IRIS: Tapping Wearable Sensing to Capture In-Store Retail Insights on Shoppers

Meera Radhakrishnan[1*] , Sharanya Eswaran[2], Archan Misra[1], Deepthi Chander[*], Koustuv Dasgupta[2]
[1] School of Information Systems,Singapore Management University,
[2] Xerox Research Center India

***Abstract*—We investigate the possibility of using a combination of a smartphone and a smartwatch, carried by a shopper, to get insights into the shopper's behavior inside a retail store. The proposed *IRIS* framework uses standard locomotive and gestural micro-activities as building blocks to define novel composite features that help classify different facets of a shopper's interaction/experience with individual items, as well as attributes of the overall shopping episode or the store. Besides defining such novel features, *IRIS* builds a novel segmentation algorithm, which partitions the duration of an entire shopping episode into atomic item-level interactions, by using a combination of feature-based landmarking, change point detection and variable-order HMM-based sequence prediction. Experiments with 50 real-life grocery shopping episodes, collected from 25 shoppers, we show that *IRIS* can demarcate item-level interactions with an accuracy of approx. 91%, and subsequently characterize item-and-episode level shopper behavior with accuracies of over 90%.**

I. Introduction

Faced with increasing online competition, retail store owners are increasingly interested in the ability to better understand the browsing behaviors and intentions of consumers inside their physical stores. A variety of technologies, such as Wi-Fi and BLE beacon based aisle-level location tracking [11], RFID based asset monitoring [10] and smartglass-based browsing monitoring [8] have been explored to capture such individual and collective in-store behavior. While these advanced technologies hold great promise, their cost makes them unlikely to be adopted widely, especially in low-margin, emerging economy markets (such as India, China or Brazil) in the near future.

In our view, solutions for capturing latent in-store individual behavior become much more practical if they can work without requiring infrastructure support, such as Wi-Fi APs, BLE beacons or in-store cameras. Accordingly, our work in this paper is motivated by the following question: *"What level of individual consumer behavior inside a retail store can we reliably infer, by appropriately mining the sensor data from readily-available personal smartphone & smartwatch devices, without requiring ANY store-level infrastructural support"?*

Driven by this objective, this paper presents our initial work on *IRIS* (**I**n-store **R**etail **Insights** on **S**hopper), an infrastructure-oblivious, mobile-cum-wearable based framework for in-store behavioral analytics of shoppers. *IRIS* is motivated by two key hypotheses: (i) A significant fraction of in-store browsing activities involve gestural interactions with objects of interest (such as picking up an item in a grocery store, retrieving and draping on a dress in a clothing store or having a coffee in the middle of a shopping episode), that a wrist-worn smartwatch should help capture; and (ii) A consumer's interest-level or familiarity level with objects of interest will also be manifested in macroscopic locomotion-related features (e.g., how long a person stood stationary in front of a product), that a smartphone can help sense. Accordingly, we believe that a combination of smartphone & smartwatch sensor data can provide unique, hitherto unexplored, behavioral insights about a consumer's in-store behavior.

*Work done while at Xerox Research Center India

Fig. 1: Typical sequence of shopper activities in a grocery store

More specifically, in this paper, we explore the use of the *IRIS* framework to understand different aspects of individual-level behavior inside *retail grocery stores*. A key contribution of our research lies in appropriately *decomposing* an entire store visit (called a "shopping episode") into a series of modular and *hierarchical* individual interactions, such as a sequence of "in-aisle" durations, interspersed with "non-aisle" activities. Each "in-aisle" segment can consist of one or more product-interaction activities, such as "picking up item" (**P**), "putting item in trolley (cart)" (**T**), or "putting item back in the aisle" (**B**). Figure 1 visually illustrates such a decomposition. This decomposition is crucial because it not only helps define the specific atomic "activities" for which we seek to extract discriminatory features and build classifiers, but also helps to conceptualize two different levels of individual-level behavior (these will be further detailed in Section III).

The insights provided by IRIS can enable new applications such as: (a) *targeted advertising:* e.g., promotions of newly launched products preferentially pushed to shoppers whose prior browsing behavior indicates a propensity to look for unfamiliar products (so-called diversity-seeking behavior); (b)

978-1-4673-8779-8/16/$31.00 ©2016 IEEE

proactive retail help: e.g., a shop assistant directed to assist the customers who exhibit an "undecided" purchase pattern (an unusually high number of items picked up from, but then returned to, the shelves); or (c) *crowdsourced store profiling: IRIS* can be built as a 3^{rd}-party mobile App, as it does not have any interaction with the store's IT infrastructure. Accordingly, crowdsourced data from a pool of shoppers using *IRIS* can be used to build typical "experience profiles" associated with the store, for use in recommendation applications.

Key Challenges & Research Questions: *IRIS'* broad goals require us to address several research questions: (a) *Shopping Interaction Recognition:* Given sensor data corresponding to a specific shopping gesture (e.g., putting an item in the cart), what discriminative features help us identify such gestures? What level of accuracy for individual-level gesture accuracy can we achieve, by intelligently combining sensor data from both smartphones and smartwatches? (b) *Accurate Episode Segmentation:* Given that a shopping episode can consist of a shopper's interaction with multiple items, and movement across multiple aisles, how do we take the sensor data for the entire episode duration and then reliably segment it into individual interaction instances (such as in Figure 1)? What are the errors in demarcating the (start, end) times of such individual interactions? (c) *Connecting Interaction-Level Observations to Overall Behavior:* Assuming that we can infer the individual-level interactions of a shopper (i.e., how many items the shopper placed in her cart, etc.), how reliably can we use such inferences to classify the overall episode-level behavioral attributes (such as whether a shopper was in a hurry or not)? Can such classification be person-independent, or do shoppers behave differently enough to warrant person-specific classifiers?

