \( s_t = [\theta_t, \dot{\theta}_t, \ddot{\theta}_t, p_t]^T \in \mathbb{R}^P \) which includes joint angles \( \theta_t \), joint velocities \( \dot{\theta}_t \), accelerations \( \ddot{\theta}_t \) and end-effector pose \( p_t \).

To collect training data for training the deep dynamics model, we use motor babbling [17], [18]. To this end, small changes are applied as the control action \( a_t \in \mathbb{R}^Q \), where \( Q \) is the number of degrees of freedom for the robot, i.e., the number of joints. Considering a naive implementation of motor babbling, the action can be sampled from an isotropic Gaussian distribution \( a_t \sim \mathcal{N}(0, \sigma^2 I) \). However, empirically we have observed that such a naive sampling approach does not effectively cover the state-action space by unnecessarily focusing on samples in the center of the distribution.

To alleviate this problem, we use a bi-modal distribution and incorporate a simple momentum term

\[
\pi \sim \mathcal{B}(0.5) \\
\mu \sim \pi \mathcal{N}(\mu I, \sigma^2 I) + (1 - \pi) \mathcal{N}(-\mu I, \sigma^2 I) \\
a_t = (\mu + a_{t-1})/2
\]

Actions sampled using the above strategy effectively cover the state-action space and generate trajectories without causing wear and tear. The result of the motor babbling phase, is a dataset for training which consists of \( N \) triplets \((s_t, a_t, s_{t+1})\) consisting of the current state, current action, and the next state. The individual matrices storing all states and actions are denoted by \( S_t, S_{t+1} \in \mathbb{R}^{N \times P} \) and \( A_t \in \mathbb{R}^{N \times Q} \).

Model Learning: The deep dynamics model is an artificial neural network that maps a state \( s_t \) and action \( a_t \) to the expected next state \( s_{t+1}^* \). Training is performed using Dropout [19] and data collected in the motor babbling process. A Euclidean loss function is used to identify the error.

The neural network generates a point estimate for any set of inputs. However, due to the non-determinism and noise underlying such tasks, it is important to reason about uncertainties when making predictions. To this end, we leverage recent theoretical insights in order to generate probabilistic predictions from a trained neural network. In particular, it was shown in [20] that neural network learning using the Dropout method [19] is equivalent to a Bayesian approximation of a Gaussian Process modeling the training data. Following this insight, we generate a set of predictions.
\{s_1, \ldots, s_T\}$ from a trained network through $T$ stochastic forward passes. The generated predictions form a possibly complex distribution represented as a population of solutions. We then extract an approximate parametric form of the underlying distribution through moment matching

$$E(s_{t+1}^*) \approx \frac{1}{T} \sum_{i=1}^{T} \hat{s}_i$$

(4)

$$\text{Var}(s_{t+1}^*) \approx \frac{1}{T} \sum_{i=1}^{T} \hat{s}_i^\top \hat{s}_i - E(s_{t+1}^*)^\top E(s_{t+1}^*)$$

(5)

Given the above distribution moments we can reason about the uncertainty underlying the prediction process. We will see in the next section, that this information can be incorporated into any distance measure used for identifying incongruent sensations.

**Perturbation as Predictive Error:** As described above and depicted in Fig. 2, we generate in every time step a prediction of the next sensory state $s_{t+1}^*$ based on the current state and action. Consequently, after executing the action $a_t$, we measure the true sensor values of the robot and compare them to the prediction, i.e., $\delta = E(s_{t+1}^*) - s_{t+1}^*$. Taking the norm of vector $\delta$ we get an estimate of the discrepancy between robot expectation and reality. Assuming a reasonably accurate model, this discrepancy is an estimate of exogenous perturbations that affected the system dynamics. To correct for the inherent model uncertainty and the probabilistic nature of the predictions, we can take an exponential according to $\Delta = ||\delta||$, with $\tau_t = \exp(\delta_t - 2 \text{Var}(s_{t+1}^*))$.

