10-2018

I4S: Capturing shopper's in-store interactions

Sougata SEN
Dartmouth College

Archan MISRA
Singapore Management University, archanm@smu.edu.sg

Vigneshwaran SUBBARAJU
*A*Star Singapore*

Karan GROVER
IIIT Delhi

Meeralakshmi RADHAKRISHNAN
Singapore Management University, meeralakshm.2014@phdis.smu.edu.sg

See next page for additional authors

DOI: https://doi.org/10.1145/3267242.3267259

Published in ISWC '18: Proceedings of the 2018 ACM International Symposium on Wearable Computers, Singapore, October 8-12, Pages 156-159.
https://doi.org/10.1145/3267242.3267259

I^{4}S: Capturing Shopper's In-store Interactions

Sougata Sen[†]*, Archan Misra[‡], Vigneshwaran Subbaraju[¶]*, Karan Grover[§]*, Meera Radhakrishnan[‡], Rajesh K. Balan[‡], Youngki Lee[⊥]*
[†]Dartmouth College, [‡]Singapore Management University,
[¶]A*STAR Singapore, [§]IIIT Delhi, [⊥]Seoul National University
sougata.sen@dartmouth.edu, {archanm, meerlakshm.2014, rajesh}@smu.edu.sg,
vigneshw1@e.ntu.edu.sg, karan13048@iiitd.ac.in, youngki.lee@gmail.com

ABSTRACT
In this paper, we present *I^{4}S*, a system that identifies item interactions of customers in a retail store through sensor data fusion from smartwatches, smartphones and distributed BLE beacons. To identify these interactions, *I^{4}S* builds a gesture-triggered pipeline that (a) detects the occurrence of "item picks", and (b) performs fine-grained localization of such pickup gestures. By analyzing data collected from 31 shoppers visiting a mid-sized stationary store, we show that we can identify person-independent picking gestures with a precision of over 88%, and identify the rack from where the pick occurred with 91%+ precision (for popular racks).

INTRODUCTION
A shopper's activity in a retail store consists of two logically distinct activities: (i) inspecting or browsing items, and (ii) eventually purchasing a selection of items. Online stores easily capture both these activities through browsing history, click streams, etc. and use the derived interest profile to offer personalized recommendations. Physical stores on the other hand rely heavily on human effort to monitor all items inspected or browsed by a customer and use this information to identify items that did not eventually translate into a sale.

In this paper, we present a sensor-based system, named *In-Store Item Interaction Identification System* (*I^{4}S*)[1], that aims to automatically capture a shopper's fine-grained interactions during a shopping episode – i.e., a store visit. More specifically, *I^{4}S* tackles the question: *from which shelves, and on which racks, in the store did shoppers pick their items?* To answer this question, *I^{4}S* utilizes the information from a set of Bluetooth Low-Energy (BLE) beacons deployed in the store (not associated with a specific product), together with the shopper's wrist-mounted smartwatch, and the shopper's smartphone.

*This work was primarily done by the authors during their affiliation with Singapore Management University.

[1]pronounced I-foresee

ISWC '18, October 8-12, 2018, Singapore, Singapore

ACM ISBN 978-1-4503-5967-2/18/10…$15.00
DOI: https://doi.org/10.1145/3267242.3267259

Broadly, *I^{4}S* involves innovative use of both the RF-sensing capabilities of the smartphone and the inertial-sensing capabilities of the smartwatch. The smartwatch's inertial sensor identifies the time instants when the *picking* gesture is performed, while the smartphone's RF-sensing capability determines the in-store location of the shopper during the picks. *I^{4}S* assumes that a separate backend repository exists that contains the location-item(s) mapping. Hence, when a pick occurs at a particular location, a lookup of this repository will directly reveal the small set of *items* present in the location.

Practical challenges: While both gesture detection and indoor localization are well-studied problems, our problem domain of fine-grained, in-store, shopper's item interaction identification gives rise to several non-standard challenges. Our analysis of the data collected in-store shows that: (i) although the picking action is transient, however, for certain picks, the duration is longer (e.g., when a user starts inspecting the item); (ii) BLE measurements on smartwatches have high packet loss and high RSSI variance during shopping episodes, making it difficult to localize such transient gestures purely using smartwatch's data; (iii) the performance of BLE-based localization is affected by the number and position of deployed beacons.