In this paper, we address these questions, by utilizing a fairly extensive set of user studies (detailed in Section IV), involving 50 distinct shopping episodes, collected from 25 individuals, across 2 different mid-sized retail grocery stores in Bangalore, India. Based on this real-world data, we make the following **key contributions:**

(i) Robust and Accurate Segmentation: We develop a novel, hierarchical segmentation algorithm to accurately delineate the (start, end) times of different item-level interaction gestures, and aisle vs. non-aisle movements, over the entire duration of a store visit. Our proposed segmentation algorithm first utilizes locomotive features to separate 'in-aisle' vs. 'non-aisle' durations (a shopper performs item-level interactions only when in an aisle), and then uses a combination of change-point detection and a lookahead-augmented Viterbi decoding process to leverage the inherent sequential nature of item-level interactions during a shopping episode (such as a P always preceding a B), and thereby identify the best *sequence* of {P,T,B} gestures (and their start and end times) embedded within an in-aisle duration. We show that this technique is both *robust* (any mis-classifications never cascade beyond the current aisle) and *accurate* (it identifies gesture start and end times with mean errors of only 4.2 seconds, and achieves an overall 92% item-level gesture recognition accuracy).

(ii) Accurate Recognition of Item-Level Interaction: We show that we can identify a variety of locomotive gestures (especially the {P,T,B} gestures mentioned before), by appropriately using inertial sensor (accelerometer & gyroscope) based features from a smartwatch and a smartphone. Using these gestures as building blocks, we also subsequently infer *item-level* interactions such as whether the shopper buys the item frequently or knows specifically what he wants, using novel high-level features. All these classifications yield accuracies of over 90%.

(iii)Accurate Prediction of Episode Attributes: As the highest level of inference, we also utilize aggregate features (the item-level interaction history, plus in the in-aisle and non-aisle movement history) to build classifiers to estimate *episode-level* attributes, (such as "was the shopper in hurry? , and "did the shopper find the items he wanted?", achieving accuracies of over 92%.

Note again that *IRIS* operates without any assumption of in-store infrastructure support or location tracking capability (no Wi-Fi, no RFID, no knowledge of store layout, etc.).

II. Related Work

Mobile phone sensing has emerged as a paradigm catering to multiple sectors such as healthcare, social networks, safety, environmental monitoring, transportation and retail. Wearable sensing has simultaneously evolved as a technology that enables human activity recognition at a finer granularity [2]. Our work utilizes a combination of such mobile and wearable sensing to uncover deeper insights into a shopper's in-store behavior.

Gesture & interaction recognition: The feasibility of mobile sensing for human activity recognition has been well explored in literature. While [7] proposes a probabilistic model based on conditional random fields to identify smoking gestures using sensor data collected from an inertial measurement unit; Khan et.al [4] implements a smartphone-based Human Activity Recognition scheme that uses a non-linear discriminatory approach together with a non-linear SVM based classifier. The trade-off between energy efficiency and classification accuracy is explored for mobile phone sensing in [13]. Unlike community-based personalized activity models [5], our work attempts to infer shopper behavior in a generalized setting where no shopper-specific training data is available.

In-store analytics: Several works have focused on human activity recognition based on images or videos. Previous work[12] proposes a finite state machine based approach to infer hand-activities in video-based retail surveillance. Additionally, the Channel State Information of Wi-Fi signals have been used to study in-store shopper behavior in [15]. The interesting problem of studying the shopping time in stores is presented in [14], where a phone-based shopping tracker uses motif groups to identify movement trajectories and transforms the problem of monitoring shopping time as a classification problem. ThirdEye [8], uses image, inertial sensor, and Wi-Fi data crowd-sourced from shoppers wearing smart glasses to track the physical browsing of shoppers. Sen et.al [9] proposes a person-independent activity recognition technique, CROSDAC, which uses smartphone based sensor (accelerometer, compass) data and Wi-Fi, to identify the shopping intent of users. Our goal is to push the boundaries of in-store behavior

analytics without relying on any special-purpose wearable or infrastructure support.

Mall-level/Shopper Behavior: There are numerous case studies on shopper/mall-level shopping behaviors which are typically confined to specific stores or demographies of shoppers [3]. Lee et.al in [6] presents an automated computing framework using smartphones designed to provide comprehensive understanding of customer behavior.

To the best of our knowledge, our work is among the first to utilize a mobile phone and smartwatch concurrently to infer item-level interactions of shoppers inside stores.

III. *IRIS*: Architecture and Key Objectives

IRIS' goal is to uncover shopper-specific and store-level behavioral attributes, both during a specific shopping episode, and via aggregated observations across a longitudinal trace of such episodes. As *IRIS* does not presuppose any support from the store (e.g., location tracking, maps, PoS data, etc.), it does not attempt to capture insights such as specific product viewed or bought by a shopper. Instead, our goal is to infer item-independent aspects of a shoppers behavior, such as number of products picked and then returned, movement speed within the store etc.

