Abstract: Enhanced robotic grasping and manipulation involves an intelligent controller changing the manipulators trajectory to avoid failure due to slippage. Tactile sensing is considered an essential feature for such dexterous robot manipulations. Predicting the tactile readings given planned actions of the robot is challenging and still remains an open problem. In this paper, we present a tactile predictive model by proposing a series of robot-action-conditioned recurrent neural network models. Moreover, we created a data-set of simple pick-and-move tasks as well as complex kinesthetic movements with a 7DOF robot manipulator while grasping a bottle-shaped object with low resolution magnetic-based tactile sensors on the fingers of the robot. The tactile predictive models are trained using the data-set and their performances on a test set are compared. The proposed architecture is inspired by state of the art video prediction models and resolves their shortcomings in a real-time manipulation tactile prediction task. The model shows good test results, suitable for use in a future slip avoidance controller during robotic manipulative movements.

Keywords: Robotic grasping and manipulation, Tactile Sensors, Action-Conditioned Predictive Models

1 Introduction

A sense of touch plays an important role in a humans capability to avoid an object slipping when grasping and manipulating. In addition to the sophisticated distributed sensing within a human finger to sense mass, friction, temperature, and geometry [1, 2], humans can also predict object slippage and incorporate it in their control policy to avoid corresponding failures [3]. The control policy may include (1) changing the gripping force—which may not be always an option, e.g. when the object is already gripped with maximum gripping force— or (2) adjusting the manipulation movements.

Contacts between robotic hands and objects during manipulative movements are too complex to analytically model. Consequently, tactile sensors can be utilised to sense these interaction forces to close the control loop and avoid slippage. Data-driven approaches can be used to capture highly non-linear and complex behaviour of interactions between a robot and its environment in videos and high dimensional images spaces [4].

While video prediction models using convolutional LSTM networks have been already exploited for future tactile data prediction [5, 6], to the authors’ knowledge, predicting future tactile data suitable for slippage detection in a complete manipulation task has not been successfully achieved. The real-world tasks demonstrated in the previous works thus far are limited to simple linear pushing motions [5], tactile signals are used for predicting grasp success [7, 8] or to obtain contact information [8].

Yamaguchi and Atkeson [9] presented an action-conditioned video prediction model to predict tactile data based on (1) a vision-based tactile sensor that generates high dimensional tactile images and, (2) planned robot movements. However, the low frame rate of the camera (e.g. circa 30 fps) limited the performance of their predictive model. Tactile sensors with discrete point-wise taxels (an individual tactile sensing node) measuring the normal and shear forces are more widely used. However, the data from the latter will make the tactile prediction problem more difficult as they do not provide the rich tactile features that camera-based sensors do.
While Tina et al. [5] proposed a tactile predictive model suitable for predicting rolling actions, that approach is not directly applicable to point-wise magnetic-base tactile data. Moreover, it may not be suitable for slippage identification. This paper proposes a novel data-driven predictive model to learn the dynamic behaviour of a point-wise tactile sensor given the planned robot movements. The predicted tactile signals will be used in future works to find the actions which minimises the slippage likelihood.

There are two types of relative movements of an object with respect to its support: (1) rolling, which involves static friction, or (2) slippage that involves kinetic friction. While rolling is the key desired movement for in-hand manipulation, slippage is the main cause of grasp failure during most of manipulation tasks. Fig. 1 illustrates these two types of contacts behaviours. Static friction \( f_s \leq \mu_s N \) where the \( \mu_s \) and \( N \) are static friction coefficient and normal force in contact is exploited in rolling an object. Rolling results in centre of mass velocity \( v_{CM} \) and acceleration \( a_{CM} \) proportional to its rotational speed \( \omega \) and acceleration \( \alpha \) respectively. In rolling motion the contact point on both objects should have the same speed, while during slippage there is relative motion between the two contact points.

The contributions of this work are: (i) we create two tactile data sets of autonomous pick-and-move tasks and human operated complex kinesthetic motions with robot hand equipped with magnetic-based tactile sensors. This data-set with several slippage and non-slippage task executions will be publicly available for future benchmarking. (ii) We develop 4 novel action-conditioned tactile predictive models capable of using low resolution magnetic tactile sensors (Xela uSkin [10]) and evaluate them on the data sets and show that they are capable of predicting tactile sequences. (iii) We show that the best performing model is capable of integration in a real time manipulation task.

