AutoTag: Visual Domain Adaptation for Autonomous Retail Stores through Multi-Modal Sensing

Carlos Ruiz
Carnegie Mellon University
Electrical & Computer Eng.
Moffett Field, California
carlosrd@cmu.edu

Joao Falcao
AiFi Inc.
Santa Clara, California
joao@aifi.io

Pei Zhang
Carnegie Mellon University
Electrical & Computer Eng.
Moffett Field, California
peizhang@cmu.edu

ABSTRACT

Autonomous checkout at retail stores could bring a large array of benefits to both consumers –no lines, better user experience– and retailers –lower operational cost, detailed insights about customer behavior. Existing approaches include self-checkout stations, on-item sensing (e.g., RFID) and infrastructure-based sensing (e.g., vision and weight). While each has their own pros and cons, the latter offers a good tradeoff between information richness and operational cost. However, several challenges currently limit their accuracy. In particular, visual item recognition is constrained by the huge amount of training data required and the domain adaptation gap that usually exists between the training –e.g., well-lit environment– and testing distributions –e.g., each store might have different lighting conditions, camera angles, etc. In this preliminary work we explore different ways to leverage multi-modal sensing (e.g., weight load cells on shelves) to automatically label frames from customers picking up or putting down items on the shelf. Then, those annotated frames could be used to continuously expand the initial visual model and tailor it to each store’s conditions.

CCS CONCEPTS

• **Computing methodologies** → **Object detection; Object identification**; • **Theory of computation** → Online learning algorithms.

KEYWORDS

Domain Adaptation; Multi-Modal; Heterogeneous Sensing; Camera; Weight; Location; Human Pose; Online Learning

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

UbiComp/ISWC '19 Adjunct, September 9–13, 2019, London, United Kingdom

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6869-8/19/09...\$15.00

<https://doi.org/10.1145/3341162.3345604>

Figure 1: An example of AutoTag in action: when other sensor modalities (e.g., weight) are very confident about what item was picked but vision is not, AutoTag finds a bounding box of the object and labels it so the visual model can be fine-tuned for the current environment via online learning.

ACM Reference Format:

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 the 2019 International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct)*, September 9–13, 2019, London, United Kingdom. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3341162.3345604>

1 INTRODUCTION

Physical retail stores are exploring autonomous checkout technologies to gain additional insights about customers –e.g. what they grabbed but didn’t end up buying–, to enhance the user experience by eliminating checkout lines and to reduce operational and labor costs. Although some companies have recently made promising progress, such as Amazon Go’s *unmanned* stores [1], there are still many challenges ahead for fully autonomous stores (Amazon reportedly uses humans-in-the-loop to generate accurate receipts [12, 20]).

Existing solutions can be broadly grouped in three areas: self-checkout, on-item sensing techniques and infrastructure-based sensing. Self-checkout, already present in many stores,

typically consists in barcode- and weighing-enabled machines where customers scan all of their items before leaving. However, this approach tends to be ineffective –sometimes customers accidentally forget to scan some of their items–, and their average receipt accuracy can be as low as 86% [3]. Another broadly deployed solution is on-item sensing, such as RFID tags [14, 17] to uniquely identify each product and prevent theft. While more effective at identification than self-checkout, these approaches suffer from a main limitation: the recurrent cost to install sensors on every item –a huge downside on already small margins for retailers–, which generally makes them only suitable for higher-end goods and not for everyday products. Finally, infrastructure-based solutions [1] offer better benefits with respect to long-term cost and sensing availability, but are currently limited in accuracy and many challenges are still being researched.

In terms of sensing modalities, the richness and recent advancements in processing techniques make vision a great source of information. Although object classifiers can achieve high accuracy, the amount of training images required –and the labor effort associated with labeling them– could have a huge impact in the feasibility of large-scale vision solutions. Furthermore, the unique conditions –such as different lighting– in each store pose an additional hurdle: tackling the domain adaptation problem (to avoid collecting and annotating data for each product in each store). Fortunately, infrastructure-based methods do not need to rely on a single sensor source, and other sensing modalities, such as weight, could help autonomous systems fine-tune their pre-trained model and boost their performance in each target store.

In this paper, we present:

- Our proposed system architecture for automatically *adapting* a visual classifier model to each store, based on multi-modal sensing (e.g., load cells on shelves).
- A discussion on different approaches and challenges for frame selection and item bounding-box estimation.

