ISACS: In-Store Autonomous Checkout System for Retail

JOÃO DIOGO FALCÃO, Carnegie Mellon University, USA
CARLOS RUIZ, AiFi Inc., USA
ADEOLA BANNIS, Carnegie Mellon University, USA
HAE YOUNG NOH, Stanford University, USA
PEI ZHANG, University of Michigan, USA

90% of retail sales occur in physical stores. In these physical stores 40% of shoppers leave the store based on the wait time. Autonomous stores can remove customer waiting time by providing a receipt without the need for scanning the items. Prior approaches use computer vision only, combine computer vision with weight sensors, or combine computer vision with sensors and human product recognition. These approaches, in general, suffer from low accuracy, up to hour long delays for receipt generation, or do not scale to store level deployments due to computation requirements and real-world multiple shopper scenarios.

We present *ISACS*, which combines a physical store model (e.g. customers, shelves, and item interactions), multi-human 3D pose estimation, and live inventory monitoring to provide an accurate matching of multiple people to multiple products. *ISACS* utilizes only shelf weight sensors and does not require visual inventory monitoring which drastically reduces the computational requirements and thus is scalable to a store-level deployment. In addition, *ISACS* generates an instant receipt by not requiring human intervention during receipt generation. To fully evaluate the *ISACS*, we deployed and evaluated our approach in an operating convenience store covering 800 square feet with 1653 distinct products, and more than 20,000 items. Over the course of 13 months of operation, *ISACS* achieved a receipt daily accuracy of up to 96.4%. Which translates to a 3.5x reduction in error compared to self-checkout stations.

CCS Concepts: • **Computer systems organization → Sensor networks**; *Real-time system architecture*; • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: autonomous checkout, retail, sensor fusion, human tracking, inventory monitoring, instant receipt

ACM Reference Format:
João Diogo Falcão, Carlos Ruiz, Adeola Bannis, Hae Young Noh, and Pei Zhang. 2021. ISACS: In-Store Autonomous Checkout System for Retail. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 99 (September 2021), 26 pages. https://doi.org/10.1145/3478086

1 INTRODUCTION

In 2020, 90% of retail sales are still happening in physical stores [9, 10, 48]. In these physical stores customers suffer from having to wait in long lines to checkout (get their receipts and pay). These long lines are the most common reason for shopping trip abandonment. Over half of shoppers are willing to spend less money in a store, or even walk away entirely, to avoid a slow checkout [17]. Having to wait in long lines affects customers' satisfaction and

Authors' addresses: João Diogo Falcão, Carnegie Mellon University, Moffett Field, California, USA, 94035, joaodf@cmu.edu; Carlos Ruiz, AiFi Inc. Santa Clara, California, USA, 95051, carlos@aifi.com; Adeola Bannis, Carnegie Mellon University, Moffett Field, California, USA, 94035, abannis@cmu.edu; Hae Young Noh, Stanford University, Stanford, California, USA, 94305, noh@stanford.edu; Pei Zhang, University of Michigan, Ann Arbor, Michigan, USA, 48109, peizhang@umich.edu.

2474-9567/2021/9-ART99
https://doi.org/10.1145/3478086

Fig. 1. ISACS deployed autonomous checkout store. This store is deployed by Carrefour SA. inline with their convenience store format.

loyalty. The checkout experience and length of the checkout line influences the shopping decisions of close to 40% of shoppers [17].

Autonomous stores have started to emerge in order to address this issue. Companies (such as Amazon, Grabango, Zippin) have started to experiment with checkout-less stores. In these stores the wait time is removed by providing a receipt without the need for a cashier to scan the products [1, 19, 54].

These autonomous stores can be grouped into three main approaches: vision-only, a combination of computer vision with additional sensors, or a combination of computer vision with sensors and human product recognition. **Vision-only** approaches for autonomous stores are most common. However, these approaches require a large number of cameras in multiple places to minimize blind spots. This approach also require high computation requirements to scale to a store-level deployment. In addition, there is also a very large human labor required to train vision systems to identify products. Tens of thousands of labelled images are required from different angles with different illuminations. Furthermore these prior approaches look at images without obstructions from humans shopping. The best approaches have only been able to reach 85% accuracy in unobstructed retail datasets [42]. This 15% loss is unacceptable for retail businesses, where stores need to operate at a maximum loss of approximately 10% [22]. In order to improve the accuracy, companies such as Amazon Go have added a human-in-the-loop to generate the receipt in the cloud. As the customer leaves a "cloud cashier" will manually review the transaction and images of the items, then the customer will be charged accordingly [1, 35, 46]. This process can generate more accurate receipts, but can cause delays that are up to hours [12]. Amazon has been leading the autonomous checkout space with several Amazon Go stores deployed, however these stores are being developed and deployed behind closed doors, and many of the learning's and details are lost to the broader community.

Other more traditional retailers have started to experiment with automation in their stores by installing *self checkout* and *scan and go* checkout systems [16, 26, 49]. In order to remove the servicing time *scan and go* technology requires the customers to carry a device from the store and scan every item they intend to purchase as they pick them up, avoiding then the time required for the employee to scan the items at the checkout station [16]. However, retailers that used *self checkout* stations did not see a reduction in the customers' wait time due to

unfamiliarity of customers to the checkout process. Furthermore, those who deployed *scan and go* technology in their stores found an added product loss of as much as 1% for every 10% of *scan and go* sales [2].

While Amazon is currently operating their Amazon Go stores it is quite hard to estimate their system accuracy, as these numbers are not public information. However prior published approaches which also combine **vision and weight** sensors can achieve up to 94% accuracy in identification by including weights of the item in addition to visual data [38]. However, this approach fails when multiple people are present at the store at the same time due to confusion from *multiple people to multiple objects matching*. Furthermore, the accuracy is limited in this approach due to the assumptions made in a traditional operating stores' setup, while a dedicated *autonomous store setup* can have high impacts on the accuracy.

In this paper, we present an In-Store Autonomous Checkout System for retail, *ISACS*, which tracks multiple shoppers throughout their shopping journey, combining a physical model of the store and customers, with live inventory monitoring and multi-human 3D pose estimation in order to produce a receipt and present it to the customer in less than 2 seconds. *ISACS* greatly reduces the computational requirements by only using weight sensors on each shelf to track items coming in and out of the shelves while identifying them. Our approach combines the 3-D physical model of the customers and their item interactions with item tracking on the shelves to match multiple people with multiple products in close proximity.

In order to fully evaluate the autonomous setup, operation, and accuracy, we fully implemented *ISACS* on a 800 sqft store with 52 cameras and 580 weight sensing shelves for 13 months in a store owned and operated by Carrefour SA, in Massy France. The store contained on average 1653 unique items, with a real stock of close to 20000 products (See Fig. 1). This store had an average traffic volume of around 140 transactions daily before COVID-19 and close to 110 daily transactions during the pandemic. We present our results and lessons-learned throughout the pre- and during COVID'19 pandemic period.

The main contributions of this paper are:

(1) An open and transparent analysis of a deployment for a in-store autonomous checkout system, where a receipt is automatically produced before the customer leaves the store, without human intervention.
(2) An algorithm that combines 3D body pose estimation, physical modelling of people and their interaction with the store, and inventory knowledge to enable multiple people and multiple products matching.
(3) Analysis and experiences of the deployment and operations of our system in an 800 square feet operational convenience store with 1653 distinct products, over 13 months in a store owned and operated by Carrefour SA, in Massy France.

The rest of the paper is organized as follows. First, we discuss the background works in people tracking, inventory management and inventory to people association in Section 2. Next, we introduce our system in Section 3. Then in Section 4 we present the association algorithm that combines 3D pose estimation with inventory management. We then describe our evaluation of *ISACS* in the real world in Section 5. In Section 6, we discuss our experiences and lessons from deploying *ISACS* in the real world, and how we improved the operations of our system over time. Finally, we conclude in Section 8.

2 AUTONOMOUS CHECKOUT BACKGROUND AND RELATED WORK

ISACS's proposes a system that generates a receipt autonomously, i.e. no human interaction, while the customer is inside the store by matching multiple people with multiple item pick up/put down events. To achieve this, *ISACS*'s system contributes to multiple fields of research, such as: multi-human tracking, object recognition and tracking, methods for identifying retail products, 3D scene reconstruction and sensor fusion. In this section we discuss existing works in these fields and how *ISACS*'s leverages these prior works to enable Autonomous retail.

2.1 Autonomous Retail Enablers

The recent advancements of computer vision research in object recognition help enable applications such as autonomous stores. These advancements can be grouped into three main approaches: vision-only, a combination of computer vision with additional sensors, or a combination of computer vision with sensors and human product recognition.

2.1.1 Vision only. These solutions face two critical challenges: data availability for training and occlusions. The latest work using deep learning techniques to do object recognition –Self-attention-based [51], Mask R-CNN [20], Center-Net [13], NormalNet [50], FoveaBox [25]– require a large amount of labelled data on every individual product for training which is highly labor intensive and quickly becomes impractical. Some solutions have alleviated this problem by automating this process either through the use of sensors [40] or semi-supervised human labelling [47].

An added challenge for vision-only approaches are objects such as fruits and vegetables. There are works such as [15] that focus on identification of these kinds of products. Due to their visual variable nature, the approach is only accurate for a small subset of fruits and vegetables. Furthermore this only identifies the item, but doesn't natively handle accurate counting or items sold by the weight –as is the case for most fruits and vegetables.

The above approaches are also very computationally intensive, which can be limiting either in the excessive hardware provisioned in the store, or with the cost of such computation in the cloud. Approaches like [18] attempt to simplify the problem by leveraging Optical Character Recognition (a lower computationally intensive approach), to identify the objects by their label. RGB-D cameras have also been used to attempt at reducing computation and increasing accuracy of object detection [21] at a prohibitive higher infrastructure cost.

Furthermore these approaches focus mainly on locating and classifying objects in 2D images [43], which provides no information about the motion of the object, and the owner of the object.

2.1.2 Vision with Additional Sensors . Works such as [14, 28, 38, 39] combine however multiple 2D views with additional weight sensors, similar to *ISACS*, to generate an understanding of which objects are being picked up or put back down. However these works are limited by only one person interacting with one item at a time and are limited to a small amount of products, preventing them to scale to a store-level deployment. Other sensors have been explored for object identification such as RFID [31, 36, 37, 53], vibration sensors for people detection and tracking [32, 33] and inertial wearable sensors [41] for product to people matching, however these solutions have not shown to be practical in terms of cost, deployment (requiring customer to wear a device) or are not accurate enough for the low margins a store needs to operate.

