

Real-Time Grocery Packing by Integrating Vision, Tactile Sensing, and Soft Fingers

Valerie K. Chen*, Lillian Chin*, Jeana Choi*, Annan Zhang*, Daniela Rus

Abstract—Although bin packing has been a key benchmark task for robotic manipulation, the community has mainly focused on the placement of rigid rectilinear objects within the container. We address this by presenting a soft robotic hand that combines vision, motor-based proprioception, and soft tactile sensors to identify, sort, and pack a stream of unknown objects. This multimodal sensing approach enables our soft robotic manipulator to estimate an object’s size and stiffness, allowing us to translate the ill-defined human conception of a “well-packed container” into attainable metrics. We demonstrate the effectiveness of this soft robotic system through a realistic grocery packing scenario, where objects of arbitrary shape, size, and stiffness move down a conveyor belt and must be placed intelligently to avoid crushing delicate objects. Combining tactile and proprioceptive feedback with external vision resulted in a significant reduction in item-damaging packing maneuvers compared to a sensorless baseline (9× fewer) and vision-only (4.5× fewer) techniques, successfully demonstrating how the integration of multiple sensing modalities within a soft robotic system can address complex manipulation applications.

I. INTRODUCTION

Picking items from clutter and placing them into ordered bins has been an important benchmark for the broader robotic manipulation community, as exemplified by the Amazon Picking/Robotics Challenge [1]. Current solutions have largely focused on vision-based segmentation for rigid grippers grasping rigid rectilinear objects [2]–[6]. While effective, these approaches require significant pre-computation, limiting their utility for “online applications,” where input is processed serially rather than upfront. Online applications describe many realistic packing scenarios such as loading a dishwasher or packing for a move. The order in which objects arrive and their material properties are unknown and must be dynamically determined.

Soft grippers offer a potential solution for online bin packing as their compliance makes them robust to changes in objects’ stiffness, shapes, and placement. This enables them to grasp objects with arbitrary material properties without the models or precise location information that their rigid counterparts would require [7]–[9]. However, sensorizing soft robotic grippers has remained challenging, especially when multiple sensor modalities are needed. A soft gripper’s deformability makes it difficult to accurately place tactile sensors and localize forces spatially along the gripper [10], [11]. Despite current efforts to create proprioceptive soft

*These authors contributed equally to this work and are listed in alphabetical order. All authors are with the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology, 32 Vassar St, Cambridge, MA 02139, USA. LC is also with the University of Texas at Austin, 2515 Speedway, Austin, TX 78712, USA.

Correspondence: zhang@csail.mit.edu

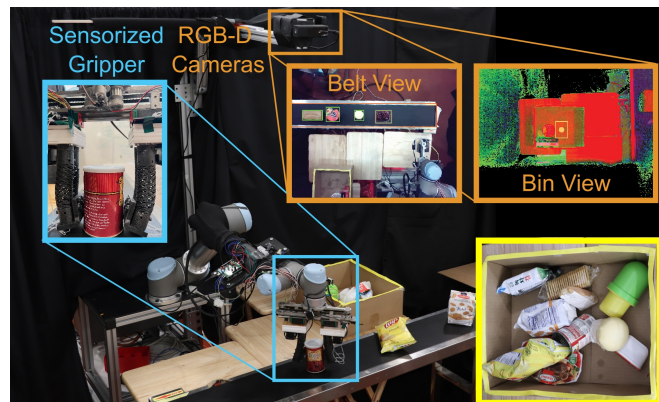


Fig. 1. Soft robotic manipulation for grocery packing system setup. Our system combines RGB-D cameras, closed-loop control servo motors, and custom soft tactile sensors to determine an optimal packing order for unknown grocery items in a safe, dynamic and online fashion.

grippers through using vision for tactile and force sensing [12]–[14] or incorporating rigid sensing elements within soft systems [15], [16], there has been relatively little exploration into applying soft robots that use contemporary sensor fusion techniques to complex online applications.

We address this gap by creating an end-to-end online bin packing system that uses multimodal sensing to grasp unknown objects and safely pack them (Fig. 1). Our system combines RGB-D cameras with soft pressure-based tactile sensors to provide the sensory feedback needed to make appropriate packing decisions for a soft gripper. This gripper, previously introduced in [17], combines parallel grasping and soft-finger grasping in a single servomotor-driven package with proprioceptive grasping feedback. These different sensing modalities complement one another to ensure an accurate and timely understanding of the properties of an object’s material, combining the global scale of vision with the localized scale of tactile sensing.

We demonstrate the power of this soft robotic system by comparing its performance against sensorless and vision-only systems in a grocery packing scenario. Grocery packing is a strong case study for online packing as groceries vary widely in shape and weight, and these characteristics may not be fully captured by pre-built models [18]. To pack groceries well, a robotic system must be able to handle delicate objects carefully and ensure that groceries at the bottom of the bin are not crushed by groceries packed above them. Initial work on food handling with a soft multimodal approach has shown success [19] but still required significant precomputation (56 hours to detect 5 classes of objects).

In our system, we write an online algorithm for bin packing unknown objects. The robot is able to detect the size and stiffness of an object in real time and determine a placement sequence that avoids crushing objects. By integrating vision and tactile sensing within a single soft gripper system, our end-to-end solution performs $9\times$ fewer item-damaging maneuvers than the sensorless baseline and $4.5\times$ fewer than using vision alone.