2 BACKGROUND & RELATED WORK

There exists a significant amount of prior work in identifying objects from their weight as well as from their visual features. However most of these works have focused individually on one sensing modality. While there are performant solutions that can identify objects solely by their weight, they fall short in tackling the inventory monitoring domain because retail products often weigh similarly, as is the case for soft drinks.

On the other hand, while vision-based object identification would likely be able to tackle the above cases –e.g., most sodas look different–, current solutions still vary wildly in accuracy due to the wide variety of objects to be identified (some stores sell thousands of different products). In addition, many items share a substantial amount of visual features

Figure 2: AutoTag system overview. Left (Brown): Weight change- and Location-based item identification pipeline. Right (Red): Vision-based item bounding box annotation pipeline. Bottom (Blue): Domain-specific dataset generator, which allows the pre-trained visual product classifier model to continuously adapt to the deployment environment, thanks to the other sensing modalities (e.g., weight).

with other products that have very distinct content (e.g., any yogurt vs. its fat-free version).

Weight-based object identification

In its simplest form, weight-based object identification uses the prior knowledge of each product’s weight distribution to determine the most likely class of an item that was picked up or put down. If multiple weight sensors are deployed, obtaining the *location* of an inventory change event is a key piece of information to help solve ambiguities, especially when multiple weight distributions are too similar [19].

While the previous approach uses the total weight of all objects on the surface, concluding the number of items on the surface by observing the total weight is challenging: often one product’s weight is a multiple of a different product’s weight. To solve this issue, a capacitive mat or textile that leverages the shape of the object imprint on its surface can be used to recognize the object [10, 11, 16]. However, such approaches tend to be impractical when the sensing material is facing the harsh environment of an operating store: liquids spilling, continuous friction and impact on the mat, permanent deformations and cost make them prohibitive.

To address the above issues, [7] combines the features extracted from the weight with visual features obtained from a camera to re-identify groceries coming in and out of a fridge. This study demonstrates the added benefit of leveraging vision with weight to improve the object identification.

Vision-based object identification

Visual object identification can be broadly grouped in two categories: feature-based and data-driven approaches.

Histogram of Oriented Gradients (HOG) [21], Scale Invariant Feature Transform (SIFT) [9] and Speeded-Up Robust Features (SURF) [2] are examples of approaches that classify an object by a predefined set of extracted features, but they do not perform well for large numbers of target classes and require hand-tuning parameters. On the other hand, Mask R-CNN [6], Fast R-CNN [5], Faster R-CNN [15] or Yolo [13], all leverage large amounts of data to generalize the characteristics of the objects to be identified. In the latter category, labelled data of the class to be identified is critical to train the model to generalize. If properly trained, these approaches have empirically shown to yield a higher accuracy which could address the retail automation market. The huge data collection effort for training the model is still a big limiting factor, and that is exactly the direction this paper wants to contribute in.

3 SYSTEM OVERVIEW

Figure 2 shows *AutoTag*'s system framework, which utilizes multi-modal sensing to fine-tune, or *adapt*, a visual model to a specific deployment environment. In particular, we focus on utilizing the weight and event location information to provide a domain-specific item label for updating/re-training the visual model.

Based on prior work [19], *AutoTag*'s pipeline is triggered when a weight change is detected, which likely corresponds to an item pick-up or put-back. From that weight change two features are extracted: the absolute weight change (used to compare against the weight distribution of each product) and the spatial distribution of the weight (used to predict location, which in turn leverages prior knowledge of item arrangement to make its prediction). In parallel, the vision pipeline also tries to identify the item, and weight- and vision- predictions are then alpha-blended to produce the final item prediction. However, visual training data is sometimes collected in different conditions than in the store –from a different perspective, distance, lighting, and so on. *AutoTag* targets cases where the weight-based prediction is very confident about what item was picked, but the cameras weren't. In such situations, processing the camera frames could lead to new training images that contain a very confident label (the output of weight-only identification) and an object bounding-box (the output of *AutoTag*'s vision pipeline).

The vision pipeline computes human pose estimation and applies a background subtraction technique to discard pixels that did not move, which avoids annotating the remaining objects on the shelf and those near the one that was picked. Using the geometric notion of where cameras are in relation to the shelves, *AutoTag* chooses the closest hand to the weight change event and performs the Product Bounding Box Estimation. After some preliminary results displayed in Figure 4, Section 5 discusses some of the challenges future work will have to tackle, especially in this Product Bounding Box Estimation module, such as successfully removing the customer's arm, handling cases where users have more than one item in their hand, etc. After estimating the tightest bounding box to the object of interest, the last step in our vision pipeline is the Bounding Box Validation, which consists in a series of sanity checks onto the generated bounding box such as ensuring a minimum and maximum bounding box size, so that the probability of malformed or incorrect training images is reduced.

