RBE 595 Final Report
Tactile-Vision Integrated Robotic Systems for Safe Manipulation of Fragile Objects

Cooper Ducharme, Farhan Seliya, Sumanth Pashuparthi
Department of Robotics Engineering
Worcester Polytechnic Institute

***Abstract*—This report introduces an end-to-end simulation framework for a tactile-vision integrated robotic system, developed for the safe manipulation of fragile objects. By combining vision-based perception and tactile sensing, the system achieves precise and adaptive manipulation, leveraging the strengths of both modalities. Vision systems provide 3D object data for spatial understanding and optimal grasp point calculation, while tactile sensors deliver real-time feedback on contact forces, surface texture, and slip detection to ensure safe handling. The system employs a dual-camera setup for 3D object reconstruction using point cloud stitching and Euclidean Clustering, alongside a Force-Brute approach to estimate grasp points. A custom tactile sensor plugin, incorporating 16 force-torque sensors in the robotic gripper, enables real-time contact force modeling and feedback. A deep learning-based grasp state assessment model classifies the grasp state as sliding, safe, or excessive force and generates contact force heatmaps to regulate a force-control loop, ensuring damage-free manipulation. Implemented in ROS2 and Gazebo Classic 11, the framework demonstrates capabilities in picking, adjusting grips to prevent micro-slips, and safely placing fragile items. Experimental results validate the system's scalability, robustness, and potential application in healthcare, logistics, and industrial automation, providing a foundation for advanced robotic manipulation.**

I. Introduction

The ability to safely manipulate fragile objects remains a significant challenge in robotics, particularly in applications that demand precision, adaptability and sensitivity to avoid damage, such as healthcare, logistics, and manufacturing. Traditional manipulation methods often struggle with fragile objects, as excessive force can eliminate slippage but also result in irreparable damage. Effective manipulation of such objects necessitates well-defined grasp criteria that consider applied force, contact properties, size, and weight to preserve safety.

While vision systems excel at providing spatial and geometric information critical for object detection, pose estimation, and grasp planning, they often neglect crucial factors such as an object's fragility and texture, potentially resulting in inappropriate grip strength and real-time force feedback necessary for delicate interactions. Fragile objects, due to their material properties and susceptibility to damage, require a careful balance between firm grasping and minimal contact force, which is beyond the capabilities of vision-based systems alone.

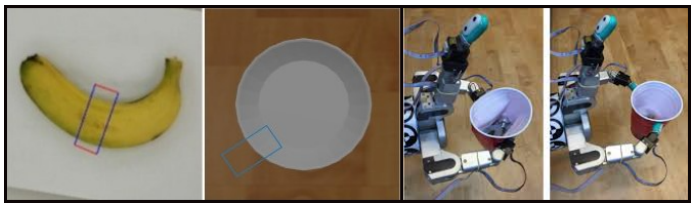

Fig. 1. (a) - GGCNN result for grasp point location (b) - Slip test on a deformable plastic cup

In the figure 1 part (a) , we observe the result of a state-of-the-art GGCNN algorithm [1], which uses a convolutional neural network to generate grasp quality maps from depth images, predicting optimal grasp locations and orientations.The vision-only system treats a banana (deformable) and a bowl (rigid) identically, applying the same force to both, which could damage the banana but not the bowl.In 1 part (b), while the vision system correctly grasps the cup, from a human perspective, it is crushed and unusable. This highlights that a vision-only system is insufficient for grasping fragile objects.

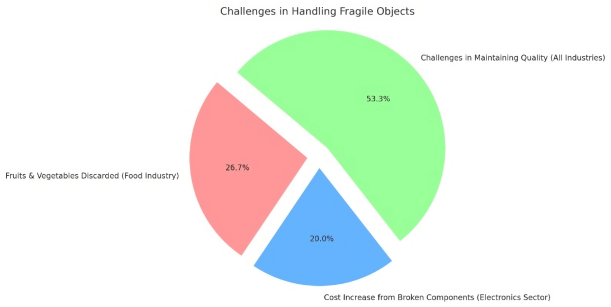

Fig. 2. Survey : Challenges and outlook in robotic manipulation of deformable objects

This outcome is further supported by a recent survey [2] on the challenges of fragile object manipulation, which reveals that the food industry discards 40% of fruits and vegetables due to inadequate manipulation techniques. Similarly, the electronics sector faces a 30% increase in costs from damaged small, fragile components during assembly. Moreover, approximately 2 80% of companies report difficulties in maintaining quality and reducing waste when

handling fragile items, underscoring the need for a robust methodology for fragile object manipulation, particularly in high-speed environments such as assembly or packaging.

To bridge this gap, the integration of vision and tactile sensing has emerged as a promising approach to to adapt to the unique properties of fragile objects and enhance manipulation outcomes. Inspired by recent advances in multi-modal sensor fusion and active perception strategies, our work aims to design a tactile-vision integrated robotic system capable of manipulating fragile objects with both precision and adaptability.

II. Literature Review

Significant amount of work has been published in the context of aiding fragile object manipulation with the combined power of vision an tactile systems. One similar work is the 3D-ViTac [3], which introduces a novel approach to robotic manipulation that combines visual and tactile data into a unified 3D representation. By integrating dense tactile sensors with vision and leveraging diffusion-based imitation learning, the system achieves significant improvements in success rates, occlusion robustness, and generalization to novel objects. However, the approach requires extensive real-world data collection and lacks a tactile simulation framework, limiting its scalability and adaptability to unknown objects. A similar framework was implemented in the paper [4]that combined visual and tactile data using Transformer-based models. By analyzing spatial-temporal embeddings of image sequences, the model predicts safe grasping thresholds, outperforming traditional CNN+LSTM models in terms of slip detection accuracy. However, the model's performance is limited by its sensitivity to noise in tactile images.