This approach effectively incorporates the confidence bounds of the prediction into perturbation estimation and the exponential scaling simplifies the process of discerning nominal behavior from abnormal behavior due to external physical influence.

Fig. 3 depicts the effect the exponential scaling at confidence bound. The ground-truth highlighted in gray was hand-labeled using a video stream as reference. We can see both predictive error functions produce elevated responses during moments of contact. Yet, by incorporating an exponential scaling as performed in $\Delta$ we can reduce spurious activations and false-positives.

**B. Perceptual Anticipation Model**

In this section, we will describe how experienced perturbations can be correlated to predictive visual cues. In turn, these visual cues can later be used to anticipate the occurrence of physical contact. For example, perceiving an approaching wall may indicate an impending collision. Still, whether or not an external perturbation will occur is also dependent on the next actions of the robot, e.g., whether or not the robot will stop its course. Hence, states and actions need to be included in the prediction process.

In our approach, a dedicated model – the Perceptual Anticipation Model – learns to predict impending perturbations from a sequence of visual input data, robot actions, and sensory states. Visual input representing the environment is given by an $R \times C \times F$ measurements, i.e., depth or color channels that may vary over time. Thus, the visual input corresponds to a tensor $X \in \mathbb{R}^{R \times C \times F}$.

Training is based on repeated physical interaction with the environment, e.g., a human or static objects. Throughout this process, a stream of visuals is recorded using either a traditional RGB camera or an RGB-D depth camera. The result is a sequence of observations $X_t$. As the same time, states $s_t$ and actions $a_t$ of the robot are recorded. In addition, at every time step, the previously introduced deep dynamics model is run, in order to generate estimates $p_t$ of perturbations currently acting on the robot.

Given the above data sets, the goal of training the Perceptual Anticipation Model is to approximate the distribution

$$P(p_{t+1}, \ldots, p_{t+k} | X_{t-j}, \ldots, X_t, s_{t-j}, \ldots, s_t, a_{t-j}, \ldots, a_t)$$

(6)

where $j$ defines a window of past time steps. More specifically, we are interested in a predictive model that generates expected future perturbations $p_{t+1}, \ldots, p_{t+k}$ conditioned on past inputs $X_{t-j}$, as well as current and previous robot states and actions.

The presented task requires both attention to spatial patterns within the visual input, as well as attention to temporal patterns and behavioral dynamics. Hence, special care has to be taken to ensure that the model architecture used for learning can effectively identify and incorporate both sources of information when making a prediction.

**Network Architecture:** In order to base all predictions on both spatial and temporal information, we employ a deep convolutional recurrent network to learn the anticipation model. The input of the network is a sequence of visual data, the robot states, and actions. The output of the network is a vector $p_t = [p_{t+1}, \ldots, p_{t+k}]^T \in \mathbb{R}^k$ that describes the likelihood of a perturbation at each of the time steps \{t+1, \ldots, t+k\}. In order to incorporate temporal dynamics, recurrent units are used according to either the Long Short Term Memory (LSTM) [21] model or the Gated Recurrent Units (GRU) [22] model. These recurrent units keep track of a hidden state. In turn, the next state of the network is calculated as a function of the hidden state and the new inputs. Note that convGRU have fewer parameters, which
reduced both training and execution time. Faster forward passes can be crucial for real-time robotics application as the one described here.

**Loss Function:** In order to train all network parameters, we use a weighted binary cross-entropy as a loss function

\[
\mathcal{L} = -\frac{1}{T} \sum_{t=1}^{T} \left( w_p \log(\hat{p}_t) + (1 - p_t) \left(1 - \log(\hat{p}_t)\right) \right)
\]

where \(w\) is weight penalty, \(p_t\) is the ground truth, and \(\hat{p}_t\) is the likelihood of perturbation generated by the network.