Key contributions: While addressing the challenges in developing *I^{4}S*, this work makes the following contributions: (i) *Design of I^{4}S*: We introduce an inexpensive approach, called *I^{4}S*, for fine-grained tracking a shopper's in-store product interactions. *I^{4}S* can be used by shopkeepers to identify items that are interacted with by shoppers, but are not purchased, and (ii) *Evaluation of* I^{4}S *in the real world:* A comprehensive real-world study shows how the different components of *I^{4}S* combine effectively to provide fine-grained tracking of item interactions. *I^{4}S* can infer person independent pick gestures with a precision of 88%, and identify the rack-level location of *picks* in popular racks with an accuracy of over 91%.

OVERALL GOALS AND SYSTEM OVERVIEW
I^{4}S's broader goal is to track all the item-related interactions that a shopper performs in a store. For this work, we focus exclusively on identifying *picks*, as picking an item is a indication of shopper's interest. This is interesting to a shopkeeper because, the checkout transactional data can reveal the items bought, but not items that interested the shopper but weren't purchased. Although it will be beneficial to identify the exact item picked, in this work we focus on identifying the *shelf*

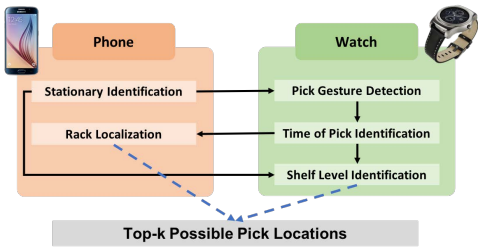

Figure 1. Overview of the system with smartwatch and smartphone

from where an item was picked. This information can be used to infer the *(category of) item* that the shopper is interested in.

Overview of Final I^4S Design

Given the twin goals of pick gesture identification and localization, we devised the I^4S system based on a gesture-triggered rack level location tracking paradigm, using the combination of rack-mounted BLE beacons, the shopper's smartphone and a smartwatch worn by the shopper. The I^4S system determines the occurrence and location of the *pick* activity as illustrated in Figure 1 and described below:

Identify when the shopper is stationary: Our empirical data showed that more than 99% picks occur when a shopper is stationary. I^4S thus uses the smartphone's inertial sensor data to determine if the shopper is relatively stationary.

Inferring pick gesture: When a shopper is stationary, I^4S uses the smartwatch's inertial sensor data to infer a *pick* gesture.

*Localize to the corresponding rack:*The racks in the store are fitted with BLE beacons. The shopper's smartphone scans for these beacons' advertisements and uses this data to determine a possible set of racks that are in the shopper's proximity.

Localize to shelf level: I^4S uses the inertial sensor data from the smartwatch, to determine the shelf level of the *pick*. To improve shelf-level localization accuracy, the smartphone data is used to infer whether the shopper is sitting or standing.

DATASET

The data collection for the study took place in a mid-sized stationary store. After obtaining IRB approval, 31 university students (14 males, 17 females) were recruited for the study. The participants were provided with a smartwatch and a smartphone running our custom data collection app. Participants wore the watch on their dominant hand and carried the phone in the front pocket of the pant. No specific task was assigned to the participants while they were in the store. We termed each such store visit as an *episode*. Similar to [6], the ground truth for these episodes was collected by shadowing the shopper.

At the start of the study, the shop was instrumented with 35 BLE beacons. All beacons were placed at the base of the racks. Overall, the store had approx 50 racks. Several aisles had 3 to 4 racks placed side by side. In case three racks – $\{R_1,R_2,R_3\}$ were present side by side, two beacons were placed at the base of racks R_1 and R_3. We set the beacons with a transmission interval of 101 ms and a transmission power of -20 dBm.

For analysis, we used data from 25 of the 31 shopping episodes. 6 episodes were omitted as participants either did not carry the smartphone (not wearing clothing with a pocket) or their data had synchronization issues. The total time taken to complete the 25 episodes was 2 hours 52 minutes. Overall, 778 picks from 43 distinct racks were observed during the 25 episodes.

Feature	No	Description
Mean	4	Average of the data from the 3-axis and their magnitude
Variance	3	Variance in the values of the axis data in the time window
MCR	3	Count of times the values cross the window's mean
Max mean	3	Compute the maximum of the means of the sub windows
Max rise	3	Divide window into sub-windows; compute maximum positive change for consecutive sub-windows
Max drop	3	Divide window into sub-windows; compute maximum negative change for consecutive sub-windows
Covariance	3	Co-variance between the axis of the sensor
Entropy	3	The spectral entropy of the axis data in the time window
Locomotion	1	The locomotion state of the user predicted by the phone

Table 1. Features extracted from the smartwatch's inertial sensors

METHODOLOGY

To identify in-store interactions, I^4S relies on inertial and BLE scan data from a smartwatch and smartphone. There are three main components to I^4S: (a) identifying the pick gesture, (b) identifying the rack from where item was picked, and (c) identifying the shelf from where item was picked.