The paper 'Self-Attention Based Visual-Tactile Fusion Learning for Predicting Grasp Outcomes' [5] presents a similar framework using self-attention-based fusion model that captures essential cross-modal interactions and focuses on key contact regions. However, the increased model complexity and limited adaptability to unusual shapes are potential limitations. With similar framework in place, 'More Than a Feeling: Learning to Grasp and Regrasp Using Vision and Touch' [6] approached the problem through a novel action-conditioned, visuo-tactile model that learns regrasping policies from raw visual and tactile data through an end-to-end, self-supervised approach.This model dynamically adjusts grip strength and position to achieve stable grasps with minimal regrasp attempts. The model outperforms vision-only approaches, especially on compliant or irregularly shaped objects. However, the model's single-step predictions and reliance on specific hardware limit its adaptability to real-world conditions and continuous control.

When investigating tactile-only systems for our application, we reviewed several research papers. One notable approach from the paper "Enabling Robot Manipulation of Soft and Rigid Objects with Vision-based Tactile Sensors" [7]presents an innovative method using low-cost DIGIT sensors. This system employs a single hyperparameter to detect touch and slip, which allows for adaptive grip force control without needing detailed information about the objects being manipulated. The method works by examining pixel-level changes in tactile images 4 to identify touch and slip events.Despite its innovative approach, the research revealed some limitations. The iterative slip detection process tends to slow down manipulation speed, and the system encounters challenges when interacting with extremely soft objects.

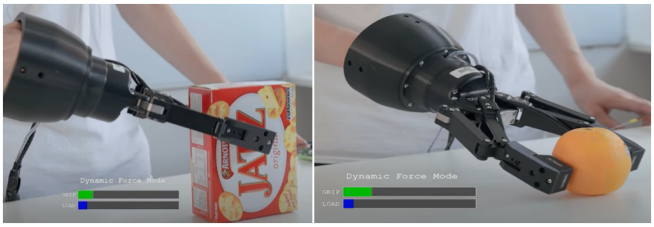

Fig. 3. Adaptive Grip Force Control using Tactile Sensors

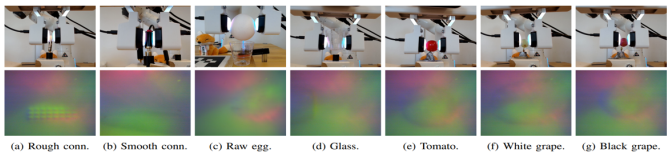

Fig. 4. Vision-based tactile sensing test on a set of soft and fragile objects

In our research on real-time grasp force control, we found a highly relevant study that addresses our primary objectives. The paper [8] presents a novel neural network architecture called C3D-VTFN, which uniquely combines camera and tactile sensor data to categorize robotic grasps into three distinct states: sliding, appropriate, and excessive 5. The method synthesizes information from two sensor streams: 5 consecutive visual frames captured by a wrist-mounted camera and 10 consecutive tactile readings from an XELA sensor. Based on the predicted grasp state, the system can dynamically adjust both grasp width and contact force. Impressively, the model was rigorously trained and validated across 16 different deformable objects and 20,000 grasp sequences, achieving an exceptional 99.97% classification accuracy. This remarkable performance establishes a robust benchmark for our ongoing research in adaptive grasp control.

Implementing a vision-tactile fusion approach for fragile object manipulation within a simulation environment presents significant challenges. Previous research in this domain has predominantly focused on real-world hardware implementations. Our work distinguishes itself as a pioneering effort to develop such a comprehensive framework entirely in a simulated setting. The grasp state assessment module integrated into our pipeline draws inspiration from the methodology mentioned in [8]

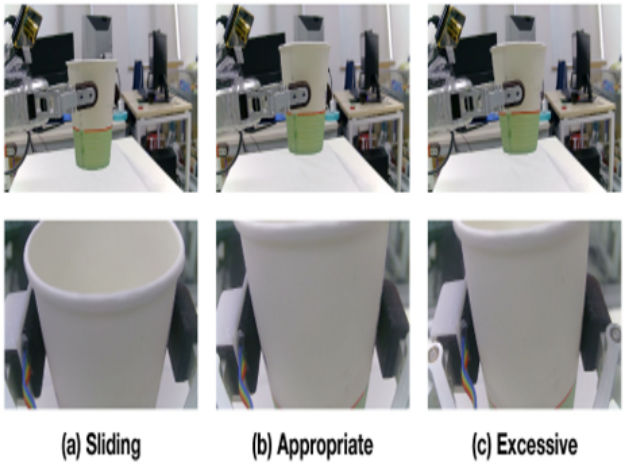

Fig. 5. Grasp State Assessment : Sliding/Safe/Excessive

III. Our Approach

The objective of our project is to design and implement an end-to-end tactile-vision integration framework to facilitate the safe, efficient, and reliable manipulation of fragile objects in high-speed environments, with targeted applications such as egg packaging and depalletizing carbonated drink packages.

