Towards Learning to Detect and Predict Contact Events on Vision-based Tactile Sensors

Yazhan Zhang **Weihao Yuan** **Zicheng Kan** **Michael Yu Wang**
Robotics Institute
The Hong Kong University of Science and Technology
Hong Kong, China

Abstract:
In essence, successful grasp boils down to correct responses to multiple contact events between fingertips and objects. In most scenarios, tactile sensing is adequate to distinguish contact events. Due to the nature of high dimensionality of tactile information, classifying spatiotemporal tactile signals using conventional model-based methods is difficult. In this work, we propose to predict and classify tactile signal using deep learning methods, seeking to enhance the adaptability of the robotic grasp system to external event changes that may lead to grasping failure. We develop a deep learning framework and collect 6650 tactile image sequences with a vision-based tactile sensor, and the neural network is integrated into a contact-event-based robotic grasping system. In grasping experiments, we achieved 52% increase in terms of object lifting success rate with contact detection, significantly higher robustness under unexpected loads with slip prediction compared with open-loop grasps, demonstrating that integration of the proposed framework into robotic grasping system substantially improves picking success rate and capability to withstand external disturbances.

Keywords: Tactile Sensors, Contact Event Detection, Video Prediction

1 Introduction

For human manipulation, neural receptors inside human fingers provide information of mechanical interaction and thus play a pivotal role in dexterous manipulations [1]. Similarly, artificial tactile sensors have been adopted and demonstrated to be effective in robotic systems for tasks including sensing object geometry [2], contact force [3, 4, 5], and detecting contact slippage [4, 6]. However, tactile sensing technology makes progress slowly for reasons including fabrication difficulties, limited resolution, multiplexing complexity, etc. [7]. Apart from difficulties in hardware development, the inherent high dimensionality of tactile signals also challenges algorithms on information interpretation.

Among multi-modality tactile signals, detection of contact events (e.g. contact making, slippage, etc.) is irreplaceable for action adaptation. Also, anticipatory control policies support dexterous object manipulation by avoiding long time delays in human nervous system [1]. To mimic human touch feedback, multiple works have put efforts in integrating tactile sensing into contact events detection [4, 8]. However, few previous work has extensively studied contact event perception. Neither thorough contact event categorization nor generalizability of analytical methods is presented in previous works. Besides, for the nature of the high dimensionality of tactile signals, model-based methods have exceptional difficulties in interpreting useful contact information from raw readings. Data-driven methods are superior in learning patterns from high dimensional data. Therefore, some works have explored interpreting contact slip with learning methods [6, 9, 10, 11]. As far as we know, thorough contact event categorization and classification utilizing deep learning frameworks have not been explored, given its important role in reactive grasp manipulations.

Two major problems that hinder the processing effectiveness of tactile readings are: 1) unavailable suitable neural network; 2) lack of properly labelled tactile sequence data. Artificial functionality of sensing contact events that provides ground truth labelling for tactile sequences is not yet available

3rd Conference on Robot Learning (CoRL 2019), Osaka, Japan.

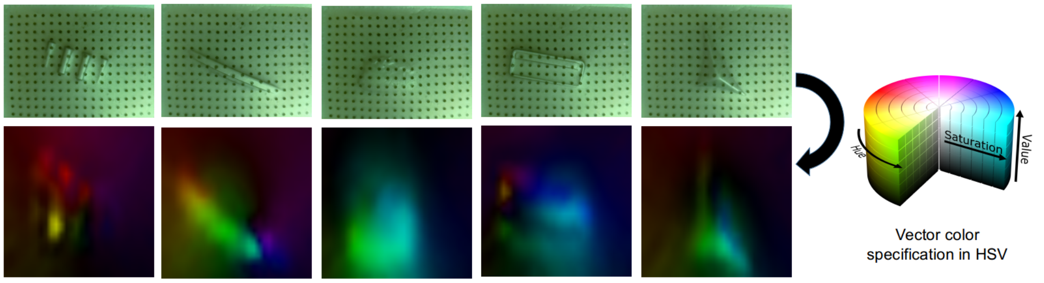


Figure 1: Examples of tactile raw images (first row) and the corresponding images of displacement vector fields (second row) collected by FingerVision. Displacement vectors are represented in HSV color specification space: Hue=Direction, Saturation=1, Value=Magnitude.