In summary, we make the following contributions:

- 1) an integrated physical soft grasping platform that merges vision, motor-based proprioception and pressure-based tactile sensing in a soft grasping system.
- 2) an online packing algorithm that takes in multiple sensor inputs to create a “well-packed” container that matches human expectations.
- 3) physical experiments with our multimodal approach and comparisons against traditional blind and vision packing methods in a realistic grocery packing scenario with irregular objects.

II. RELATED WORK

Current research on packing has focused on minimizing unoccupied volume or runtime for a given number of rigid objects [6], [20]. These works often rely on knowing the packed objects’ and bin’s geometries beforehand, with many requiring significant offboard pre-processing [5], [21], [22]. Indeed, current solvers may reach intractable run times when instructed to pack as few as six objects [23]. For online applications, where objects are not known beforehand and may be deformable or fragile, these methods are insufficient.

Sensor fusion may provide an effective way to achieve online packing. In particular, visual and tactile sensors provide complementary ranges of data, focusing on global and localized scales respectively. When these different modalities are combined, significant manipulation milestones can be accomplished, such as improving grasp reliability [24]–[26], task accuracy [27], [28], and scene/object understanding [29]. With few exceptions [30], the majority of previous work on sensor fusion for robot manipulation relies upon machine learning for at least one portion of the pipeline [25], [26], [31]. While powerful for learning various grasping policies, these methods require detailed prior knowledge of the items, such as CAD models [32], extensive dataset building [25] or built learned representations [26], which again makes these methods ill-suited for online applications.

Although the sensorization of soft grippers is an active area of research, there has been relatively little overlap with contemporary sensor fusion techniques. For example, the multimodal approach in [19] solely uses its tactile sensors to determine grasp pose rather than combine the tactile information with visual information. One major challenge is the soft gripper’s deformability, making it difficult to accurately place tactile sensors and localize forces spatially along the gripper. Significant focus has thus been placed on obtaining accurate proprioception as an intermediate step before more integrated sensor fusion [10]. The most popular

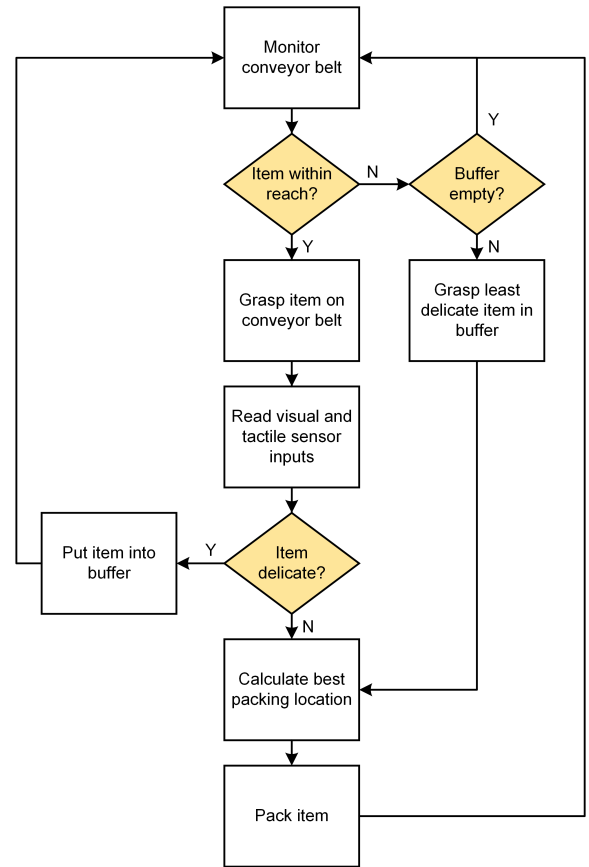


Fig. 2. A flowchart of our grocery packing algorithm with multimodal sensing.

combination of sensor modalities is in the use of vision for tactile sensing, where the high resolution of a camera or time-of-flight sensor is used to track the deformation of a soft surface to get tactile information [12], [13]. Others incorporate rigid elements to provide proprioception within their soft structure, occasionally supplementing this with further tactile sensors [15], [16]. We build on this approach and our previous work [33] by choosing a strategy where we incorporate proprioceptive feedback from the rigid servo motors that drive our soft gripper with the soft tactile sensors and external vision system for our multimodal approach.

III. PROBLEM STATEMENT AND APPROACH

In this project, we aim to build a robot system that can pack a bin in an “online” fashion, where objects are unknown and must be processed one at a time rather than all at once. The goal is for this robotic system to avoid placing “delicate” objects under non-delicate objects. In other words, we seek to avoid crushing objects under the weight of future objects. This is consistent with the qualitative human intuition for what “good” bin packing would look like. We define four areas to our system:

- 1) the robot, which has a gripper that can grasp one object at a time;



Fig. 3. Calibration set. These 25 items were used to characterize the vision system and determine the priority packing heuristic. The objects that are considered “delicate” are the apple, clementine, grapes, mozzarella, peach, pear, and the pound cake.

- 2) the conveyor belt, where objects are delivered to the robot sequentially;
- 3) the bin, where objects are to be packed;
- 4) a buffer zone, where objects may be placed temporarily before being packed.

Our approach to packing is summarized in Fig. 2. Briefly, the conveyor belt is monitored for new objects. If an object is detected on the belt, it is grasped and analyzed by the robot’s internal and external sensors. If this item is not delicate, it will be packed directly in the bin. If the item is delicate, it will be placed in a buffer in the hopes of being packed after less delicate objects coming down the line. When there are no other objects to pack on the belt, the robot will then pack things from the buffer into the bin.