Pick Gesture Detection

We first describe the smartwatch's data processing pipeline.

Framing: We use the smartwatch's accelerometer and gyroscope data to infer the *pick* gesture. The pre-processed data is divided into frames of length w, with 50% overlap between frames. Every instance of the frame is represented by [*Time*, $Accel_x$, $Accel_y$, $Accel_z$, $Gyro_x$, $Gyro_y$, $Gyro_z$]. We empirically found that for pick inference, $w = 2$ seconds has an optimal balance between false positives and false negatives. From the data we found that a pick gesture lasted for approximately 4 sec (varied from 2 to 10 sec). Since $w = 2$ is used, multiple frames together represents a complete pick gesture.

Feature extraction: For each frame, we compute both time and frequency domain features, as described in Table 1. Features are computed for every sensor axis. In addition to standard features, some features based on empirical observation have also been used – e.g., we found that only 4 out of the 778 picks occurred when the user was moving. We thus used the user's locomotion state (derived from the smartphone) as a feature.

Classification: Once the features are extracted and labeled using ground truth data, a classifier is used to determine whether the frame represented a picking gesture. Various classifiers were tested and we found that the Random Forest classifier has the best inference accuracy. We thus use it in our study.

Smoothing: Since a *pick* lasts for more than $w = 2$ seconds, temporal information from consecutive frames could be useful in smoothing the data. More specifically, if we observed s consecutive $w = 2$ second frames, and if s' frames ($\forall s' \in s$) are identified as picking, then we declared that a pick occurred. This smoothing filtered out random hand movements.

Localize the Pick to a Rack

From the smartphone's BLE scan log, frames of size w_b ($w_b/2$ seconds before and after pick P_i) are extracted. We use $w_b = 4$

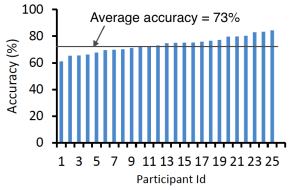

Figure 2. Pick inference accuracy per user

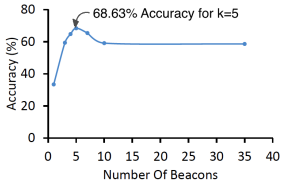

Figure 3. Location accuracy for top-k beacons

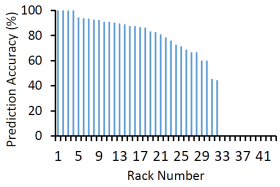

Figure 4. Localization accuracy per rack

seconds. Let B be the set of beacons deployed in the store. During w_b, a tuple (FP_i) for P_i of size m ($m \mid m \leq count(B)$), is generated. Entry in FP_i for jth beacon (b_j) during P_i is $\{b_j, \overline{RSSI_{b_j}}\}$. If a beacon is not heard in w_b, the RSSI value for the beacon is set to a very small value. Finally, FP_i's label is extracted from ground truth. For P_i, the generated tuple is the fingerprint for the *rack* from where item is picked. By repeating this step for all $P = 778$ picks, we created a fingerprint map (FP) of dimension $\{\{m$+label$\}$ x $P\}$.

To identify the rack, we used a RADAR like approach [1]. For each P_i in P, we compute FP_i's Euclidean distance from all other $\{FP\text{-}FP_i\}$ fingerprints. P_i is assigned a *probabilistic* (top-k) location, where a location's probability is computed as the inverse of the Euclidean distance between FP_i and the k^{th} closest fingerprint. To improve the prediction, Viterbi smoothing technique is applied to remove unlikely Rack traversals. For the Viterbi implementation, we used a depth of $h = 4$. The path with the highest probability is selected as the most likely path and the node at Level h in the most likely path is inferred as the rack from where item was picked.

Localize the Pick to Shelf Level

To identify the shelf from where the item was picked, we extract all the frames which are labeled as *picking*. The class label of the frames is changed to the shelf-level position of the hand during the frame. We label shelves as: L1→if shelf (or hand) is 0 to 30 cm from ground, L2→if shelf (hand) is 30 cm to 60 cm from ground, $\cdots$ L6→if shelf is 150 cm to 180 cm from ground. The updated frames are passed through a Random Forest Classifier to infer the shelf.

EVALUATION

In this section, we present the evaluation of I^4S. Specifically, we evaluate the accuracy of (a) inferring pick gestures, (b) localizing the pick to a rack, and (c) localizing pick to a shelf.