2 Related Works

Analytical model-based approaches have been applied for processing different classes of tactile information [9, 11] in grasping and manipulation. This includes spike train analysis [12], tactile force derivative thresholding for slip detection [13], Hidden Markov Model for slip prediction [14], and friction cone for slip detection [15, 16]. These approaches are limited to the type of the sensors, gripper, or the task and cannot be generalised to a new tactile based manipulation system.

Tactile based deep neural networks are used for grasp policy learning [17], slip detection [18], tactile and visual data fusion for grasping [19], and tactile Reinforcement Learning for grasping [20]. While these research items contain models to process existing tactile information for a desired objective, we are interested in developing a model to predict the raw tactile signal itself.

Zapata-Impata et al. [6] proposed convolutional LSTM (convLSTM) networks to classify direction of slippage with image representation of BioTac sensor introduced in [21]. This approach generates one-step-ahead prediction by using only the history of tactile images and regardless of the planned manipulative movements.

Action-conditioned models for tactile prediction have generally been applied to visual based tactile sensors [4, 5]. The most relevant being the convolutional dynamic neural advection model (CDNA) [4] which was applied in [5] for the prediction of tactile images for use in model predictive control of a rolling task. They generated a data set of robot actions and tactile readings with a single object, by specifying a target tactile image, they were able to perform object rolling tasks to move the object to a target location.

The application of CDNA to visual tactile data is intuitive in the tasks presented in [5] as a mask of the object is clearly defined within the tactile image allowing the motion of that mask to be applied with convolutional kernels. Although specifying a goal tactile image for this rolling task can be done by making contact between robot’s finger and marble at the desired state and storing that reference image signal, the reference tactile signal of a more typical robot manipulation task is not so easily defined. Furthermore, the application of CDNA for our problem presents two key issues for use with low resolution magnetic tactile sensors. Firstly, the low resolution data means there is no outline of the object within the Xela uSkins 4x4 taxel matrix, which is required for the CDNA model. Second, conversion of the Xela tactile forces to a point spatial representation such as that used in [8] could be an acceptable input for use with CDNA. However, this representation has issues with extremely poor resolution due to the small image size used by CDNA, hence, it fails to produce
Figure 1: (a)-(b) rolling motion manipulation with static friction for tasks in [5], (c) slipping motion with kinetic friction, (d) Experimental setup for in hand manipulation with tactile prediction including xela tactile sensors and proximity sensor [23].

Using the FingerVision [22] visual tactile sensor, [7] proposed a new tactile video prediction model called PixelMotionNet for grasp success rate improvement. However, the model cannot predict abrupt changes in tactile images due to the limited frame rate of camera and the absence of proprioceptive data in the input [7]. To address the findings of this recent work, we use a tactile sensor with a high frequency (48Hz) relative to FingerVision (15Hz), and introduce proprioceptive data in the form of task space robot states and actions as input to our models.

Our main research question is to find a suitable multi-modal predictive model for tactile signals which can be later used to predict future slippage during planned robot movements. This prediction can be later used in a Model Predictive Control system to avoid grasp failure and enhance manipulation.

3 Hardware setup and data collection

We created a data-set of 290 trials of manipulative movements with variable length—71 by autonomous and 219 by kinesthetic motions. In each trail, a 7 degree of freedom (DOF) Panda arm, manufactured by Franka Emika, grasps and moves about a single bottle shaped object (Fig. 1 (d)). We mounted a pair of Xela uSkin magnetic-based tactile sensors on the Panda’s jaw gripper to sense the contact forces during grasping and manipulation. This data-set includes synchronised tactile signals, robot state, and proximity sensor measuring slippage values. We also mounted a proximity sensor to the robots end effector, so that the line of view was the top of the object allowing us to observe motion changes in object relative to the end effector. This signal will be useful to identify slippage directly during robot motions. Sample signals of our data-set are shown in figure 2.

The tactile sensor contains 16 sensing elements arranged in a square grid, each outputting shear x, shear y and normal forces with the directions, as shown in figure 1 (d). The limiting factor for frequency of data collection is the tactile sensor, which has the reading frequency at 48Hz while proximity and robot states can be read at 400 and 1000 Hz respectively. We use ROS to synchronise the data collection of tactile sensors, robot states, and the proximity sensor together.
We collected data in two different ways: (1) Controlled motions of single bottle object. This data set has variance in motion start and end position, trajectory length, end effector orientation and variance of the velocity and acceleration of the Cartesian trajectories. The objective of this data set is to test a model’s ability to predict future tactile data on realistic “pick and place style” trajectories whilst maintaining low speeds and Cartesian trajectory profiles. (2) Kinesthetic motions of a single bottle object. There are no constraints to robot motion in this data set. The object is grasped and then moved about the robots work space via kinesthetic driving by a human operator, initial grasp width was also controlled by the operator. The tactile variance in this data set is far greater than in data set A. This data set is used to test models on more extreme and aggressive motions that may be more realistic in a real world manipulation and control scenarios.