2.1.3 Vision, Sensors and Human Product Recognition . Given the high accuracy required to operate a profitable convenience store, companies like Amazon have leveraged humans-in-the-loop in their cashier-less stores [1, 35, 46]. This approach is quite restrictive: delaying the delivery of the receipt prevents shoppers who are budget conscious from verifying their receipt before leaving the store, affecting the lower end of the social economic spectrum. American consumer patterns are changing, and discount retailers are seeing an increased adoption from a larger population sample that is more budget conscious [52]. COVID-19 has further pushed this behavior and driven shoppers' behavior towards a more targeted spending [7]. Furthermore, the majority of shoppers make purchases based on in-store discounts and the "smart-shopper feelings" towards pricing, discounts and promotions, act as a major component of the emotional response affecting shoppers' behavior to favor in-store price confirmation [8, 34, 44]

2.2 Retail Technology

There are several works in the computer vision, robotic and sensing fields that have pushed forward the retail technology domain. Robots that grab different kinds of objects are faced with similar challenges in identifying

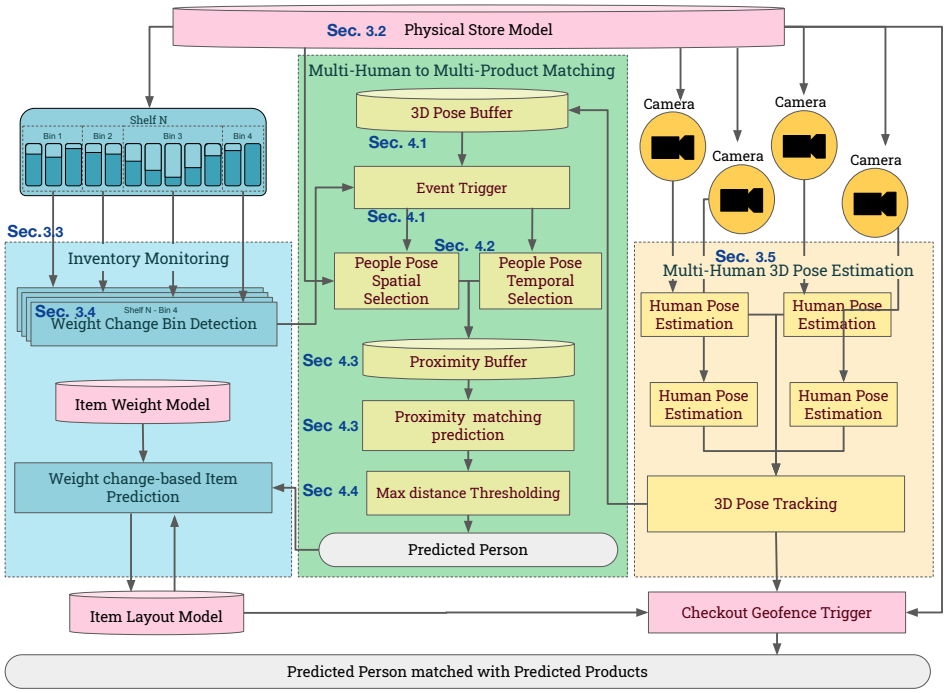


Fig. 2. *ISACS* System overview. Left: Weight sensing pipeline for inventory monitoring (blue). Middle: Multi-human to Multi-product matching pipeline (green). Right: Computer vision pipeline for Multi-human 3D Pose Estimation (yellow). Models of the store, Items and Item location (pink). The corresponding sections of each component are noted as Sec. X.Y

objects [24]. These solutions provide visual understanding of the objects in a *fast* manner even withing a cluttered environment [29]. Others provide a multi-view approach for object pose understanding, in this case for the purpose of robotic manipulation [27].

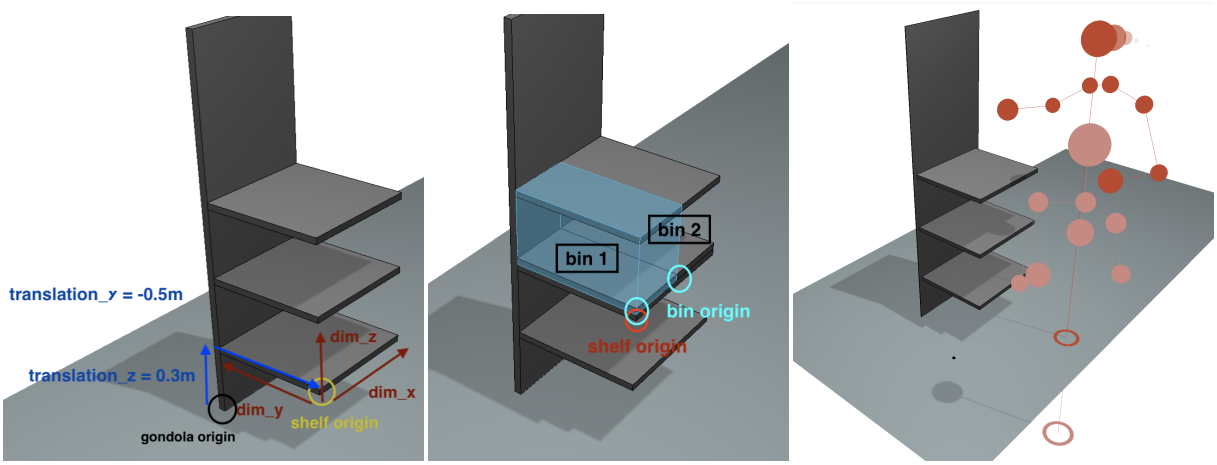
Counting and tracking people without interfering with their normal behavior is a relevant problem that spans beyond the Autonomous Checkout domain. There are works [5, 6, 11, 30] that focus on leveraging cameras to track people continuously across multiple camera views [6], focused on dense environments [30] or on the ability to reconstruct the 3D motion of people [45] in order to understand their behavior.

While each of these works alone does not address the autonomous checkout domain they share common challenges in visual understanding of the scene (store), objects (products) and people (shoppers).

3 *ISACS* OVERVIEW

ISACS is the first deployed, fully operational autonomous checkout system with in-store receipts. *ISACS* 1) accurately produce receipts in a real-world convenience store setting, while the shopper is in the store (a key and legal requirement for operational stores) and 2) does not require human-in-the-loop during checkout. In this section, we provide an overview of our system.

Figure 2 shows *ISACS* system framework. *ISACS* combines a physical model of the store, multi-human 3D pose estimation, and live inventory monitoring to improve the accuracy of the match between the selected products and the customers. *ISACS* is constantly doing cross-view multi-human 3D pose estimation for every camera in the store. As a customer walks in the field of view of more than one camera he/she gets an anonymous ID assigned to him/her. When the customer picks up a product from a shelf, the *inventory monitoring* and *multi-human*



(a) Shelves (b) Bins (c) Multi-human 3D Pose Tracking

Fig. 3. Physical store model showing the origin and dimensions of a gondola (a), shelf (a) and bin (b) and 2 people being tracked (c). *ISACS*'s uses this this information to accurately map the inventory changes to the shopper.

to multi-product matching pipelines are triggered. With the knowledge of the physical model of the store, the system is able to understand the 3D location of the product being picked up –*where*–. From there the *ISACS*'s multi-human to multi-product matching pipeline determines the probability of each person picking up the object –*who*– by looking at the physical distance of each hand over a small window of time during the interaction –*when*–. The inventory monitoring section will compute the probability of the product(s) being picked up –*what*–, as well as the quantity, based on an updated knowledge of the inventory on the shelves and people. A similar process occurs when a customer puts back the objects. Finally as the customer enters a virtually geo-fenced defined area (checkout area), a receipt is then computed based on each event and delivered to the terminal for payment.

3.1 Assumption

Our intention was to operate a normal convenience store with minimal impact to the way customers behave in a normal store, in order to validate *ISACS*'s ability to operate with real world constraints. However we made the assumption that **the last customer who removes a product from a store's shelf and leaves with it is the customer charged for that product.** While this naturally occurs in most transactions, it does not happen for all transactions. In our deployment we educated the customers about this limitation by adding visual diagrams/instructions explaining this.

3.2 Physical Store Model

ISACS relies on a centimeter accurate physical model of the store to produce an accurate matching between the movement of products and people. In order to do so, the system needs to know where each shelf and camera are located, in 3D, and their orientation.

We define 4 types of spatial entities: a gondola, a shelf, a bin and geo-fenced areas. We can define all of these by their origin location, dimension and rotation (X_o, X_d, θ) (See Fig. 3). We assume that gondolas can only be rotated over the z (height) axis; shelves can only rotate on the y (horizontal, for tilt) axis; while a bin and a geo-fence cannot rotate. We mathematically define G^i as gondola, S^i shelf, and B^i bin with index i. And G^i_c, S^i_c, B^i_c as the

center of these volumes. We formalize the location of the gondolas/shelves/bins, such that (note that G can be swapped with S and B to denote a shelf or a bin):

$$G^i = \left\{ X : \begin{array}{c} |X_x - G^i_{x_o} + G^i_{x_d}/2| < G^i_{x_d}/2 \\ |X_y - G^i_{y_o} + G^i_{y_d}/2| < G^i_{y_d}/2, \quad \forall X \in \mathcal{R}^3 \\ |X_z - G^i_{z_o} + G^i_{z_d}/2| < G^i_{z_d}/2 \end{array} \right\} \tag{1a}$$

$$G(X) = \begin{cases} i & X \in G^i \\ 0 & \text{otherwise} \end{cases} \tag{1b}$$

$$G^i_c = \left(G^i_{x_o} + G^i_{x_d}/2, G^i_{y_o} + G^i_{y_d}/2, G^i_{y_o} + G^i_{y_d}/2 \right) \tag{1c}$$

A gondola represents a physical fixture that contains multiple shelves stacked vertically. A shelf contains multiple sensing platforms and multiple bins. A bin is a virtual concept that is defined by the space that an integral number of plates a product occupies on a shelf. Each bin contains only 1 product assigned to be placed inside it. This means that changing the placement of the products implies a simple substitution of the old bin location –containing the products being changed–, with the new bin location with respect to the shelf. Geo-fences are also defined for regions of interest, such as entry, exit and checkout areas. These areas along with the shelves and cameras seldom change.

In addition *ISACS* requires the knowledge of the location of the cameras and their orientation. This can be obtained via traditional methods for multi-camera calibration. These methods produce a map of relative position and orientation of the cameras without scale. To further obtain the scale and align the two world coordinate systems, *ISACS* uses a visible *anchor vector* in the form of a sheet of paper with a ArUco pattern, with a fixed size of 20cm, laid on the floor during the calibration step [3].

3.3 Inventory Monitoring

ISACS uses live inventory monitoring, achieved through weight sensors, in order to compute the list of products that a person is shopping. *ISACS* uses a similar approach as *FAIM* [14], that is, each shelf is instrumented with multiple plates, suspended over 2 weight sensors each, along the depth axis (See Fig. 4). Depending on the width of the shelf, each shelf contains 4, 6, 9 or 12 sensing plates. Each sensing plate is 12cm-wide and each sit contiguous to each other. This one-sized narrow sensing plate allows for ease of manufacture-ability as well as low cost given that they will not have much weight sit on top of it, allowing for a lower maximum capacity load cell thus higher weight resolution without expensive ADCs (Analog-to-Digital Converters).