Pick Gesture Inference

Inferring the picking gesture frame: To classify the *pick* gesture, we tested two cross-validation methods: (a) 10 fold cross-validation (10F-CV), and (b) leave-one-user-out cross validation (LOO-CV). For 10F-CV, I^4S could accurately distinguish *pick* frames from *non-pick* ones in 92.85% cases, with a precision and recall of 92% and 81.5% respectively in detecting *pick* gestures. To more carefully understand the cross-individual differences, we performed LOO-CV. Figure 2 shows the user-wise *accuracy* of I^4S in identifying pick or non-pick. Overall, the significant drop in accuracy (avg. accuracy of 73%) suggests that users exhibit unique picking styles.

Inferring the entire pick gesture: We empirically observed that picking gesture usually lasts for ≈4 seconds (3 frames). To infer the entire *pick* gesture, we used 5 temporally consecutive frames (2 extra frames as buffers), i.e $s = 5$. If $s' \in s$ frames were classified as *pick*, we declared that a picking gesture was taking place. We evaluated the performance for $s' = 2$ and $s' = 3$, and found that a tighter pick inferring criteria ($s' = 3$) has a higher precision (88.8%), but lower recall (missing many picks). A more permissive criteria ($s' = 2$) has more false positives, but misses far fewer actual picks. Table 2 shows the performance for the two values of s'. The overall accuracy for $s' = 2$ is **89.18%**. This indicates that although the frame level classification is noisy, it can be corrected by smoothing across multiple frames, leading us to choose $s' = 2$ in our system.

Smoothing threshold	Accuracy	Precision	Recall
2 out of 5	89.18%	78.9%	87.5%
3 out of 5	87.59%	88.8%	67.4%

Table 2. Effect of smoothing on *pick* gesture inference

Localize Picks to a Rack

Ideally, to identify the rack from where item is picked, every window that has been inferred as *pick* by the *gesture recognizer*, should be localized. However, to understand the performance of the localization in-store, during the pick gesture, we extract BLE scan data for the windows when actual picks occurred, rather than based on the gesture recognizer's output.

We tested a commonly-used strategy of computing location using the RSSI readings from a subset of k 'stronger-signal' beacons heard during w_b. Figure 3 shows the performance for different values of k. From the figure we see that we achieve the best accuracy (68.63%) for $k = 5$. For $k = 5$, we note the top-3 closest racks predicted and observe that in 80.84% cases, the correct rack is amongst the top-3 chosen racks, indicating that using the shopper's movement history might be useful in identifying the correct rack. We used data from the window w_b when the pick occurred, as well as data from the 3 windows of length 4 seconds immediately preceding the pick. For each window, we independently computed the location probabilities and used a depth=4 Viterbi decoder to estimate the pick location. Based on this path-smoothing approach, the rack prediction accuracy during pick increases to 85.47%.

The 85.47% location accuracy is, however, skewed by popularity. Figure 4 shows the accuracy distribution across the 43 distinct racks. From the figure we can see that for 11 racks, the accuracy is 0. On closer inspection, we found that these racks had less than 5 picks in the dataset, indicating that the loss of accuracy was due to insufficient training data. From the dataset we observed that 346 picks (or 44.5% of total picks) occurred from just the 5 most popular racks. On running our localization algorithm for just these 346 picks (while allowing the prediction to be any of the 43 total racks), we are able to

	Accuracy	Precision	Recall
Personalised Pick Detection (cv)	92.85%	92%	81.5%
Genralised Pick Detection (sm)	89.18%	78.9%	87.5%
Localization for Top-5 racks	91.32%	NA	NA
Shelf Level Identification	89.07%	NA	NA

cv: Cross Validation results; *sm*: with smoothing performed

Table 3. Summary of the performance of various components of I^4S

achieve an accuracy of **91.32%**. This shows that with more data for all racks, the prediction performance can be improved.

Localize Picks to Shelf Level

Similar to *rack localization*, for *shelf-level localization*, we used the frames from the actual-pick time-windows. We used the data from the smartwatch's inertial sensor to build the classifier, with *Shelf Level* being the predicted output.

The accuracy of a 10-fold cross validation for a 6-way shelf classification is 77.12%. We observed that several picks that occurred from lower shelves were classified as upper shelves. We hypothesize that the hand trajectory in sitting and picking from a lower shelf might be similar to standing and picking from an upper shelf. To test the hypothesis, we used the sensor data from the phone to infer if the shopper was standing or sitting. This {sitting—standing} attributes became an additional feature. On performing cross validation, we found that this feature increased the shelf level classification accuracy to **89.07%**, thereby vindicating our hypothesis.