The data-set contains realistic tactile-object slip as well as non-slip and drop instances relevant to real world manipulation tasks–90% of kinesthetic movement cases are continued to cause grasp failure.

4 Deep Action-conditioned Predictive Models

In this work, we present and test a variety of recurrent neural network architectures for the action-conditioned predictive model of tactile data sequences. The problem of action-conditioned tactile prediction can be defined mathematically as, a model $\tau$ must predict a sequence of future tactile states $D_{t+1:T+H_\tau}$ given a sequence of previous robot actions $R_{t−C:T}$, previous tactile states $D_{t−C:T}$ and a sequence of future robot actions $R_{t+1:T+H_\tau}$, where $C$ is the length of the state history time window, $r$ is the current time in the trial and $H_\tau$ is the prediction horizon, for all presented models use the same time windows for history and prediction horizons ($C = 9, H_\tau = 10$). A robot action, $r \in \mathbb{R}^9$, is the end-effector task space position and orientation (Euler angles) with respect to the robot base at time while a tactile sample is $d \in \mathbb{R}^{16 \times 3}$.

For each collected trial in the data set, we stack samples of robot action $r$ and tactile data to create data sequences $R = \{r_t−C, ..., r_{t+H_\tau}−1\}$ and $D = \{d_t−C, ..., d_t\}$. The future frame sequence $\hat S$ is computed by the following transformation function given the input sequence $S = \{\{r_t−C, d_{t−C}\}, ..., \{r_t, d_t\}, \{r_{t+1}, ..., r_{t+H_\tau}\}\}

\hat S = \{\hat d_{t+1}, ..., \hat d_{t+H_\tau}\} = \tau(S, \theta_e)

(1)

Where $\theta_e$ is the model parameters. The ground truth $\hat S$ is defined as $\hat S = \{D_{t+1}, ..., D_{t+H_\tau}\}$. We can then define the loss as:

$$L = \frac{1}{C + H_\tau} \sum_{k=t−C}^{t+H_\tau} |d_k - \hat d_k|$$

(2)

In the following, we present a natural language processing inspired model (SATPM), three 1D action-conditioned tactile predictive models (ATPM/D1/D2) and an action-conditioned tactile video predictive model (ATVPM). Both the ATPM and the ATVPM models are inspired by the CDNA architecture presented in [4] and [8]. By providing the model context frames of previous time steps, the two chain LSTM states will encode the objects dynamics in the robots grasp allowing for more informed predictions into the future.
4.1 Split Action-conditioned Tactile Predictive Model (SATPM)

The first action-conditioned tactile predictive model we implemented is an LSTM model (Fig. 3) inspired from natural language processing techniques to use an auto-regressive recurrent model used for character level text translation [24]. This model flattens $r$ so that $d \in \mathbb{R}^{48}$. This model includes an encoder LSTM module with zero-initialised cell where hidden states use the concatenated history of the tactile and robot trajectory data from $C_t$. The final state of this layer is the initial state for the decoder LSTM layer (Fig. 3). In the decoder module from $t+1$ to $t+H_p$ at each time step the predicted tactile data is concatenated with the planned robot movements at the corresponding time horizon as the input for the next cell. While in NLP translation application, the encoder LSTM uses the source language sentence to create a context for the decoder LSTM for mapping to the translated sentence, here we use the encoder LSTM to create a context from the recent past tactile and robot state for the decoder to predict future tactile readings by having concatenated previous tactile signal prediction and future robot trajectory in the input.

4.2 Action-Conditioned Tactile Video Prediction Model (ATVPM)

By converting the 16x3 tactile feature vector into a 4x4 image of 3 channels for each of the forces, we are able to ensure that the topological structure of the tactile data is encoded in the input to a model suitable for camera-based tactile prediction [5] and similar models.

First, we up-scale the 4x4x3 tactile feature matrix to a 32x32x3 image and apply convolution to the tactile data in the form of two convolutional LSTMs and a single convolution layer (Fig. 4). This model draws inspiration from the PixelMotionNet model in overall model design, using two convLSTMS and individual convolution layers to combine the skip connections with the output of the LSTM.