The following sections will fill in the details of this framework. Sec. IV will characterize the hardware components that provide the sensing and grasping information, Sec. V will describe how these components are integrated algorithmically, and Sec. VI will give the experimental results of how our packing approach fares in a grocery packing scenario.

IV. HARDWARE ARCHITECTURE AND CHARACTERIZATION

In our described approach, we require a system that can (a) visibly locate objects, (b) grasp these objects, and (c) gain tactile information about the objects. We achieve these goals through three major hardware components: (1) two external RGB-D cameras, (2) a previously introduced soft gripper [17] with added proprioceptive motor feedback, and (3) new pressure-based tactile sensors attached to the fingers of this gripper. These components are all integrated on a robot arm in a packing scenario for evaluation.

A. External Vision

To sense visually, our system uses two RGB-D cameras: an ASUS Xtion 2 to detect the locations and sizes of objects on the conveyor belt and an ASUS Xtion PRO LIVE to determine the best packing location in the bin.

To locate the objects on the conveyor belt, we perform color segmentation via OpenCV to threshold out the conveyor belt’s uniform black color. Once the item has been

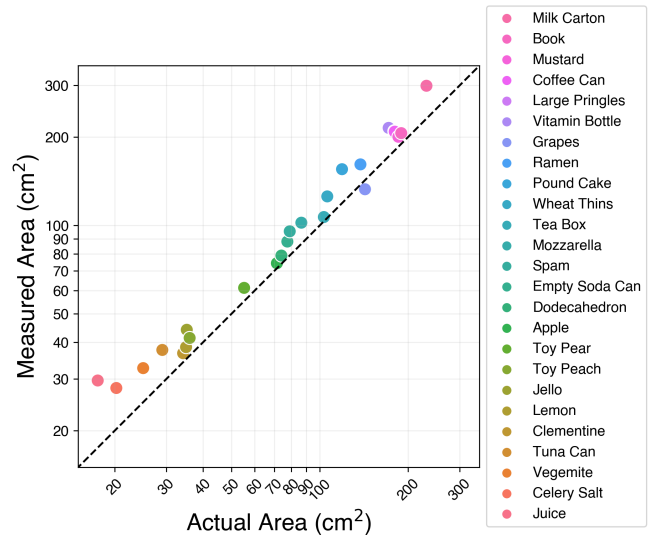


Fig. 4. Log-log plot comparing the measured area of an object viewed from the top against the vision system’s calculation of the object’s area.

segmented, we fit a bounding box to the object’s contours to estimate its size. Given the average time to plan and execute robot trajectories from the neutral position to the conveyor belt and the speed of the conveyor belt, we determine a meeting waypoint for object and robot, as well as the time the object will reach that point. We then use MoveIt! with RRT-Connect for motion planning of the robot arm to move and grasp the object at that point.

To evaluate how accurate the external vision system is in its size estimation, we compare the camera’s area measurement of the 25 items in Fig. 3 against a ground truth manual measurement of the object. Specifically, we use the camera to compute the bounding box area of each object. We see a good match between the actual and estimated size of each test object, with a tight correlation ($r = 0.987$) to the line $y = x$ (Fig. 4).

Once an object is grasped, the vision system then identifies a favorable packing location. First, the vision system locates the packing box in the RGB image via color segmentation. The mask of the bin is then applied to the registered depth image, leaving only the region of packing interest. Object dimensions recorded previously during the detection and grasping tasks are translated into pixel coordinates. A kernel of ones with dimensions of these translated length and width values is convolved with the depth image of the bin to create a heatmap of packing locations. We perform this operation twice: once with the kernel reflecting the current object orientation, and once with the kernel rotated 90 degrees. The final packing orientation is determined by the highest score of the two heatmaps. Although this approach is not the most optimal, it bypasses the computational intractability and training requirements found in contemporary algorithms [23], [34], allowing us to perform online packing without significant pre-computation.

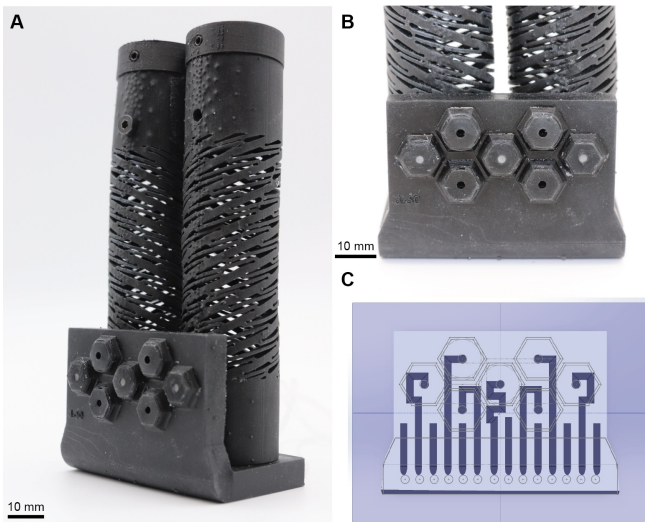


Fig. 5. (A) Soft finger assembly with hexagonal tactile sensors. (B) Close up on the tactile sensor array. (C) Cross sectional render of the tactile sensors.