as a tool. Therefore, we design a scheme taking advantage of the human sense of touch to label tactile sequences. Towards the goal of tactile signal interpretation, we propose a neural network to process spatiotemporal readings and collect 6650 trials of contact sequence as a dataset. Raw tactile images are retrieved from FingerVision (first named in [12]) tactile sensor we developed [13] (see Figure 3(a)) that outputs deformation images of the elastomer. A robot can employ the model in predicting and detecting contact events, preventing the robot from lifting target objects without enough contact and guaranteeing stable grasping by applying extra gripping force when unstable contact events appear. This work presents the first attempt to classify contact events into explicit 7 categories, and experimental results demonstrate that incorporating predictions and detections of contact events can substantially improve the grasping capability.

Contributions: 1) Collect a dataset containing vision-based tactile sequences with careful labelling by human expert is collected. 2) Propose a network capable of predicting and detecting tactile contact events that is critical to robotic grasping systems.

2 Related Works

Contact event detection. Previous studies on human contact event detection have proven its importance for the interaction with environment. During a reaching and picking operation, four types of mechanoreceptors in human hand respond to contact events with different firing patterns, cooperatively extracting spatiotemporal features associated with mechanical contact events [1]. Most of the previous works on contact events studies focused on slip detection. Analytical methods with hand-crafted features have been presented in literatures. Heyneman et al. [14] proposed two features based on spectral analyses extracted from dynamic tactile sensors that could be used to discriminate hand/object and object/world slip. Yuan et al. [4] presented entropy feature of the deformation fields from vision-based tactile sensor Gelsight to segment sensing area into slipping and stable regions with properly selected thresholds.

Since the generalizability of these model-based methods with hand-crafted features were not tested in large number of repetitions and on different contact properties, data-driven methods, by comparison, are more appealing. Su et al. [10] proposed to classify tactile signals into slip and stable categories with a lightweight multilayer perceptron (MLP) using BioTac tactile sensor [15]. However, classification performance was not adequate for robotic operations (accuracy around 80%). SVM [16], random forest [17], Long-short-term-memory (LSTM) network [18] and convolutional LSTM (ConvLSTM) [19] have been applied to various tactile sensors [9, 11, 20] to generate slip/nonslip classifications. In this work, we provide a finer categorization of contact events by borrowing insights of neuron firing patterns triggered by distingshed contact events during human manipulations [1]. Based on these categories, we propose a classification network and conduct extensive evaluations.

Video prediction. We are interested in predicting future frames of tactile image sequences in an unsupervised learning fashion. Currently, the state-of-the-art models of video prediction are PredNet [21] with a inter-frame difference feed-forward operation, network in [22] with pixel-transformation-based module called Dynamic Neural Advection (DNA), etc. ConvLSTM [19] units are building blocks of these two models, which extract spatial and temporal features simultaneously. While both two models successfully predict future frames with more natural looks and fewer de-

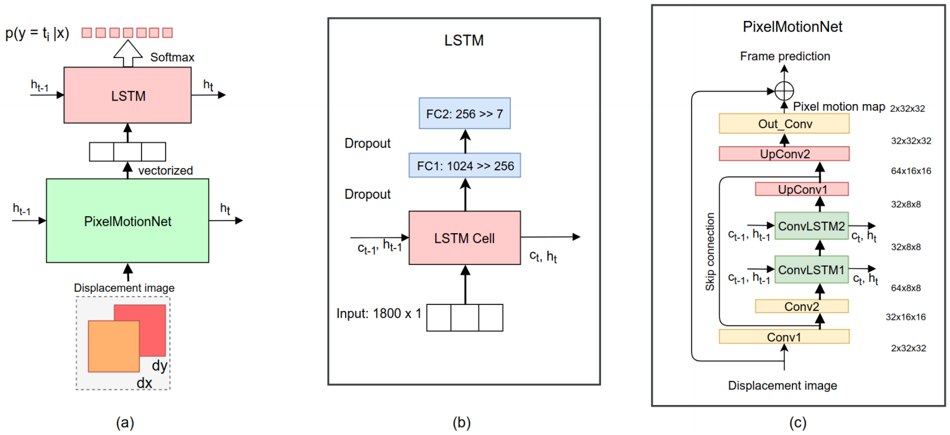


Figure 2: Network structure for contact events prediction. (a) Network structure.

fects, they are more suitable for videos involving only rigid body objects. In our work, we propose a pixel motion network ("PixelMotionNet") that explicitly predicts the velocity of each individual pixel's value that is added afterwards to the current frame to obtain a future frame. This network shows superior performance on tactile sequence prediction by comparison.