After every block of N_{th} (with $N_{th} \geq 1$) new, autonomously annotated bounding boxes are generated, the model can be updated using the regular training procedure.

4 PRELIMINARY RESULTS

This section presents the experiments carried out to understand, compare and discuss different potential solutions to improve the performance of visual classifiers with the assistance of other sensing modalities, such as weight.

Experiment Setup

For our experiments, we instrumented 5 shelves with 12 4" weight sensing plates each and installed 4 720p cameras to the left, right, above and in front of the shelves. We went to a local convenience store, purchased everything on one of their 5-shelf fixtures and replicated their layout in our lab store. In order to train the item prediction model:

- We placed each product on a turntable, while recording from 3 cameras at different angles and distances. As displayed in Figure 3, we use background subtraction (with some human-aided fine-tuning) to crop images of each product and randomly place them on different backgrounds (we use the MS-COCO dataset [8] due to its large variety) and train a classifier (Faster R-CNN [15] using ResNet-50 and FPN) to recognize all 33 different product classes in our inventory. For higher accuracy and lower convergence time, we initialized the model weights by pre-training the network on [22]'s dataset and trimming the last –classifier– layer.
- We weighed every item and fitted a Normal distribution for each class.

Figure 3: An example of the data generation process for initial training of the visual model. We collected videos of each item from three different angles as shown in (a). The view from one of camera angles is illustrated in (b). Background from images like (b) is removed to obtain images with transparency like (c). Randomly sampled products from (c) are overlaid onto random backgrounds from the COCO Dataset [8], as in (d). Finally, a Faster R-CNN [15] model was trained with 30k images like (d).

- We annotated the item arrangement, i.e., the mapping between location –weight plate IDs– and products (e.g., Cling Wrap is on plates (6, 7) on shelf 5).

Finally, we designed our experiments to try to simulate a real shopping experience where customers have a shopping list they intend to purchase. With 8 participants and 5 trials per subject, we obtained a total of 176 events. They were given a randomly generated shopping list with 3 to 6 items for each trial.

Metrics. The main area this paper tackles is item identification. We define the average item identification accuracy, *Avg. ID accuracy* for short, as:

$$\text{Avg. ID accuracy} = \frac{\# \text{ correct items predicted}}{\# \text{ events}} (\%) \quad (1)$$

Preliminary results on real-world experiments

The Avg. ID accuracy using weight-sensing only (absolute weight change + event location) was 91.5% [18]. While this is not a bad result, it would still generate many incorrect receipts and customer complaints. Fusing weight with our pre-trained vision model (alpha-blending the predictions with $\alpha_{\text{weight}} = 0.6$) leads to 93.2%, a promising improvement that proves the value of multi-modal approaches. However, after reviewing the images we feel that these results could be even better with *AutoTag*'s multi-modal based continuous learning approach. As this is still ongoing work, we want to focus the attention of the remainder of the paper in the next section, which discusses some of the challenges we have faced trying to adapt the visual model from the source (well-lit items on white background) to the target (blue-ish and darker) distributions.

5 CHALLENGES FOR FUTURE WORK

This section proposes a few different approaches for the Product Bounding Box Estimation module and discusses some of the challenges ahead for multi-modal based continuous

learning systems like *AutoTag*. Figure 4 highlights some of the current limitations of visual models: perspective and lighting conditions have a big impact on performance (for instance, a distant front view like 4(a) barely finds any objects). This indeed reinforces our motivation for *AutoTag* as an autonomous approach to *adapt* to a particular environment. Furthermore, 4(c) illustrates how challenging estimating an item's bounding box is in the real-world: customers might carry multiple items in their hands, so how can we determine which object to label based on the weight prediction? This is important because annotating the wrong object(s) would quickly degrade the performance of the pre-trained model.

We propose a few methods to find an item bounding box:

- A) “Model-based”: relies on the existing visual model to recognize objects –although with low confidence or misclassified–; then, the proposed bounding box the model outputs could directly be used as the annotated bounding box for training (with the correct label from the weight sensors).
- B) “Empirical”: within a circle with center at the wrist and radius the wrist–elbow distance of the customer, find the biggest blob after the background subtraction and select the tightest bounding box to enclose it.
- C) “Hand”: combining “empirical” with a hand pose estimation model [4, 23], a more accurate bounding box could be estimated by masking out the customer's hand and arm, and understanding what they just grabbed.