B. Proprioceptive Gripper

To grasp objects with proprioceptive feedback, we upgrade the soft gripper introduced in [17] with new servomotors for closed loop feedback. To summarize briefly, the original gripper offered “multiplexed” manipulation, enabling us to grasp objects using parallel jaw grasps, soft finger grasps or a combination of the two. This combination allows us to pick up objects that could not be picked up by single grasping modes alone. This is important in grocery packing applications, where objects may not be of uniform shape.

For this work, we upgraded the servo motors from HiTec HS-5585MH servos to Dynamixel MX-28T servos. These Dynamixel servos provide feedback on the servo’s position, speed, and estimated load, allowing us to detect when an object is grasped and automatically combine grasping modes. We also print the soft fingers out of a flexible polyurethane (FPU 50, Carbon Inc.) rather than laser cutting them [35].

For each grasp, the servo drives the parallel jaws closed with a constant velocity. This servo will halt its motion when either the maximum travel distance is reached or a sharp spike in the estimated load is measured, indicating object contact. Next, the servos driving the soft fingers will move inwards, creating an enclosing grasp around the object. This combination of parallel jaw and soft grasping creates a stronger and more conforming grasp to the object.

To release an object, the jaw servo opens outwards with a constant velocity until it measures zero estimated load, indicating no more contact with the object. This controlled release is necessary to ensure that the gripper does not open too wide when releasing objects within the packing bin and bump the wide fingers against the bin’s edge.

C. Tactile Sensing

To provide tactile information to our grasps, we use our previous work on printing internal air-based sensors [36] to manufacture arrays of tactile sensors. In our prior work, we

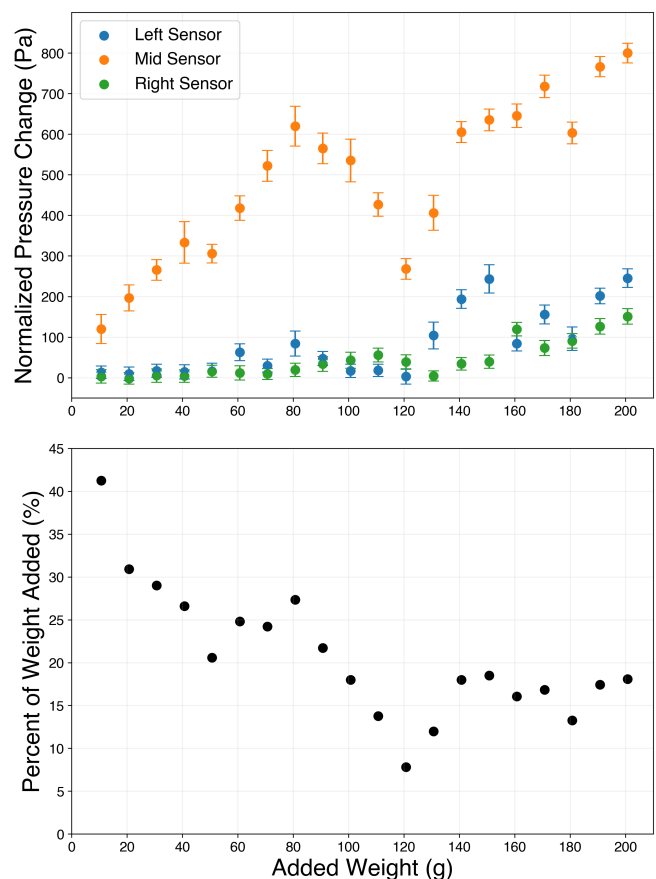


Fig. 6. (Top) Sensor response of the leftmost, center and rightmost tactile sensors for weights ranging from 10 g to 200 g in 10 g increments. Error bars represent standard deviation across three trials. (Bottom) The total pressure measured by the sensors as a percentage of the expected amount of pressure from the given amount of weight placed on top of the sensors.

printed air channels within a structure. When the structure was pressed or bent, we could measure the resulting change in pressure of the internal air channel to get a sense of the contact forces on the surface. While that previous paper demonstrated the potential of that technique, no real-world applications with contact were explored.

We apply the sensorization technique to create external tactile sensors for measuring grasping forces (Fig. 5A,B). Specifically, we print hollow, thin-walled hexagonal prisms out of an elastomeric polyurethane (EPU 40, Carbon Inc.). Each hexagon serves as a single “bubble” sensor that will react to a grasped object. These prisms rest on a thick panel that also serves as the end cap for the gripper’s soft fingers. Empty fluidic channels run from the inner cavity of each sensor to an outer edge for measurement access. These channels were designed to have equal lengths, so each channel should have a similar response to a given force (Fig. 5C). Excess resin is removed from the printed part by aspirating with vacuum to create open channels. The resin exit holes are then sealed with Gorilla Super Glue Gel. Silicone tubing is used to connect the closed volumes to differential pressure transducers (HSCDRRN160MDAA5, Honeywell) which are read through a 24-bit analog digital

converter (ARD-LTC2499, Iowa Scaled Engineering).

To characterize the soft tactile sensors’ accuracy in measuring grasping forces, we conducted a series of experiments where weights were placed on top of the hexagonal sensors as they rested on a table. It is difficult to measure grasping forces *in situ*, so this proxy measurement was chosen instead. The weights placed ranged from 10 g to 200 g in 10 g increments. This overlaps with the range of previously measured gripper forces from [17], which reported grasping force estimates of 0.75 N to 2 N. Three trials were conducted for each weight class.