3 Learning Contact Event Detection and Predicting

In this work, we propose a LSTM-based neural network architecture as a spatiotemporal video sequence classifier for vision-based tactile data. In this section, a pipeline for tactile image sequence processing is introduced first and then description of the neural network structure is given.

3.1 Tactile Image Preprocessing

Instead of directly feeding the raw tactile images into neural networks [23], we preprocess raw images (see Figure 1) to acquire displacement fields. The sensor surface undergoes different deformations while making contact with different objects. By tracking the motions of the markers and then applying interpolation (for smoothness), the displacement vector fields are retrieved. In Figure 1, 2-D displacement vector fields are color-coded in HSV space to facilitate better visualization, in which color and intensity represent direction and magnitude of vectors, respectively.

we extract tactile displacement images iteratively and then stack samples within a time window of t_w as a sequence. Let $D[n]$ be the tactile image at n^{th} sample in the sequence S, N_s be the length of the tactile sequence and f_s be the sampling frequency. $D[n]$ has two channels, $dX[n]$ and $dY[n]$, which are projection components of V_d along X and Y axes respectively, on a grid of size $N_h \times N_w$. Then the collected sequence S can be represented as

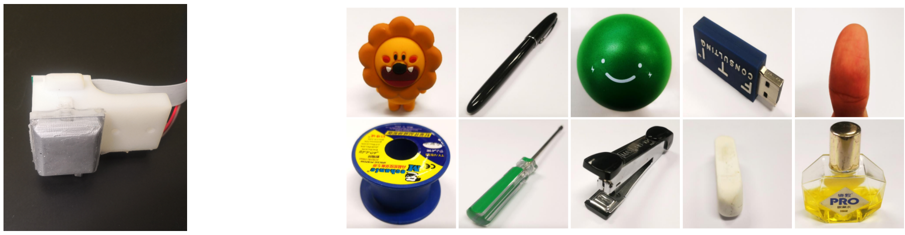
$$S = \{D[n] \mid D[n] = (dX[n], dY[n])^T, n = 0, 1, ..., N_s\} \tag{1}$$

where

$$dX[n]\{i,j\} = \langle \vec{V}_d\{i,j\}, \vec{e}_1 \rangle \vec{e}_1$$
$$dY[n]\{i,j\} = \langle \vec{V}_d\{i,j\}, \vec{e}_2 \rangle \vec{e}_2$$

$i = 0, ..., N_w$ and $j = 0, ..., N_h$ and $\vec{e}_1$, $\vec{e}_2$ denote the orthonormal bases for the X and Y axes on the sensor's coordinate system respectively.

In our configuration, the system runs at frequency $f_s = 30$ Hz, and time window $t_w = 1$ s, therefore each sequence contains 30 samples of deformation images. To further reduce the forward-propagation time of the network, we resample the original sequences with a stride of 2, thus the length of each resampled sequence $N_s = 15$. After interpolation, entries of each $dX[n]$ and $dY[n]$ are of shape 30×30. Succinctly, we have $D[n] \in \mathbb{R}^{C_h \times N_h \times N_w}$, where C_h denotes the number of channels ($C_h = 2$ in our case).



(a) (b)

Figure 3: FingerVision sensor (a) and objects used for data collection (b).

3.2 Network Architecture

The collected dataset is in the form of $I = \{S^1, ..., S^K\}$, with K being the number of samples. Corresponding to each sequence, label $y^i \in \mathbb{R}^C$ is in the one-hot encoding, where C denotes the number of classes. The proposed network in Figure 2 consists of two subnetworks: Contact event detection network and video prediction network.