ISACS does not use computer vision to further improve the object identification. From prior work results [14], we have observed that the small added value of computer vision comes at the expense of very high computation needs, which limits this approach from scaling to a store level deployment. Furthermore this computer vision addition only provides extra information in the case the prior knowledge of the location of the objects is not deterministic. In other words computer vision for inventory monitoring only plays a role when objects are either misplaced or sitting in the same sensing plates as other different products. From experience we have observed that misplacements seldom occur in an organized convenience store. **So *ISACS* enforces that each sensing plate has only 1 product type sitting on it**, placing a physical barrier (metal separator, see Fig. 5) between products, and that store employees properly re-stock and organize the store once a day.

Inventory monitoring in this context is measured as the quantity of products sitting on a shelf at any point in time. This means that when a customer or employee picks up a product, or puts it down (whether in the right or wrong place), *ISACS* tracks all of these changes and updates its inventory correspondingly. This leads to *ISACS*

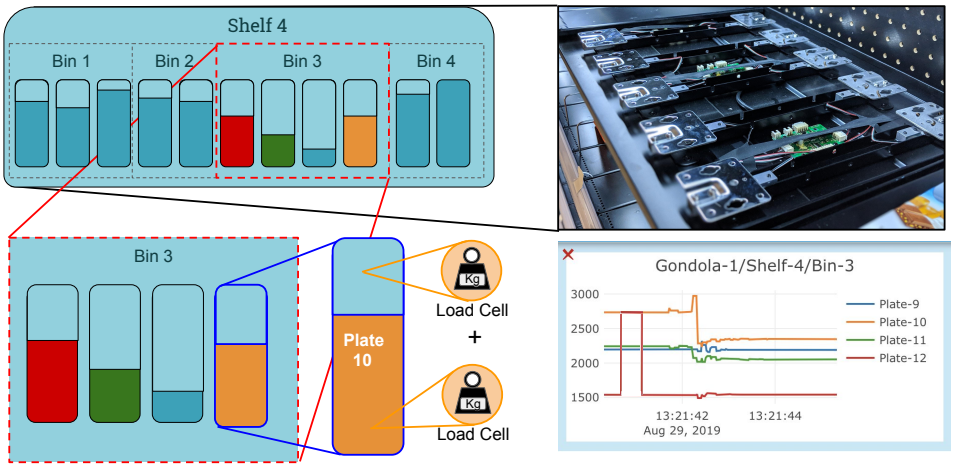


Fig. 4. *ISACS*'s weight sensor shelves showing a deeper view into how a shelf is divided into multiple bins, each with multiple plates, each with 2 load cells. Top Right image shows a real image of the shelves used with the load cells exposed for clarity. Bottom Right image shows a plot of the resulting measurements used in *ISACS*.

Fig. 5. Typical gondolas in the store showing each bin separated by metal dividers (red arrows) containing only 1 type of product per bin with multiple items inside.

accommodating normal shopping behaviors, such as, picking up a product to further inspect it and putting it down without getting charged for that product.

3.4 Weight Change Bin Detection

The triggering point that initiates the association process between a product and a person is the moment the person picks up, or puts down, a product from a shelf. *ISACS* detects this event by continuously processing the

signals coming from the sensing plates. These plates are aggregated into *bins* and processed together as such (See Fig. 3b).

In prior work we have computed the mean and variance of the weight values over a sliding window, and classified it as *stable* or *active*. This way we can observe the weight values at the beginning and end of the *activity* and compute the difference in weight. Furthermore, we have observed from previous studies that the smaller the bin width the higher accuracy we obtain for item identification. So we made every bin in the store contain one product only, and cover the minimum consecutive amount of plates that is sufficient to support the product. This makes the event detection more robust to **multiple people** interacting with the same shelf simultaneously by treating each interaction separately. Collision cases still occur if 2 or more customers pick from the exact same *bin* simultaneously. This means that upon a weight change detection occurring, if there are multiple hands within the associated bin, the system has no valid way to detect which hand is the correct one and will decide in favor of the first one that arrived at that location. We have never observed such a case with real shoppers outside of our intentional tests.

Prior work defines mathematically $w^n_{g,s,p}$ as the weight on the p^{th} weight plate on shelf s, gondola g at discrete time n. We also defined ($\mathcal{L}$) as the Item Layout Model such that:

$$l_{g,s,p} = \{i \in \mathcal{I} \mid \text{product } i \text{ is stocked at plate } p \text{ on shelf } s \text{ and gondola } g\} \tag{2a}$$

$$\mathcal{L} = \{l_{g,s,p}\}, \quad \forall g, \forall s, \forall p \tag{2b}$$

And $|l_{g,s,p}|$ as the total number of items at plate p on shelf $s$and gondola g. *ISACS* further defines a bin b as a set of consecutive weight plates $\{p^{g,s}\}$, defined according to the physical store model and constrains the location of the products such that $|l_{g,s,p}| = 1$. Given this we then compute the bin's aggregated weight as:

$$w^n_{g,s,b} = \sum\nolimits_{p \in b} w^n_{g,s,p} \tag{3}$$

Then, the bin's aggregated moving mean and variance are, respectively, $\mu^n_{g,s,b}$ and $v^n_{g,s,b}$:

$$\mu^n_{g,s,b} = \frac{1}{2\,\mathrm{N}_w + 1} \sum_{t=n-\mathrm{N}_w}^{n+\mathrm{N}_w} w^t_{g,s,b} \tag{4}$$

$$v^n_{g,s,b} = \frac{1}{2\,\mathrm{N}_w + 1} \sum_{t=n-\mathrm{N}_w}^{n+\mathrm{N}_w} |w^t_{g,s,b} - \mu^t_{g,s,b}|^2 \tag{5}$$

where N_w is the sliding window half-length in samples, which corresponds to 0.5s in our implementation ($2\,\mathrm{N}_w + 1 = 61$).

An event is detected according to Equations 6:

$$\text{Event begins on gondola } g \text{ and shelf } s, \text{ bin } b: \quad v^t_{g,s,b} > \varepsilon_v, \forall t \in [n_i,\ n_i + \mathrm{N_h}) \tag{6a}$$

$$\text{Event ends on gondola } g, \text{ shelf } s, \text{ bin } b: \quad v^t_{g,s,b} \leq \varepsilon_v, \forall t \in (n_e - \mathrm{N_l},\ n_e] \tag{6b}$$

$$\text{Temporal consistency:} \quad n_e > n_i \tag{6c}$$

where $\mathrm{N_h}$ and $\mathrm{N_l}$ correspond to the minimum length the weight variance has to exceed or fall short of the threshold ε_v in order to detect the initial and ending timestamp of an event. Based on some initial experiments we empirically set the values to $\mathrm{N_h} = \mathrm{N_l} = 30$ (0.5s) and $\varepsilon_v = 0.01\,kg^2$. Once an event has been detected, the Weight Change Bin Detection module determines the event weight difference $\Delta\mu$ for every bin for which the event occurred.

3.5 Multi-human 3D Pose Estimation

Understanding customer motion throughout the store is done through images captured from cameras placed on the ceiling of the store, given this is the place with the most unobstructed view into the people in the store. In Section 6 we discuss the camera placement and its impact in the accuracy of the system. In this section we describe the Multi-human 3D pose estimation approach used by *ISACS*.

The multi-human 3D pose estimation pipeline is broken down into 2 main components: human pose estimation and 3D tracking. The human pose estimators compute the set of keypoints: $\{k_{head}, k_{neck}, k_{shoulderL}, k_{shoulderR}, k_{elbowL}, k_{elbowR}, k_{wristL}, k_{wristR}\}$, $k_n \in \mathcal{R}^2$ for all humans present in an image using the same approach as in [4, 6]. These 2D key points are computed for all cameras in the store and passed on to the 3D tracking component. At this stage the 3D tracking uses a temporal and spatial affinity metric to match the multiple keypoints across different views. Finally by using the camera placement provided by the Physical store model the 3D pose tracker triangulates the multiple views into separate consistent 3D people [6]. (See Fig. 3c) *ISACS*'s human pose estimator was trained using the CMU panoptic dataset [23] and enhanced with data collected and annotated from several other deployments provided by AiFi Inc. These deployments included a diverse set of racial, gender, age and clothing. However they were collected from real stores. This means that intentional odd patterns in clothing or masks were not contemplated in these datasets.

We've measured the accuracy of our tracking model in two ways: consistency of shopper identification across its shopper journey (between entering and exiting), and location accuracy of the joints predicted. By counting the number of people coming into the store and coming out of the store, and looking at the consistency of the trajectory we were able to determine that across the 2 weeks dataset (Detailed in Sec. 5.3) we only had 12 incorrect ids tracked out of 1874 transactions, this is approximately 99.36% accuracy in consistent identity tracked. The location of the joints however was measured empirically by our testers. This measurement was done by placing a hand in the center of every bin in the store and verifying if the visualization of the 3D projection matched the center location of the bin. In this case the spatial accuracy of our tracking solution was within 1 bin length of the expected bin. This translates to an average accuracy of approximately 10cm when not occluded. All measurements with respect to shoppers mentioned in this paper are relative to this tracking system. Section 4 further details the impact of the measurements taken by the tracking system in the matching of people to products taken.

While *ISACS* mainly leverages the approach presented in [6], it improves that approach by adding physical contextual information to the tracker. By leveraging the *Physical Store Model* we improved the people detection accuracy by initializing and eliminating people tracks only at the entrance and exit of the store. This prevented the tracker from generating contextual impossible 3D tracks, such as, someone appearing in the middle of a store or on top of a gondola. It further allows for a person to be missed in some frames, due to occlusions or inaccurate model detection, and

Even our human pose estimation model provided confidence level for the prediction of each joint, our system did not fully leverage this information. In the case the joint was predicted with a confidence level of above 80% *ISACS* would use that prediction. However, a more promising approach would be to use the confidence level of each joint prediction when doing the matching between products and humans. This information could be useful when deciding between two competing hands picking a certain object.

4 MULTIPLE HUMAN TO MULTIPLE PRODUCTS MATCHING

The previous section described the inventory monitoring and multi-human 3D pose estimation pipelines that *ISACS* relies on. Here, we detail how *ISACS* aligns the inventory triggering mechanism with the pose information (Section 4.1) and further leverages the pose information and the physical store model to select the correct people interacting with the shelves (Sections 4.2) and how this selection is then combined with the proximity of the event to emit a prediction of people interacting with the shelf (Sections 4.3-4.4).

4.1 Visual Event Timing

After an event is detected and triggered by the Weight Change Bin Detection module *ISACS looks* at who was interacting with that shelf at the event time. However, timing between the sensor readings and the video requires synchronization. While all the cameras are able to synchronize via NTP (Network Time Protocol), the custom weight sensing hardware is not. *ISACS* addresses this issue by making use of a 3D Pose Buffer.

4.1.1 3D Pose Buffer. In order to properly match the timing of the event with the right set of estimated 3D people in the store, *ISACS* keeps a running buffer of 3D Pose estimated people. This buffer ensures that a fast motion of pick up/put down can still be detected an accounted for.