To equally distribute the weights’ effect across the sensors, a sheet of cardboard (0.7 g) was placed over all 7 hexagonal bubbles. A measurement was taken with no weights on the cardboard to use as a normalization across all other data points. Due to space constraints to mount the electronics to the gripper, the sensor performance could only be recorded for 3 hexagonal bubbles. We thus chose to measure the leftmost hexagon, the center hexagon, and the rightmost hexagon (Fig. 6-top).

To compensate for this incomplete measurement, we compared the measured pressure change in the channels against the total weight placed on the sensor channels (Fig. 6-bottom). For example, in an ideal world, if 10 g were placed across all 7 sensors, each of which has a surface area of 42.4 mm², the sensors would report $\frac{1}{7} \frac{0.01 \text{ kg} \cdot 9.81 \text{ m/s}^2}{42.4 \cdot 10^{-6} \text{ m}^2} = 331 \text{ Pa}$ in total. Again, in an ideal world, this pressure would be distributed equally across the 3 sensors being measured, so the 3 sensors would, in total, measure $\frac{3}{7} = 43\%$ of the total pressure. In real life, however, the entire sensor is built out of an elastomer and not all of the force of the weight will be applied directly to the sensors. We notably see a large bias towards the middle sensor as that is where most of the weight was loaded. Across our 3 measured sensors, we measure a total of 136 Pa for a weight of 10 g. This is about 41% of the expected total pressure, which is very close to the expected 43%. We report the equivalent percentage for all of the weight classes.

Subsequent measurements do not match nearly as well as the original 10 gram measurement. This implies that as more weight is loaded onto the hexagons, the entire elastomeric structure itself takes on the weight rather than the individual hexagons. Thus, unlike the 43% we expected for the 3 hexagons to carry, we see that they carry approximately 20 – 25% of the weight. The higher the weight is, the more focused the compression is of the overall structure, leading to the large dip from 100-120 g. This dip corresponds to when we switched from having multiple smaller weights to one large 100 g weight which had a more focused compression effect. More experiments are needed to tease out a tighter relationship between weight positioning and measured response.

Nevertheless, despite this variation, we do see that there is an overall trend for larger weights leading to a larger response in the pressure sensors. The sensors are able to track very slight 10 g changes, provided that the force does not overly compress the entire elastomeric structure. Even

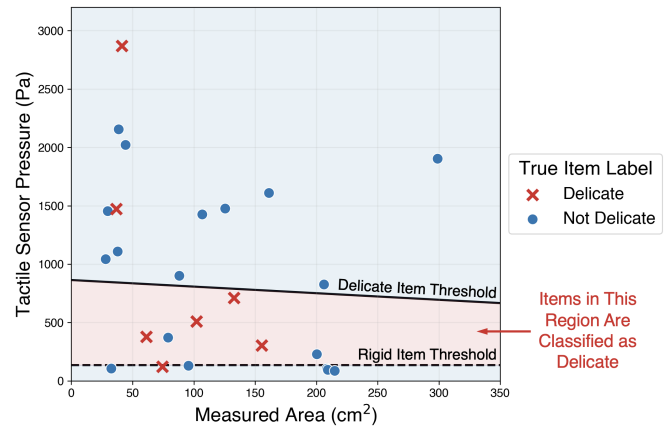


Fig. 7. Using the calibration set of grocery items, we determine two thresholds for classifying whether an item should be packed immediately or not. This graph defines a region of interest for delicate objects, enabling us to create a packing priority score metric.

on sensors that are not directly compressed, heavy weights can be detected and differentiated from lighter weights. This indicates that the tactile sensors provide information about the grasp condition and can be used in more complex grasping operations.

V. PACKING ALGORITHM

Now that we have detailed the individual components in our system architecture, we next describe how they are all integrated algorithmically into a single program. The previous section expanded on the outline from Fig. 2, discussing how the vision system detects objects on the belt, how the modified proprioceptive gripper can pick up objects securely, and how the tactile sensors can read grasp information. In this section, we discuss how the packing system determines if an object is “delicate,” which is the key decision point that determines whether we place an object into the buffer or into the bin.

To evaluate this, we introduce the core metric of the *packing priority score*. This score determines how “delicate” an object is as a combination of the vision and tactile readings. An object’s fragility is extremely qualitative; however, it is one of the main metrics that a human uses to decide on how to pack objects [37], [38]. We attempt to make this metric more quantitative by proposing that an object’s size and stiffness can be leveraged to measure fragility. This is a similar methodology to our previous work on sorting objects for recycling [39]. Size can be determined through our system’s vision sensors, while stiffness can be determined by our system’s tactile sensors. Once we have a scoring equation, we can use this score to determine whether non-delicate objects should be packed in the bin directly or kept in the buffer area until no items are found on the conveyor belt. This then allows the system to pack the bin with buffer items in descending order by priority score.

Using the item set shown in Fig. 3, we perform three grasps on each of the 25 objects. For each grasp trial, we calculate the average tactile output across all six sensors

(three sensors for each finger) and plot them against the measured area of each object (Fig. 7). We manually add the labels for “delicate” vs. “non-delicate” objects to this graph to see if we can classify delicate vs. non-delicate objects.

From this plot, we make two major observations. First, we see a general trend that delicate objects tend to apply lower amounts of force to the tactile sensors. We believe that this is due to their compliance preventing a larger force resistance against the gripper, which is consistent with the lower grasping force seen in the soft grasping mode in [17]. This leads to a decision boundary to separate most delicate items from non-delicate items. Specifically, the line has intercept 864 Pa and slope -0.564 Pa/cm^2 , as shown by the solid line in Fig. 7.

