2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Learning incipient slip with GelSight sensors: Attention Classification with Video Vision Transformers

Amit Parag[1], Edward H. Adelson[2], and Ekrem Misimi[1]

***Abstract*— An important aspect of robotic grasping is the ability to detect incipient slip based on real-time information through tactile sensors. In this paper, we propose to use Video Vision Transformers to detect the onset of slip in grasping scenarios. The dynamic nature of slip makes Video Vision Transformers well-suited for capturing temporal correlations with relatively small datasets. The training data is acquired through two GelSight tactile sensors attached to the generic finger grippers of a Panda Franka Emika robot arm that grasps, lifts and shakes 30 everyday objects in order to induce slip. We further conducted an ablation study by considering 5, 4, 3, and 2 frames prior to slip onset, revealing consistent prediction accuracy. Our approach demonstrates the capability to predict slips well in advance, even up to the 5^{th} frame before the onset. This underscores the predictive capability of our approach, indicating its effectiveness in slip detection well before of its occurrence. This advance prediction capability may be a valuable tool for undertaking preemptive corrective actions, such as implementing a more secure gripper closure. We evaluate the efficiency of our approach to predict onset of slip on 10 previously-unseen objects and achieve a zero-shot mean prediction accuracy of 99%.**

I. Introduction

Dexterous manipulation and grasping is characterized by grasp planning, force and compliance control and real-time adaptability. An important aspect in grasp planning then becomes detection of onset of slip since slip often results due to application of insufficient force or a flawed grasping strategy. To anticipate and respond preemptively to slip, access to information about the contact state between the gripper and the object is crucial. This information can be obtained through various sensing modalities, including visual and tactile sensors or a combination of both.

Research in this area, as discussed more comprehensively in [1]–[4], has focused on developing sophisticated tactile sensors using accelerometers [5], force transducers [6], pressure-sensitive tactile arrays [7], [8], piezoelectric polymer films [9], [10], elastomer-embedded cameras [11], bio-mimetic optical tactile sensors [12], carbon nanotube-polymer composites [13], and strain gauges [14]. The feedback from these sensors is then analyzed through slip detector algorithms. Typically, slip detector algorithms tend to rely on spectral analyses to extract relevant features for slip classification [7], [10], [14] or use band-pass filters [15] or optical flow algorithms [8], [16] or force measurements to predict the onset of slip [17]. In [15], [18] grasp strategy in tactile sensors were implemented with slip detection through vibration measurements.

[1]SINTEF Ocean, Norway
[2]MIT, USA

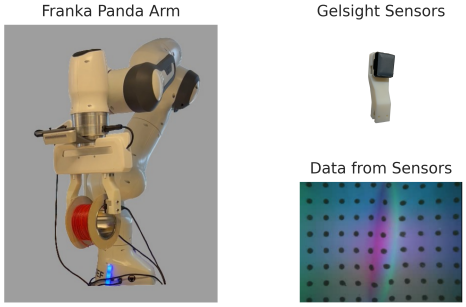


Fig. 1: Experimental Setup. The GelSight sensor is attached to the 3D printed gripper finger pads of the Panda Franka Emika Arm. The image on the lower right is the corresponding tactile data gathered from the sensor while it is gripping the object.

While tactile sensors have successfully detected slip, data-driven strategies, as demonstrated in [19], [20], have shown comparable or superior results. Support Vector Machines were employed in [12] for slip detection, and this method was extended to a multi-fingered gripper in [21]. A combination of visual and tactile information trained deep neural networks for slip classification in [22].

The effective detection of slip events is hindered by its inherently temporal nature, posing a challenge in capturing the dynamic characteristics associated with slipping. This has led to sophisticated methodologies to enhance slip detection accuracy. Models based on Recurrent Neural Networks (RNNs) [23], [24] and Long Short-Term Memory (LSTM) [24]–[28] or networks such as Temporal Convolutional Networks [29], [30] have also been leveraged to capture the temporal dependencies inherent in slipping phenomena. Additionally, physics-informed deep architectures with an entropy based strategy [31] have been successfully used to deal with the inherent disorder and unpredictability associated with slip. In [32] the authors use a deep neural network to learn a distribution over contact forces which is then used for slip classification. It was also shown in [33] that data driven methods can successfully anticipate slip in deformable objects based on data from soft force sensor.

The dynamic nature of slip events demands an approach capable of capturing intricate temporal correlations effectively during complex movements. Slip events often involve both spatial and temporal aspects, requiring a holistic approach for accurate detection. Despite their merit, previous work is predominantly based on inducing slip by gripper opening. In this paper, we introduce a set of intricate robot motion such as shaking and wiggling, to induce slip in a more

realistic robot grasping scenario. Additionally, we propose to use Video Vision Transformers and their attention mechanism [34], [35] for end-to-end learning for slip detection based on high fidelity feedback from GelSight sensors [36], [37].

