Towards Identification of Packaged Products via Computer Vision

Convolutional Neural Networks for Object Detection and Image Classification in Retail Environments

Klaus Fuchs, Tobias Grundmann, Elgar Fleisch
D-MTEC, Auto-ID Labs ETH/HSG
ETH Zurich
Zurich, ZH, Switzerland
{fuchsk, tobiasgru, efleisch}@ethz.ch

ABSTRACT

Identification of packaged products in retail environments still relies on barcodes, requiring active user input and limited to one product at a time. Computer vision (CV) has already enabled many applications, but has so far been under-discussed in the retail domain, albeit allowing for faster, hands-free, more natural human-object interaction (e.g. via mixed reality headsets). To assess the potential of current convolutional neural network (CNN) architectures to reliably identify packaged products within a retail environment, we created and open-source a dataset of 300 images of vending machines with 15k labeled instances of 90 products. We assessed observed accuracies from transfer learning for image-based product classification (IC) and multi-product object detection (OD) on multiple CNN architectures, and the number of images instances required per product to achieve meaningful predictions. Results show that as little as six images are enough for 90% IC accuracy, but around 30 images are needed for 95% IC accuracy. For simultaneous OD, 42 instances per product are necessary and far more than 100 instances to produce robust results. Thus, this study demonstrates that even in realistic, fast-paced retail environments, image-based product identification provides an alternative to barcodes, especially for use-cases that do not require perfect 100% accuracy.

https://doi.org/10.1145/3365871.3365899**CCS CONCEPTS**

• Computer vision • Field studies

KEYWORDS

Computer vision, Product identification, CNN

ACM Reference format:

Klaus Fuchs, Tobias Grundmann and Elgar Fleisch. 2019. Towards Identification of Packaged Products via Computer Vision: Convolutional Neural Networks for Object Detection and Image Classification in Retail Environments. In *Proceedings of ACM The 9th International Conference on the Internet of Things (IoT 2019). ACM, Bilbao, Spain, 8 pages.* https://doi.org/10.1145/3365871.3365899

1 Motivation

Although the computer vision field has advanced significantly over the recent years, object detection and image classification of packaged products in retail environments still remains at its infancies, thereby limiting the development of novel, more natural human-product interactions. Ever since the introduction of AlexNet [21], the consequential development of (deep) convolutional neural networks (CNN) [13, 48] and the increased affordability and availability of ever more performant hardware has advanced object detection and image classification across many domains, leading to novel, advanced applications. For example, the detection of road signs and traffic situations through computer vision has enabled the development of autonomous cars and the real-time translation of image-encoded words has eased the life of travelers [27, 34]. Further application areas of computer vision include guidance of robots, interpretation of satellite images, analysis of medical images (e.g. from x-ray scans), photomicrographs in microscopy, and industrial inspection [6]. Surprisingly, published research on image classification and object identification of packaged food products still remains at an early stage, despite the almost daily frequency with which consumers interact with packaged products. This is counterintuitive, as first retailers have recently started to introduce related technologies, such as computer vision supported self-checkout [1, 45, 46], albeit on small scales featuring only small and compact store layouts, requiring strictly fixed shelf maps (planograms), supporting only a limited product portfolio, and relying on fixed, stationary cameras.

One barrier towards public research on computer vision on packaged products has been the lack of publicly available labelled datasets, as producers have been hesitant to release images of their own products into the public. Therefore, research on identification of packaged products relies on relatively few, rather small and quite old datasets [9, 17]. Existing studies on identifying packaged products via computer vision indicate promising potential [8, 17, 43, 44], but they rely on such limited datasets and are conducted under resource-intense lab conditions, and do therefore not prove real-world applicability of computer vision based product identification. Although standards on product identifiers (e.g.

October 22–25, 2019, Bilbao, Spain

ISBN 978-1-4503-7207-7/19/10…$15.00
https://doi.org/10.1145/3365871.3365899

GTIN [10]) allowed the aggregation of public ingredient databases on packaged products (e.g. Openfoodfacts.com with over 800 000 product entries [31]), there does not yet exist a similarly large aggregation of labelled product images. Therefore, in contrast to other datasets in related fields (e.g. FoodNet101 with 101 000 labelled images of composed dishes [32]), no large-scale dataset of labelled images of packaged products for training of computer vision models has been published yet. In addition, the sheer magnitude of different product classes compounds this issue, as supermarkets offer thousands of products that may even be changing in their package appearance over time.