A. Types of individual and store-level insights

One of our long-term goals is to use microscopic *gestural-level* insights obtained during a consumer's interaction with a single product as a "building block", to help build progressively deeper insights about both a shopper's short-term and longer-term behavioral attributes. In this view, the item-specific insights gained by looking at a set of sensor data *frames* (a relatively small duration lasting a few seconds) can be viewed as elements of a periodic table of in-store shopping behavior; these elements are then combined in hierarchical fashion to discover the higher-level individual and store-level attributes. More specifically, we categorize the insights into three broad bins:

• *Item-Level Insights (Individual):* These insights describe aspects of an individual shopper's behavior with a specific product (or product type). For example, based on the time that the user inspects the product, i.e., the interval between a 'P' (pick) and the corresponding 'T' (in trolley) activity, we hope to learn if this is a "familiar" product (that the shopper regularly buys without much additional thought) or an "unfamiliar" one. Similarly, an observation of multiple 'P' (picks) and 'B' (put backs), before an eventual 'T (trolley), might indicate that the shopper had no apriori brand affinity, but instead compared multiple brands before picking a specific item.

• *Episode-Level Insights (Individual & Store):* These insights are obtained at the shopping episode-level (an episode comprises multiple item-level interactions) by aggregating individual item-level labels/features. These insights can capture the episode-level behavior of the shopper (e.g., a relatively small number of in-trolley ('T') actions, coupled with shorter "non-aisle" durations, might indicate that the "shopper was in a hurry"). Moreover, the insights can also describe properties of the store itself (e.g., unusually slow movement during "non-aisle" segments might indicate that the store was overly crowded).

• *Longitudinal Insights (Individual & Store):* These insights are obtained by aggregating observations across a large collection of episodes (independent store visits), observed over a period of weeks and months. At an individual-level, they can help reveal the shopper's *persona*— for example, that the "shopper is always hurried during a weekday visit" or that "the shopper always shops in bulk". At a store-level, they can help reveal the stores macroscopic properties for example, that "store X has more (or less footfall) during specific times or days".

In this paper, given the absence of longitudinal data, we focus only on item and episode-level behavior of shoppers.

B. The IRIS Architecture

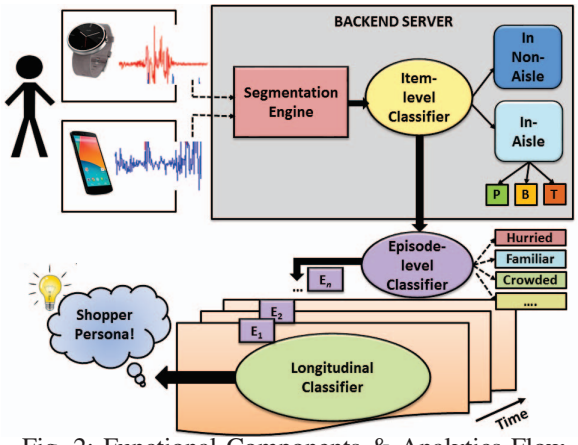


Fig. 2: Functional Components & Analytics Flow

Figure 2 illustrates the device and backend components of the *IRIS* framework, as well as the typical flow of the analytics pipeline. Each individual shopper carries an on-body smartphone and smartwatch, whose sensor streams capture the individual's physical movement and gestural activities, over an entire shopping episode. At the backend, this entire stream is first run through a *Segmentation Engine*, which splits up the entire shopping episode duration into different segments (time chunks), each corresponding to a single movement or gestural activity. Each individual chunk is then fed into a hierarchical *"Item-level"* classifier, which attempts to first classify each chunk as either "in-aisle" vs. "non-aisle", and subsequently separately classifies different gestures within an "in-aisle" segment into one of multiple interaction-related labels (e.g., {P, B, T} gestures). This collection of gestures and movement patterns (from the Item-level classifier) is then collectively analyzed by the *Episode-level* classifier, to help discern episode-level labels (e.g., "was the shopper in a hurry?"). Finally, the *Longitudinal Classifier* operates at longer time scales, analyzing (a) multiple episodes of the same shopper to determine *"persona-level"* attributes, and (b) episodes from multiple in-store shoppers to determine *"store-level"* attributes.

IV. Data Collection

We first describe our process of collecting real-world shopping behavioral data. We conducted a user study with 25

middle-aged volunteers (15 female, 10 male) recruited from Xerox Research Centre. Each participant was asked to visit one of the two different retail grocery stores in Bangalore, India (one large and spacious, the other much more cramped for space) and purchased items from a given shopping list. We collected 50 shopping episodes from the grocery stores at different times of the day. Each episode lasted, on average, for about 20 minutes and belonged to one of 3 distinct types: (i) *Engineered List* (20 episodes), (ii) *Clocked* (20 episodes) and (iii) *Discretionary* (10 episodes).

Engineered List: The participants were given a list of 14 grocery items which consisted of 4 Frequent-Choice (*FC*), 4 Infrequent-Choice (*IC*), 3 Frequent-Specific (*FS*) and 3 Infrequent-Specific (*IS*) items. The items were categorized based on general consensus after a small survey. For example, egg and bread were frequent items, while dishwashing soap and Schezwan sauce were infrequent items; "select a juice of your choice" is an example of a FC item; while "Tropicana Orange Juice–1 gallon" exemplifies a Specific item. The participants were asked to shop for the items in the same order as in the list.

Clocked: The objective here was to emulate "hurried" behavior. Hence, we paired up the participants, gave each a list of 10 items and engaged them in a shopping competition. The participants were informed that the person clocking the least overall time, while buying all the items listed, would be declared the winner. To control for differences in familiarity with the shop, all the participants were familiarized with the shop and its aisles before the episode started. All items in the list were open-ended (Choice), and selected "randomly" (by picking ingredients from arbitrary common recipes).