Contact Event Detection Network. The event detection network is a long-short-term-memory network (LSTM). Input to this network is sequences of tactile images $S_v = \{D_v[1], ..., D_v[N_s]\}$, where $D_v[n] \in \mathbb{R}^{M \times 1}$ and $M = 1800$. To estimate the probabilities of a sequence S_v that belongs to certain class, the last hidden state vector at position $n = N_s$ is fed into two cascaded Fully-Connected (FC) layers and a Softmax layer. To avoid overfitting in the training phase, two Dropout [24] regularization layers with possibility of 0.5 are added, as depicted in Figure 2(b). Loss of the network is a reduced multi-class cross entropy between ground truth encodings and the corresponding predicted probability vectors.

Video Prediction Network. The network PixelMotionNet is composed of convolution/upsampling and ConvLSTM modules with a skip connection and an additive operation, as illustrated in Figure 2(a). The model predicts the value expectation of each pixel in the next frame depending on spatiotemporal features propagated from previous image frames, inspired by the pixel transformation module presented in [22, 25]. Upsampling layers recover feature maps back to the original size for the additive operation between the predicted pixel motion map and the current frame. Future frame prediction can be seen as a small modification by the pixel motion prediction map on the current frame. Different from [22], correlations between value changes of pixels in this study are only constrained by the sizes of the convolution kernels, instead of a spatial extent parameter. For multiple-frame prediction scenario, estimated next frames are circulated as new inputs back to the prediction network iteratively.

The problem of spatiotemporal sequence prediction is to predict the most likely future sequence with length N_p given the previous N_{in} observations. For the PixelMotionNet, the loss function used during the training phase is L^2-norm of the difference between the ground truth images and the predicted ones. Suppose the future frame sequence is $\hat{S}$, $\hat{S}$ is computed by the following transformation function given an input sequence $S = \{D[1], ..., D[N_{in}]\}$

$$\hat{S} = \{\hat{D}[1 + N_{in}], ..., \hat{D}[N_p + N_{in}]\} = \mathcal{T}(S, \theta_v) \tag{2}$$

and the loss is given by

$$\mathcal{L} = \frac{1}{N_p} \sum_{k=1}^{N_p} (\hat{D}[k + N_{in}] - D[k + N_{in}])^2 \tag{3}$$

where N_{in} and N_p are lengths of the input frames and future frames, θ_v is the learned parameters of the prediction network, $\mathcal{T}$ is the transformation function of the PixelMotionNet.

4 Data Collection

Tactile image sequences span across temporal and spatial dimensions. When the tactile sensing area undergoes certain contact events, the tactile images evolve accordingly. In previous works,

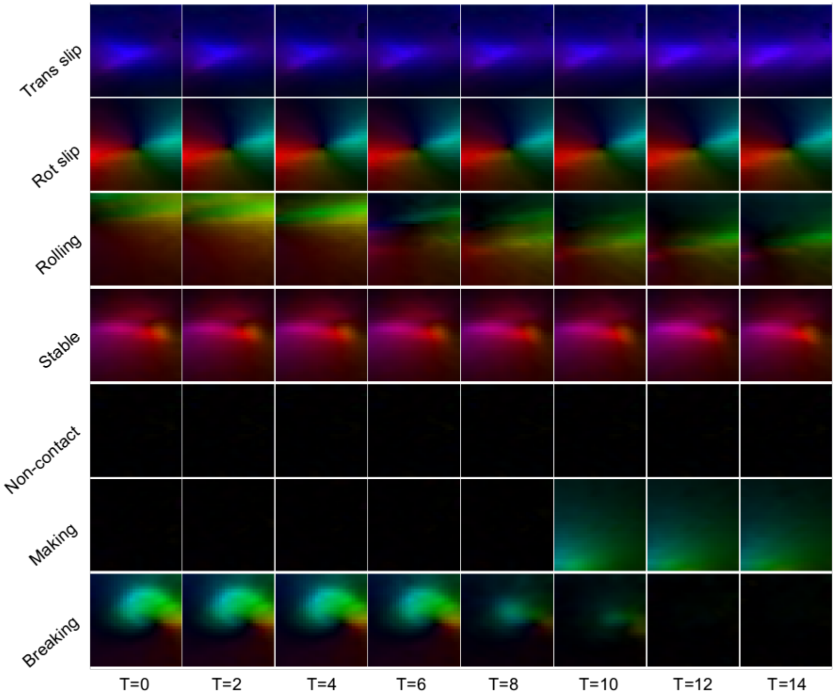


Figure 4: Example network input image sequences with different labels. Extra resampling with stride of 2 is applied for visualization and saving space.

automatic and outcome-associated contact events labelling methods were adopted. In [10], an IMU-based slip detecter was used to automatically annotate tactile readings during grasping into slip or nonslip. In [6, 9, 23], external cameras monitor relative motions between grippers and objects, or extra contact making/breaking detection networks are pre-trained as grasp success/failure discriminators. Object dropping from gripper is a consequence of unstable contacts, and the labelling of the tactile image with success or failure of grasping is reasonable only for stability estimation. However, dynamic grasping adjustment on the fly requires more temporal accuracy on labeling time windows during when contact events happen.