After overcoming the existing barriers and given the expected adoption of mixed reality headsets over the next years, the simultaneous location and mapping (SLAM) of packaged products [2] via wearable headsets will allow for novel human-product interactions, for example by identifying contexts from shelves or areas in the supermarket. Today, consumers and retail employees rely on barcode scanning with handheld devices (e.g. smartphones) to identify packaged products, e.g. for checking nutrients or a price. The existing optical scan is limited since it is only capable of identifying one item at a time, requires close proximity, and the line-up of scanner and the product barcode, which is usually on the backside. Thus, the user must actively trigger the process, requiring effort and high saliency on the user side. Similarly for recently suggested consumer-oriented augmented reality applications via markers tagged on the shelf, users must actively align their device with the shelf's marker to allow for the correct identification [3]. Given the fast-paced retail environment in which consumers pass thousands of products within few minutes, neither shelf-based markers, nor barcodes allow for convenient, hand-free, passively triggered human-product interaction. Therefore, a new product identifier based on the visual layout of each packaged product in order to support automatic identification (e.g. via the cameras of wearable mixed-reality headsets) of packaged products is needed. CNN-based feature extractors are a strong candidate as they are able to transform matrix data of a product image into a one-dimensional vector. Recent progressions in deep learning and representation learning have created many such feature extractors [23], which are generic and multi-purpose which can be applied to a range of objects [18]. Given that a robust CNN-based product detection can be achieved, this would then allow for novel handsfree human-product interaction. First, detecting products from wearable cameras does not require any additional store-installed hardware (e.g. fixed cameras) or up-to-date planogram interfaces, as the identification solely relies on the video feed (e.g. from the wearable headset of consumers or employees). Second, the identification of multiple products within single frames of the video feed at the same time becomes possible. Third, given a sufficient object detection accuracy (i.e. high mean average precision (mAP)), the retrieval of product dimensions becomes feasible. Finally, also the relative positioning of user and products can be approximated via spatial computing [19]. With these advantages, the display of product-related information or services can be achieved. For example, supporting consumers in finding the healthiest product within a shelf while remaining hands-free by simply viewing through a mixed-reality headset, as suggested by El Sayed [7] and Microsoft [16]. Also, inventory analysis could become more effective and efficient via automated robots or employees wearing head-sets that allow for detection of misplaced items, false labels or any other deviation from the current planogram. Also, store planning, employee training and even theft prevention could be enhanced or supported by computer vision-based identification of packaged products.

Figure 1: Vending machine and potential use-case of wearable mixed reality headsets.

Therefore, this paper aims to contribute towards the development of image classification and object detection on packaged product data under realistic 'in-the-wild' conditions.

2 Related Work

Convolutional neural networks (CNNs) have been applied in many different applications, but still require context-specific adaptation in order to enable image classification and object detection. Compared to traditional classification methods, CNNs do not rely on pre-defined mappings (e.g. planograms), markers (e.g. barcodes, QR codes) or heuristically hand-designed algorithms for detection of objects in images but on learning how to classify from data. Thus, making them a natural choice for implementation of a computer vision-based detection of packaged products in a retail environment [2]. The choice of a certain CNN architecture and its hyper-parameters are context-, task- and design-dependent and cannot necessarily directly be transferred to each new context or dataset without adaptation and consequential testing on actual, realistic context-specific data. There are no guidelines in terms of how to choose an optimal training dataset of images for an CNN, however the more data and the more variance present in a dataset, the higher the likelihood of successfully classifying an unknown instance of a known object. Therefore, labelled training data collected under realistic conditions such as the ones collected and assessed for this study becomes a valuable tool for development of computer vision-based solutions.

After collection of sufficient training data, a CNN is then validated and tested on previously unseen data in order to assess the network’s accuracy and ultimately ability to support a certain task, e.g. image classification (i.e. which product displayed on an image patch) and object detection (i.e. which products are located where on the image and what are their dimensions). With enough training data and a suitable architecture choice, CNNs can be taught to extract features on multiple levels, and in some applications such as face recognition even achieves above human performance [33]. In terms of image classification, recent accuracy rates range above 84% [42] on the Imagenet dataset [36]. For object detection tasks, recent results show accuracy (i.e. mean average precision as defined by the COCO challenge [25]) ranks around 60% [30] on the Open Images dataset [22]. Object detection extends image classification by the added probabilistic step of locating the position of the object, which increases the complexity of achieving high accuracy compared to classifying a perfectly masked image.