To introduce robot state and action to PixelMotionNet general architecture, we take inspiration from the CDNA video prediction model, concatenating the tiled robot data with the output of the first convLSTM layer. Unlike CDNA however, the convolution kernels and the image masks are not incorporated in this model as they would have a negative impact on performance due to their focus on finding and applying motion to specific objects within the image, which in our scene does not exist.

This model takes the same data structure as the tactile feature vector prediction models and so the mathematical notation remains the same, however the data sequences for tactile data $D_t$ are now $d \in \mathbb{R}^{32 \times 32 \times 3}$. The models presented in this subsection sequence through $\{ t - c : t \}$ in time-steps $t$. Once all the context data has been fed to the model, it predicts the future tactile sequence $S$ from time-step $t + 1$ to $t + H_p$, where the predicted tactile data will be the input for to the model for the next time-step.
Figure 4: Network structure for video prediction model ConvATPM. Here, we sequence through a given sequence in time steps \( k \), the robot state and action \((r_0, r_k)\) are tiled and concatenated with the tactile image \( I_{k-1} \). The model outputs the tactile prediction at time \( k \). The return connection only applies when \( k > C \).

Figure 5: The models concatenate the tactile data, with tilled robot state and actions \((r_0, r_k)\), it is then pushed through the two LSTMs, concatenated with the tactile input (skip connection) and two linear layers resulting in the predicted tactile data sample \( d_k \) at time \( k \). Model ATPMD1 uses the tactile data \( d_{k-1} \) and the 1st derivative of the tactile data \( d'_{k-1} \), whilst ATPMD2 extends this to the 2nd derivative of the tactile data \( d''_{k-1} \). Due to the recurrent nature of the models, ATPMD1 and ATPMD2 must also predict the future tactile derivatives included in their inputs \( d_k' \) and \( d_k'' \) to be used as input once \( k > C \).

4.3 Action-Conditioned Tactile Predictive Models (ATPM)

In our final action-conditioned tactile predictive model, we first tile the robot state \( r_0 \) and the robot action \( r_k \) into a 48 feature vector and concatenate it with the flattened tactile data, \( d \in \mathbb{R}^{48}, d_{k-1} \), where \( k \) is the time step in the sequence. It is then fed through two LSTMs in an attempt to encode the relative momentum of the object within the robots grasp. The output is concatenated with the tactile input through a skip connection. Inspired by the CDNA [4] and PixelMotionNet [8] models, this enables the LSTMs to focus on calculating motion of the tactile feature vector based on the robots trajectory, and removes the layers requirement to also store the current state of the tactile data. The skip connection and the LSTM output are then fed through two linear layers to result in the predicted tactile feature vector at time \( t \).

The three models contain the same architecture, only differing in the input, for model ATPMD1, the 1st order derivative of the tactile data is included in the data sequence \( D \) so that \( D = \{d_{t-C}, d'_{t-C}, \ldots, d_{t+H_p}, d'_{t+H_p}\} \). Likewise for model ATPMD2 we also introduced the 2nd derivative of the tactile data as input \( D = \{d_{t-C}, d'_{t-C}, d''_{t-C}, \ldots, d_{t+H_p}, d'_{t+H_p}, d''_{t+H_p}\} \). These models sequence through the input data in the same way as the ATVPVM model, once reaching the end of the context data, the model returns its previous tactile prediction as input for the next time step.
Figure 6: This figure shows the ATPM network applied to a kinesthetic (a,b,c) and a controlled motion trial (d,e,f). The plots for ground truth \( d_i, C \) and \( [d_{i+1, C+5}, d_{i+10}] \) are all plotted at time step \( i \).

5 Results and Discussion

In this section, we present the results obtained by the different models described above, trained using the collected data sets. We also show the models integration and implementation in a real time, online tactile prediction problem.

We implemented our models in Pytorch [25], using ADAM with the suggested hyper-parameters to train and find the optimal weights [26]. For training, the data sets were split 85:7:8 training, validation and testing percentages, respectively. The models MSE on the raw tactile values are used for comparison. For the online experiment, the model was integrated into a ROS controlled system and the pytorch model was implemented on the CPU. The system ran on a computer with Intel(R) Core(TM) i7-8700k CPU @ 3.7GZ. We trained on a combination of both data sets, and tested with full trial cases from each. The performances of the designed models are summarised in the table 1.

We notice that the longer the prediction window, the larger the divergence between the prediction and the ground truth is.

This has been expected as the uncertainty in prediction propagates and its uncertainty magnitude increases along the time. Furthermore, the prediction for the normal forces is not as good as the shear x and shear y forces. One aspect of this is the models inability to predict the initial grasp forces as the finger joint values are not given to the model, however, this is a small aspect of the data and will not completely describe the models increased prediction error on the normal forces.

