**Product Detection Using Deep Learning**

**Introduction**

Object detection using computer vision has been an area of interest with the development of deep learning algorithms. Object detection is defined as the classification of semantics objects in images and videos using the combination of shapes and edges based on an existing library of classified objects. Its application has already extended into areas including manufacturing, automatous driving, traffic detection. The usage of object detection has greatly enabled faster and more automatous decision-making process.

**Opportunity**

One of the applications of object detection is in retail settings. There are two potential application use cases. Currently, retailers already use closed-circuit television (CCTV) to monitor the status in stores. The use of CCTV is mainly to prevent theft and deter vandalism in its store locations. For most of retailers, theft detection is already part of its security prevention model with security personnel actively monitoring CCTVs. However, some challenges of the traditional method include:

* High costs – the costs associated with the hiring of security personnel to monitor the cameras
* Lack of accuracy – since it is impossible to focus on all screens at a given time, many thefts are not even detected
* Delay in response – in most cases, thefts are detected after the fact as a response to a lost item

Many other security measures to prevent theft also exist such as the usage of electronic article surveillance (ESA). However, ESAs’ effectiveness has been challenged by research where "based on the available evidence it is difficult to determine the effectiveness of tags as a theft reduction measure, albeit there is suggestive evidence that more visible tags are associated with greater reductions in theft than less visible tags"[[1]](#footnote-1). There are costs associated with implementation as ESAs require tags as well as sensors.

Moreover, another potential application of computer vision in retail settings is customer insights. One of potential shortcomings of the existing customer insights generation for retailers is that it is only limited to products customers already purchased. Retailers have a complete list of products purchased by all customers, obtained as part of the POS systems. Analysis such as Market Basket Analysis[[2]](#footnote-2) can performed to understand what product is likely to be purchased with another product to optimize store placement, discount strategy, and product combination. However, the currently data cannot support the analysis to examine important metrics such conversion rate since there is no method to obtain information on customer’s “Path of Purchase”. This metric is something that is readily used in ecommerce where conversion ratios such as “probability to put in a basket”, “probability to purchase” can all be determined. And because these metrics can be collected, optimization strategies such as A/B testing can be performed to increase conversion that can have a material impact on the top line of the company. Moreover, further optimizations on store placement can be possible with insights on customers time at sections or products.

Thus, computer vision in retails settings offers an opportunity to both reduce cost and increase in effectiveness in preventative measures, while also providing an opportunity to optimize product and placement strategy to increase revenue.

**Competitors/Existing Technology**

*Everseen*

As one of the leading innovators in computer vision, Everseen’s technology already has numerous retail applications. Everseen’s retail application ranges from logistics, which tracks the movement of goods through autonomous processes, salesfloor, which alerts sales personnel on the floor when a product is not converted, and checkout, which identifies items not being scanned properly at the self-checkout counters. The application of checkout is already been implemented at Walmart[[3]](#footnote-3).

While Everseen offers an effective solution at checkout, not all theft occurs during the checkout phase. The opportunity to expand the surveillance throughout the store will effectively minimize all thefts that occur on the store premise.

Moreover, Everseen’s algorithm should be based on matching product movement across the screen with items recorded in the system. When a customer moves a product across the screen, the algorithm detects the movement and register in the system. However, if there is no corresponding input on the self-checkout counter that matches the timestamp, the algorithm will send an image to the unidentified product to the store clerk to address the missed scan item.

*StopLift*

Similar to Everseen, StopLift is also focused on detecting missing scanned items at the checkout counter. StopLift was recently acquired by NCR, in a move to boost NCR’s presence in intelligent monitoring system.

*Digital Mortar*

Digital Mortar provides store traffic-based analytics solution. Digital Mortar offers an array of solution from:

* People counting, which counts the number of people in an area that can generate traffic flow graphs
* Display measurement, which can identify hot spots on the floor
* Queue management, which can actively manage the staff-to-customer ratio to optimize queue time
* Shopper journey, which creates pathing information on how the shopper move through the area

Digital Mortar’s gap lies in the inability to identify objects. While the algorithm is capable of capturing a shopper’s journey throughout the store, it is unable to detect whether if a customer lingered in front of a product (demonstrate interest level) ultimately converted the purchase. It is also unable to recognize at a product level, thus only capable of providing insights based on heatmapping. The system requires configuration of the store layout to effectively identify the products/services shoppers are interested in. The algorithm can provide information such as a section (i.e. produce, diary) that generates high traffic, but is unable to identify the specific product (i.e. a specific fruit or a specific brand of milk)

**Technical Challenges**

*The lack of existing image library*

One of the technical challenges of this project is the lack of image library on a product level. For most image libraries available open source, the focus remains on detecting the class of objects (i.e. vehicle) rather than on product level (i.e. Tesla and Ford). One of the most prominent object libraries, Colombia Object Image Library[[4]](#footnote-4) (COIL-100), only focuses on object level. The same applies for the majority of image libraries[[5]](#footnote-5) where the object classes are on an object level rather than a product level. This creates the next technical challenge.