We achieve this in three stages 6

- **Optimal Grasp Location Identification:** Implemented an extended active vision system using a dual-camera setup to accurately determine optimal grasp points.
- **Real-Time Force Sensing and Feedback:** Integrated tactile sensors to continuously measure and provide feedback on contact forces during manipulation.
- **Grasp State Evaluation and Force Regulation:** Developed a custom visuo-tactile deep learning model to classify grasp states (sliding, stable, excessive) and ensure safe manipulation through precise force control.

A. Simulation Setup

Our experimental framework was implemented in the Gazebo simulation environment, chosen for its robust physics engine and seamless integration with ROS (Robot Operating System). The simulation environment was designed to replicate real-world manipulation challenges while maintaining computational efficiency.

The primary components of our setup 7 include a Franka Emika Panda robot, positioned adjacent to a standard wooden table surface. The robot model was enhanced with tactile sensors integrated into its gripper fingers, enabling precise force feedback during manipulation tasks. This modification was crucial for achieving the delicate balance required when handling deformable objects.

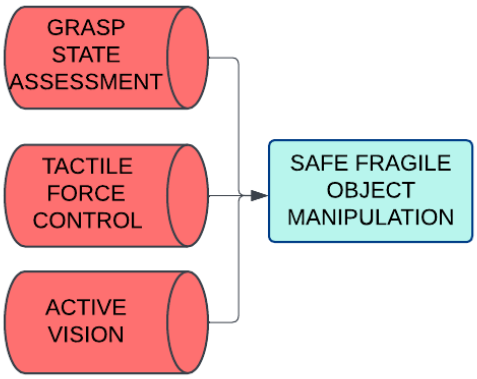

Fig. 6. Project Workflow

The sensing infrastructure consists of two strategically placed cameras providing complementary views of the workspace. The first camera captures a frontal perspective, while the second offers a top-down view, ensuring comprehensive visual coverage of the manipulation area. This dual-camera setup enables robust object detection, pose estimation, and deformation monitoring during manipulation tasks.

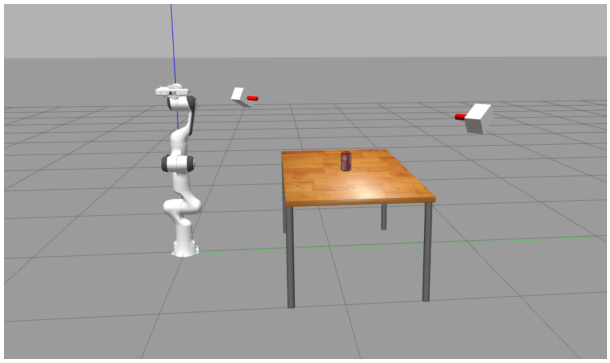

Fig. 7. Environment Setup in Gazebo

For testing our integrated tactile-vision approach, we utilized a deformable coke can as the target object, simulated with appropriate material properties and deformation characteristics. This choice presents a practical challenge common in real-world robotics applications, where objects may be both fragile and deformable, requiring careful handling based on both visual and tactile feedback.

B. Active Vision

Our active vision system employs a dual-camera **??**setup with point cloud processing to achieve robust object perception. The system processes and merges point clouds from two RealSense cameras, implementing a comprehensive pipeline for object detection and scene understanding. This approach enables accurate 3D reconstruction of deformable objects while maintaining real-time performance.

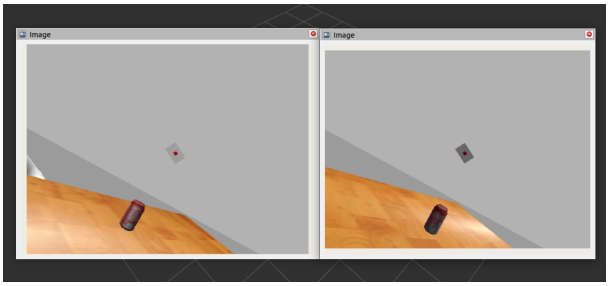

Fig. 8. Both Camera Views

The primary components of our vision pipeline include point cloud transformation, filtering, and segmentation. Each camera's point cloud data undergoes several processing stages:

- Point cloud transformation to world coordinates using TF2 for consistent spatial reference
- Voxel grid downsampling with 1cm leaf size for computational efficiency
- PassThrough filtering to define the region of interest
- RANSAC-based plane segmentation for table surface removal
- Euclidean clustering for object identification

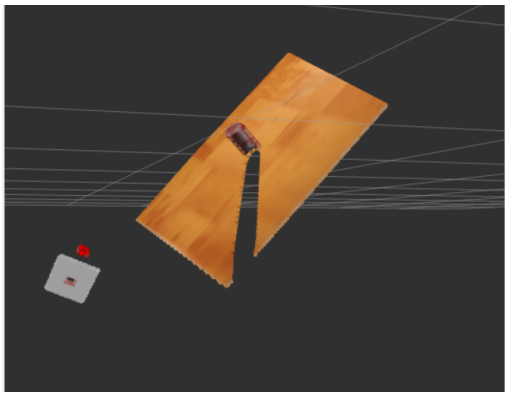

Fig. 9. Unprocessed Raw Pointcloud from Camera1

For point cloud preprocessing 9, we implemented a sophisticated filtering pipeline. The PassThrough filter removes points outside the defined workspace boundaries, significantly reducing computational load. The Voxel Grid filter downsamples the point cloud while preserving the overall geometric structure, with leaf size optimized for our application. RANSAC segmentation effectively isolates the table surface, allowing for reliable object detection.