We define mathematically p_i^n as the set of 3D keypoints (Section 3.5) defining the i^{th} person at discrete timestamp n. We also defined ($\mathcal{P}$) as the location of all people in the store at any given time, such that:

$$p_i^n = \left\{ \begin{array}{c} k_{\text{head}}, k_{\text{neck}}, \\ k_{\text{shoulderL}}, k_{\text{shoulderR}}, \\ k_{\text{elbowL}}, k_{\text{elbowR}}, \\ k_{\text{wristL}}, k_{\text{wristR}} \end{array} \right\}, \quad k_m \in \mathcal{R}^3 \tag{7a}$$

$$\mathcal{P} = \{p_i\}, \quad \forall i \tag{7b}$$

And $|\mathcal{P}^n|$ as the total number of people in the store, at any given time n.

4.1.2 Event Trigger. We define the moment an event is detected from Weight Change Bin Detection (Section 3.4) as n_i and the moment it is triggered n_t, such that $n_t > n_i$. These 2 times are never exactly the same given the delay created by the detection mechanism.

4.2 People Pose Selection

Once detected an event, in order to accurately match multi-humans with multi-products *ISACS* assigns different likelihood to the 3D Poses in the buffer, based on the timing and physical distance of each 3D Pose to the triggering bin's location.

From the *3D Pose Buffer* ($\mathcal{P}$, Eq. 7b) and the time the event was detected n_i (Eq. 6b) the **People Pose Temporal Selection** estimates the likelihood of a person on $\mathcal{P}$ belonging to that event. Such that:

$$\text{Event Time Window:} \quad \mathcal{E} = \{n : n_i - \epsilon_t < n < n_t\}, \forall n \in \mathcal{P} \tag{8a}$$

$$P_T\left(p \mid \mathcal{E}, \mathcal{P}\right) = \left\{ \begin{array}{cc} 1/|n_i - n| & \forall n \in \mathcal{E} \\ 0 & \text{otherwise} \end{array} \right. \tag{8b}$$

This means that the closest a person is to the detected timestamp n_i, the higher the likelihood that that person is the one that created the event. *ISACS* considers the event to have a duration of ϵ_t. Although an event is detected using the weight sensors, the motion of a pickup/put back spans a larger time window than the measure sensing window. We empirically set the value of $\epsilon_t = 3.5$ seconds based on initial experiments.

The **People Pose Spatial Selection** further estimates the same likelihood using the spatial proximity of the detected person to the events location (Section 3.2), such that:

$$P_S(p \mid B, \mathcal{P}) = \left\{ \begin{array}{cc} 1 & B(p_i^n(\text{wrist}^*)) = b, \forall i, \forall n \in \mathcal{P} \\ 1 - \frac{min^*(|p_i^n(\text{wrist}^*) - B_c^b|)}{\sum_{i \in \mathcal{P}} |p_i^n(\text{wrist}^*) - B_c^b|} & B(p_i^n(\text{wrist}^*)) \neq b, \forall i, \forall n \in \mathcal{P} \end{array} \right. \tag{9}$$

This Equations assigns a probability of 1 to a certain person (p) in case the location of either hand of a person is estimated to be inside the bin (b) that generated the event. It further assigns a normalized decaying likelihood to the closest hand of each person inside the buffer.

Extruded Bins. When calculating the spatial probability of each pose for each event in the entire dataset, we have observed that $P_S(p|B, \mathcal{P}) = 1$ rarely occurred. This happened because when the event takes place the hands of the person are inside the shelf –which most of the time is occluded by the upper shelves (except on the top shelf)– leading the 3d Pose estimation to incorrectly predict the position of the hands. Therefore we have extruded the size of the bins towards the front of the shelf by $\epsilon_x = 10$cm. This ensured that we could capture the position of the hand while it was coming in and out of the bin, and consider it inside the volume. We show the results of these approaches in Section 5.

4.3 Proximity Buffer and Matching Prediction

In this stage of the pipeline *ISACS* filters out all estimated people from the 3D Pose Buffer that have a likelihood of 0 based on Equations 8 and 9. This creates a much smaller buffer, the **Proximity Buffer**, which contains the estimated people that are close to the event in both time and space. Finally the Eqs. 8 and 9 are combined, such that:

$$P^i_{\text{pose}} = P_S(p \mid B, \mathcal{P}) \cdot P_T(p \mid \mathcal{E}, \mathcal{P}) \tag{10}$$

This Equation creates a list of likelihoods for every 3D Pose inside the Buffer ($\mathcal{P}$, *Eq.7b*).

Hand inventory information during put backs. During put backs, *ISACS* leverages an extra piece of information: the inventory in the hands of the people. This means that once a Put Back occurs, at this stage, *ISACS* removes from the likehood model p^i_{pose} all i's that do not have the products for which the *Inventory Monitoring pipeline* has predicted, leaving only those who could have placed that item in the shelf, in the first place. We show the results of the approaches in Section 5.

4.4 Maximum Distance Thresholding

The last step in making a prediction on *Who* interacted with the shelves is a thresholding on maximum distance. This means that predictions where $|p^i(k_{wrist*}) - B_c| > \epsilon_s$ are ignored.

This threshold exists to remove events that are created based on ambient noise on the sensors or any faulty signal artifact. In this case if a sensor triggers an event but there is no person close to the shelf we can safely assume that this is a false trigger and *ISACS* ignores the event. We've defined empirically $\epsilon_s = 2$meters, given that most people will not be physically able to pick any items from the shelves if they are standing 2 meters away from it.

5 REAL-WORLD EVALUATION AT AN OPERATING CONVENIENCE STORE

In this Section we present our implementation of *ISACS*, the experimentation setup, the metrics used to evaluate the performance of our multiple human to multiple product matching algorithm, as well as the natural experimentation results. We further demonstrate how the store setup (proper placement of the products/shelves) can affect the performance of the system, and the steps taken to address these issues.

5.1 System Implementation Designed for Retailers

We approached this deployment with the intention of ensuring that the store would be designed by retailers and deployed by retailers, so that we would observe the real life implications of *ISACS*. We have partnered with Carrefour SA to deploy *ISACS* in a real setting. Due to privacy limitations the store was deployed inside of Carrefour's Headquarters, in France, with access only to employees or visitors of the company. The employees were informed and gave consent to be recorded in this store.

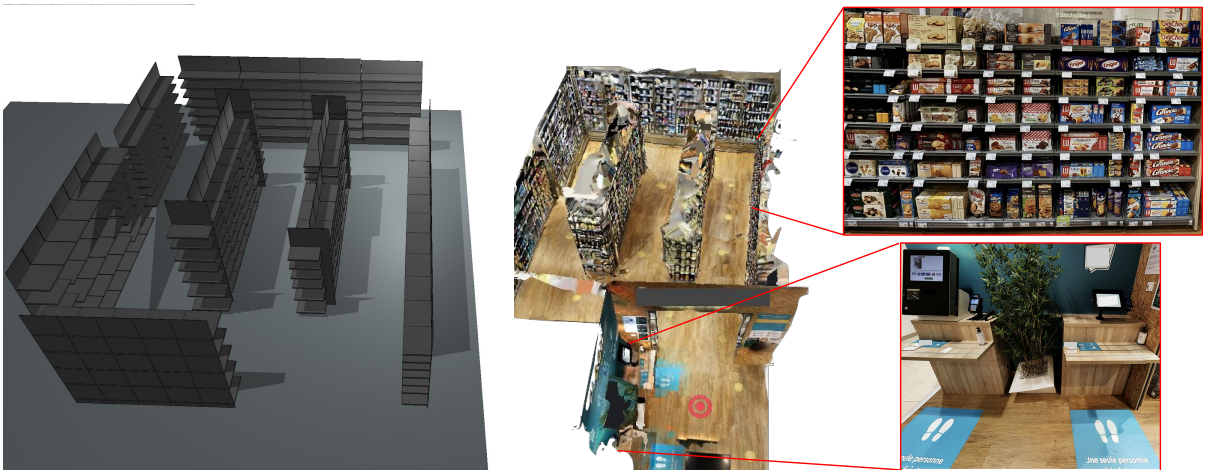


Fig. 6. 3D physical store model (left) and 3D rendering of the store, with a focus on the payment terminal and 3 gondolas showing the density, distribution and product assortment (right). This store has 52 gondolas, with 6 to 10 shelves each, 6 large refrigeration units with double open doors and 7 shelves each. These were decided and deployed by Carrefour which demonstrates the ability of *ISACS* to scale to a store level deployment.

The store was designed in line with Carrefour's convenience store format. This means the number of products and variety truly reflects a real store, as well as its disposition in the store.

5.1.1 Store Layout. The store has 52, 0.5 meter wide, gondolas with 6 to 10 shelves each. There are 6 large refrigeration units with double open doors and 6-7 shelves each. This store is divided into several sections: fresh food, condiments, alcoholic drinks, regular drinks, snacks, candy, chips and convenience items. (See Fig.6).

All shelves are instrumented with the weight sensors defined in Sec. 3. In the ceiling there are 52 IP cameras, laid out according to Sec.3. All cameras and sensors are connected via Ethernet to a cabinet that holds our servers. These servers are not connected to the internet, and only connect to the retailer's POS (point of sale), so that *ISACS* can provide the receipt to the payment terminal.

5.1.2 Store Entry and Exit. This is an open store setting, where there is no gate/door to enter the store, so anyone can get inside without any limitations. In order to present a receipt to a customer that shops in an autonomous store there has to be a place where a matching between the sensed person and the real person occurs. Therefore we have placed 2 payment terminals at the exit. One for self checkout, in case the person does not wish to be a part of the autonomous checkout experience, and a second one with a autonomous terminal. This terminal operates just like the self checkout providing similar payment methods, but automatically shows the receipt as the person approaches the terminal, using *ISACS*.

In our system we leverage the fact that customers are used to going to a checkout station at the end of their shopping journey to match the predicted receipt with their payment method. *ISACS* uses a virtual geo-fence around the checkout area to match the sensed customer with the real customer. It then presents the receipt in the payment terminal inside that area. This way the association is done at the moment of payment, reducing the need for any entry blocking mechanism. Similar to a self-checkout station where the customer does not need to scan the products. In most cases, when the cart is correct, the shopper chooses their method of payment (credit/debit card, loyalty card, apple/android pay, etc.), the shopper would then execute the payment and leave with the purchased items. In the cases where the predicted receipt was incorrect, the shopper would be able to

ask for help (there was always a person by the payment terminals to assist shoppers). The employee providing help would enter a correction mode, and scan any missing items, or remove any extra items. This correction is then registered with ISACS and used for accuracy measurements.

5.2 Allowed Customer Shopping Journeys

There is a vast list of shopping journeys that customers can take inside a store. From taking a cart, or basket, and filling it up with items, or just carrying them around. Coming in alone or with their family. These are just a few of the options possible. *ISACS* focuses on the journeys present in convenience stores which are the ones in which the customer most values speed. This means that each customer is tracked individually and is charged by the items that he, or she, picks up from the shelves and does not put them back on any shelf. Passing items between people is outside the setting of *ISACS*. This means that unless the customer leaves items in the store in a place that is not a sensored shelf (i.e. floor or give them to someone else) *ISACS* will correctly predict what items the customers are carrying out of the store. In this context people tend to either carry the objects in their hands or put than into a small bag.