**IV. EXPERIMENTAL RESULTS**

We conducted a human-robot interaction study to investigate the validity of the introduced approach. In all of the following experiments the prediction rate was 5Hz. System state dimension \(P\) is 28 and degrees of freedom \(Q\) is 7. Mean and standard deviation for action sampling is \(\mu = 0.1, \sigma = 0.05\). Dimension of environment data: \(R, C = 64\) and \(F = 1\) (gray-scale frame channel or depth channel). The parameters of DPM are set to: \(k = 10, w = 3, j = 9\).

**A. Interaction Dataset**

For training the anticipation model, we have created an interaction dataset involving 10 participants. Each participant interacted with the robot for about ten minutes. This resulted in a data set of 3000 data points per person. The participants were instructed to move towards the robot and touch its arm at any location. Throughout this process, the arm of the robot was continuously performing a forward-backward movement.

In addition, in 40-50% of the interactions, the participants were instructed to pretend approaching the robot and then turning back without any contact. Incorporating such *feigning moves* on the part of the human interaction partner, as well as a constant movement of the arm, ensures that the robot has to constantly monitor its state and the environment in order to appropriately update its belief about impending physical contact. Data recorded from 9 of the participants was used for training, while 1 participant was used for validation. Data from a separate set of 3 participants was used as test data to evaluate the generalization capabilities of the model.

Two different versions of the training set were generated. First, we created a setup in which depth images were used as input. In the second setup, we used grayscale video images as input.

**B. Example Interaction**

Fig. 4 shows an example interaction between the robot and a human subject applying forces on its arm. After training the perceptual anticipation model, we provide current observations in form of images (or depth images) as input. In turn, the network generates estimates for the next \(k\) time steps which reflect the likelihood of a future perturbation at \(t + k\). The moment of contact is depicted in Fig. 4 by a vertical line. We can see that the predictions for \(p_{t+1}, p_{t+5}\) and \(p_{t+10}\) are aligned along a diagonal. The predictor with a larger horizon, i.e., looking ten time steps into the future, activates early to indicate a high likelihood of a contact. The predictors with a shorter horizon activate at later time steps, based on the horizon they have been trained on. Note the attenuation of the activations, once the robot is outside the envelope for which the predictors have been trained to fire.

**C. Visualizing Network Saliency**

To better understand the decision process by which the perceptual anticipation model generates predictions, we visualized the underlying saliency using the method introduced in [23]. Fig. 5 shows an example of saliency maps at different moments in time during a human-robot interaction. Regions highlighted in red or yellow correspond to pixels with a strong influence on the output of the network. Originally, the network is not focusing on any particular region within the image. However, as the human subject starts walking in direction of the robot, the network approximately starts focusing on pixels around the body of both the human subject, as well as the right arm of the robot. As the human comes closer, the saliency scores for each pixel become larger, as indicated by the yellow coloring. At the onset of a contact, the saliency scores attenuate.

**V. CONCLUSIONS**

In this paper, we showed how Deep Predictive Models and self-supervised learning can be used for anticipating impending physical impact to the robot body. By learning a mapping between observed visual cues and future perturbations, robots can anticipate upcoming hazardous, noxious or unwanted states. To this end, we introduced a deep network architecture that combines spatial, temporal, and probabilistic reasoning in order to generate predictions over a horizon of next time steps. The network learns to focus on visual features that are most indicative of future exogenous perturbations. For future work, we aim at extending the framework, such that the predictions can be conditioned on entire control policies of the robot and not only a single next action. Furthermore, we will incorporate prediction framework into reinforcement learning, in order to also autonomously learn preventive motions for avoiding perturbations.

**References**

Fig. 4: Predictions of the likelihood of physical perturbation generated from a perceptual anticipation model which was trained on grayscale video input. The predictors fire at different moments ahead of the actual moment of contact.

Fig. 5: Saliency map visualization of $p_{t+5}$ predictor. It anticipates any interaction 1s into the future. The images show, highlighted in red, the pixels the network is paying attention when generating an output. In the beginning, almost no region is active, but as the human starts approaching the robot, the network focuses on the human and also on the robot arm. Once the subject is in contact with the robot, the activations attenuates.