The easiest to implement is method A), since the pre-trained model would already output bounding box proposals. However, it runs the risk of never learning if the model never finds bounding boxes (as in Fig. 4(a)). Method B) could help narrow down the bounding box region by leveraging human pose information, however it would still fail in the challenging cases like Fig. 4(c). For those cases, perhaps an approach like method C), which tries to understand the position of the fingers over time could help determine the right item to

(a)

(b)

(c)

Figure 4: Challenges for *AutoTag*: in some perspectives, like (a), the model barely finds objects, while this is not the case for some other views and lighting conditions such as (b). Also note how, customers shopping normally might carry multiple items in their hand, which makes the Product Bounding Box Estimation very challenging, i.e., which item was just picked? Mistakes might make the model even worse

annotate. This is still an open problem but future work will hopefully provide successful answers to these challenges.

6 CONCLUSION

In this paper, we introduce *AutoTag*, a system that uses multi-modal sensing to guide the adaptation of a pre-trained visual object classifier to a target deployment environment. This is particularly useful for the autonomous retail domain, since stores have up to tens of thousands of different items and collecting data for each set of conditions (lighting, perspective, distance...) could make vision approaches infeasible. Given that each store might also have slightly different conditions (camera setup, store layout, number of products...), a global, unique visual model would not perform as well as a continuously updated one that is tailored to each individual store. Leveraging detections from other sensing modalities (e.g., weight change or location) with high confidence, the true label is obtained, and *AutoTag*'s visual pipeline finds the corresponding bounding box in the camera frames so that the model can be updated. In addition, we illustrated some challenges when deploying these systems in real-world scenarios and proposed potential solutions.

REFERENCES

- [1] Amazon. 2019. *Amazon.com: Amazon Go*. <https://www.amazon.com/b?node=16008589011> Accessed: 2019-04-10.
- [2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *European conference on computer vision*. Springer, 404–417.
- [3] Adrian Beck. 2011. Self-scan checkouts and retail loss: Understanding the risk and minimising the threat. *Security Journal* 24, 3 (2011), 199–215.
- [4] Ali Erol, George Bebis, Mircea Nicolescu, Richard D Boyle, and Xander Twombly. 2007. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding* 108, 1-2 (2007), 52–73.
- [5] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.
- [6] 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.
- [7] Rena Kamoda, Mayumi Ueda, Takuya Funatomi, Masaaki Iiyama, and Michihiko Minoh. 2012. Grocery re-identification using load balance feature on the shelf for monitoring grocery inventory. In *Computers workshop (CwC)*. Citeseer, 13.
- [8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
- [9] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.
- [10] Christian Metzger. 2005. Indirect Object-sensing Technology to Prevent Out-of-stock at Retail-level. In *Proceedings of the Workshop on Smart Object at the International Conference on Ubiquitous Computing, Tokyo*. Citeseer.
- [11] Christian Metzger, Jan Meyer, Elgar Fleisch, and Gerhard Tröster. 2007. Weight-sensitive foam to monitor product availability on retail shelves. In *International Conference on Pervasive Computing*. Springer, 268–279.
- [12] 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.
- [13] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.
 - [14] 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.
 - [15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.
 - [16] Mahsan Rofouei, Wenyao Xu, and Majid Sarrafzadeh. 2010. Computing with uncertainty in a smart textile surface for object recognition. In *2010 IEEE Conference on Multisensor Fusion and Integration*. IEEE, 174–179.
 - [17] George Roussos. 2006. Enabling RFID in retail. *Computer* 39, 3 (2006), 25–30.
 - [18] 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 Conference on Systems for Built Environments*. ACM.
 - [19] Albrecht Schmidt, Martin Strohbach, Kristof Van Laerhoven, Adrian Friday, and Hans-Werner Gellersen. 2002. Context acquisition based on load sensing. In *International Conference on Ubiquitous Computing*. Springer, 333–350.
 - [20] 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.
 - [21] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter Tu. 2007. Shape and appearance context modeling. In *2007 IEEE 11th international conference on computer vision*. Ieee, 1–8.
 - [22] Xiu-Shen Wei, Quan Cui, Lei Yang, Peng Wang, and Lingqiao Liu. 2019. RPC: A Large-Scale Retail Product Checkout Dataset. *arXiv preprint arXiv:1901.07249* (2019).
 - [23] Chi Xu and Li Cheng. 2013. Efficient hand pose estimation from a single depth image. In *Proceedings of the IEEE international conference on computer vision*. 3456–3462.