5.3 Experimental Setting and Dataset Characteristics

Due to the nature of the operating store, data privacy and infrastructure requirements –approximately 1.5TB/operating day of local storage capacity– we have not been able to evaluate our most recent approaches against all of the experiment. However we recorded a total of 2 weeks of data and have evaluated our multi-people matching with multiple objects in its multiple versions. In these 2 weeks of data there were 1874 transactions, each with approximately 5 interactions per transaction. There are a total of 7840 pick ups and 2093 put backs. These are real shoppers entering the store purchasing products for their own use, so no instructions were given to them and the number of items, duration of the trip and behavior is completely natural.

For the same data privacy and store requirements reasons this paper does not address the impact of *customer education* throughout the length of the experiment. It was impossible for *ISACS* to track recurring customers given that every new customer that entered the store and became visible by the cameras was "initialized" as a new "anonymous track" without any association to an account. At the moment of payment, the "anonymous track" that stood in the geo-fenced area in front of the payment terminal was then sent to the retailer's payment system and no storing of the payment identity is ever made in the *ISACS* system.

This dataset was collected on the first 2 weeks of October 2020, using version 6 being operated (See Table1). Every interaction of the shoppers with the shelves was recorded and manually labelled with the following information: timestamp, which person interacted, how many people surrounding the event, distance of the people to the event, product(s) interacted with, quantity of products picked/put-down, basket prior to interaction, and final basket. The final basket was further validated with the POS information provided by the retailer.

During these two weeks the shoppers were not instructed any differently than before, leading to their natural behavior which was present for the totality of the experiment (13 months).

5.4 Multi-human to Multi-product Matching Metrics

In this paper we evaluate the accuracy of the matching algorithm between multiple people and multiple products. We define average matching accuracy as, Avg. Match Accuracy for short, as:

$$\text{Avg. Match Accuracy} = \frac{\text{\# correct person predicted}}{\text{\# events}}(\%) \tag{11}$$

Its complement, the average matching error, can then be easily defined as:

$$\text{Avg. Match Error} = \frac{\text{\# incorrect person predicted}}{\text{\# events}}(\%) \tag{12}$$

Furthermore *ISACS* performance is dependant on the behavior of the customer, this means that there is a performance difference in the case a customer does multiple sequential events or multiple customers do a single event. In case of matching errors the latter would affect multiple receipts however the first would only affect 1 receipt. Therefore we further measure accuracy by looking at the final receipt accuracy and final item accuracy as such:

$$\text{Avg. Receipt Accuracy} = \frac{\text{\# correct predicted receipts}}{\text{\# receipts}}(\%) \tag{13a}$$

$$\text{Avg. Item Accuracy} = \frac{\text{\# correct items receipts}}{\text{\# items purchased}}(\%) \tag{13b}$$

It is worth noting that the metrics in Equation 13 are not only dependent on the matching algorithm but are also affected by the accuracy of the Multi-Human 3D Pose Estimation and the Inventory Monitoring pipelines. This is therefore reflective of the overall *ISACS* framework performance. In the next subsection, we analyze our experiment results and the dependency of *ISACS*'s Avg. Match Accuracy on different customer behaviors.

Ground truth was collected through the POS (point of sale) of the store. Furthermore, in order to ensure the quality and accuracy of the data there was an employee at all times present at the POS verifying the shoppers transactions and educating the customers. Given there were only 2 registers at the exit, and the small average basket size in this store, it was possible to maintain reliable transaction data.

5.5 Experiments at a Convenience Store Level Deployment

In order to fully evaluate the system in a real-world setting, we recorded the "store level" shopping behavior described above, and present the results here. We want to understand how the density of people around a pickup and a put down and the distance of the people affect the matching performance for *ISACS*.

One of the parameters that affects the performance of the matching algorithm is the amount of people surrounding the event as it happens. We define the number of people surrounding a event as the number of people within a radius of 2 meters of the center of the bin being interacted with. In Figure 7a we can see that when a person picks up items the association with the *head* (in blue) of that person suffers the most with the increase in density around the event. This accuracy drops from 82.35%, when there are 2 people, to 40% when there are 5 people around. However when leveraging the hand keypoints for the matching (in yellow) we observe a smaller effect of the density and a clear increase in accuracy, from 92.11% with only 2 people to a 80% with 5. The effect of density also affects the ability to proper estimate the hand location using the 3D Multi-Human Pose Estimation as it increases the natural occlusions in a small area. Furthermore the moment an event occurs is when the hand of the customer is mostly occluded by the upper shelves as it is inside the bin. Because this phenomenon occurs almost in all shelves, except in the top ones, the Extruded Bins only (Section 4.2) approach yields a higher accuracy with 97.27% with only 2 people and 90% with the highest density observed, 5 people. During pick ups, the information provided by the inventory at hand does not add value to the decision, hence the accuracy observed with *ISACS* (in green) equals the Extruded Bins only approach (in red).

It is important to note that the number of events that occur at higher densities is much lower. There were only 10 pick ups and 5 put backs with 5 people around; and only 4 put backs with 4 people around. This happened during a demonstration of the store, and it is not a common occurrence in the dataset.

In Figure 7b we can see the equivalent metric for when people are putting items down. We can see that the accuracies follow an equivalent pattern to Fig. 7a, however in this case *ISACS*'s performance gets an accuracy increase due to the use of the inventory at hand (Section 4.3) as a another source of information for deciding the person. *ISACS*'s accuracy got to 100% in most cases except in the case of 3 people where it got to 90%. These

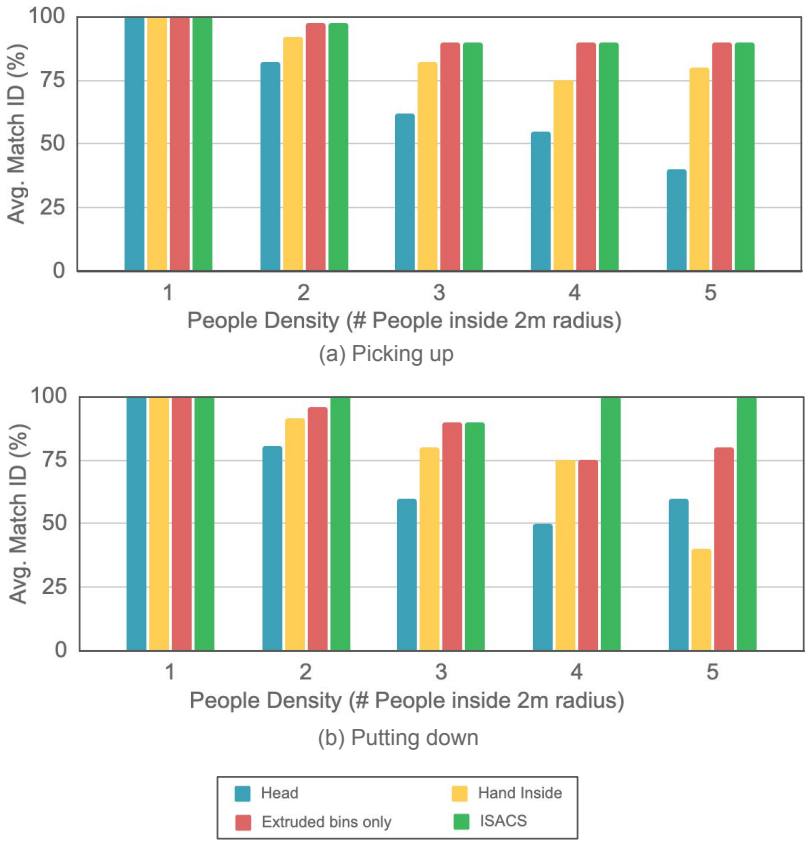


Fig. 7. Multi-human to Multi-product Matching accuracy for different people densities around a: (a) *pick up event*, (b) *put down event*. The number of people are counted within a radius of 2 meters around the bin location of the event. In (a) *ISACS*'s performance decreases with the increase in density. However in (b) *ISACS*'s performs better by leveraging the information about the inventory the people have.

failure cases occurred when 2 or 3 people are shopping together and pick the same products and then 1 of them puts them back. In this case the inventory at hand adds no value, leading to the same accuracy as the Extruded Bins approach.

Furthermore we observe on Figure 7b the *head* approach (in blue) performed at 60% while the *hands inside* only performed at 40%. This is observed only across 5 events. After investigation this occurs because in 1 case that hands of the person putting down the item are not seen by any angle when inside the shelf while the heads are clearly visible. The *Extruded bins* approach solves this situation by matching the right person slightly before the hand is occluded, preforming at 80% with 5 people around.

However it isn't only people density that affect the performance of the matching algorithm. The distance at which people stand from the event when multiple people are interacting can also significantly impact performance. As shown in Fig. 8a, **when multiple people interact with the shelves we consider the distance of the closest person to the event that *did not* perform the pick up or put down**. What we observe is that when people are very close to each other, closer than 50cm, we see a significant out-performance from *ISACS* (85% under 20cm and 87.4% between 20-50cm), when compared to the *head* (25% under 20cm and 77.3% between

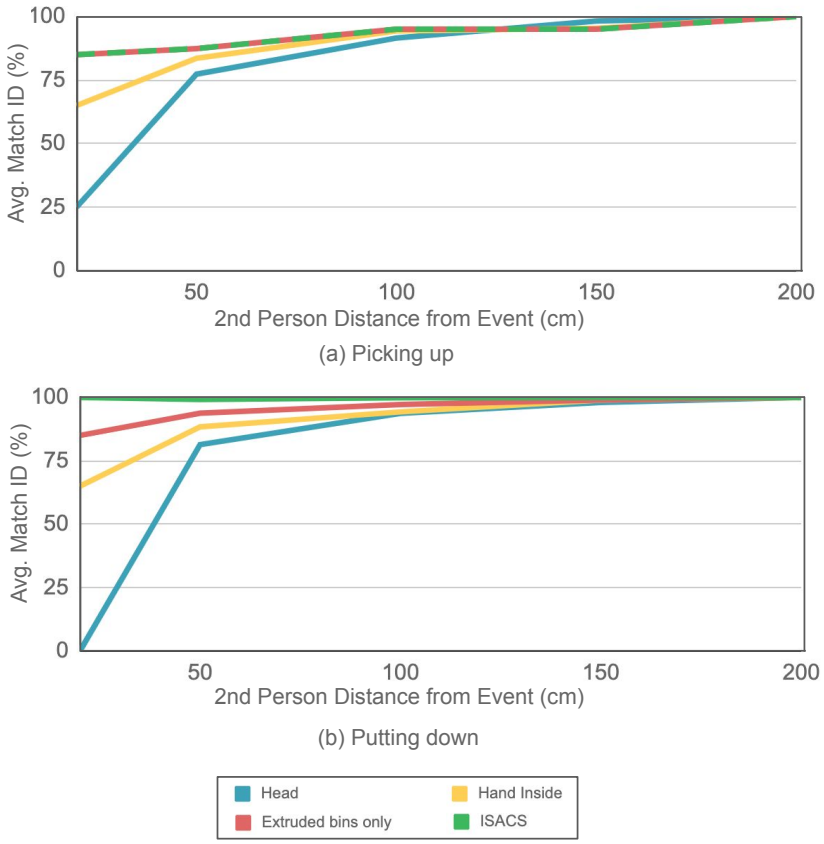


Fig. 8. Multi-human to Multi-product Matching accuracy for multiple people when: (a) *picking up*, (b) *putting down* close to each other. The distance is measured from the center of the *bin* being interacted with to the head of the closest person *not* interacting with the bin. In (a) *ISACS*'s accuracy increases as people get further apart. However in (b) *ISACS*'s accuracy is not affected by distance due to the added inventory information at the hand of the person.