Discretionary: The objective here was to capture behavior in situations where a shopper could choose not to buy an item, due to a variety of factors (such as budget constraints, product unavailability, or deficient quality). The items in the list were chosen to elicit some of these factors. Sample items included fruits that were out of season, items with budget constraints which were not feasible, "greens that needed to be fresh enough", "red coffee mug with a design they liked", etc. The shoppers were unaware of our study objectives; the traces thus capture the natural behavior of shoppers who earnestly look for a preferred item but may be unable to find it.

A. Sensor Data Collection

Each participant was given a smartphone (running Android v4.3 or above) and a smartwatch (Android Moto 360). The phone was placed in the right-side pant pocket facing front, and the watch was worn on the dominant hand (all our participants were right-handed). The devices were pre-installed with our custom data collection apps for the smartphone and smartwatch. The apps recorded data from the sensors listed in Table I, at the maximum permitted sampling frequencies of 200 Hz (phone) and 25Hz (watch). Some sensors were only exclusive to a single device–e.g., the magnetometer was unavailable on the watch, whereas the heart rate sensor was unavailable on the phone. Ambient sensing (temperature, light and audio) was more reliable on the watch since the phone was placed inside the pocket.

TABLE I. List of sensors monitored

Sensors	Purpose	Device
Accelerometer	Speed and patterns in walking and hand activities	Watch, Phone
Gyroscope	Rotational and angular information during walking and hand activities	Watch, Phone
Magnetometer	Directional information during walking	Phone
Step Counter	Number of steps directly obtained from Google Fit API	Watch, Phone
Battery Temperature	Distinguish between zones using ambient temperature (e.g., freezer section)	Watch
Light Sensor	Ambient lighting in the store	Watch
Audio Sensor	Ambient noise in the store	Watch
Heart Rate	To study if specific browsing behavior causes excitement	Watch

B. Ground Truth Collection

The ground truth of a shopping episode was collected by having a person *shadow* the shopper (without the shopper's knowledge). The shadower used an app on his own device, which enabled him to both record micro-activity labels of the shoppers ("Picking", "In Trolley", "Enter Aisle", etc.), and to record audio notes, along with the timestamps (all three devices, i.e., shopper's phone & watch, and shadower's phone, were time-synchronized). Other non-activity related information, such as the shopper's *familiarity level* with the store or the *crowdedness* of the store were captured via a survey filled in at the end of each episode. To ensure uniformity in ground truth annotation, an item-level interaction was assumed to start after the preceding "Trolley" label (where the user was pushing a trolley), and continued till the subsequent "Trolley" label; the interval itself could contain multiple labels, such as "pick", "put back", etc. Note that all our studies (and analyses) make the assumption that the shopper always uses a trolley, although we believe that the technique can be extended to other modes (e.g., a shopping basket).

V. Classifying under Perfect Segmentation

As the first step in investigating *IRIS*, we first seek to extract the discriminatory features of smartwatch & smartphone sensors, and understand their classificatory power, to help infer various shopper-experience related item-level and episode-level properties. More specifically, in this section, we assume that, via some as-yet unknown mechanism, we have perfect knowledge of the (start, end) times of each item-level interaction (e.g., the "P", "B, "T", "in-aisle" or "out-of-aisle" activities), and investigate two questions via a supervised classification approach: (1) How accurately can we classify each of the distinct item-level interaction activities, and what features aid this classification? (2) Given knowledge of such item-level behavior, how accurately can we infer *episode-level* properties, and what features (defined over the aggregated item-level interactions) aid this classification?

A. Item-level Shopper Experience Attributes

We start by trying to identify the following four item-level attributes (based on the shopper's interaction with that specific item), as insights on these four attributes help reveal a shopper's buying preferences and habits: • *Frequent Item:* An item that the shopper buys frequently or routinely and is familiar with. • *Infrequent Item:* An item that the shopper is less familiar with because he does not buy it as often. • *Specific Item:* An item for which the shopper has *a-priori* knowledge

TABLE II. Features for Item-level classification

(1)	Mean number of picks
(2)	Variance in number of picks
(3)	Mean hold time, i.e., duration between picking and putting back
(4)	Variance in hold time
(5)	Mean duration of time between picking an item for the first time and putting in trolley (W1)
(6)	Variance in W1
(7)	Mean Duration between entering an aisle to putting item in trolley (W2)
(8)	Variance in W2
(9)	Mean Duration between walking in non-aisle to entering an aisle (W3)
(10)	Variance in W3
(11)	For each time window W1, W2 and W3, following features from phone accelerometer: mean & variance in magnitude, spectral entropy & energy.

of the specific brand & product detail. • *Choice Item:* An item for which the shopper does not have an *a-priori* product in mind, but instead needs to view alternative products and make a choice.

Table II lists the various features that we used to classify these 4 labels. The features have a hierarchical structure as follows. Initially, different statistical features (similar to that used in [13]) are used to identify each interaction/movement activity as "P", "B", "T", "in-aisle" and "non-aisle". While the phone-based features help identify the walking/gait-related patterns (e.g., "in-aisle" or non-aisle), the watch-based features help identify the gestural interactions ("P", "B", "T"). Subsequently, features (1-10 in Table II), defined over the interaction and movement activities, help classify the *item-level* aspects of shopper experience.

Features 1-10 were defined to help exploit several intuitive properties of human behavior that we visually observed across shopping episodes. For example, for either a Specific (*S*) or a Frequent (*F*) item, we can expect the shopper to perform a smaller number of picks (P), exhibit smaller hold time (H), as well as have smaller durations of the time windows W1, W2 & W3. In contrast, for Choice (*C*) or Infrequent (*I*) items, shoppers will likely exhibit a larger number of pick (P) and put back (B) gestures and a longer duration of window W1 (as they evaluate multiple items before converging on a selection). Moreover, for Infrequent items, shoppers will likely spend more time and effort to locate the item, resulting in larger durations of windows W2 and W3. Note that the analysis of *F* vs. *I* is performed by considering only those users who were familiar with the store, to avoid the confusion on whether a shopper's item-level behavior was due to unfamiliarity with the item or the store's layout.