The two processed point clouds 10 are merged using an optimized alignment algorithm. This merger involves centroid alignment, height offset adjustment, and rotation compensation to create a comprehensive 3D representation of the scene. The system employs statistical outlier removal and precise voxel filtering (2mm) to maintain high-quality object reconstruction while managing computational load.

Our implementation utilizes several advanced features:

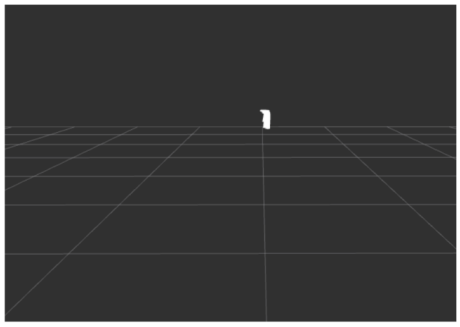

Fig. 10. Processed and world transfromed point cloud from camera1

- Real-time transformation broadcast system for coordinate frame alignment
- Adaptive statistical outlier removal for noise reduction
- Dynamic parameter adjustment based on scene complexity
- Multi-threaded processing for parallel point cloud analysis
- Memory-efficient point cloud storage and manipulation

The system architecture integrates several key components:

- Point Cloud Library (PCL) for efficient 3D data processing
- ROS2 middleware for distributed computing and message passing
- Custom-developed algorithms for point cloud merging and registration
- Optimized data structures for real-time processing

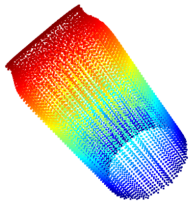

Fig. 11. Merged Point Cloud

For enhanced object detection and reconstruction, we implemented a multi-stage processing pipeline:

1) Initial point cloud acquisition and transformation
2) Noise removal and downsampling
3) Plane segmentation using RANSAC
4) Cluster extraction for object identification
5) Point cloud merging with centroid alignment 11
6) Final filtering and optimization

The camera placement was optimized to maximize coverage while minimizing occlusions. Camera 1 provides a frontal view of the workspace, while Camera 2 offers a complementary perspective, ensuring complete object visualization. This dual-viewpoint approach enables robust 3D reconstruction even in challenging scenarios.

C. Grasp Point Detection Method

The core functionality of grasp point detection is implemented using a brute force search method in the point cloud data. The algorithm systematically searches through pairs of points to find optimal antipodal grasping positions 12based on the following criteria.

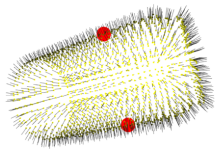

Fig. 12. Grasp Point Genrated on the Merged Point Cloud

The implementation evaluates each pair of points against three key parameters:

$$D = \|p_1 - p_2\| \quad \text{(distance between points)}$$

$$N = \vec{n_1} \cdot \vec{n_2} \quad \text{(normal alignment)}$$

$$C = \sum_{k \in \{x,y,z\}} (|p_{1k} - c_k| + |p_{2k} - c_k|) \quad \text{(center of mass proximity)}$$

For each point pair, these parameters are combined into a final score:

$$score = D + |N + 1| + 10C$$

The points with the lowest score are selected as grasp points, ensuring:

- Appropriate gripper width (D between minimum and maximum thresholds)
- Opposing surface normals (N close to -1)
- Balance near center of mass (minimizing C)

The results are visualized with the grasp points marked in red and connected by a green line, overlaid on the yellow point cloud.

D. Tactile Sensing

Our tactile sensing system represents a novel adaptation of force-torque sensors in the Gazebo Classic 11 simulation environment, addressing the inherent limitations of direct tactile sensing capabilities compared to Gazebo Sim 9 16. The implementation transforms the traditional pendulum-based force-torque sensor into a sophisticated planar array configuration, enabling high-resolution force measurement across the gripper surface.

1) Sensor Array Architecture: The sensing system comprises a 4x4 grid of force-torque sensors 14, each independently calibrated and configured:

- Individual sensor dimensions: 19mm x 19mm x 5mm
- Mass distribution: 0.2kg per element for optimal sensitivity
- Grid spacing: 0.021m between sensor centers
- Total array coverage: 85mm x 85mm sensing area
- Unified cover plate: 1mm thickness for force distribution

2) Physics Engine Configuration: Critical simulation parameters were precisely tuned to ensure stable and accurate force measurements:

- Core Parameters:
  - Step size: 0.0005s for precise contact detection
  - Real-time factor: 1.0 for synchronized simulation
  - Update rate: 50Hz matching robot control frequency
- Solver Configuration:
  - Type: Quick step ODE solver
  - Iterations: 500 for convergence
  - SOR parameter: 1.2 for optimal relaxation
  - Dynamic MOI rescaling: Enabled
- Contact Dynamics:
  - CFM (Constraint Force Mixing): 0.0001
  - ERP (Error Reduction Parameter): 0.1
  - Maximum correction velocity: 50 m/s
  - Contact surface layer: 0.002m

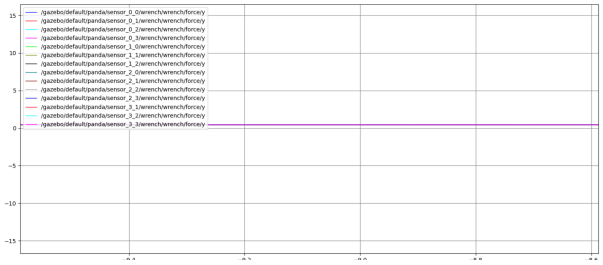

Fig. 13. Stabilised Force readings after changing parameters

3) Custom Build: The custom build integrates our sensor array with the Franka Panda robot's sophisticated hand system 15, requiring careful consideration of mechanical properties and dynamic responses.