20-50cm) and *hands inside* (65% under 20cm and 83.6% between 20-50cm) approach. During pick ups, there is no advantage of using the inventory information of what the person has in their hands (Section 4.3), therefore this approaches performance equals the one of the *Extruded bins only*.

On the other hand, during put downs, *ISACS*'s performance is not affected by the distance of the closest person, performing consistently above 99.2% (See Figure 8b).

6 LESSONS IN DEPLOYING STORES WITH REAL CUSTOMERS

We have deployed and evaluated *ISACS* in an operating convenience store covering 800 square feet with 1653 distinct products, and more than 20000 items in a store owned and operated by Carrefour SA, in Massy France.

Over the course of 13 months, we gathered data and experience that allowed us to iterate over the hardware and software of the deployment, to withstand the realities of an operating convenience store and improve the autonomous store setup for higher accuracy. In this section we describe the design decisions taken to reduce the impact of setup or operations, based on our deployment experience as well as the lessons learned through multiple iterations of this project.

(a) Consecutive shelves stacked too close to each other.

(b) Items overflowing to neighboring sensing plates.

(c) Items falling behind the shelf into neighboring sensing plates.

Fig. 9. Unexpected Challenges from real-world deployment due to normal retailer operations.

6.1 Unexpected *ISACS*'s Operation Challenges

When deploying *ISACS* we realized that the Item Weight Model was challenging to obtain. Getting the operators of the store to weight every single item was not feasible. So we have initially approximated the weight distribution of the items by weighing 3 items of each type and generating a distribution for each type. This worked well for pre-packaged goods. However fresh food, condiments and light weighted items became difficult to identify and count accurately due to a much **higher variance in the weight distribution**. In order to accurately count and predict these items we have manually identified these *high variance* items and changed their weight distribution model accordingly. Furthermore we have trained the operator to identify these items and label them as such. Up until version 5 *ISACS* had a fixed variance threshold applied to every *Weight Change Bin Detection* (described in Sec. 3.2). After the first COVID lockdown, the *high variance* issue started to became more apparent given that the retailer started to sell more items that had higher variance, such as cups with pieces of cut fruit. These quickly became the most selling items, severely impacting the accuracy of the system.The ability to set a particular product as a *high variance* product was introduced to the restocking application in version 6 (See Table 1). This meant that employees prior to restocking would define products that had higher than 10% variance across units as *high variance* products.

Another requirement of *ISACS* is an understanding of the placement/location of the products in each shelf ($\mathcal{L}$, Section 3.4). This information is commonly available for large retailer which standardize their store layouts for store operations efficiency. However in this experimental store we were faced with a **continuously changing items' layout**. We observed fluctuations, in version 1 & 2 (See Table 1 and Fig. 10), of the accuracy of *ISACS* due to the changing of the items' layout without the respective change of the location model, $\mathcal{L}$ model in the system. This was solved by training the operator to change the $\mathcal{L}$ model by using a simple mobile application that allowed them to switch out a product, or place it in a different shelf by simply scanning it on the phone. The mobile application used did not change the system's performance. Ultimately "synchronization" between the location model, $\mathcal{L}$, and the physical reality was the impacting factor to the system. Once trained, the restocking employees were able to ensure an accurate sync between $\mathcal{L}$ and reality.

As time passes and the employees operate the store (restock and change product locations) more regularly we observed another decrease in accuracy. This was due to fact that the employees in order to increase the number of items available for sale, increasing the density, started to hit certain limitations of the system. Such as, placing items so close to each other that the physical separators between *bins* would actually not prevent **the object from laying partial weight on neighboring sensing plates** (See Figs. 9b-9c). This lead to incorrect weight measurement and inaccurate item counting. Further, the proximity of the shelves, in height, was decided based

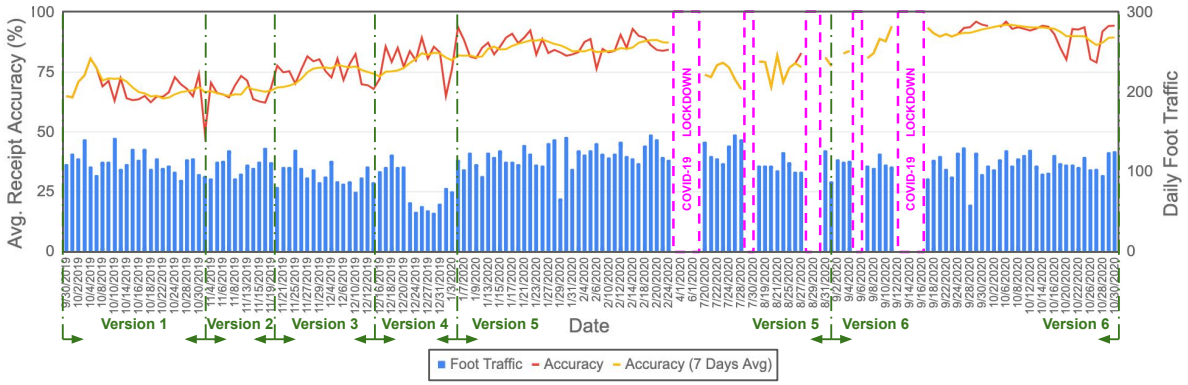


Fig. 10. Receipt accuracy across 1 year of *ISACS*'s deployment. The store was locked down due to COVID-19 from 24-Feb-2020 up to 20-Jul-2020, as well as, 4 smaller periods representing in light purple. *ISACS*'s achieved up to 96.4% Receipt accuracy over the lifetime of the deployment.

on the minimum distance possible so that the customer can see the product but also the retailer can fit as many as possibly vertically, notice how vertically *snug* the objects are in Fig. 6. This lead to situations where objects ended up *squeezed* in between 2 shelves, leading to a **increase in weight measurement due to the weight of the upper shelf** (See 9a). These issues were resolved with proper employee training and monitoring. You can observe these accuracy swings and steadily increases as we address the issues in Fig. 10.

Finally, *ISACS* does not take into account the restocking procedures. This initially meant that as an employee would restock the store, the system would be tracking them attempting to understand what they would pick up and put back. However employees behavior is far from a normal shopper: employees pick up a full shelf worth of goods to look at the expiration date, put back items that are not yet in the system to be monitored and also re-arrange the products orientation in the shelf to have the branding facing the client in a neat way. For the purpose of measuring the system we removed the carts generated for the employees.

6.2 Uncontrolled Customer Behavior

Ultimately the biggest challenge that *ISACS* faces is the uncontrolled behavior of the customers. There is a particularly challenging effect when deploying such a store: **the curiosity effect**. When customers are faced with this store for the first time they try to understand it, by stressing the system to the limits. **Pressing the sensors up and down, trying to hide the objects, running through the store, shopping extremely close to each other are just a few of the example behaviors we have observed.** This effect takes place more strongly in the beginning of the experiment. As time passes and the customers get used to the experience we have observed that those behaviors almost completely disappear.

In order to understand the impact of the accuracy presented in Section 5 we have measured the behavior of the shoppers, across the following dimensions:

- Average number of products in a receipt
- Average density of shoppers when interacting with a shelf
- Average distance of close-by shoppers to the interacted shelf

These variables affect *ISACS* performance both in people matching as well as inventory detection and counting. Figure 11 shows the number of people present –density– when an event (pick up/put down) occurs. We can observe that **most events occur with under 3 people close-by**. Figure 12a presents the distance of the close-by shoppers when an event occurs. This is showing that if an event has more than 1 person close-by, that person

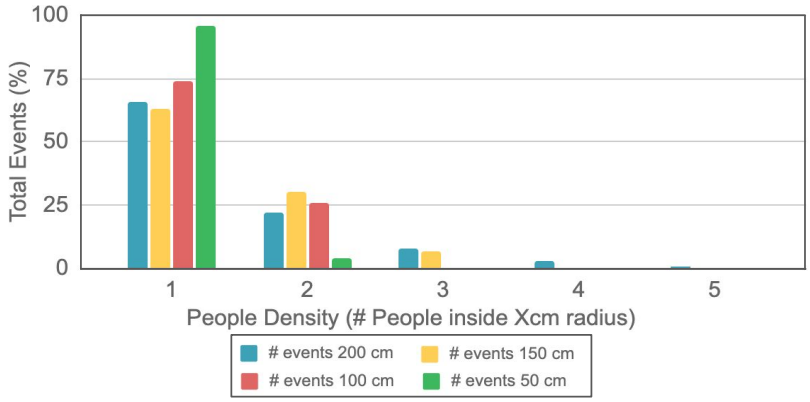
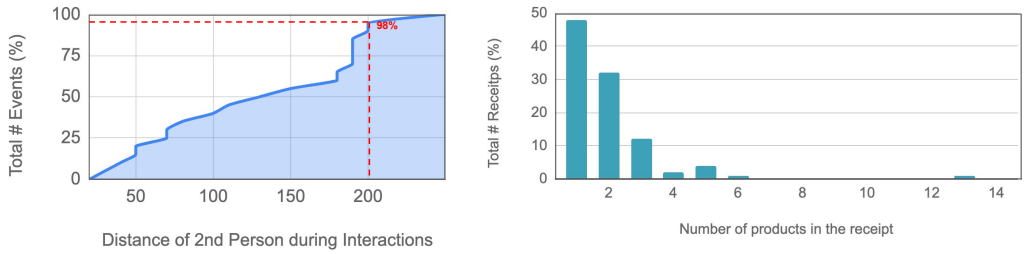


Fig. 11. People density around a pick up or put down event during 1 year of operating *ISACS*. People density is measured as the % of events found with people within 200, 150, 100 and 50 cm (each color add up to 100%). Note that most interactions are with a single person, up to a maximum of 5 people within 2 meters.



(a) (b)

Fig. 12. Deployment statistics of a 1 year real world operations. (a) Average distance between people when multiple people interact with the shelves at the same time (98% of the multiple people interaction happen under 2 meters). (b) Number of products per receipt. Average number of products taken is 1.83

will be no further than 2 meters and as close as 20 centimeters. Notice the red line showing the threshold picked for 3, where people outside the radius of *2 meters* are not considered to be matched by *ISACS*.