Video Vision Transformers (ViViTs) inherently integrate spatial and temporal information, allowing them to analyze both the spatial configuration of the gripper and the object and the temporal evolution of the interaction.

The self-attention mechanisms within ViViT allows it to discern nuanced patterns while also facilitating end-to-end learning: in principle this allows the model to directly map raw input (video frames from the GelSight sensors) to labelled classes - onset of slip or wiggle - we define *wiggle* more concretely in II-D. We experimentally demonstrate that this allows us to achieve a comparable accuracy of 99% with [31] albeit without the need for intricate handcrafted features. Our transformer-based model for slip onset prediction is validated on a dataset of 10 previously-unseen objects and achieves a zero-shot mean prediction accuracy of 99%. This end-to-end learning capability significantly simplifies the model architecture and training process. Furthermore, in slip detection scenarios, acquiring exhaustive labeled datasets is usually challenging due to the dynamic and often unpredictable nature of slip events and this is where the efficiency of ViViTs in capturing relevant spatio-temporal information from limited data mitigates the sample dependency concern.

The rest of the paper is organized as follows - in Sec II-A we describe the capabilities and features of tactile sensors used for our the experiments. In Sec II-D we describe our hardware setup and data acquisition process. In Sec III we show the results of our experiments before concluding in Sec IV.

II. Experimental Setup

A. GelSight Sensors

GelSight sensors, initially introduced in [38], [39], have emerged as a compelling solution for applications in robotics where tactile information is required and especially in detection of incipient slips. GelSight sensors are vision based optical tactile sensors with soft sensor surface and a high-resolution sensing of geometry [36]. The sensors are equipped with a compliant elastomer interface and embedded optical structures that can efficiently capture the deformation of the elastomer surface. The resulting high-resolution 3D geometry of the contact surface can be reconstructed from the camera images. Additionally, GelSight sensors with markers can be used to provide information about both normal and shear forces.

The soft, compliant elastomer pads are coated with a reflective material, such as silicone rubber. Under contact with an object, the elastomer surface deforms, causing a local change in the reflected intensity. This pattern is then captured using a high-resolution camera positioned above the pad, which observes the reflection of light off the surface of the pad. Additionally, the feedback from the GelSight can also be used to estimate the magnitude and distribution of forces applied to the pad by measuring the degree of deformation.

B. Hardware Setup and Data Acquisition

The hardware setup we use for our experiments is shown in Fig 1. We use 3D printed grippers to mount the GelSight sensors. We perform our experiments on a variety of 40 objects with different texture and shapes, some of which are shown in Fig 2. The objects typically have the maximum width less than the maximum opening distance of the GelSight mounted grippers - in case of objects such as bowl made of jute, bottom right in Fig 2, the two grippers were made to grip these objects from one end.

The objects are initially grasped and then lifted to undergo a *wiggling* motion and subsequently placed back. The feedback from the sensors is then collected at a frame rate of 25 and at a resolution of 320×240.

During the wiggling motion, the gripper shakes the objects in the following sequence:

- **Perpendicular Shake**: This type of shake involves applying a perturbation perpendicular to the surface or direction of motion. It induces movement in a direction perpendicular to the end-effector's current orientation or path.
- **Rotation Shake**: Rotation shake applies a perturbation that causes the end-effector to rotate around its axis. This rotation can be clockwise or counterclockwise, altering the orientation of the end-effector.
- **Tangential Shake**: Tangential shake induces movement along a tangent to the end-effector's path or surface. It applies a perturbation parallel to the direction of motion, causing lateral movement without altering the orientation.
- **Vertical Shake**: Vertical shake involves applying a perturbation that induces movement along the vertical axis of the end-effector. This perturbation causes the end-effector to oscillate or move vertically, either upward or downward, relative to its current position.

The entire sequence from initial grasping, lifting, *wiggling*, and placing the object back typically takes 18 seconds to complete - note that each shake is carried out 3 times[1]. Each entire sequence is then saved as a video file. This entire sequence is then repeated multiple times for multiple different objects. This forms our base dataset. The motivation behind our particular choice of experiment is twofold:

- Analyze the efficiency of the gel pad when the sensors are gripping an object while performing complex movements.
- Induce and analyze slip cases and learn incipient slip during different robot movements.