Today, there exists a wide range of established CNN architecture with varying levels of complexity, able to combine image classification and object detection. MobileNet [12, 38] is a CNN aimed to support implementations on mobile devices with limited computational capabilities. Next, ResNet [4, 11] (presented at ILSVRC2015) by Kaiming He et al. features heavy batch normalization and 152 layers, making it computationally more intense than MobileNet. Finally, Inception v4 [40] was demonstrated to outperform ResNet at ILSVRC2017, albeit featuring again higher complexity. Each CNN architecture type differs in the required computational effort involved in predicting objects within the video frame and classifying them accordingly. Given the motivation for computer vision and the challenges around implementing them within a realistic, fast-paced retail environment, we therefore assess the potential of proven CNNs (i.e. MobileNet, ResNet, Inception v4) to enable image classification and object detection within a realistic retail environment, using cloud-based computing infrastructure as well as consumer devices.

3 Setup Design

The goal of this study is to contribute to the research of image-based product identification in retail environments.

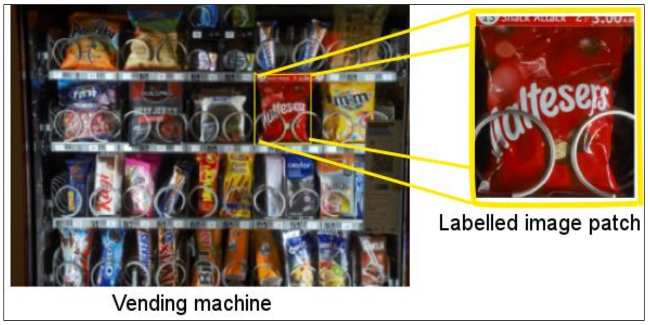

Figure 2: Vending machine picture with a labelled, rectangular image patch (e.g. Maltesers snack 100g). (Product classes: 90 in total, of which N=39 with over 100 labelled image patches).

3.1 Vending Machine

For the purpose of this study, we chose vending machine from Selecta (Figure 1), the European market leader (125’000 machines worldwide), as the study location and source of data. The main reasoning behind this decision was to focus on a retail setup that had a limited number of products for which the research team had to manually image patches accordingly. The majority of Selecta machines have an equally assorted product collection with typical snacks (e.g. chocolate bars, crisps) and beverages (e.g. Coca Cola, Red Bull). Further, we aimed to focus on vertically displayed packaged products (which is the wider established norm in retail environments). In addition, as vending machines are present in many regions globally, this study could potentially be reconstructed in other regions again. Last, but not least, the conduction of a mixed-reality headset mediated user study, which is not part of this paper was conducted, leveraging the object detection and image classification developed in this study (Figure 7).

The selection of the vending machine as focus allows for a certain level of generalizability, since the study could potentially be reproduced and applied in similar form across vending machines internationally. Nevertheless, the generalizability towards other retail layout with non-vertical product representation and thousands of products remains a limitation, requiring further work that goes beyond of the focus of this study.

3.2 Research Questions

In order to assess the potential of CNNs to enable image classification and object detection of packaged products in the vending machine, we assess the accuracy and requirements for image classification and object detection separately in this study. Still, since classification and detection are tasks that build upon each other we decided to create one large object detection dataset and evaluate both, classification and detection tasks, via labelling image patches from the same sample (Figure 2).

Image classification operations are the backbone of object detection operations as they classify an object that was found to be potentially interesting within an image in a first step. Especially due to the absence of publicly available labelled training data for packaged products, and since labelled product images are potentially expensive to acquire or generate for the millions of products that exist in the world, we decided to evaluate how many labelled image instances of a product are required to achieve suitable performance. Thus, to succeed in object detection the first corner stone is a successful image classification performance thus we pose the following two research questions (RQ).

RQ1. Can current CNNs yield a sufficiently high accuracy for image-based product classification in a realistic retail environment?

RQ2. How many instances of product images are required to achieve 90% (95%) accuracy in image-based product classification in a realistic retail environment?

Based upon the results of the image classification tasks we are going to address the same objective for object detection (which includes image classification as a subsequent task).

RQ3. Can current CNNs in combination with object detection networks (ODNs) yield a sufficiently high enough mean average precision for image-based product detection?

RQ4. How many instances of product images are required to achieve 90% (95%) mean average precision (mAP) in image-based product detection in a fast-paced realistic retail environment?

In the context of this study, 'fast-paced' as part of the research questions aims to take into account realistic conditions in a retail environment. Namely, that consumers or employees expect instant results (i.e. within 1 second), as they are potentially passing thousands of products on their journey through a retail store. The duration of one second proved well-suited in multiple user tests, when equipping a user with a HoloLens device in front of the vending machine type used in this study (Figure 1 and Figure 7). With different CNN architectures featuring varying complexity levels and accuracy rates, the trade-off between latency, accuracy, number of available training images, computational environment (i.e. mobile device or cloud backend) are the focus of this study. This allows us to use an alternative option to reduce *k* (the number of images used for training) for the object detection task by turning the object detection task into a video object detection task with motion and spatial information where we try to detect the object in i.e. one second by allowing the neural net to detect a product on as many images as it can process in this time. We can use spatial information from algorithms such as SLAM [2, 29, 37], video object tracking [20, 24] or optical flow to track [15, 28] the position over multiple frames. The prediction scores from those multiple frames are then average-pooled to choose the maximum-confidence prediction across all frames (within one second). In this case, the latency of the neural net architecture becomes another parameter to influence the mAP.