In this work, labelling of data is handled by a human expert. The reason why human involved in this procedure is that automatic data collection requires a heuristic rule system or pretrained discriminator to help label tactile squences, which is unavailable or asks for human prior to guide another labeling process to traine the discriminator beforehand. In comparison, labeling with the aid of human sense of touch is superior in temporal accuracy and more direct. When collecting each tactile sequence, one hand of the expert holds object and drives motion that associates to a targeted contact event on tactile sensing area. At the same time, the other hand triggers the corresponding labelling action for this sequence. According to [1], for human, the time gap between tactile sensing of external stimuli and execution of the muscle contraction is roughly 100 ms, therefore, roughly 3 frames could be mislabelled with our annotation method. We classify contact events into 7 sets with unique contact behaviors spatiotemporally: 1) Translational slip: object sticks with the sensor surface and slips translationally; 2) Rotational slip: object sticks with the surface and slips rotationally; 3) Rolling: object rolls on the surface with contact maintained (for round edge) or experiences short contact breaking and remaking (for flat surface); 4) Stable: object moves with relatively small/no motion on the surface; 5) Noncontact: no object is making contact with the surface; 6) Making contact: object makes contact within the sequence; 7) Breaking contact: object breaks contact the current sequence.

To guarantee the generalizability of the network trained on the dataset to wide range of objects with different geometry, hardness, elasticity, texture, we selected 10 objects as shown in Figure 3. Apart from properties of the selected objects, we applied forces and drove motions randomly by hand while collecting the data. Force intensity is in the range of 0∼20Nand the duration is 0.5∼2s for each interaction. For each class, we collected around 500 sequences. With skipping resampling (stride

Table 1: Best performances and properties of models.

Model	Acc(%)	Prec (%)	Rec (%)	F1 (%)	T_f (ms)	N_{in}	End Epoch
ConvLSTM	87.86	88.01	88.12	87.92	**2.7**	11	95
CNN+LSTM	92.14	92.19	92.13	92.1	32.7	10	78
LSTM	**98.50**	**98.63**	**98.39**	**98.50**	9.4	12	55

2) and removal of out-of-class samples, we generated 6650 tactile sequences in total. Examples of tactile sequences with labels are shown in Figure 4. The collected dataset is publicly available at https://sites.google.com/view/tactile-event-corl2019/home.

5 Experiments

In this section, we present settings of the network training and baseline comparisons first, then describe the implementation of reactive grasping experiments with integration of the proposed network.

5.1 Network Training and Baseline Comparison

We train separately the sub-networks considering better convergences and more convenient model evaluations. After separate training, our cascaded network takes in tactile image sequences and predicts jointly the probabilities of each class, as shown in Figure 2(a). All networks run on a computer with Intel i7-6700K CPU and NVIDIA GTX 1080Ti GPU, with batch size of 16, Adam optimizer [26] and early stopping mechanism to prevent overfitting. To regularize the networks in the training phase, an extra weight decay with value of 0.05 is added. Considering the relatively small dataset size, small initial learning rates 4×10^{-5} and 6×10^{-5} are employed and decreased with training epoch for classification and video prediction network, respectively. For both training, dataset was split into training set and validation set with ratio of $9 : 1$ sampled randomly with a fixed random seed. After split, resulting support for each class is evenly spread.

For classification network, two additional baselines are selected. One is ConvLSTM, of which we flatten the output and then feed the vector to two FC layers. The another is a cascaded Convolution neural network (CNN) with a vanilla LSTM (CNN+LSTM). Three models are all built with relatively shallow structure bearing the goal to achieve real-time capability for decentralized processing of tactile units.