During the entire motion, it is crucial to note that the object may slip at any time, particularly when the grippers are executing the wiggling sequence. To prevent ambiguities in terminology, we will denote wiggling as the sequence of

[1]Videos of the experiment and the code are available here :https://github.com/amitparag/Incipient-Slip-Detection.git

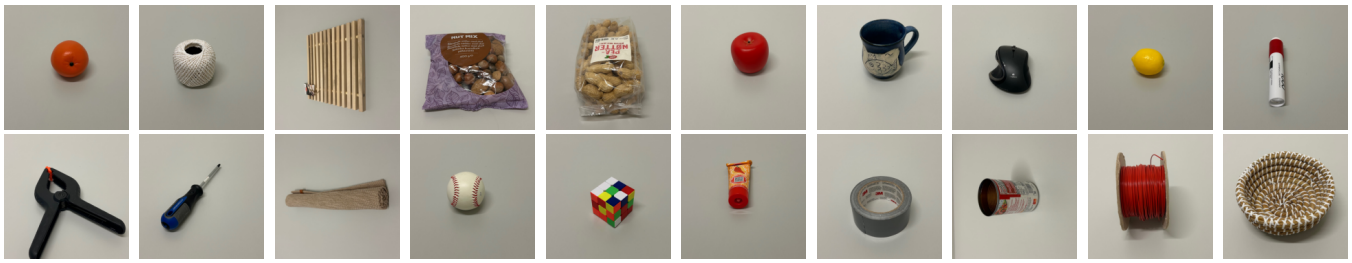


Fig. 2: A subset of objects used to generate the datasets. GelSight sensors attached to 3D printed grippers, were used to grasp, lift and wiggle the objects. The corresponding feedback from the sensors were recorded to form the training and testing datasets during the supervised learning phase.

Fig. 3: Evaluation objects for our approach. The previously-unseen objects from which the data was collected for validation of the trained neural networks. The model achieves a zero-shot mean slip onset prediction accuracy of 99%.

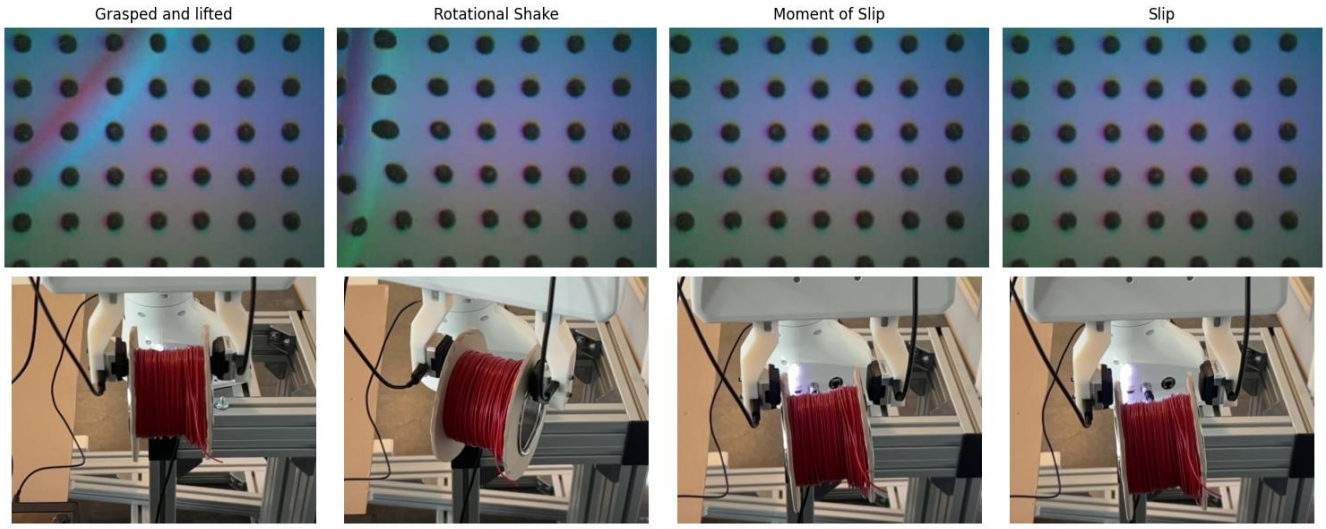


Fig. 4: Illustration of different reading on the sensors during different movements of the gripper. The top row displays the readings from the sensor on the left gripper while the bottom row displays the corresponding physical movement of the robot. In the first two figures of the top and the bottom column, we see the change in the readings as the grippers starts rotation. This change is due to slight movement of the object on the surface of the sensors. The images in the third column shows the moment of slip - we see that there is no contact between the object and the sensor on the right while there is a slight contact between the sensor on the left and the left end of the object. The images in the final column show when sensors are no longer in contact with the object. Images are zoomed in for better clarity.

events that may involve some relative motion between the object and the elastomer surface without any loss of contact between the object and gel pad. Consequently, slip implies that the object underwent relative motion with respect to the elastomer surface before eventually losing contact completely.