Figure 3 shows the values of these features for each of these classes averaged across all episodes we collected, in order to gain insight into the dataset w.r.t these features. We see that the data reflects certain intuitive or expected trends. For example, compared to *S* items, *C* items have a higher mean duration for windows W1 and W2 (features 5 & 7); similarly, *I* items tend to exhibit longer durations of non-aisle movement (feature 9). To understand the ability of these features in classifying these product-level attributes, we trained J48 decision tree classifiers, along with Correlation Feature Selection (CFS) to identify the most dominant (discriminatory) features. Note that we trained 3 different classifiers, two binary classifiers (one each to distinguish between *S* vs. *C* and *F* vs. *I*) and one quaternary classifier (to distinguish between the 4 composite labels (*FS*), (*IS*), (*FC*) and (*IC*)).

TABLE III. Item-level classification with ground truth. Column 3 uses indices from Table II

	Precision	Recall	Dominant Features
Frequent	0.997	1.0	(1), (3), (5),(6)
Infrequent	1.0	0.998	(1), (3), (6)
Specific	1.0	0.999	(2), (1), (4), (5)
Choice	0.999	1.0	(2), (1), (4), (5)
Freq-Spec	0.993	0.999	(2), (1), (4), (5)
Freq-Choice	1.0	1.0	(2), (1), (4), (5)
Infreq-Spec	1	0.997	(2), (1), (4), (5)
Infreq-Choice	1	0.998	

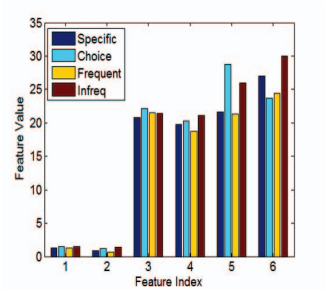


Fig. 3: Values of dominant item-level features listed and indexed in Table II

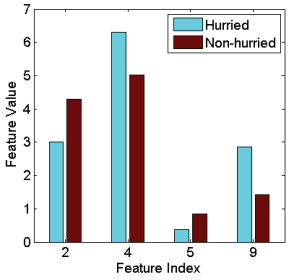


Fig. 4: Values of dominant episode-level features listed and indexed in Table IV

Table III tabulates the results obtained via 10-fold cross validation. We note that we get almost 100% accuracy (both precision and recall values are over 99% for all labels)! This is a very encouraging result, especially given that our dataset contains labels aggregated from 25 users, who we expect have diverse shopping styles and preferences. **These results suggest that the behavioral markers of shoppers are distinct enough (between {*S*, *C*, *F*, *I*} products) for us to robustly identify them from a combination of smartwatch and smartphone sensor data.**

B. Episode-level Shopper Experience Attributes

We next focus on inferring the individual-specific episode-level characteristics, such as whether the shopper was in a hurry (or not) or whether the shopping experience was productive (i.e., did the shopper find most of the items he was looking for?). Following the approach used previously, we used J48 binary classifiers and features 1-10 (listed in Table IV) to study whether a shopper was "hurried" or not. The data from 20 hurried ("Clocked") episodes were combined with 20 non-hurried ("Engineered List") episodes to perform the *HU* vs. *NH* analysis. Figure 4 shows the values of these features averaged across all episodes from our data set, correlating as expected with the feature set. With these features, a J48 decision tree binary classifier yielded an overall precision and recall of 99% each, which are tabulated later in Table VII for comparison. Features (2), (4) & (9) were the most dominant.

TABLE IV. Feature set for determining hurriedness (* marks the dominant features)

(1)	Mean duration in an aisle (seconds)
(2)*	Variance of duration in an aisle
(3)	Mean duration in a non-aisle (seconds)
(4)*	Variance of duration in a non-aisle
(5)*	Mean step rate in an aisle
(6)	Variance in step rate in an aisle
(7)	Mean step rate in a non-aisle
(8)	Variance in step rate in a non-aisle
(9)*	Mean hold time (seconds)
(10)	Speed of picking item (mean magnitude of watch accelerometer during pick)

Summary: Our results in this section indicate that *IRIS* can indeed very reliably (with accuracies usually above 99%) infer item-level and episode-level aspects of a shopper's in-store behavior. However, there is a big caveat: our high accuracy has been demonstrated (thus far) only under the assumption that the overall sensor data has been reliably *segmented*–i.e., the (start, end) times of each activity label are correctly known. We next develop novel techniques to perform such automated and accurate segmentation.

VI. Automatic Segmentation

The supervised learning discussed in Section V assumed the use of ground truth labels to demarcate the time segments corresponding to different activities. As a key contribution of this paper, we now describe how to automatically deduce the (start, end) times of various labels through a combination of (i) landmarking based on significant sensor features, to distinguish between non-aisle and aisle zones (ii) Viterbi decoding to predict the sequence of hand activities and (iii) improving the precision of this hand sequence prediction by estimating the likelihood of an item being found using survival analysis models, and utilizing this information to bias the transition probabilities in a time-dependent markov model.