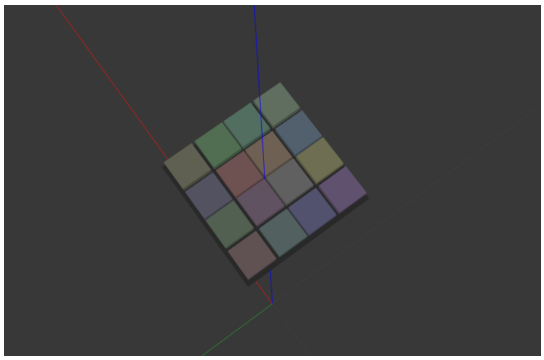

Fig. 14. Custom Built Tactile Sensor

4) Gripper Integration: The Franka hand system presents specific integration requirements:

- Hand Assembly:
  - Total mass: 0.73kg
  - Center of mass: (0, 0.0015244, 0.0275912)m
  - Principal moments of inertia:
    - Ixx: 0.00278560230025 kgm²

    - Iyy: 0.000400033405336 kgm²
    - Izz: 0.00256378041832 kgm²
- Finger Properties:
  - Mass: 0.1kg per finger
  - Center of mass offset: (0, 0.0145644, 0.0227941)m
  - Inertial parameters:
    - Ixx: 3.01220925051e-05 kgm²
    - Iyy: 2.95873808038e-05 kgm²
    - Izz: 6.95125211657e-06 kgm²

5) Mechanical Integration: The sensor array integration considers the robot's kinematic chain:

- End Effector Frame:
  - Position: 0.21m offset from link 7
  - Orientation: RPY (3.14159, -1.57079, -0.785398)
  - Mass: 0.01kg (dummy inertial)
- Hand Mounting:
  - Fixed joint connection
  - Orientation: -0.785398 rad about Z-axis
  - Offset from link8: 0.0584m in Z direction

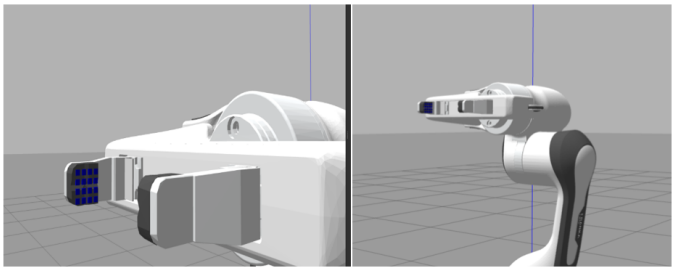

Fig. 15. Sensor Integrated on Franka Panda Arm

Feature	Gazebo Sim 9	Gazebo Classic 11
Tactile Sensor Plugin Available	✅	❌
Force-Torque Sensor Available	✅	✅

Fig. 16. Issue faced with tacile sensor plugin

This comprehensive configuration enables accurate force sensing while maintaining the manipulator's dexterity and control capabilities, providing a robust platform for investigating grasp strategies and manipulation tasks.

E. Heat Map Generation

Our heat map implementation provides real-time visualization of force distribution across the tactile sensor array. The system processes force measurements from the 4x4 sensor grid and generates an intuitive color-coded representation for monitoring grasp conditions.

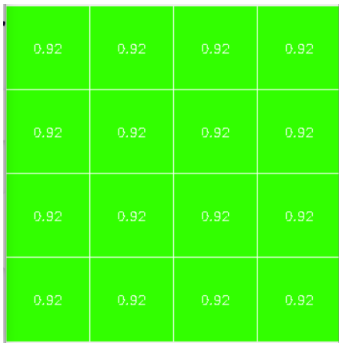

Fig. 17. Heatmap at the time of Grasp(0.75 Normalisation value)

Fig. 18. Heatmap at pre grasp condition(just before applying force to hold) at 0.5 Normalisation Value

1) Force Data Processing: The heat map generation pipeline includes:

- Force array initialization: 4x4 matrix storing force values from individual sensors
- Fixed scale normalization:
  - Minimum force threshold: -1.0N
  - Maximum force threshold: 2.5N
  - Linear normalization to [0,1] range
- Update frequency: 50Hz for real-time visualization

2) Visualization Schema: The heat map 17 18 employs a specialized color mapping inspired by the XELA sensor visualization:

- Color gradient transitions:
  - Dark Blue (0,0,139) → 0.0 force
  - Cyan (0,255,255) → 0.25 force
  - Green (0,255,0) → 0.50 force
  - Yellow (255,255,0) → 0.75 force
  - Red (255,0,0) → 1.0 force
- Display Features:
  - Resolution: 400x400 pixels
  - Cell size: 100x100 pixels per sensor
  - Force value overlay with white grid lines
  - 180-degree text rotation for readability

F. Grasp State Assessment

Our grasp state assessment system implements a deep learning approach to classify grasp conditions based on

temporal force patterns. The system processes sequences of tactile data to make real-time predictions about grasp quality.The force distribution heatmap generated serves as the input to the network.