We define the *moment of slip* as the last frame in the feedback from the sensors before the object loses contact with either of the GelSight sensors as we see in the third column of Fig 4. This moment marks a critical event in the motion process and is eventually used in hand labelling frames to create a dataset. Therefore, even if the objects slips slightly, as long as it remains in contact with the sensors, we categorize it as undergoing wiggling. This distinction is essential for accurately analyzing the motion dynamics and ensuring precise control over the object manipulation.

In our experiments, slip typically occurs when the gripper is undergoing one of the movements in the wiggling sequence. From each video, a sequence of 5 successive frames are extracted. For the onset of slip cases, we use the last 5 frames leading up to the moment of slip: the frame on the corner top left of Fig 5 is the last frame before the slip happens, the frames before that correspond to onset of slip - note that the *moment of slip* frame itself is not used, and only 5 frames before that are used to create the onset of the slip dataset. For the cases where the object does not slip, we extract 5 frames from random starting positions in the wiggling sequence in the corresponding video. Typically,

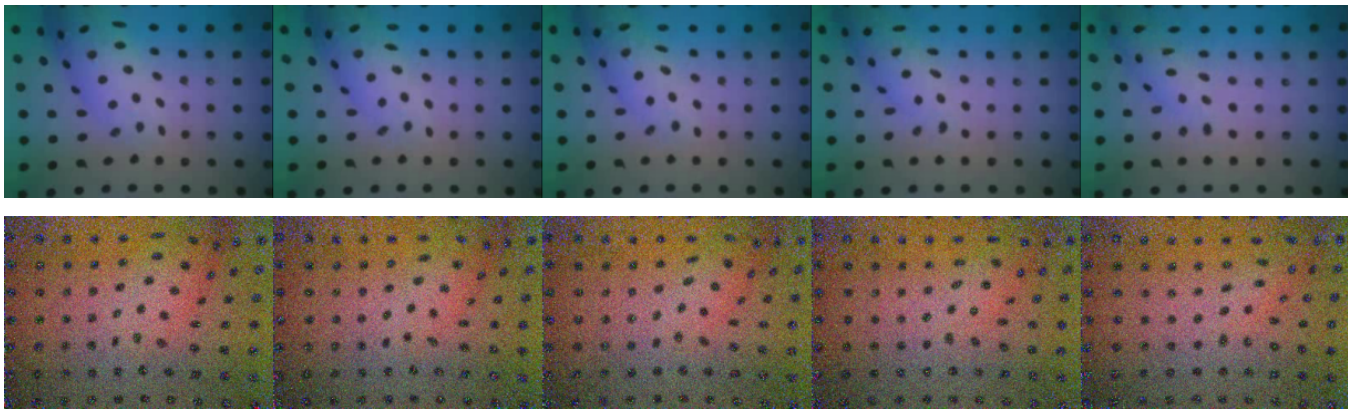


Fig. 5: Illustration of data transformation for augmentation. The top row displays a sequence of 5 frames extracted from a video corresponding to the onset of slip. The 5 frames (from left to right) are the last five frames before the *moment of slip* - i.e. before slip happens. The bottom row shows the corresponding frames after undergoing various transformations - horizontal flipping, swapping of red and blue channels, and the addition of noise. The transformed video, along with the original, are then used for training.

the patterns of deformations on the gel pad can look nearly identical in videos of onset of slip and wiggle.

Overall, we record 449 videos of wiggle and 142 videos of slip.

C. Data Augmentation

The original raw dataset of recordings from GelSight sensors was then transformed and appended to the initial dataset to create an augmented dataset. Each video in the original dataset underwent three steps sequentially to create the transformed video:

1) Add Gaussian noise with a standard deviation of 15.
2) Swap red and blue channels.
3) Horizontally flip the frames.

Fig 5 shows the resulting frames in transformed video. The resulting transformed videos, along with the raw videos, were then used for training and testing. The dataset was split in the 90%-10% ratio for training and testing, respectively.

For validation, the trained models were evaluated using data from 10 previously-unseen objects that were not included in the training set, depicted in Fig 3.

D. Video Vision Transformers

The transformer architecture was proposed by [40] as an encoder-decoder network based on the attention mechanism. Attention mechanism for neural networks, proposed by [41], consists of three main components: a set of *queries*, Q, a set of *keys*, K and a set of *values*, V. By using matrix multiplication, it is possible to compute the attention values of multiple queries at the same time, using the following formula:

$$\text{Attention}(Q, K, V) = \text{Softmax}(Q \cdot K^T) \cdot V$$

In [40] was proposed a scaled dot product attention mechanism and the multi-head attention, performing h linear projections of the queries, keys and values to d_k, d_k and d_v dimensions, respectively. The vision transformer, proposed by [42], is a self-attention-based architecture for image classification and is an adaptation of [40] for image inputs.