For the video prediction network, quantitative evaluation of the proposed PixelMotionNet, state-of-the-art models PredNet [21] and ConvLSTM [19] are presented. Models are trained only on a subset of 4 classes in the dataset without rolling, making contact, and breaking contact data, considering the degradation of video prediction performance when encountering these abrubtly varying events with hardly observable temporal coherences. Mean square error (MSE) and Structural Similarity Index (SSIM) [27] are metrics used for evaluation.

5.2 Performance Evaluation

Following performance evaluations are all on the validation set.

Classification network. Performances of LSTM and baseline models on tactile sequence classification are summarized in Table 1. Here evaluation metrics including forward propagation time T_f, accuracy, average precision, average recall, average F1-score over all classes, N_{in} with which models achieved the best performances, and end epochs during training are given.

From the experimental results, LSTM outperforms the other two baseline models in all aspects except for the forward propagation time. ConvLSTM is superior in model size and acceleration on forward propagation for its weights sharing convolution layer compared to the fully connected structure in the vanilla LSTM. LSTM network overall achieves an accuracy peak of 98.50% costing short forwarding time with $N_{in} = 12$.

Video prediction network. Quantitatively, from the results in Table 2, PixelMotionNet is superior in both metrics compared to the other two baselines. We notice that the farther the network predicts, the larger the divergence between the prediction and the ground truth is. Furthermore,

Table 2: Evaluation of models on predicting future frames, w.r.t. the future frame index.

Model	Metrics	1	2	3	4	5
ConvLSTM	MSE	0.642±2.2930	0.940±3.600	1.310±5.7200	1.560±7.4700	1.800±9.4300
	SSIM	0.950±0.0100	0.940±0.0200	0.970±0.0300	0.850±0.0400	0.830±0.0400
PredNet	MSE	0.304±0.1580	0.259±0.1660	0.341±0.2110	0.483±1.0900	0.649±1.3900
	SSIM	0.976±0.0009	0.974±0.0040	0.967±0.0040	0.956±0.0059	0.940±0.0069
PixelMotionNet	MSE	**0.024±0.0005**	**0.061±0.0049**	**0.107±0.0142**	**0.179±0.1060**	**0.301±0.1680**
	SSIM	**0.990±0.0001**	**0.988±0.0002**	**0.979±0.0006**	**0.970±0.0018**	**0.957±0.0029**

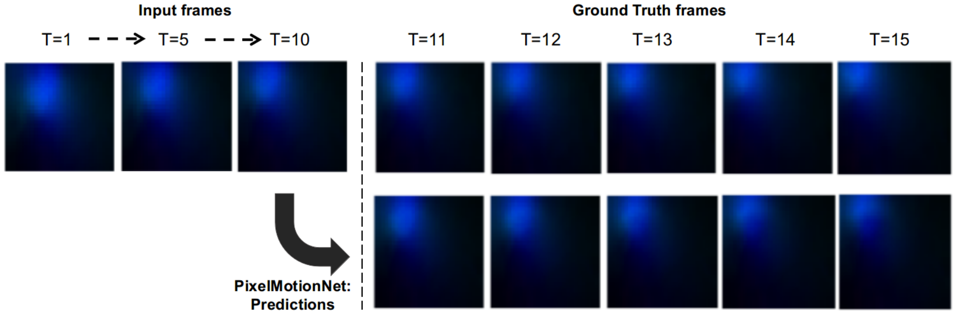


Figure 5: Tactile sequence prediction by PixelMotionNet: ground truth sequence vs. predicted sequence.

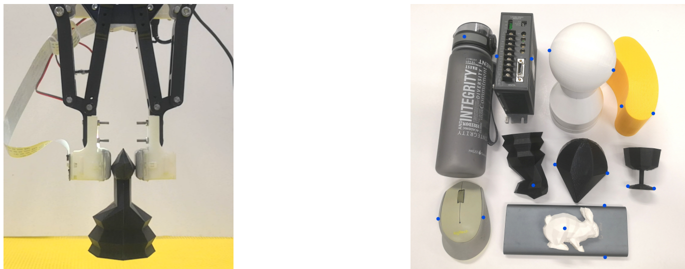
variations of predictions rise as future frame index increases, which is reasonable since the pixel value expectations of the future frames are less predictable as the network predicts further into the furture. Qualitatively, in Figure 5, the predicted future frames are illustrated aside with the ground truth frames to show how well the prediction network performs on the validation set. The PixelMotionNet captures the motion of force concentration and pixel values successfully (the result is more than copying the last image of the input sequence). It can be also seen that as the models predict more frames into the future, blurrier the images become, which is consistent with the quantitative analysis.