Second, we notice that some items fail to yield significant tactile sensor readings, staying close to the x -axis. These items are mostly rigid, non-delicate items. Unlike compliant items which deform to adapt their shape to the gripper and increase the contact patch, rigid items do not yield, instead maintaining minimal contact with the gripper. To separate these items from true delicate items, we introduce a threshold for rigid items at 134 Pa, as shown by the dashed line in Fig. 7.

With these two constraints, we now can evaluate the packing priority score. When conducting the packing algorithm, we measure the size of the object A using the bounding box technique described in Sec. IV-A and the grasping pressure P_g on the object using the tactile sensors described in Sec. IV-C. The priority score p thus becomes a weighted version of these area and pressure readings. Based on our thresholds, we determine $P_g < 864 - 0.564A$ and $P_g > 134$. Thus, the equation to determine the priority score is $p = 0.564A + P_g > 864$. In words, items with a priority score of less than 864 are deemed delicate, while items with a priority score of 864 or greater are deemed not delicate. With this, we have successfully achieved a closed-form formula we can use in online evaluations to determine whether an object is “delicate.”

VI. GROCERY PACKING EXPERIMENT

Finally, we describe how we adapt our general packing algorithm to a grocery packing case study. Our experimental setup consists of a UR5 robot arm outfitted with the proprioceptive gripper described in Sec. IV-B. In front of the robot is a conveyor belt running at constant speed (0.1 m/s) on which items are manually loaded. Three small tables are supplied adjacent to the robot to provide the buffer. To the side of the robot is a cardboard box, which serves as the bin. For ease of detection, this bin has colored markers along its boundary.

To perform online bin packing, the vision system monitors for objects on the conveyor belt. Once the vision system sees an object, the robot arm grasps the object using a fixed grasp pose perpendicular to the belt movement. The system then evaluates the grasped object’s priority score by measuring its area and averaging tactile sensor readings over a one second period. If the priority score is outside of the delicate object

threshold, the robot will directly pack it. Otherwise, the robot will place the object into the buffer zone. When no objects are present on the conveyor belt, the robot will grasp objects in the buffer zone in descending priority score order.

A. Task Evaluation

To evaluate our packing system, we conducted three grocery packing experiments with a new set of 15 objects to pack (Fig. 8-top). Unlike the prior set of calibration objects (Fig. 3), all of these were chosen to be realistic grocery items, including some very out-of-distribution objects like kale in a produce bag. Each object was given two binary labels: delicate/not delicate, and heavy/not heavy¹. We defined that a “bad pack” occurred when a heavy object was placed on top of a delicate object. This would result in damage to the delicate object, for example when the heavy soup can crushes the pack of chips. This was a consistent definition with the ambiguous qualities that defined “good” grocery packing from humans [37], [38]. We performed three trials on these objects. For each trial, we chose 10 objects randomly and randomized the order which they were presented. To compare across baselines, the sequences selected were:

- **Trial 1:** kale, ice cream, crackers, seaweed, pot roast, baking soda, muffin, chips, gum, soup
- **Trial 2:** bread, kale, stroopwafels, pot roast, muffin, cheese, chips, sprinkles, gum, crackers
- **Trial 3:** cheese, muffin, crackers, pot roast, soup, chips, stroopwafels, Pringles, ice cream, seaweed

We compared our multimodal approach against (1) a baseline experiment where objects were dropped in exactly the same spot every time, and (2) an algorithm that used only our vision system to determine where to pack objects. More specifically:

- **Baseline:** Vision and proprioceptive outputs are used only to ensure objects are grasped. All objects are packed immediately and in the center of the box.
- **Vision-Only:** Vision and proprioceptive outputs are used to ensure objects are grasped. In addition, vision provides an estimate of object shape (bounding box area) that is used to calculate the location where the object is packed. Items with object size greater than a threshold² are considered “large” and are packed immediately, while smaller items are packed later. Items in the buffer are packed by order of decreasing size.

We did not explore a baseline of only tactile sensing as that approach would not be able to detect objects on the conveyor belt.

B. Results

Overall, the results of the grocery packing scenario demonstrates that additional sensing greatly improves the packing performance of the robotic system (Fig. 8). As expected,

¹Specifically, the delicate items were the bread, chips, crackers, kale, muffin, seaweed, and stroopwafels. The heavy objects were the baking soda, gum, ice cream, pot roast, soup, sprinkles, and stroopwafels.

²experimentally set to be 80 cm^2 based on the items in Fig. 3












	Trial 1	Trial 2	Trial 3
Baseline	 7 bad packs	 7 bad packs	 4 bad packs
Vision	 6 bad packs	 2 bad packs	 1 bad pack
Multimodal	 0 bad packs	 2 bad packs	 0 bad packs

Fig. 8. (Top) The 15 objects used to evaluate the packing system. (Bottom) Results of packed bins for each of the three trials for baseline, vision, and multimodal experiments.

the baseline experiment results in poor packing, with an average of six potentially damaging occurrences of a heavy item dropped on a fragile item (“bad packs”) per trial. The vision experiment produces improved packing performance, with an average of three bad packs per trial. Meanwhile, the multimodal experiment averaged less than one bad pack per trial, a $9\times$ improvement over the baseline system and a $4.5\times$ improvement over the vision system.