The Video Vision Transformer (ViViT) [34], shown in Fig 6, is a transformer-based model for video classification, and is based on the vision transformer [42]. ViViT offers two different methods for embedding the videos, *uniform frame sampling* and *tubelet embedding*, and four different video vision transformer models, *spatio-temporal attention*, *factorized encoder*, *factorized self-attention* and *factorized dot-product attention*. For our experiments, we selected the *spatio-temporal attention* mode with uniform frame sampling, as we were interested to capture the spatio-temporal changes prior to and during the onset of the slip.

III. Experimental Results

We conducted the experiments through supervised training of ViViT with specific training parameters. The ViViT model was trained on video resolutions of $(240, 320)$ with a patch size of $(40, 40)$.

We conducted experiments with different spatial and temporal depths. Our experiments showed that deeper ViViT models yielded results similar to a ViViT model with just 2 spatial layers followed by 2 temporal layers and only 8 neurons in the hidden layers. In this paper, we present the results of this particular lightweight model architecture.

We also compare the performance of ViViT with that of the widely used VideoResNet50 model. We use the pretrained VideoResNet50 model [43] which are based on the VideoResNet50 for image classification introduced in [44].

We use the cross-entropy loss, defined as:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right), \quad (1)$$

where N is the batch size, y_i is the ground truth label, and $\hat{y}_i$ is the predicted probability of the positive class. This loss function is suitable for classification tasks as it penalizes the model based on the difference between predicted and actual class probabilities.

To mitigate overfitting, L_2 regularization was applied with a weight decay parameter (λ) of $1e-4$. Regularization helps prevent the model from becoming too complex by penalizing

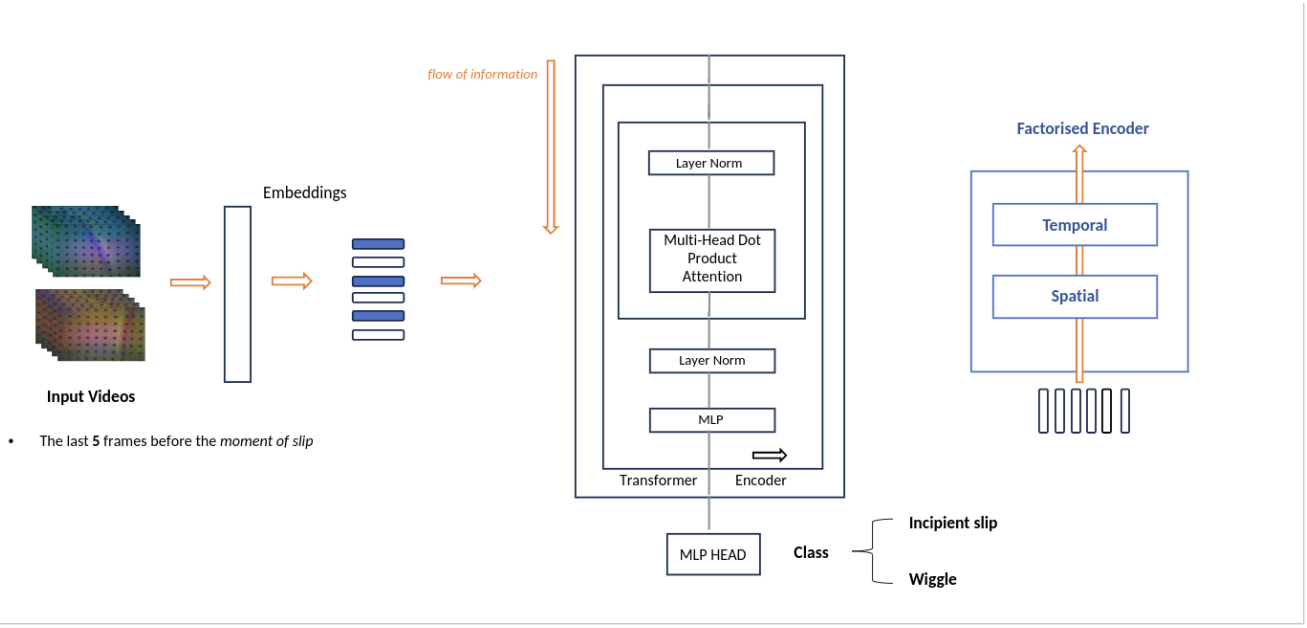


Fig. 6: Illustration of the basic architecture of the ViViT - video vision transformer model and the schematic of the learning process. The arrows represent the direction of flow of information. The blue figure on the right represents the first variant of the video vision transformer model, where temporal layers follow the spatial layers.

large weights in the model's parameters, thus improving generalization to unseen data.