*The need to create a product a product image library*

Since building the algorithm requires a comprehensive product image library, it is necessary to create one that served this purpose. Typically speaking, 10,000 training images for one product is necessary to yield an accurate result. Since a product does not look the same from any perspective, it is also important to factor in controlled conditions to ensure a sufficient training set. This includes controlling the angle, position, lighting as well as other factors that the product may experience in a retail setting. (i.e. A product does not look the same to a computer viewed from a top view compared to a bottom view). This process must be done manually since it is difficult to gather enough training sets with differing view. Typically speaking, images of a product only have a few angles and typical other angles are of low-quality. While the process is resource and time intensive, it is a critical step to generate the necessary training sets for the algorithm.

*The need to create a fast algorithm*

With the need for almost-instantaneous detection speed for theft prevention, it is important for the algorithm to have a fast response rate. However, the trade-off is that since the algorithm must detect on a product level, the computation requirement is also substantially higher. (i.e. It requires more computation power to differentiate between a can of Pepsi or Coke than to classify it is a can).

**Algorithm Design**

Logistic Regression

With the spread of monitors in supermarkets, supermarkets are getting tons of videos everyday. It is not satisfying only using monitors as deterrence or checking theft by human power. These videos can be used to implement an algorithm to give real-time warning of theft behavior based on detecting and categorizing goods on shelves from time to time. The track of certain good can be followed and a warn signal can be given if the track does not end in a basket or cart. As a result, a classification model is needed to give high accuracy detection of goods with quick response time and limited memory size usage by training these videos. The main characteristic of this task is all inputs are videos (or pictures) and the scale is extremely large, hence logistic regression and neural networks should be considered.

Logistic regression is an approach in this Git, it gives the quickest response with 98% accuracy in a binary classification. However, as the training set getting bigger, logistic regression becomes harder to converge and accuracy decreases significantly, which means it cannot provide satisfying result in practice, with thousands of goods and millions of pictures. As a result, logistic regression is not the ideal method to solve this problem, therefore this Git focuses on implementation of a specific neural network – Xception.

Xception is the proper approach to detect and categorize goods. It is a variety of neural network focusing in figure recognition. It has two main characters. First, it utilizes Depthwise Separable Convolution, a method to significantly decrease the number of calculations in the convolutional layer. As inputs from the monitoring system is getting sharper, which means the size of each single input is getting bigger, this structure can efficiently prevent the increase in response time as the bigger the size is, the more efficient this structure is. Second, it allows “skip connection”, which allows double or triple layers skip. This guarantee the model can still converge and be trained with millions of input pictures and it can reduce response time. At the same time, this structure allows a deeper network to be trained and thus the accuracy of recognition can be increased as well.

To realize the function of goods detection and categorization, the key optimization factor is accuracy. Response time and memory size are satisficing metrics. Xception is the model can converge with huge data input and bring the highest accuracy while ensuring the response time and memory size needed is acceptable for industry usage, thus Xception is chosen for this Git.

**Image Library Creation**

A product image library must be created to generate

**Placement of Cameras**

**Camera Field of View**

1. [**^**](https://en.wikipedia.org/wiki/Electronic_article_surveillance#cite_ref-18) Sidebottom, Aiden; Thornton, Amy; Tompson, Lisa; Belur, Jyoti; Tilley, Nick; Bowers, Kate (2017-05-30). [*"A systematic review of tagging as a method to reduce theft in retail environments"*](https://crimesciencejournal.springeropen.com/articles/10.1186/s40163-017-0068-y). Crime Science. **6** (1): 7. [*doi*](https://en.wikipedia.org/wiki/Digital_object_identifier):[*10.1186/s40163-017-0068-y*](https://doi.org/10.1186%2Fs40163-017-0068-y). [*ISSN*](https://en.wikipedia.org/wiki/International_Standard_Serial_Number) [*2193-7680*](https://www.worldcat.org/issn/2193-7680). [↑](#footnote-ref-1)
2. Add notes [↑](#footnote-ref-2)
3. https://www.businessinsider.com/walmart-tracks-theft-with-computer-vision-1000-stores-2019-6 [↑](#footnote-ref-3)
4. http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php [↑](#footnote-ref-4)
5. https://arxiv.org/pdf/1810.08293.pdf [↑](#footnote-ref-5)