All bad packs performed during the multimodal trials embody intricacies of real-world grocery packing. More specifically, packing the bag of stroopwaffels on top of other delicate items exemplified one such intricacy because the stroopwaffels were considered both delicate and heavy. This meant that prioritizing the safety of this item could mean potentially damaging other objects; however, packing it first could result in it sustaining damage itself.

Meanwhile, bad packs in the vision trials reflected the fact that large objects do not necessarily mean less delicate objects. For Trial 2, the bread was packed earlier in the process for both the baseline and vision trials than in the multimodal

trial. Since the bread is so large, this essentially guaranteed that the bread would be crushed under subsequently packed items. Overall, our results demonstrate the importance of having a more detailed knowledge of an object than pure vision alone can provide.

VII. DISCUSSION AND FUTURE WORK

In this work, we have shown a soft robotic system that leverages multimodal sensing input to pack groceries in an online fashion with a human-legible “well-packed” metric. To achieve this, we synthesized an external vision system, a set of fluidic-based tactile sensors, and a proprioceptive soft gripper into an integrated packing system. Unlike previous packing methods, we avoid creating models of the items to be packed. Instead, we embrace the variety of shapes, sizes and stiffness and perform significantly better than baseline systems with minimal pre-computation. Our system combines the robust safe handling of soft grippers with the richness of a multimodal sensor suite to outperform traditional vision-only based approaches in this complex task.

For the future, there is significant work that could be explored with more robust tactile and proprioception sensing. In this work, we did not use the fact that the gripper’s proprioception also provides size information about the object. Combining this information with vision could provide a tighter size estimate than the relatively simple approach of bounding box size. Our current work has also shown that positioning is extremely important for tactile sensor feedback. Exploring different configurations of soft sensors could better ensure contact between the soft fingers and target objects, turning our sensitivity into an advantage.

ACKNOWLEDGMENTS

The authors would like to thank John Romanishin for making the gripper more mechanically robust, Ryan Truby for early iterations of the tactile sensor design, and James Bern for his generous time and support. This work is supported by the National Science Foundation (EFRI grant #1830901), Amazon Robotics, and the Gwangju Institute of Science and Technology. LC is supported under the National Science Foundation Graduate Research Fellowship grant #1122374 and the Fannie and John Hertz Foundation.

REFERENCES

- [1] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, *et al.*, “Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching,” *The International Journal of Robotics Research*, vol. 41, no. 7, pp. 690–705, 2022.
- [2] A. Makhal, F. Thomas, and A. P. Gracia, “Grasping unknown objects in clutter by superquadric representation,” in *2018 Second IEEE International Conference on Robotic Computing (IRC)*, 2018, pp. 292–299.
- [3] M. Danielczuk, M. Matl, S. Gupta, A. Li, A. Lee, J. Mahler, and K. Goldberg, “Segmenting unknown 3d objects from real depth images using mask r-cnn trained on synthetic data,” in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 7283–7290.
- [4] P. Schmidt, N. Vahrenkamp, M. Wächter, and T. Asfour, “Grasping of unknown objects using deep convolutional neural networks based on depth images,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 6831–6838.