The batch size used during training was 16 for ViViT and 8 for VideoResNet50. A larger batch size allows for more samples to be processed simultaneously, potentially leading to faster convergence and better generalization. The learning rate for ViViT was set to $3e-4$ and for VideoResNet50 was $1e-3$, with an Adam optimizer. Adam optimizer adapts the learning rate for each parameter, providing faster convergence compared to traditional stochastic gradient descent.

Additionally, a learning rate scheduler with a step size of 30 epochs and a gamma of 0.1 was utilized to adjust the learning rate during training epochs. This scheduler gradually reduces the learning rate over time, allowing the model to fine-tune its parameters effectively. The ViViT model was allowed to train for 75 epochs and the pretrained VideoResNet50 model was fine tuned in 10 epochs

A. Adding Noise

To evaluate the robustness of our approach, we introduced various levels of noise to the previously-unseen objects' validation dataset and assessed the performance of the Video Vision Transformer (ViViT). The noise was added at different intensities, simulating real-world scenarios where sensory inputs might be corrupted or distorted.

We observe that the ViViT model maintained a fairly high level of accuracy across low - medium noise levels, indicating its resilience to noisy inputs. At higher noise intensities, the ViViT consistently performed poorly. This robustness at small noise intensities is particularly advantageous in real-time slip detection tasks where environmental conditions may vary.

B. Comparison with VideoResNet50

The ViViT model performs comparably to the pretrained VideoResNet50 model in terms of accuracy and performance. However, the ViViT model has significantly fewer learnable parameters compared to that of VideoResNet50 $(31,657,410)$, making it more computationally efficient as is evident in the corresponding training times - ViViTs took 8 minutes to train for 75 epochs while the pretrained VideoResNet50 took 81 minutes to train for 10 epochs.

Additionally, the ViViT model achieves an inference rate of 58.29 Hertz, while VideoResNet50 only achieves an inference rate of 3.43 Hertz. This shows that self-attention mechanism in ViViT allows it to effectively integrate spatial and temporal information, resulting in superior performance in detecting incipient slip events, while requiring fewer parameters and less computational resources.

C. Training on fewer frames

Lastly, we investigated the impact of training the ViViT model with fewer frames to assess its efficiency in capturing temporal correlations with limited data. We train the ViViT model on datasets that consisted of videos of decreasing number of frames - 5, 4, 3 and 2 frames. By reducing the number of frames used for training, we aimed to simulate scenarios where acquiring extensive labeled datasets is challenging due to practical constraints. Surprisingly, our experiments demonstrated that even with a reduced number of frames, the ViViT model maintained high accuracy in predicting the onset of slip. We also observe that training is slightly more unstable with dataset of 2 frames. This initial instability is however soon corrected in the later training epochs. We are not fully sure why this happens.

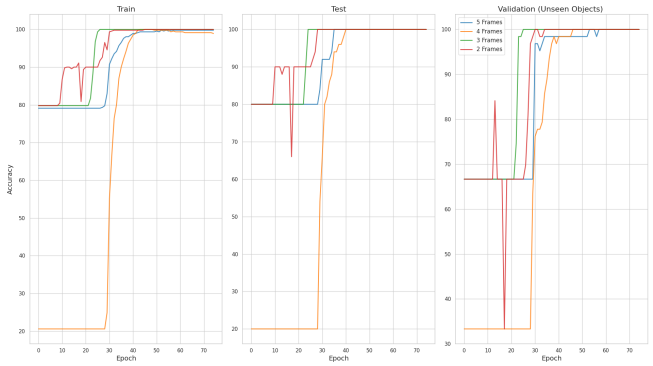
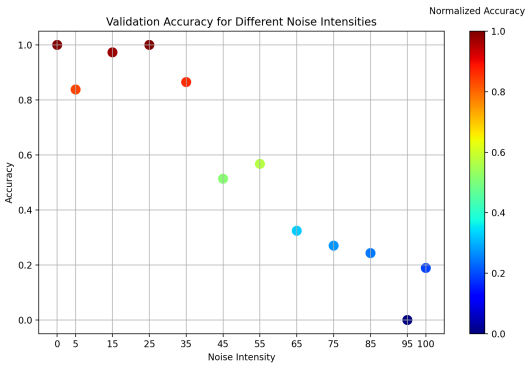


Fig. 7: Accuracy scores across 75 epochs during training, testing, and validation phases for slip and wiggle videos of various lengths (5 frames, 4 frames, 3 frames, 2 frames). The plot illustrates the performance of the ViViT model over the training period, showing how the accuracy evolves over time on different datasets.



a: Accuracy

Fig. 8: Different accuracies during validation on previously-unseen objects' data. The neural network used in this experiment was trained on the 5-frames dataset. Noise is added to the previously-unseen validation dataset at different intensities.