6.3 Hardware Resilience

There are three types of hardware equipment that need to operate continuously in order to for *ISACS* to perform at its best: cameras, weight sensors and servers. While cameras and servers are quite a mature product easily available and maintainable the weight sensing shelves are not.

6.3.1 Weight Sensing Shelves. This store contained 580 sensing plates, each of them connected to a control box, at the bottom of each gondola, which in turn was connected to the local network via Ethernet.

Given the lack of maturity of these custom sensing shelves, we faced some initial challenges in operating them. Particularly we observed that some shelves mistakenly would swap the sensing plate id with a different

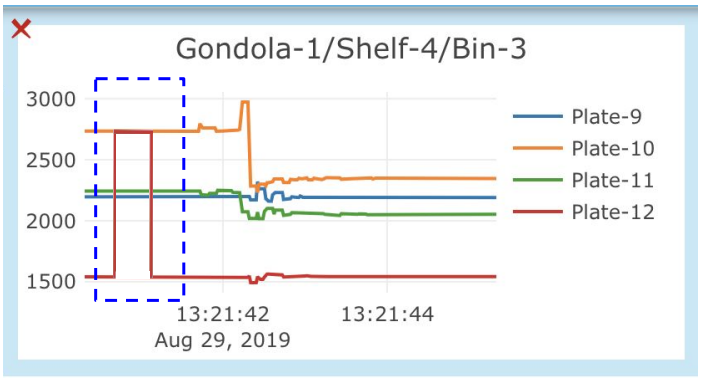


Fig. 13. *ISACS*'s weight sensor signal demonstrating the swapping of plate IDs being reported. Notice the inside the blue dashed box that plate 12 (in red) jump to the same line of plate 10 (in orange). This occurs because plate 10's id started to report as if it was plate 12's causing incorrect weight changes detected.

plate, for a few seconds, and then return to the original id (See Fig. 13). This would occur more often inside the refrigeration units given **the sensitivity of shelves to the temperature fluctuations**. The result of this issue would be 'ghost' weight detections of the products siting on those sensing plates. If by chance a customer was close-by he or she would get those objects attributed to them. This was observed before version 5 (See Table 1). This issue was caused by a design flaw in the sensing HW, where the identifier of the plate *–plateID– was being measured as an analog pin on the controller for each plate, based on a combination of resistor dividers. This meant that depending on the tolerances of the resistors the* plateID *would change if the resulting voltage would dip, or spike, due to the change in resistance caused by temperature effects.*

Upon realization of this issue, 2 solutions were developed: the *Max distance thresholding* –Sec. 3, and a packet reconstruction technique. The first approach removes any association of a person who would be too far to actually have been able to pick the products, while the second one would attempt to revert the ID swap before the signal reached the Weight Change detection module. This was accomplished by observing the continuity of the incoming signals and matching a instantaneous discontinuity with another discontinuity on sensing plate of the same shelf. These solutions were applied in Version 5 (See Table 1). A more proper solution, but highly costly, would have been to redesign the sensing HW to store its internal ID digitally and not be subject to environmental factors, and subsequently swap all shelves.

Furthermore, during the first 3 weeks of the deployment, prior to opening of the store, the shelves required significant attention and fixing due to the transport impact on the shelves. Many internal connections were loose, and some control boxes were not consistently working. However after the first iteration of shelf hardware fixing, these shelves remained consistently working for the following years with very minimal maintenance.

6.3.2 Camera Selection and Placement. In placing cameras we have to consider 2 main aspects: people tracking and people interaction with shelves. For people tracking it is best to place cameras in the ceiling, given this is the place with the most unobstructed view into the people in the store. An interesting trade off to consider when instrumenting an autonomous store is the camera specification, that is, which king of visual sensor to use. Whether a structure light sensor,a time of flight sensor, stereo sensors or a simple 2D RGB sensor. With the recent advances in computer vision 2D RGB sensors are becoming more and more powerful. Coupling this with their ease of use and accessibility we have decided to use them for our application.

The cameras setup of this store went through 5 iterations as we experimented with the best performing setup. As detailed in 4 *ISACS* uses the head and hands of the person to make its matching prediction between the

Table 1. Version history of *ISACS* throughout the 13 months of deployment, showing how multiple approaches of camera selections and the equivalent software changes required to accommodate the hardware changes.

ISACS	Hardware	Software	Date Start	Date End
Version 1	34 RGB cameras	Head keypoint tracking	09-30-2019	10-30-2019
Version 2	v1 + 15 Top Down Intel RealSense D435	v1 + Hand Depth Processing	10-30-2019	11-19-2019
Version 3	Angled 34 RGB	v1 + Keypoint tracking with Hand prediction	11-19-2019	12-12-2019
Version 4	52 RGB Cameras 34/15 (Angled/Top down)	v3 + Top Down RGB Keypoint tracking for hands	12-12-2019	1-5-2020
Version 5	v4 + plateID reconstruction	v4 + event distance threshold	1-5-2020	8-31-2020
Version 6	Version 5	v5 + High variance items	8-31-2020	10-30-2020

customers and the products. With the advancements in the state of the art on people tracking we upgraded the camera setup throughout the following versions:

Version 1 - Camera Setup. Initially we deployed cameras with that would cover every volume of the store, when empty, with at least 3 angles. This was a requirement for tracking people accurately through the triangulation technique present in [6]. This resulted in 34 cameras spread out throughout the ceiling of the store.

Version 2 - Depth and RGB Setup. As we started to observe the need for better matching between people and products, we realized the need for better understanding of the movement of the hands of the customers. Therefore we deployed 2 types of cameras: depth cameras –Intel RealSense d435–, and regular RGB cameras. The depth cameras were intended to enhance the hand position of the person as they interacted with the shelves. These were placed top-down in the ceiling aiming just in front of each gondola, in order to have the most unobstructed view of the place of interaction with the shelf. While the others followed the principle layed out in *Version 1.*

Version 3 - Angled RGB Camera Setup. Given the added complexity in depth cameras –added computation, network requirements, physical device stability– we decided to remove them and compute the hand location based solely on RGB 2D images. This was achieved using the approach in [6] with cameras placed on the ceiling angled between 20-70 degrees. This angled produced the required side images of people to which the models available were trained with.

Version 4 - Top Down and Angled RGB Camera Setup. We finally reached our last version by adding top down RGB cameras to our system, given that better models and more available data emerged and allowed the tracking modules to perform better with those angles, giving an particularly useful point of view. Top down cameras are the most useful to understand *who* is interacting with a shelf given that they suffer the least from occlusions.

6.3.3 Servers. *ISACS* runs on a server with 4 NVidia Tesla T4 GPUs. These hardware accelerator cards are used for running human detection from all of the available cameras in the store. In order to process the required frames for Multi-Human 3D tracking the GPUs are fully occupied. These servers stay in a cabinet inside the store, to allow for very low latency and provide the receipts instantly as the customer approaches the payment terminal. Due to the stability of server technology, these servers did not require any maintenance during the entire duration of the experiment.

6.4 *ISACS* Iterative Deployment Process

We have iterated through several approaches during the length of the experiment. These can be seen in Table 1 and Figure 10. This figure shows the impact of the changes in approach in the accuracy of the receipts.

It is worth noting that the daily fluctuation of accuracy occurs due to the uncontrolled nature of the human behavior, both from the employees as well as from the customers. When the store setup was properly re-mapped ($\mathcal{L}$, correctly estimated), or a new approach was deployed, we can only confidently observe the improvements over a multi day trend. The orange line represents the 7 day trailing accuracy. In this line we can observe the upwards trend of the accuracy given every new iteration of the approach.

The process of upgrading the system was lead by the previous versions' results. Each version would run for a minimum of 1 month to validate the accuracy impact of that version's changes. Any upgrades to the system were done during closed hours of the store. This meant that the deployment process would not impact the normal operations of the store.

Over the course of 1 year of operation, *ISACS* achieved a receipt daily accuracy of up to 96.4%. Which translates to a 3.5x improvement over prior reported self-checkout accuracies.

7 PRIVATIZED RETAIL SYSTEMS DISCUSSION

Amazon has shown to the world how autonomous stores can look like through their Amazon's Go technology. These stores provide an opportunity for the customer to shop in the store by downloading the Amazon application scanning the app to enter and "just walk out", with their receipt arriving within the next few hours [12].

While this has validated the customer experience of "just walking out" it unfortunately has been all done under closed doors, preventing the broader scientific community from leveraging the learnings of such stores and expanding more broadly beyond Amazon. Furthermore there are several media articles speculating how Amazon operates their Go's stores [46]. These [35] claim that Humans are leveraged in the cloud to perform the final checkout for the customer, essentially displacing the cashier workload to a remote one.

Furthermore, Amazon Go's stores favors a particular demographic group: upper-class, tech-savvy, and younger generation. The requirement of an application to enter and shop and the delayed receipt upon exit becomes quite restrictive for a more broad community to adopt. Shoppers who are more budget conscious usually come with the mindset of how much they will spend for their shopping and require validation before leaving the store. American consumer patterns are changing and discount retailers are seeing an increased adoption from a larger population sample which is more budget conscious [52]. To serve the broader community *ISACS*'s presents an in-store receipt, despite the delayed receipt by Amazon. This is driven by the fact that in-store discounts and the "smart-shopper feelings" towards pricing act as a major component of the emotional response affecting shoppers' behavior to favor in-store price confirmation [8, 34, 44].

8 CONCLUSION

In this paper, we presented *ISACS* (In-Store Autonomous Checkout System) for Retail. Utilizing a centimeter accurate physical model of the store, multi-human 3D pose estimation and live inventory monitoring through weight sensors, *ISACS* is able to match multiple people interacting with multiple products, maintain an accurate store setup and provide a receipt within 2 seconds. *ISACS* has a receipt accuracy of up to 96.4%, without a human in the loop, which is a 3.5x reduction in error compared to the 86% accuracy reported for self-checkout stations.

This paper demonstrates the performance of *ISACS* in a convenience store where most transactions are done at an individual level. Therefore, the assumption present in the paper of: *the last customer who removes a product form a store's shelf and leaves with it is the customer charged for that product* is reasonable. However, it would be possible to relax this constraint by focusing on events that occur in the entire store (i.e., floor or between people), rather than solely on the shelves. This however presents an added challenge of where to focus the sensors and cameras in order to properly identify these events, something that we intend to address in our future work.

ISACS is the first fully autonomous system deployed in a convenience sized store that leverages computer vision and sensors to provide a in-store receipt, without relying on humans-in-the-loop.

ACKNOWLEDGMENTS

This work is partially supported with funding from AiFi Inc. and partially supported by the National Science Foundation (under grant NSF-CMMI-2026699). In addition, we like to thank AiFi Inc and Carrefour for their help in deploying and running the store during the deployment.