5.3 Experiments in Real-time Grasping

Since what we concern more in practice is how well the contact event prediction and detection networks help robotic manipulation, we directly perform real-time grasping experiments integrating the proposed network instead of evaluating it on a test dataset. In the grasping experiments, we install the FingerVision sensor on a Robotiq 2-finger gripper mounted on a UR10 robotic arm, as shown in Figure 6(a). This section is supplemented with a video document.

Contact detection experiment. To evaluate the performance of our framework in detecting contact making, we test the grasping success rate with and without our tactile contact detection on 10 objects with different shapes, sizes, and materials, as presented in Figure 6(b). We first choose the grasping sites and measure the required gripper openings with ruler manuly, with which the objects can be narrowly grasped and lifted. Then we add a small noise with standard deviation $\sigma_n = 1$ mm to simulate noises in non-contact measurements, e.g., vision, and test if the gripper can grasp and lift the objects. For open-loop operation, the gripper closes until it reaches target opening, while for close-loop grasp, the gripper adjusts its opening until the network indicates contact making event. 10 trials are executed on each object in both groups. The number of successful trials for each object and average success rate are summarized in Table 3. The success rate of these 10 objects is 46% without contact detection while 98% with the contact detection, reflecting that the contact detection facilitates the grasping significantly.

Stable grasp under slippage. Complementary to the above experiments, a further ability test of the proposed framework to help stabilize grasped object under external disturbance by predicting and detecting slip occurrence is performed. In this experiment, a gripper with the tactile sensor holds the object, then weights are loaded on top of the object one by one to trigger slip on the contact surfaces, as illustrated in Figure 7. Different outcomes under increasing loads are given in Figure 7(b) and (c), without and with slip prediction by our framework respectively. With same initial gripper opening, while gripper lost stable contact after 3 weights were loaded without slip



(a) Experiment scene

(b) Grasping objects

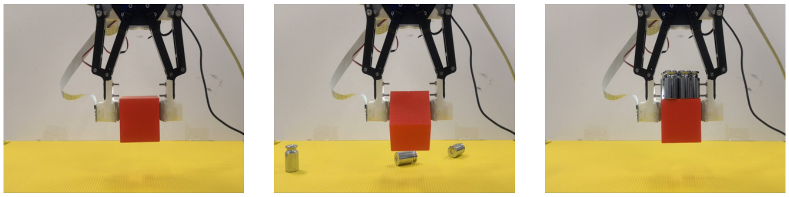
Figure 6: Real-time grasping experiments are performed on 10 distinct-shape objects. Grasping sites are denoted by blue dots in (b).

Table 3: Grasping Success Rates

Detection	Box	Ball	Chess	Yellow	Bottle	Cup	Cone	Rabbit	Mouse	Power	Average
Without	2	3	6	6	2	6	5	6	5	5	46%
With	10	10	10	9	10	10	10	10	9	10	**98**%

prediction, it maintained stable grasp with all 7 weights loaded by actively controlling the gripper's opening when slip prediction is provided. Qualitatively, slip prediction mechanism enhances the ability of the grasping system under external disturbances.

Overall, the experimental results suggest that the proposed framework is able to cover multiple phases of robotic grasping and enhance grasping system performance substantially.



(a) Initial state

(b) Slip occurs

(c) Stable grasp

Figure 7: Slip happens in (b) without slip prediction while grasp remains stable with slip prediction in (c) under increasing load. (Better shown in video)

6 Conclusion and Discussion

Humans manipulate objects with smooth transitions between contact phases by predicting and detecting contact events. In this paper, we try to endow the robot with similar capabilities. To this goal, we develop a contact event prediction and detection network, consisting of classification and sequence prediction subnetworks. We collect a contact sequence dataset of size 6650 with careful labelling by a human expert. Taking a separate training and evaluation scheme, the results show that the subnetworks outperform baselines on inference tasks given tactile sequences. Jointly, we integrate the networks together and implement real grasp experiments, of which the results show that the proposed framework grant robotic grasping system with new skills and improve overall performance.