This suggests that ViViT's self-attention mechanism enables it to effectively leverage the temporal information present in a smaller subset of frames, making it suitable for applications where data collection is limited or costly. Moreover, training with fewer frames resulted in faster convergence and reduced training times, further emphasizing the efficiency and effectiveness of ViViT-based approaches for slip detection tasks.

D. Discussion

One limiting factor in our experiments is the need for exhaustive hand-labeled datasets. Furthermore, we noted the fragility of the sensor pads in inducing slip for the intricate range of objects and wiggling gripper motions implemented in this paper. While the optical cameras efficiently capture the deformations of the elastomer surface, the gel pad itself may be susceptible to wear-out during complex motion. We observed that the fragility seems to be dependent on the properties of the object in question - hardness, surface texture and sheer forces acting on the sensors during the wiggling sequence. In our experiments, we found that the gel pad on the sensors started to scrape off after an average of 8-10 grasp experiments. Addressing these potential limitations through the development of more durable gel materials or protective measures could enhance the reliability of elastomer-based sensors for real-world robotic manipulation with a large variety of objects.

IV. Conclusion

In this paper, we have shown the effectiveness of using Video Vision Transformers in detecting incipient slip in grasping scenarios by leveraging high-fidelity feedback from GelSight sensors. Our approach combines the advantages of ViViTs, which inherently integrate spatial and temporal information through self-attention mechanisms, with the capabilities of GelSight sensors to capture subtle surface deformations and force distributions. Our experiments showed promising results: the lightweight ViViT model achieves high accuracy in predicting the onset of slip with relatively low computational overhead making it suitable for online slip detection and potential for undertaking a preemptive corrective action.

In our future research, we aim to integrate generative models with ViViT in order to enhance the accuracy of slip onset prediction. This stems from the understanding that the determination of the optimal number of frames for supervised training is primarily heuristic in nature. Generative models offer the potential to predict future frames based on current GelSight data, which can subsequently be employed by ViViT to anticipate slip occurrence.

Consequently, this approach removes the necessity of predefining the number of frames utilized during training. For instance, the ViViT model may undergo training using three-frame videos to forecast slip, while the generative model can be trained to produce "n" frames into the future given the input of three current frames. The ViViT model, once trained, can then be leveraged to classify the forecasted frames, allowing the grippers additional time to undertake and adapt its actions accordingly. This is necessary since the latency in closing the current grippers is usually quite large, especially when 3D printed grippers are used.

Acknowledgment

The work is supported by the BIFROST (313870) project financed by the Research Council of Norway.

References

[1] R. A. Romeo and L. Zollo, "Methods and sensors for slip detection in robotics: A survey," *Ieee Access*, vol. 8, pp. 73 027–73 050, 2020.
[2] M. R. Cutkosky and W. Provancher, "Force and tactile sensing," *Springer Handbook of Robotics*, pp. 717–736, 2016.
[3] R. S. Dahiya, G. Metta, M. Valle, and G. Sandini, "Tactile sensing—from humans to humanoids," *IEEE transactions on robotics*, vol. 26, no. 1, pp. 1–20, 2009.
[4] Z. Kappassov, J.-A. Corrales, and V. Perdereau, "Tactile sensing in dexterous robot hands," *Robotics and Autonomous Systems*, vol. 74, pp. 195–220, 2015.
[5] R. D. Howe and M. R. Cutkosky, "Sensing skin acceleration for slip and texture perception." in *ICRA*, 1989, pp. 145–150.
[6] Y. Yamada and M. R. Cutkosky, "Tactile sensor with 3-axis force and vibration sensing functions and its application to detect rotational slip," in *Proceedings of the 1994 IEEE International Conference on Robotics and Automation*. IEEE, 1994, pp. 3550–3557.

[7] E. Holweg, H. Hoeve, W. Jongkind, L. Marconi, C. Melchiorri, and C. Bonivento, "Slip detection by tactile sensors: algorithms and experimental results," in *Proceedings of IEEE International Conference on Robotics and Automation*, vol. 4. IEEE, 1996, pp. 3234–3239.

[8] J. A. Alcazar and L. G. Barajas, "Estimating object grasp sliding via pressure array sensing," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 1740–1746.

[9] D. Goger, N. Gorges, and H. Worn, "Tactile sensing for an anthropomorphic robotic hand: Hardware and signal processing," in *2009 IEEE International Conference on Robotics and Automation*. IEEE, 2009, pp. 895–901.

[10] B. Heyneman and M. R. Cutkosky, "Slip classification for dynamic tactile array sensors," *The International Journal of Robotics Research*, vol. 35, no. 4, pp. 404–421, 2016.