A. Differentiating Aisle and Non-aisle zones

The key observation used in landmarking aisle and non-aisle zones is that when a shopper moves into an aisle to look for an item, there is a marked difference in the walking speed hand movement, as he slows down after entering an aisle of interest. The inter-step interval (i.e., the duration between consecutive steps) is higher inside an aisle than in non-aisle; moreover, while a shopper mostly pushes the cart (or carries a basket) in non-aisle, he has a lot more variations in the hand movements due to various browsing-related actions. Further, the inter step interval for a shopping episode (Figure 5) reveals that an aisle zone always begins from the foot of a peak until the peak; similarly, a non-aisle zone spans from the peak to the foot. However, the number of peaks spanned, i.e., duration for each zone is variable. Accordingly, using peak and valley detection, we identify all peak-points (t_{peak_i}) and foot-points (t_{foot_i}) of all ramps. In order to determine the duration of the zones, we use change point detection analysis using a binary random forest classifier trained to identify aisle and non aisle regions, using statistical features from phone accelerometer, watch accelerometer and watch gyroscope listed in Table V. A sliding window size of 10 seconds was used. The precision and recall of this classifier model is 0.888 and 0.875, respectively. The reasoning behind the change point detection algorithm is that the classification probability will drop when the test set contains mixed data, i.e., data from across different categories. Accordingly, we first gather the features within the window corresponding to the first ramp, $w = [t_{foot_1}, t_{peak_1}]$ and compute the probability $Pr(aisle|featureset(w))$ using the binary classifier. Next we increase the window size to include subsequent peaks, one peak at a time, until the classification probability drops. Suppose the accuracy dropped for the window $[t_{foot_j}, t_{peak_i}]$, the region $[t_{foot_j}, t_{peak_{(i-1)}}]$ is marked as "Aisle". Similarly, next the features in window $w = [t_{peak_{(i-1)}}, t_{foot_i}]$ is used to compute $Pr(nonaisle|featureset(w))$, and the window size is incremented to include subsequent foots until the probability drops, say at t_{foot_k}; the region $[t_{peak_{(i-1)}}, t_{foot_{(k-1)}}]$ is then marked as "non-aisle".

TABLE V. Feature set for classifying aisles/non-aisle zones and hand/non-hand activities

Feature	Aisle vs Non-Aisle	Hand vs Non-Hand
Mean phone accelerometer magnitude	✓	✓
Spectral entropy of phone accelerometer magnitude	✓	✓
Mean watch accelerometer across x,y,z axes	✓(only y,z axes)	✓
Spectral entropy of Watch accelerometer across x,y,z axes	✓(only y,z axes)	✓
Mean watch gyroscope along x,y,z axes	✓(only x-axis)	✓
Spectral entropy of watch gyroscope along x,y,z axes	✗	✓
Variance in step rate	✓	✗

Accuracies for segmenting aisle and non-aisle regions are as shown in Table VI. The possible reason for higher false positives in classification of non-aisle is because of the "walk-and-browse" characteristic, i.e., the time instances when a shopper continues to walk after entering the aisle, without necessarily slowing down or picking items to check items. The average offset in time between an actual segment and predicted segment is around 5 seconds.

B. Identifying Hand Activities

There are two parts to solving the problem of identifying hand activities, which are defined as either a Pick (P), Put Back (B) or In Trolley (T). The first is to identify if any hand activity occurred, and if so, the next is to identify which of these three actions it was. The first part is straightforward by analyzing the gyroscope data from the smart-watch. Figure 6 shows the gyroscope data, after performing quaternion rotation with respect to a common origin [111], and fitting it to a spline curve [7]. The value plotted is the normalized product of pitch, roll and yaw. The figure also shows the ground truth in terms of the times when a hand activity did occur. We observe that the peaks are a good indicator of a hand activity, with negligible false negatives, but there are a significant number of false positives, resulting from arbitrary hand movements. To address this, we first run a peak detection algorithm to identify the peaks and then eliminate bulk of the false positives by filtering out those peaks that occur during 'Non-aisle' segment (as described in Section VI-A). For each remaining peak, we compute the features in the window corresponding to the width of the peak (full-width at half-maximum), and feed it to a random forest binary classifier to compute the probability that it is a hand activity based on a combination of watch gyroscope, watch accelerometer and phone accelerometer features (Table V). This process yields a precision of 95% and recall of 98% in identifying a hand activity.

The next step after identifying the existence of a hand activity, is to predict if it is a P, B or T. We propose using a Viterbi decoding approach on a Hidden Markov Model in order to leverage the inherent sequential nature of gestures in a shopping episode. The state transition probabilities between P, B and T are computed from the experimental data. The trellis diagram corresponding to the Viterbi decoding is shown in

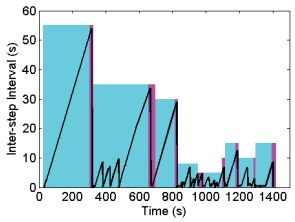


Fig. 5: Inter-step interval and corresponding aisle (dark blue) and non-aisle (light pink) zones

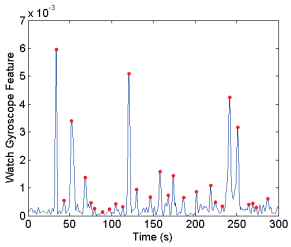


Fig. 6: Watch gyroscope peaks indicating potential hand activities. The red dots show the actual hand activities from ground truth.

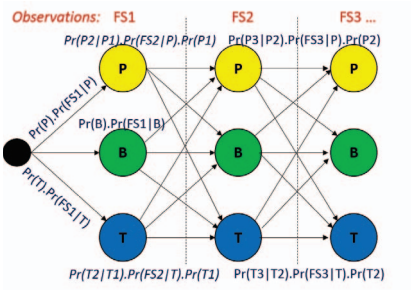


Fig. 7: Trellis diagram corresponding to the Viterbi decoding of hand action sequences.