- [5] F. Wang and K. Hauser, "Stable Bin Packing of Non-convex 3D Objects with a Robot Manipulator," in *2019 International Conference on Robotics and Automation (ICRA)*, May 2019, pp. 8698–8704.
- [6] Y.-D. Hong, Y.-J. Kim, and K.-B. Lee, "Smart Pack: Online Autonomous Object-Packing System Using RGB-D Sensor Data," *Sensors*, vol. 20, no. 16, p. 4448, Jan. 2020.
- [7] J. Shintake, V. Cacucciolo, D. Floreano, and H. Shea, "Soft robotic grippers," *Advanced Materials*, vol. 30, no. 29, p. 1707035, 2018.
- [8] L. Chin, J. Lipton, R. MacCurdy, J. Romanishin, C. Sharma, and D. Rus, "Compliant electric actuators based on handed shearing auxetics," in *2018 IEEE International Conference on Soft Robotics (RoboSoft)*. IEEE, 2018, pp. 100–107.
- [9] S. Li, J. J. Stampfli, H. Xu, E. Malkin, E. V. Diaz, D. Rus, and R. J. Wood, "A vacuum-driven origami "magic-ball" soft gripper," *2019 International Conference on Robotics and Automation (ICRA)*, pp. 7401–7408, 2019.
- [10] H. Wang, M. Totaro, and L. Beccai, "Toward Perceptive Soft Robots: Progress and Challenges," *Advanced Science*, vol. 5, no. 9, p. 1800541, 2018.
- [11] B. Shih, D. Shah, J. Li, T. G. Thuruthel, Y.-L. Park, F. Iida, Z. Bao, R. Kramer-Bottiglio, and M. T. Tolley, "Electronic skins and machine learning for intelligent soft robots," *Science Robotics*, vol. 5, no. 41, 2020.
- [12] A. C. Abad and A. Ranasinghe, "Visuotactile sensors with emphasis on gelsight sensor: A review," *IEEE Sensors Journal*, vol. 20, no. 14, pp. 7628–7638, 2020.
- [13] N. Kuppuswamy, A. Alspach, A. Uttamchandani, S. Creasey, T. Ikeda, and R. Tedrake, "Soft-bubble grippers for robust and perceptive manipulation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2020, pp. 9917–9924.
- [14] A. Zhang, R. L. Truby, L. Chin, S. Li, and D. Rus, "Vision-based sensing for electrically-driven soft actuators," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 509–11 516, 2022.
- [15] A. M. Gruebele, M. A. Lin, D. Brouwer, S. Yuan, A. C. Zerbe, and M. R. Cutkosky, "A Stretchable Tactile Sleeve for Reaching Into Cluttered Spaces," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5308–5315, July 2021.
- [16] R. A. Romeo, M. Gesino, M. Maggiali, and L. Fiorio, "Combining Sensors Information to Enhance Pneumatic Grippers Performance," *Sensors*, vol. 21, no. 15, p. 5020, July 2021.
- [17] L. Chin, F. Barscevicius, J. Lipton, and D. Rus, "Multiplexed manipulation: Versatile multimodal grasping via a hybrid soft gripper," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 8949–8955.
- [18] G.-N. Zhu, Y. Zeng, Y. S. Teoh, E. Toh, C. Y. Wong, and I.-M. Chen, "A bin-picking benchmark for systematic evaluation of robotic-assisted food handling for line production," *IEEE/ASME Transactions on Mechatronics*, vol. 28, no. 3, pp. 1778–1788, 2022.
- [19] J. H. Low, P. M. Khin, Q. Q. Han, H. Yao, Y. S. Teoh, Y. Zeng, S. Li, J. Liu, Z. Liu, P. V. y Alvarado, *et al.*, "Sensorized reconfigurable soft robotic gripper system for automated food handling," *IEEE/ASME Transactions On Mechatronics*, vol. 27, no. 5, pp. 3232–3243, 2021.
- [20] R. Shome, W. N. Tang, C. Song, C. Mitash, H. Kourtev, J. Yu, A. Boularias, and K. E. Bekris, "Towards robust product packing with a minimalistic end-effector," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9007–9013.
- [21] M. Agarwal, S. Biswas, C. Sarkar, S. Paul, and H. S. Paul, "Jampacker: An efficient and reliable robotic bin packing system for cuboid objects," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 319–326, 2021.
- [22] A. Yasuda, G. A. G. Ricardez, J. Takamatsu, and T. Ogasawara, "Packing planning and execution considering arrangement rules," in *2020 Fourth IEEE International Conference on Robotic Computing (IRC)*, 2020, pp. 100–106.
- [23] F. Wang and K. Hauser, "Robot packing with known items and nondeterministic arrival order," *IEEE Transactions on Automation Science and Engineering*, pp. 1–15, 2020.
- [24] M. A. Lee, M. Tan, Y. Zhu, and J. Bohg, "Detect, reject, correct: Crossmodal compensation of corrupted sensors," 2020.
- [25] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, p. 3300–3307, Oct 2018. [Online]. Available: <http://dx.doi.org/10.1109/LRA.2018.2852779>
- [26] Y. Bekiroglu, R. Detry, and D. Kragic, "Learning tactile characterizations of object- and pose-specific grasps," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 1554–1560.
- [27] K.-T. Yu and A. Rodriguez, "Realtime state estimation with tactile and visual sensing for inserting a suction-held object," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1628–1635.
- [28] D. De Gregorio, R. Zanella, G. Palli, S. Pirozzi, and C. Melchiorri, "Integration of robotic vision and tactile sensing for wire-terminal insertion tasks," *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 2, pp. 585–598, 2019.
- [29] J. Bohg, M. Johnson-Roberson, M. Björkman, and D. Kragic, "Strategies for multi-modal scene exploration," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 4509–4515.
- [30] M. A. Lee, B. Yi, R. Martín-Martín, S. Savarese, and J. Bohg, "Multimodal sensor fusion with differentiable filters," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 444–10 451.
- [31] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 8943–8950.
- [32] S. Wang, J. Wu, X. Sun, W. Yuan, W. T. Freeman, J. B. Tenenbaum, and E. H. Adelson, "3d shape perception from monocular vision, touch, and shape priors," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1606–1613.
- [33] L. Chin, M. C. Yuen, J. Lipton, L. H. Trueba, R. Kramer-Bottiglio, and D. Rus, "A simple electric soft robotic gripper with high-deformation haptic feedback," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 2765–2771.
- [34] R. Hu, J. Xu, B. Chen, M. Gong, H. Zhang, and H. Huang, "Tap-net," *ACM Transactions on Graphics*, vol. 39, no. 6, p. 1–15, Nov 2020. [Online]. Available: <http://dx.doi.org/10.1145/3414685.3417796>
- [35] R. L. Truby, L. Chin, and D. Rus, "A recipe for electrically-driven soft robots via 3d printed handed shearing auxetics," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 795–802, 2021.
- [36] R. L. Truby, L. Chin, A. Zhang, and D. Rus, "Fluidic innervation sensorizes structures from a single build material," *Science Advances*, vol. 8, no. 31, p. eabq4385, 2022.
- [37] K. Renae, "8 tips for bagging groceries, according to someone who does it every day," <https://www.insider.com/best-way-to-bag-groceries-2018-12>, Dec 2018, accessed: 2023-10-30.
- [38] I. of Food Technologists, "Tips for bagging groceries safely," <https://www.ift.org/career-development/learn-about-food-science/food-facts/food-facts-food-safety-and-defense/bagging-groceries>, accessed: 2023-10-30.
- [39] L. Chin, J. Lipton, M. C. Yuen, R. Kramer-Bottiglio, and D. Rus, "Automated recycling separation enabled by soft robotic material classification," in *2019 2nd IEEE International Conference on Soft Robotics (RoboSoft)*. IEEE, 2019, pp. 102–107.