1) Network Architecture: The assessment system utilizes a C3D (3D Convolutional Neural Network) architecture as described in [8]. The proposed model processes a sequence of visual deformation frames captured using a wrist-mounted camera and a sequence of tactile force reading frames captured using a XELA tactile sensor. This approach 19 achieved a classification accuracy of 89.97%, tested on 16 different deformable objects and 20,000 grasp sequences.
However, directly employing this methodology for our purpose posed significant challenges. The primary limitation was the inability to model deformations in simulation, rendering the vision-based architecture ineffective for our use case. To address this, we redesigned our framework 20 to rely solely on tactile feedback for grasp state assessment.

- Model Configuration:
  - Input sequence length: 10 tactile frames
  - Frame resolution: 4x4 (matching sensor grid)
  - 3 convolution layers and FC layer dimensions: 32 (base) $\rightarrow$ 128 (hidden)
  - Output classes: 3 grasp states - sliding, appropriate and excessive
  - Data Processing:
    - Image Normalization:
      - $\mu = [0.485, 0.456, 0.406]$
      - $\sigma = [0.229, 0.224, 0.225]$
    - Sequence Buffering: Implemented a 10-frame window for temporal analysis.
  - Model Training: Optimized using a batch size of 32 and Adam optimizer with a learning rate of 0.001.

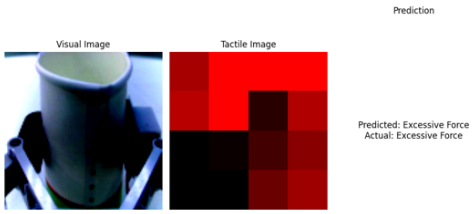

Fig. 19. Baseline Architecture Result (Including Visual inputs)

2) State Classification: The system classifies grasps into distinct categories:

- Classification states:
  - Sufficient: Optimal grasp force
  - Excessive Force: Risk of object damage
- Processing Pipeline:
  - Frame acquisition at sensor update rate
  - Sequence accumulation in 10-frame buffer

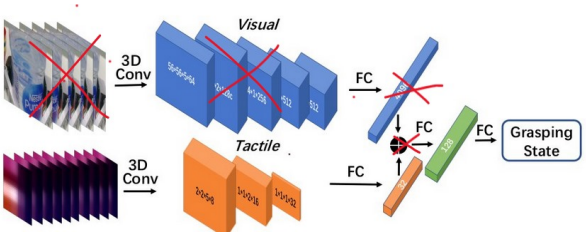

Fig. 20. Revised Model Architecture - No visual pipeline

  - Batch processing when buffer fills
  - State publication on ROS2 topic 'grasp_state'

The integrated system provides both visual feedback through the heat map and analytical assessment via the grasp state classifier, enabling comprehensive grasp quality monitoring during manipulation tasks. The heatmap for safe contact and Excessive force /Damage will be highly different which will be learned by our grasp state assessment model that helps us assess grasp state and predict the appropriate force control strategy.

3) Issues with This Approach: The current approach for tactile data-driven grasping was based on previous data collection using the XELA tactile sensor. The tactile data is gathered through hardware experimentation, ensuring diverse data variations to improve robustness. This method achieves high accuracy during testing, likely due to the quality and variability of the collected data. However, a significant limitation is observed in its application within simulation environments, where real-time accuracy decreases considerably, highlighting the challenges of transferring this methodology from real-world hardware to simulated scenarios.

4) Solution Incorporated: A feasible approach for tactile data-driven grasping in our case, starting with data collection using the custom built tactile sensor. Unlike the hardware-based method, data is collected in a simulated environment **??**, leading to lower data variation. Despite this limitation, the method achieves mid-to-high accuracy levels. The key advantage of this approach lies in its high real-time accuracy in simulation, making it well-suited for simulated environments and potentially more adaptable to virtual testing scenarios.

G. Force Control Strategy

In our application, once the grasp state assessment readings are obtained, a simple force control strategy is deployed **??** based on the predicted grasp state. If the predicted state is "Excessive," the grasp width is gradually increased by 0.1 cm, starting from an initial width of 0.04 m. This adjustment continues until the model predicts the state as "Sufficient." To ensure stability, the model waits for ten consecutive timestamps of "Sufficient" predictions before activating the Franka arm's inbuilt attach function

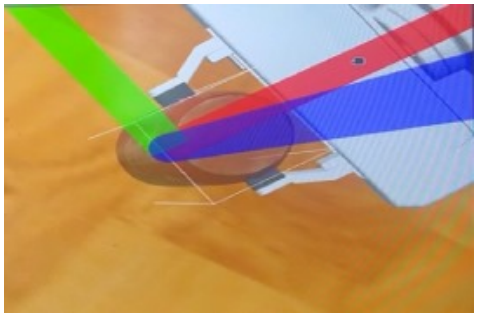

Fig. 21. Sample Data collection in Simulation

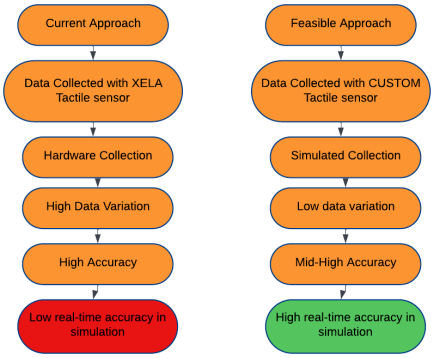

Fig. 22. Challenge Vs Solution (Real-time Grasp State Assessment)

to securely grasp and lift the object. This process is visually represented through a heat map: red indicates an excessive force state, yellow signifies borderline damage, and green represents a safe condition for manipulation. This approach has been trained and tested specifically for our object of interest—a deformable coke can. The positive results from this experiment suggest that the method can be effectively extended to other deformable objects with a high success rate.