Figure 7. The emission probability is defined as $Pr(FS|l)$, where $l = P, B, T$, and $FS = [f_1, f_2 ... f_n]$ is the set of features from watch gyroscope and watch accelerometer (features 3, 4, 5, 6 in Table II), which are the observations in our HMM. The emission probability is obtained as: $Pr(FS|l) = \frac{Pr(l|FS) * Pr(FS)}{Pr(l)}$, where $Pr(FS) = \prod_{i=1}^{n} Pr(f_i)$, since sensor features are independent. The probabilities $Pr(f_i)$ and $Pr(l)$ can be obtained from the distribution of the empirical data. The probability $Pr(l|FS)$ is obtained from the random forest ternary classifier, which is trained to distinguish between P, B and T using the features in FS (with an average precision and recall of 0.926 and 0.927 respectively).

One salient aspect about this decoding approach is that it avoids onset of cascaded prediction failures. This is because, the length of the predicted sequence is limited to each aisle segment, i.e., the sequence is predicted independently for each aisle segment, since the activities within each aisle-segments are independent of other segments, and this helps contain prediction errors. The performance of classification is shown in Table VI.

TABLE VI. Accuracy of automatic segmentation in identifying Aisle, Non-aisle, P, B and T

	Aisle	Non-aisle	P	B	T
Precision	0.9775	0.9051	0.9863	0.9149	0.8200
Recall	0.9669	0.9376	0.9863	0.9053	0.8367

C. Survival Analysis

We see that the prediction accuracy for T is lower than the other two activities, and is often mispredicted as B. In order to improve the accuracy, we use the likelihood of finding the item as an indicator of whether the action would converge in a Put Back or a Trolley. This probability is obtained by using the Cox Proportional Hazards model [1], given the time elapsed since the search began, and the number of picks as a covariate. Since this value is not constant, we treat each discrete value of number of picks as a separate covariate and derive a different hazard function for each case. Our analysis shows that the family of survival functions obtained this way has 81.3% accuracy in predicting the likelihood of an item being found.

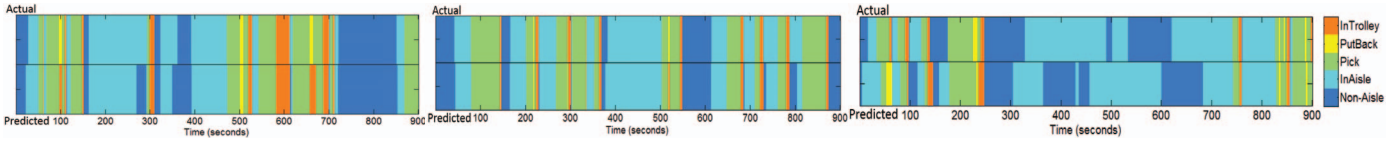
If the survival function indicates that the item is not likely to be found (< 0.5), then we bias the sequence prediction towards a B (by multiplying the state transition probabilities for the transitions into T by the likelihood of finding item), or else towards a T. This is done by making use of the fact that as the number of B for an item increases, and the item is likely to be found eventually, the likelihood of a T increases, i.e., there arises a time-dependent Markov chain. We retain the Markov property by conditioning the states based on the number of prior Pick-Put Back actions during that item-episode. In other words, we compute a family of transition probability matrices $\{TPM_i\}$, where each matrix TPM_i gives the transition probabilities between Pick, Back, Trolley given there have occurred i Pick-Put Back prior transitions. **Using this approach the precision and recall of prediction of T improved by 7.6% and 4.2%, respectively. to 0.8830 and 0.8646, respectively; the precision and recall of prediction of B improved to 0.9226 and 0.9337**.

VII. Putting it all together: Attribute Classification with Automatic Segmentation

Finally, we re-ran the supervised learning classification experiments described in Section V, with the same set of features, but with labels obtained from our automatic segmentation approach instead of ground truth. We compared accuracies with (a) classification with ground truth labels and (b) classification with a brute-force approach for automatic segmentation. The basic idea behind this brute force approach is to use a regular classifier to determine which label a time window belongs to. Accordingly, we split our data into windows of 10 seconds. We then use the binary classifier trained with the features in Table V, as discussed in Section VI-A to determine if each window belongs to Aisle or Non-Aisle. Next, for each predicted aisle segment, we split it into 3 second windows, compute the features in Table II for these windows, and use the ternary classifier discussed in Section VI-B to determine if that window belongs to a P, B or T. We decided these window sizes of 10 seconds and 3 seconds after some trial and error, selecting that window which gave the highest accuracy. This brute-force approach only yielded an average precision and recall for Aisle/Non-aisle of 71.3%, 73.5%, respectively; and for P, B, T classification, 50.5% and 38.8%,respectively.

Next we re-ran the attribute classification after automatic segmentation. Table VII shows the average classification accuracy for item level and episode-level attributes, using our automatic segmentation method and brute force approach.

We see that our segmentation yields very good accuracy. Interestingly, we see that the accuracy is higher for Frequent vs. Infrequent and Hurried vs. non-hurried, than the other classification. This is most likely because the dominant features of



(a) (b) (c)

Fig. 8: Trace of actual and predicted lables in the first 15 minutes of a shopping episode for (a) unhurried, familiar shopper (b) hurried, familiar shopper and (c) unfamiliar, unhurried shopper

TABLE VII. Comparison of Classification Accuracy with different approaches

Attribute	P	R	P_gt	Rgt	P_bf	R_bf
Freq - Infreq	0.921	0.926	0.99	0.99	0.653	0.666
Specific-Choice	0.88	0.89	0.99	0.99	0.553	0.644
Hurried-Non hurried	0.916	0.922	0.99	0.99	0.693	0.714

these attributes involves non-aisle and picks which are more accurately predicted, than those that involve trolley and put back labels.