By plotting the models predictions for taxel values, we are able to see the models performance on predicting the changes in tactile values.

<table>
<thead>
<tr>
<th>Model</th>
<th>( D_{normalised} )</th>
<th>( D )</th>
<th>( d_{i+1} )</th>
<th>( d_{i+5} )</th>
<th>( d_{i+10} )</th>
<th>( D_{shearX} )</th>
<th>( D_{sheary} )</th>
<th>( D_{normal} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATPM</td>
<td>0.085</td>
<td>91.3</td>
<td>69.5</td>
<td>86.7</td>
<td>110.1</td>
<td>61.6</td>
<td>61.3</td>
<td>151.0</td>
</tr>
<tr>
<td>ATPMD1</td>
<td>0.0086</td>
<td>92.9</td>
<td>70.0</td>
<td>88.0</td>
<td>112.0</td>
<td>62.6</td>
<td>64.2</td>
<td>151.0</td>
</tr>
<tr>
<td>ATPMD2</td>
<td>0.0090</td>
<td>97.0</td>
<td>73.5</td>
<td>92.5</td>
<td>136.9</td>
<td>64.5</td>
<td>68.1</td>
<td>158.4</td>
</tr>
<tr>
<td>ConvATPM</td>
<td>0.090</td>
<td>97.2</td>
<td>70.1</td>
<td>93.3</td>
<td>138.8</td>
<td>63.4</td>
<td>65.2</td>
<td>162.1</td>
</tr>
<tr>
<td>SATPM</td>
<td>0.0317</td>
<td>105.7</td>
<td>81.4</td>
<td>102.1</td>
<td>124.0</td>
<td>92.3</td>
<td>81.9</td>
<td>212.6</td>
</tr>
</tbody>
</table>

Table 1: Evaluation of action-conditioned models on predicting future tactile states. The performance metrics shown are L1Loss values
As shown across figure 6, although the loss increases at larger prediction time windows, qualitatively, the model is still able to predict shifts in the tactile data even at $t + 10$, despite predicted and ground truth values not being equivalent. This may not hurt the performance of a classifier to be tested in our future work as the slippage will be classified based on the changes of the signal not absolute value of tactile signal. As shown by the gradient of predictions from $t + 1$ to $t + 10$ the system is clearly predicting beyond just repeating the tactile value of the last context frame.

The convolutional model, although not performing poorly does not produce increased results over the ATPM models. Due to the increased computational load of convLSTM’s and converting the raw tactile data to images, there is no reason to use this model over the more simple ATPM.

To investigate the models performance metrics within the context of a real time control loop we ran the model during several pick and place tasks. Once integrated, the ATPM was able to make tactile predictions within the 48 Hz system loop, proving the model can be applied to robot control without decreasing the frequency of the ROS callback loop.

This is interesting as the predicted shifts can be used to identify slippage in future works. The same observation exists for shear Y and normal force in Fig.7b and Fig.7c, respectively.

By performing the same task for multiple trials, we observed that the object had slightly different in hand rotational slippage across repeated trials, although this is not an issue with more extreme cases of slipping (especially those seen in the kinesthetic data set), the small slip shifts in the controlled motion trails tested here may indicate a lack of repeatability and resolution in the tactile sensors.

Our action-conditioned predictive tactile models show a good performance on predicting the behaviour of the tactile signals based on the planned robot movements. Nonetheless, our data-set was collected all on a single object. To test the presented models ability to generalise across objects, we would like to collect and test on objects with different weights, materials, and geometries. In doing so, we may need to introduce other modalities to the models, such as visual input, or extending the context data into more complex features to include more information about the individual object currently grasped. The model slightly outperforms on the real robot experiment w.r.t. the data-set. The sampling frequency may have slight variation across different executions which causes this issue. Future works includes predicting tactile signals in real-time, classify it as slippage or non-slippage, and close the control loop to avoid slippage.

6 Conclusion

Modelling the tactile dynamic behaviour of point-wise tactile sensors is challenging as the contacts are highly nonlinear. In this letter, we proposed a novel data-driven action-conditioned tactile predictive model. Our model predicts future tactile data based on past robot and tactile states and planned robot trajectories. The proposed multi-modal model is tested in a real time system and proved its success to predict tactile force dynamics before being observed in the ground truth signal. This action-conditioned tactile predictive model will be the base for a Model Predictive Control that enhances manipulation by controlling predictive slippage.
References