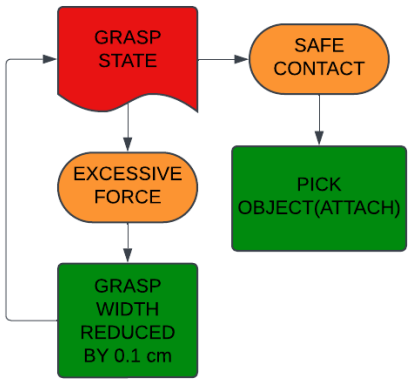

Fig. 23. Force Control Strategy

IV. Experimental Result and Evaluation

We validated our framework in simulation, successfully regulating the grasp width and force to securely pick the object, as indicated by the heat map transitioning to green 24 as mentioned earlier.

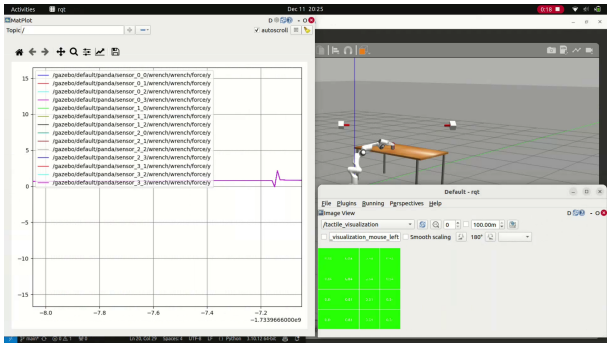

Fig. 24. Working Result

This observation is further supported by 25 . In the simulation without force feedback, we observed the grasp width continuously decreasing in a linear fashion, as anticipated. However, when force control was applied, the grasp width stabilized after reaching a certain point, corresponding to when the grasp state was predicted to be safe. Correspondingly, stabilizing force readings 26.

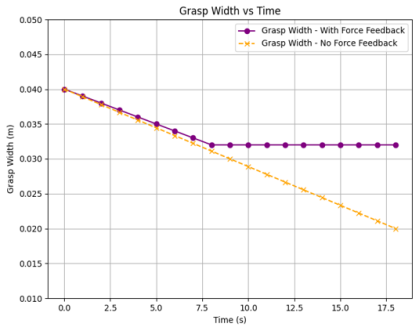

Fig. 25. Grasp width vs Time

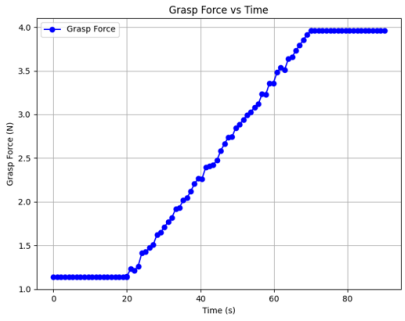

Fig. 26. Grasp Force Vs Time

V. Future Scope

The next phase of the project will involve extending the current framework to real hardware for practical testing, allowing us to evaluate its performance under real-world conditions. This will include integrating the tactile-vision system with robotic arms, such as the Franka Emika, to assess its real-time grasping capabilities. Additionally, we will incorporate a visual deformation module to enhance the accuracy of grasp state assessments, combining both tactile and visual feedback for more precise predictions. To further improve system performance, we will be developing a more robust force control feedback mechanism, ensuring that the applied force is continuously adjusted based on real-time tactile data. Efforts will also focus on improving the system's robustness and adaptability in dynamic, real-world environments, where variations in object properties and environmental conditions can influence performance. This will involve refining the force control strategy and enabling the system to handle a wider variety of objects, with the ultimate goal of developing a generalized solution for safe and efficient manipulation of fragile objects.

VI. Conclusion

In conclusion, we successfully developed an end-to-end visual-tactile fusion pipeline for the safe manipulation of fragile objects. The proof of concept model leveraged the power of active vision through a dual-camera setup, a custom-built tactile sensor plugin for real-time force feedback, and a bespoke grasp state assessment module to ensure safe grasping of fragile objects. While the proof of concept demonstrated significant potential, the accuracy of real-time simulation was limited by insufficient experimental data. Future work will focus on expanding the framework to real hardware, improving force control feedback, and enhancing the system's robustness and adaptability for real-world applications.

VII. Acknowledgments

We would like to express our sincere gratitude to Prof. Berk Calli for his invaluable guidance, insights, and support throughout the course of this project. His mentorship played a pivotal role in shaping our understanding and execution of this work. We are grateful for his constructive feedback and assistance, which greatly enhanced the quality of our project. Lastly, we extend our heartfelt thanks to our course colleagues for their encouragement and helpful discussions, which significantly contributed to the success of this project.

References

[1] Douglas Morrison and Peter Corke and Jürgen Leitner. "Closing the Loop for Robotic Grasping: A Real-time, Generative Grasp Synthesis Approach." *arXiv preprint* arXiv:1804.05172, 2018. Available at: https://doi.org/10.48550/arXiv.1804.05172.