Figure 8a shows a sample trace of predicted and actual labels for the first 15 minutes of a shopping episode for a shopper who was not in a hurry and was familiar with the store; Figures 8b and 8c show similar traces for unhurried-familiar, and unhurried-unfamiliar shopper, respectively. Interestingly we can see that these traits of a shopper are revealed to some extent in the traces. For instance, a hurried shopper (Figure 8b) has fewer put backs, and an unfamiliar shopper (Figure 8c) spends longer durations without interacting with items. We also observe that the classification accuracy of our framework varies with such profiling. For instance, the accuracy of Put Back and Non-aisle for hurried shopper is lower than average (87% and 84%), which can be reasoned that the gestures are performed in a hurry, and the shopper walks fast in both aisle and non-aisle when in a hurry.

VIII. Conclusion & Future Work

This paper presents the design and initial prototype of *IRIS*, a framework for obtaining behavioral insights about a shopper's in-store interactions and behavior, utilizing only sensing data available from the shopper's personal smartphone and wearable device (smartwatch). Results show that, given a trace of an entire shopping episode in representative retail stores, *IRIS* is able to (i) delineate the (start, end) times of different in-store interactions, and (ii) utilize various shopping-related features to characterize such individual in-store interactions – both with very high (approx. 90%) accuracy. Such interactions reveal novel insights into the familiarity & premeditated choice attributes for each item, the level of hurriedness of the shopper and her familiarity with the store, simply by exploiting behavioral patterns captured by the mobile and wearable sensors.

There are a variety of additional approaches & possibilities that we're addressing in ongoing work. For example, we plan to incorporate physiological sensor data (e.g., smartwatches contain embedded heart rate or GSR sensors) to additionally infer (or even *predict*) a shopper's in-store browsing intent and product-specific reactions. As a preliminary effort, we observed that using the mean and variance of heart rate values (captured by our smartwatch) allowed us to obtain a classification accuracy of 78% for item-level interactions. Moreover, in environments where additional infrastructure is available, *IRIS* can be augmented to provide finer-grained information. For example, if BLE beacons are deployed to facilitate fine-grained, in-aisle location tracking, *IRIS* can also associate the customers "experience" with a specific product (indexed by its location on a specific aisle & shelf).

Acknowledgement

This work was supported partially by Singapore Ministry of Education Academic Research Fund Tier 2 under research grant MOE2011-T2-1001 and partially by the National Research Foundation, Singapore under its Interactive Digital Media (IDM) Strategic Research Programme. All findings and recommendations are those of the authors and do not necessarily reflect the views of the granting agency, or Singapore Management University.

References

[1] Norman E Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review*, pages 45–57, 1975.
[2] Oscar DL and Miguel AL. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials*, 15(3):1192–1209, 2013.
[3] H Hu and CR Jasper. A qualitative study of mall shopping behaviors of mature consumers. *J. Shopping Center Research*, 14(1):17–38, 2007.
[4] AM Khan, A Tufail, AM Khattak, and TH Laine. Activity recognition on smartphones via sensor-fusion and kda-based svms. *IJDSN*, 2014.
[5] ND. Lane, Y Xu, H Lu, S Hu, T Choudhury, AT Campbell, and F Zhao. Enabling large-scale human activity inference on smartphones using community similarity networks (csn). In *Proc. of UbiComp '11*.
[6] S Lee, C Min, C Yoo, and J Song. Understanding customer malling behavior in an urban shopping mall using smartphones. In *MSCSS'13*.
[7] A Parate, M-C Chiu, C Chadowitz, D Ganesan, and E Kalogerakis. Risq: Recognizing smoking gestures with inertial sensors on a wristband. In *Proc. of MobiSys'14*.
[8] S Rallapalli, A Ganesan, K Chintalapudi, VN Padmanabhan, and L Qiu. Enabling physical analytics in retail stores using smart glasses. In *Proc. of MobiCom'14*.
[9] S Sen, D Chakraborty, V Subbaraju, D Banerjee, A Misra, N Banerjee, and S Mittal. Accommodating user diversity for in-store shopping behavior recognition. In *Proc. ISWC'14*.
[10] Longfei Shangguan, Zimu Zhou, Xiaolong Zheng, Lei Yang, Yunhao Liu, and Jinsong Han. Shopminer: Mining customer shopping behavior in physical clothing stores with passive rfids. In *In Proc. of Sensys'15*.
[11] V Subbaraju, S Sen, A Misra, S Chakraborti, and RK Balan. Using infrastructure-provided context filters for efficient fine-grained activity sensing. In *Proc. of PerCom'15*.
[12] H Trinh, Q Fan, J Pan, P Gabbur, S Miyazawa, and S Pankanti. Detecting human activities in retail surveillance using hierarchical finite state machine. In *Proc. of ICASSP'11*.
[13] Z Yan, V Subbaraju, D Chakraborty, A Misra, and K Aberer. Energy-efficient continuous activity recognition on mobile phones: An activity-adaptive approach. In *Proc. of ISWC'12*.
[14] C-W You, C-C Wei, Y-L Chen, H-H Chu, and M-S Chen. Using mobile phones to monitor shopping time at physical stores. *IEEE Pervasive Computing*, (2):37–43, 2011.
[15] Y Zeng, PH Pathak, and P Mohapatra. Analyzing shopper's behavior through wifi signals. In *Proc. of WPA '15*.