REFERENCES

[1] Amazon. 2019. *Amazon.com: Amazon Go.* https://www.amazon.com/b?node=16008589011 Accessed: 2019-04-10.
[2] ECR An. 2018. Self-checkout in Retail: Measuring the Loss. (2018).
[3] Gwon Hwan An, Siyeong Lee, Min-Woo Seo, Kugjin Yun, Won-Sik Cheong, and Suk-Ju Kang. 2018. Charuco board-based omnidirectional camera calibration method. *Electronics* 7, 12 (2018), 421.
[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *arXiv preprint arXiv:1812.08008* (2018).
[5] Marco Carraro, Matteo Munaro, Jeff Burke, and Emanuele Menegatti. 2018. Real-time marker-less multi-person 3D pose estimation in RGB-Depth camera networks. In *International Conference on Intelligent Autonomous Systems*. Springer, 534–545.
[6] Long Chen, Haizhou Ai, Rui Chen, Zijie Zhuang, and Shuang Liu. 2020. Cross-View Tracking for Multi-Human 3D Pose Estimation at over 100 FPS. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3279–3288.
[7] McKinsey & Company. 2020. *Redefining value and affordability in retailâĂŹs next normal.* https://www.mckinsey.com/industries/retail/our-insights/redefining-value-and-affordability-in-retails-next-normal Accessed: 2021-05-15.
[8] B. C. Cotton and Emerson M. Babb. 1978. Consumer Response to Promotional Deals. *Journal of Marketing* 42, 3 (1978), 109–113. http://www.jstor.org/stable/1250544
[9] Statista Research Department. 2020. *Retail e-commerce sales worldwide from 2014 to 2023(in billion U.S. dollars).* https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/ Accessed: 2020-11-15.
[10] Statista Research Department. 2020. *Total retail sales worldwide from 2018 to 2022(in trillion U.S. dollars).* https://www.statista.com/statistics/443522/global-retail-sales/ Accessed: 2020-11-15.
[11] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. 2019. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7792–7801.
[12] Meiling Du. 2018. *Examining the User Experience of Amazon Go Shopping âĂŤ Just Walk Out.* https://blog.prototypr.io/examining-the-user-experience-of-amazon-go-shopping-just-walk-out-bfeda3c9ce39 Accessed: 2019-11-15.
[13] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. 2019. CenterNet: Keypoint Triplets for Object Detection. In *The IEEE International Conference on Computer Vision (ICCV)*.
[14] João Falcão, Carlos Ruiz, Shijia Pan, Hae Young Noh, and Pei Zhang. 2020. FAIM: Vision and Weight Sensing Fusion Framework for Autonomous Inventory Monitoring in Convenience Stores. *Frontiers in Built Environment* 6 (2020), 175.
[15] Frida Femling, Adam Olsson, and Fernando Alonso-Fernandez. 2018. Fruit and vegetable identification using machine learning for retail applications. In *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 9–15.
[16] Teresa Fernandes and Rui Pedroso. 2017. The effect of self-checkout quality on customer satisfaction and repatronage in a retail context. *Service Business* 11, 1 (2017), 69–92.
[17] Forrester. 2018. *Consumers Cringe At Slow Checkout.* https://www.digimarc.com/docs/default-source/digimarc-resources/forrester-study.pdf Accessed: 2019-11-15.
[18] Petia Georgieva and Pei Zhang. 2020. Optical Character Recognition for Autonomous Stores. In *2020 IEEE 10th International Conference on Intelligent Systems (IS)*. IEEE, 69–75.
[19] Grabango. 2016. *Eliminate Lines and Save People Time.* https://grabango.com/ Accessed: 2019-04-10.
[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.
[21] Tomáš Hodaň, Xenophon Zabulis, Manolis Lourakis, Štěpán Obdržálek, and Jiří Matas. 2015. Detection and fine 3D pose estimation of texture-less objects in RGB-D images. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4421–4428.
[22] Mahdi Hussain. 2019. *Convenience Store Loss Prevention - The Complete Guide.* https://petrooutlet.com/blog/posts/convenience-store-loss-prevention-guide/ Accessed: 2019-11-15.
[23] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. 2017. Panoptic studio: A massively multiview system for social interaction capture. *IEEE transactions on pattern analysis and machine intelligence* 41, 1 (2017), 190–204.
[24] Ellen Klingbeil, Deepak Rao, Blake Carpenter, Varun Ganapathi, Andrew Y Ng, and Oussama Khatib. 2011. Grasping with application to an autonomous checkout robot. In *2011 IEEE international conference on robotics and automation*. IEEE, 2837–2844.

[25] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. 2020. FoveaBox: Beyound Anchor-Based Object Detection. *IEEE Transactions on Image Processing* 29 (2020), 7389–7398.

[26] Ronald B Larson. 2019. Supermarket self-checkout usage in the United States. *Services Marketing Quarterly* 40, 2 (2019), 141–156.

[27] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. 2018. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 683–698.

[28] Lizheng Liu, Bo Zhou, Zhuo Zou, Shih-Ching Yeh, and Lirong Zheng. 2018. A smart unstaffed retail shop based on artificial intelligence and IoT. In *2018 IEEE 23rd International workshop on computer aided modeling and design of communication links and networks (CAMAD)*. IEEE, 1–4.

[29] Ming-Yu Liu, Oncel Tuzel, Ashok Veeraraghavan, Yuichi Taguchi, Tim K Marks, and Rama Chellappa. 2012. Fast object localization and pose estimation in heavy clutter for robotic bin picking. *The International Journal of Robotics Research* 31, 8 (2012), 951–973.

[30] Xiaobai Liu. 2016. Multi-view 3D human tracking in crowded scenes. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 3553–3559.

[31] Xiaochen Liu, Yurong Jiang, Kyu-Han Kim, and Ramesh Govindan. 2020. Grab: Fast and Accurate Sensor Processing for Cashier-Free Shopping. *arXiv preprint arXiv:2001.01033* (2020).

[32] Mostafa Mirshekari, Jonathon Fagert, Shijia Pan, Pei Zhang, and Hae Young Noh. 2020. Step-Level Occupant Detection across Different Structures through Footstep-Induced Floor Vibration Using Model Transfer. *Journal of Engineering Mechanics* 146, 3 (2020), 04019137.

[33] Mostafa Mirshekari, Shijia Pan, Jonathon Fagert, Eve M Schooler, Pei Zhang, and Hae Young Noh. 2018. Occupant localization using footstep-induced structural vibration. *Mechanical Systems and Signal Processing* 112 (2018), 77–97.

[34] Bryan (Forbes) Pearson. 2017. *How Retailers Can Maximize The Power of Coupons.* https://www.forbes.com/sites/bryanpearson/2017/03/15/research-reveals-how-retailers-can-maximize-the-power-of-coupons/?sh=4873e6532f01 Accessed: 2019-11-15.

[35] Recode. 2019. *Amazon's store of the future has no cashiers, but humans are watching from behind the scenes.* https://www.recode.net/2017/1/6/14189880/amazon-go-convenience-store-computer-vision-humans Accessed: 2019-04-10.

[36] Yacine Rekik, Evren Sahin, and Yves Dallery. 2008. Analysis of the impact of the RFID technology on reducing product misplacement errors at retail stores. *International Journal of Production Economics* 112, 1 (2008), 264–278.

[37] George Roussos. 2006. Enabling RFID in retail. *Computer* 39, 3 (2006), 25–30.

[38] Carlos Ruiz, Joao Falcao, Shijia Pan, Hae Young Noh, and Pei Zhang. 2019. AIM3S: Autonomous Inventory Monitoring through Multi-Modal Sensing for Cashier-Less Convenience Stores. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 135–144.

[39] Carlos Ruiz, Joao Falcao, Shijia Pan, Hae Young Noh, and Pei Zhang. 2019. Autonomous Inventory Monitoring through Multi-Modal Sensing (AIM3S) for Cashier-Less Stores. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 395–396.

[40] Carlos Ruiz, Joao Falcao, and Pei Zhang. 2019. AutoTag: visual domain adaptation for autonomous retail stores through multi-modal sensing. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 518–523.

[41] Carlos Ruiz, Shijia Pan, Adeola Bannis, Ming-Po Chang, Hae Young Noh, and Pei Zhang. 2020. IDIoT: Towards Ubiquitous Identification of IoT Devices through Visual and Inertial Orientation Matching During Human Activity. In *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 40–52.

[42] Bikash Santra and Dipti Prasad Mukherjee. 2019. A comprehensive survey on computer vision based approaches for automatic identification of products in retail store. *Image and Vision Computing* 86 (2019), 45–63.

[43] Silvio Savarese and Li Fei-Fei. 2007. 3D generic object categorization, localization and pose estimation. In *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 1–8.

[44] Robert M Schindler. 1989. The excitement of getting a bargain: some hypotheses concerning the origins and effects of smart-shopper feelings. *ACR North American Advances* (1989).

[45] Zheng Tang, Renshu Gu, and Jenq-Neng Hwang. 2018. Joint multi-view people tracking and pose estimation for 3D scene reconstruction. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[46] TechCrunch. 2019. *Inside Amazon's surveillance-powered, no-checkout convenience store.* https://techcrunch.com/2018/01/21/inside-amazons-surveillance-powered-no-checkout-convenience-store/ Accessed: 2019-04-10.

[47] Ervin Teng, João Diogo Falcão, Rui Huang, and Bob Iannucci. 2018. Clickbait: click-based accelerated incremental training of convolutional neural networks. In *2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. IEEE, 1–12.

[48] Malvina Vega. 2020. *15+ Retail Statistics 2020 and The Future of Shopping.* https://review42.com/retail-statistics/ Accessed: 2019-11-15.

[49] Denis Vuckovac, Pascal Fritzen, Klaus Ludwig Fuchs, and Alexander Ilic. 2017. From Shopping aids to fully autonomous mobile self-checkouts-a field study in retail. (2017).

[50] Cheng Wang, Ming Cheng, Ferdous Sohel, Mohammed Bennamoun, and Jonathan Li. 2019. NormalNet: A voxel-based CNN for 3D object classification and retrieval. *Neurocomputing* 323 (2019), 139–147.

[51] Wenyong Wang, Yongcheng Cui, Guangshun Li, Chuntao Jiang, and Song Deng. 2020. A self-attention-based destruction and construction learning fine-grained image classification method for retail product recognition. *Neural Computing and Applications* 32, 18 (2020), 14613–14622.
[52] Cale G. Weissman. 2019. *Why discount stores are one of the fastest growing retail sectors.* https://www.modernretail.co/retailers/why-discount-stores-are-one-of-the-fastest-growing-retail-sectors/ Accessed: 2021-05-15.
[53] Jian Zhang, Yibo Lyu, Thaddeus Roppel, Justin Patton, and CP Senthilkumar. 2016. Mobile robot for retail inventory using RFID. In *2016 IEEE international conference on Industrial technology (ICIT)*. IEEE, 101–106.
[54] Zippin. 2020. *Zippin. Checkout-free technology.* https://www.getzippin.com/ Accessed: 2019-04-10.