In our work, we attempt to predict and detect contact events in the absence of visual modality and proprioceptive input. However, we observe that for contact making and breaking, PixelMotionNet is incapable of capturing changes that happen in relatively short time, which leads to an inaccurate prediction. On one hand, this stems from limited frame rate of hardware set up. On the other hand, vision and proprioceptive signals could potentially alleviate this problem by building a robotic perception system based on multi-modality sensor fusion. This raises an interesting direction as our future work.

Acknowledgments

This work is supported by the Hong Kong Innovation and Technology Fund (ITF) ITS-018-17FP.

References

[1] R. S. Johansson. Sensory control of dexterous manipulation in humans. In *Hand and brain*, pages 381–414. Elsevier, 1996.

[2] M. K. Johnson and E. H. Adelson. Retrographic sensing for the measurement of surface texture and shape. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1070–1077. IEEE, 2009.

[3] K. Vlack, T. Mizota, N. Kawakami, K. Kamiyama, H. Kajimoto, and S. Tachi. Gelforce: a vision-based traction field computer interface. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1154–1155. ACM, 2005.

[4] W. Yuan, R. Li, M. A. Srinivasan, and E. H. Adelson. Measurement of shear and slip with a gelsight tactile sensor. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 304–311. IEEE, 2015.

[5] D. Ma, E. Donlon, S. Dong, and A. Rodriguez. Dense tactile force distribution estimation using gelslim and inverse fem. *arXiv preprint arXiv:1810.04621*, 2018.

[6] J. Li, S. Dong, and E. Adelson. Slip detection with combined tactile and visual information. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7772–7777. IEEE, 2018.

[7] R. S. Dahiya, G. Metta, M. Valle, and G. Sandini. Tactile sensing-from humans to humanoids. *IEEE Trans. Robotics*, 26(1):1–20, 2010.

[8] S. Dong, D. Ma, E. Donlon, and A. Rodriguez. Maintaining grasps within slipping bound by monitoring incipient slip. *arXiv preprint arXiv:1810.13381*, 2018.

[9] F. Veiga, H. Van Hoof, J. Peters, and T. Hermans. Stabilizing novel objects by learning to predict tactile slip. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5065–5072. IEEE, 2015.

[10] Z. Su, K. Hausman, Y. Chebotar, A. Molchanov, G. E. Loeb, G. S. Sukhatme, and S. Schaal. Force estimation and slip detection/classification for grip control using a biomimetic tactile sensor. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pages 297–303. IEEE, 2015.

[11] K. Van Wyk and J. Falco. Slip detection: Analysis and calibration of univariate tactile signals. *arXiv preprint arXiv:1806.10451*, 2018.

[12] A. Yamaguchi and C. G. Atkeson. Combining finger vision and optical tactile sensing: Reducing and handling errors while cutting vegetables. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pages 1045–1051. IEEE, 2016.

[13] Y. Zhang, Z. Kan, Y. A. Tse, Y. Yang, and M. Y. Wang. Fingervision tactile sensor design and slip detection using convolutional lstm network. *arXiv preprint arXiv:1810.02653*, 2018.

[14] B. Heyneman and M. R. Cutkosky. Slip classification for dynamic tactile array sensors. *The International Journal of Robotics Research*, 35(4):404–421, 2016.

[15] N. Wettels, V. J. Santos, R. S. Johansson, and G. E. Loeb. Biomimetic tactile sensor array. *Advanced Robotics*, 22(8):829–849, 2008.

[16] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[17] A. Liaw, M. Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[18] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[19] S. Xingjian, 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 *Advances in neural information processing systems*, pages 802–810, 2015.

[20] B. S. Zapata-Impata, P. Gil, and F. Torres. Learning spatio temporal tactile features with a convlstm for the direction of slip detection. *Sensors*, 19(3):523, 2019.

[21] W. Lotter, G. Kreiman, and D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[22] C. Finn, I. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in neural information processing systems*, pages 64–72, 2016.

[23] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine. The feeling of success: Does touch sensing help predict grasp outcomes? *arXiv preprint arXiv:1710.05512*, 2017.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15 (1):1929–1958, 2014.

[25] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[26] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.