[11] 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)*. IEEE, 2015, pp. 304–311.

[12] J. W. James, N. Pestell, and N. F. Lepora, "Slip detection with a biomimetic tactile sensor," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3340–3346, 2018.

[13] M. Vatani, E. D. Engeberg, and J.-W. Choi, "Force and slip detection with direct-write compliant tactile sensors using multi-walled carbon nanotube/polymer composites," *Sensors and Actuators A: physical*, vol. 195, pp. 90–97, 2013.

[14] R. Fernandez, I. Payo, A. S. Vazquez, and J. Becedas, "Micro-vibration-based slip detection in tactile force sensors," *Sensors*, vol. 14, no. 1, pp. 709–730, 2014.

[15] 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)*. IEEE, 2015, pp. 297–303.

[16] B. Galvin, B. McCane, K. Novins, D. Mason, S. Mills *et al.*, "Recovering motion fields: An evaluation of eight optical flow algorithms." in *BMVC*, vol. 98. Citeseer, 1998, pp. 195–204.

[17] M. Kaboli, K. Yao, and G. Cheng, "Tactile-based manipulation of deformable objects with dynamic center of mass," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2016, pp. 752–757.

[18] A. Ajoudani, E. Hocaoglu, A. Altobelli, M. Rossi, E. Battaglia, N. Tsagarakis, and A. Bicchi, "Reflex control of the pisa/iit softhand during object slippage," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 1972–1979.

[19] 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)*. IEEE, 2015, pp. 5065–5072.

[20] F. Veiga, B. Edin, and J. Peters, "Grip stabilization through independent finger tactile feedback control," *Sensors*, vol. 20, no. 6, p. 1748, 2020.

[21] J. W. James and N. F. Lepora, "Slip detection for grasp stabilization with a multifingered tactile robot hand," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 506–519, 2020.

[22] 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)*. IEEE, 2018, pp. 7772–7777.

[23] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, no. 64-67, p. 2, 2001.

[24] T. Wang and F. Kirchner, "Grasp stability prediction with time series data based on stft and lstm," in *2023 International Conference on Advanced Robotics and Mechatronics (ICARM)*. IEEE, 2023, pp. 587–593.

[25] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

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

[27] B. S. Zapata-Impata, P. Gil, and F. Torres, "Learning spatio temporal tactile features with a convlstm for the direction of slip detection," *Sensors*, vol. 19, no. 3, p. 523, 2019.

[28] A. Begalinova, R. D. King, B. Lennox, and R. Batista-Navarro, "Self-supervised learning of object slippage: An lstm model trained on low-cost tactile sensors," in *2020 Fourth IEEE International Conference on Robotic Computing (IRC)*. IEEE, 2020, pp. 191–196.

[29] M. Meier, F. Patzelt, R. Haschke, and H. J. Ritter, "Tactile convolutional networks for online slip and rotation detection," in *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*. Springer, 2016, pp. 12–19.

[30] A. Grover, C. Grebe, P. Nadeau, and J. Kelly, "Under pressure: Learning to detect slip with barometric tactile sensors," *arXiv preprint arXiv:2103.13460*, 2021.

[31] X. Hu, A. Venkatesh, G. Zheng, and X. Chen, "Learning to detect slip through tactile measures of the contact force field and its entropy," *arXiv preprint arXiv:2303.00935*, 2023.

[32] P. Griffa, C. Sferrazza, and R. D'Andrea, "Leveraging distributed contact force measurements for slip detection: a physics-based approach enabled by a data-driven tactile sensor," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 4826–4832.

[33] E. Judd, B. Aksoy, K. M. Digumarti, H. Shea, and D. Floreano, "Slip anticipation for grasping deformable objects using a soft force sensor," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 10 003–10 008.

[34] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.

[35] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[36] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[37] S. Dong, W. Yuan, and E. H. Adelson, "Improved gelsight tactile sensor for measuring geometry and slip," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 137–144.

[38] 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*. IEEE, 2009, pp. 1070–1077.

[39] M. K. Johnson, F. Cole, A. Raj, and E. H. Adelson, "Microgeometry capture using an elastomeric sensor," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 4, pp. 1–8, 2011.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017.

[41] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[42] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and e. a. Gelly, Sylvain, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[43] H. Fan, T. Murrell, H. Wang, K. V. Alwala, Y. Li, Y. Li, B. Xiong, N. Ravi, M. Li, H. Yang, J. Malik, R. Girshick, M. Feiszli, A. Adcock, W.-Y. Lo, and C. Feichtenhofer, "PyTorchVideo: A deep learning library for video understanding," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, https://pytorchvideo.org/.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.