Overall Performance of I^4S

Table 3 summarizes the performance of each component of I^4S. Each component can be tuned based on the overall application requirement. Since each component is mutually exclusive, the overall performance of the system can be calculated as: P(Shelf Estimation) = $P(Pick)$ * $P(Rack)$ * $P(Shelf)$. If we use the values highlighted in the previous subsections for each of the parameter ($P(Pick) = 0.8918$, $P(Rack) = 0.9132$, and $P(Shelf) = 0.8907$), we can identify the *precise shelf*, where the pick occurred, with 72.53% accuracy. Tolerating a ± 1 error in the shelf estimation increases this accuracy to 76.72%.

RELATED WORK

Understanding a shopper's movement patterns via mobile sensors to study consumers behavior in retail spaces has been investigated in works by You et al. [8] and Lee et al. [3]. Alternately, analysis of in-store customer behavior using an infrastructure-based video monitoring techniques has been performed by Krockel et al. [2]. The system *ThirdEye* tracks different elements of physical browsing using images, inertial sensors and Wi-Fi data captured from a smartglass and a smartphone [5]. Although, *ThirdEye* can detect that the shopper is looking at an item, the use of images increases both privacy and power concerns. An alternate approach of using the Channel State Information of Wi-Fi signals to infer a shopper's location within a store has been studied by Zeng et al. [9]. Shangguan et al. proposed an RFID-based system to infer comprehensive shopping behaviors in a clothing store setting [7]. However, their system cannot create an individual-level shopper profile. Radhakrishnan et al. proposed a framework that uses a smartphone's and a watch's sensor data to recognize item-level gestural interactions and overall in-store behavior of the shoppers [4].

Several previous works provide anonymized individual-level shopping behavior information [4, 5, 6, 8]. However, they do not provide information about finer-grained browsing behaviors such as the number of item 'interactions' at each location, something I^4S is designed to achieve.

CONCLUSION

In this paper we describe the design of I^4S, a system which can identify the items that a shopper picked during shopping; some of which might not have been purchased. Through a user study, we show the possibility of identifying pick with a precision of 88%, the rack of interest amongst top-5 racks with 91% accuracy and the shelf of interest with 89% accuracy.

ACKNOWLEDGMENTS

This research was supported by the National Research Foundation, Prime Ministers Office, Singapore under its IDM Futures Funding Initiative, the Singapore Ministry of Education Academic Research Fund Tier2 under research grant MOE2014-T2-1063 and National Science Foundation under award number CNS-1329686. All findings and recommendations are those of the authors and do not necessarily reflect the views of the grant agencies.

REFERENCES

1. P. Bahl and V. N. Padmanabhan. RADAR: An in-building RF-based user location and tracking system. In *INFOCOM 2000*. IEEE.
2. J. Krockel and F. Bodendorf. Intelligent processing of video streams for visual customer behavior analysis. In *Proceedings of ICONS 2012*, Vol. 1.
3. S. Lee, C. Min, C. Yoo, and J. Song. Understanding Customer Malling Behavior in an Urban Shopping Mall Using Smartphones. In *ACM Conference on Pervasive and Ubiquitous Computing (UbiComp '13 Adjunct)*.
4. M. Radhakrishnan *et al.* Iris: Tapping wearable sensing to capture in-store retail insights on shoppers. In *International Conference on Pervasive Computing and Communications (PerCom'16)*. IEEE.
5. S. Rallapalli *et al.* Enabling Physical Analytics in Retail Stores Using Smart Glasses. In *20th Annual International Conference on Mobile Computing and Networking. 2014 (MobiCom'14)*.
6. S. Sen *et al.* Accommodating user diversity for in-store shopping behavior recognition. In *ACM 18th International Symposium on Wearable Computers (ISWC'14)*.
7. L. Shangguan *et al.* ShopMiner: Mining Customer Shopping Behavior in Physical Clothing Stores with COTS RFID Devices. In *13th ACM Conference on Embedded Networked Sensor Systems. 2015 (SenSys '15)*.
8. C. You, C. Wei, Y. Chen, H. Chu, and M. Chen. 2011. Using phones to monitor shopping time at physical stores. *IEEE Pervasive Computing* 10, 2 (2011).
9. Y. Zeng, P. H. Pathak, and P. Mohapatra. Analyzing shopper's behavior through wifi signals. In *2nd workshop on Workshop on Physical Analytics (WPA'15)*.