[2] Jihong Zhu and Andrea Cherubini and Claire Duneand David Navarro-Alarcon and Farshid Alambeigi and Dmitry Berenson and Fanny Ficuciello and Kensuke Harada and Jens Kober and Xiang Li and Jia Pan and Wenzhen Yuan and Michael Gienger. "Challenges and Outlook in Robotic Manipulation of Deformable Objects." *arXiv preprint* arXiv:2105.01767, 2021. Available at: https://doi.org/10.48550/arXiv.2105.01767.

[3] Binghao Huang and Yixuan Wang and Xinyi Yang and Yiyue Luo and Yunzhu Li. "3D-ViTac: Learning Fine-Grained Manipulation with Visuo-Tactile Sensing." *arXiv preprint* arXiv:2410.24091, 2024. Available at: https://doi.org/10.48550/arXiv.2410.24091.

[4] Yunhai Han and Kelin Yu and Rahul Batra and Nathan Boyd and Chaitanya Mehta and Tuo Zhao and Yu She and Seth Hutchinson and Ye Zhao. "Learning Generalizable Vision-Tactile Robotic Grasping Strategy for Deformable Objects via Transformer." *arXiv preprint* arXiv:2112.06374, 2021. Available at: https://doi.org/10.48550/arXiv.2112.06374.

[5] Cui Shaowei and Rui Wang and Junhang Wei. "Self-Attention Based Visual-Tactile Fusion Learning for Predicting Grasp Outcomes." *IEEE Robotics and Automation Letters*, vol. PP, no. 99, pp. 1-1, 2020. DOI: https://doi.org/10.1109/LRA.2020.3010720.

[6] Roberto Calandra and Andrew Owens and Dinesh Jayaraman and Justin Lin and Wenzhen Yuan and Jitendra Malik and Edward H. Adelson and Sergey Levine. "More Than a Feeling: Learning to Grasp and Regrasp using Vision and Touch." *IEEE Robotics and Automation Letters (RAL)*, vol. PP, no. 99, pp. 1-1, 2018. DOI: https://doi.org/10.48550/arXiv.1805.11085.

[7] Michael C. Welle and Martina Lippi and Haofei Lu and Jens Lundell and Andrea Gasparri and Danica Kragic. "Enabling Robot Manipulation of Soft and Rigid Objects with Vision-based Tactile Sensors." *IEEE International Conference on Automation Science and Engineering (CASE2023)*, 2023. DOI: https://doi.org/10.48550/arXiv.2306.05791.

[8] Cui Shaowei and Rui Wang and Junhang Wei. "Grasp State Assessment of Deformable Objects Using Visual-Tactile Fusion Perception." *arXiv preprint* arXiv:2006.12729, 2020. DOI: https://doi.org/10.48550/arXiv.2006.12729.

[9] Roberto Calandra and Andrew Owens and Manu Upadhyaya and Wenzhen Yuan and Justin Lin and Edward H. Adelson and Sergey Levine. "The Feeling of Success: Does Touch Sensing Help Predict Grasp Outcomes?" *arXiv preprint* arXiv:1710.05512, 2017. DOI: https://doi.org/10.48550/arXiv.1710.05512.

[10] Zhuangzhuang Zhang and Zhinan Zhang and Lihui Wang. "Digital twin-enabled grasp outcomes assessment for unknown objects using visual-tactile fusion perception." *Robotics and Computer-Integrated Manufacturing* 84(102601):1-17, December 2023. DOI: https://doi.org/10.1016/j.rcim.2023.102601.

[11] Di Guo and Fuchun Sun and Bin Fang and Chao Yang. "Robotic grasping using visual and tactile sensing." *Information Sciences* 417, July 2017. DOI: https://doi.org/10.1016/j.ins.2017.07.017.

[12] Ankit Goyal and Jie Xu and Yijie Guo and Valts Blukis and Yu-Wei Chao and Dieter Fox. "RVT: Robotic View Transformer for 3D Object Manipulation." *arXiv preprint* arXiv:2306.14896, 2023. Available at: https://doi.org/10.48550/arXiv.2306.14896.

[13] Sulabh Kumra and Shirin Joshi and Ferat Sahin. "Antipodal Robotic Grasping using Generative Residual Convolutional Neural Network." *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020. arXiv preprint arXiv:1909.04810v4. Available at: https://doi.org/10.48550/arXiv.1909.04810.

[14] Sabhari Natarajan and Galen Brown and Berk Calli. "Aiding Grasp Synthesis for Novel Objects Using Heuristic-Based and Data-Driven Active Vision Methods." *Frontiers in Robotics and AI*, vol. 8, pp. 696587, July 2021. DOI: https://doi.org/10.3389/frobt.2021.696587.

[15] Yongkyu Lee and Shivam Kumar Panda and Wei Wang. "Measure Anything: Real-time, Multi-stage Vision-based Dimensional Measurement using Segment Anything." *arXiv*, December 2024. DOI: https://doi.org/10.48550/arXiv.2412.03472.

[16] Hu Cao and Guang Chen and Zhijun Li and Jianjie Lin and Alois Knoll. "Lightweight Convolutional Neural Network with Gaussian-based Grasping Representation for Robotic Grasping Detection." *arXiv*, January 2021. DOI: https://doi.org/10.48550/arXiv.2101.10226.