3.3 Convolutional Neural Networks (CNNs)

There are many CNN architectures available for image classification and object detection. With Inception Resnet V2 [41], Resnet 50 V2 [11] and Mobilnet V2 [39] as classification networks, we use a subset of popular architectures of differing complexity. The corresponding, implemented object detection networks (ODNs) are listed below:

1. **Inception.** Inception Resnet V2 [41] for classification, with Faster RCNN [35] for object detection
2. **Resnet.** Resnet 50 V2 [11] for classification, with SSD and Focal Pyramid Networks (Retinanet) [26] for object detection.
3. **Mobilenet.** Mobilnet V2 [39] for classification, with SSD [47] for object detection.

We train all neural networks with finetuning from existing checkpoints, for image classification those checkpoints are from Tensorflow hub and are pretrained with the ImageNet 2012 dataset [36] and for object detection those checkpoints are from the Tensorflow object detection API [14] which are pretrained with the COCO 2014 dataset [25].

3.4 Image Datasets

To ensure consistency across multiple $k \in [0,100]$ images, we chose a subset of N=39 products from the vending machine, for which the total labelled dataset includes at least 100 (training) + 20 (test) instances per product class. Thus, we can evaluate the CNN architectures for any *k* smaller or equal to 100 for the N=39 products. For product classification we excluded the 20 instances per class as a holdout dataset. Similar for product detection, the holdout dataset including randomly sampled 20% of the images that guarantee there are at least 100 instances of each product in the training set. This means the test set for object detection was unbalanced in number of classes but as well included at least 20 instances of every product.

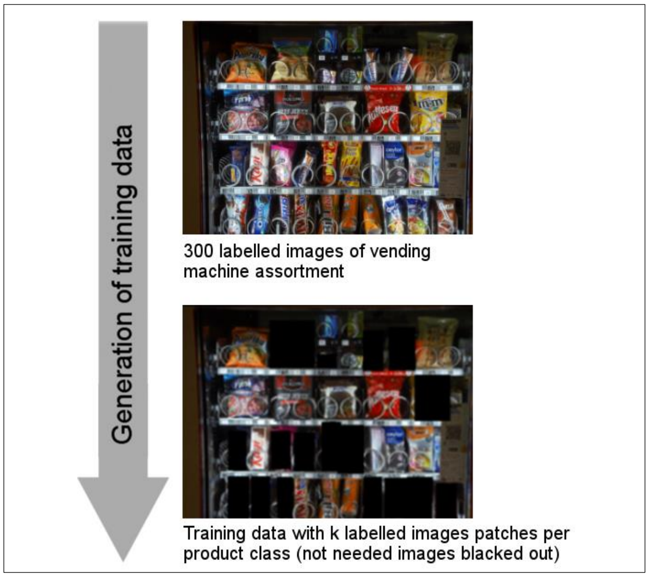
Generation of training data
300 labelled images of vending machine assortment
Training data with k labelled images patches per product class (not needed images blacked out)

Figure 3: Generation of training data with *k* images per product class for the training of the neural networks.

For each *k* smaller than 100, subsamples of the entire training dataset were chosen. For the image classification, the images were simply excluded from the dataset (Figure 3**)**. For the object detection task, the differing frequency of certain product instances that were present multiple times in vending machine, makes this impossible (e.g. RedBull is present four times in every Selecta vending machine). We randomly chose instances in the training set to be blacked out until every labelled product exists exactly *k* times across all training images (Figure 3**).** Training images without labelled instances (i.e fully blacked out) were excluded. For every *k*, the training dataset looks slightly different, since there are different images blacked out, the holdout set is however the same for all datasets. During training, we further randomly subsampled from the training set by cutting out random image patches of the training images, to create training data with varying box sizes and positions of the products on the image.

3.5 Technical Devices

The devices used to record images and test the latency are a Microsoft Hololens and a OnePlus 6t Android smartphone. For training and inference the Google Cloud with P100 GPU instances and TPU v2 instances was used.

4 Results

To address the research questions RQ1 to RQ4, we compare the performance of the three CNNs that were introduced in the previous chapterm, i.e. Inception Resnet V2 (‘Inception’), Resnet 50 V2 (‘Resnet’), Mobilenet V2 (‘Mobilenet’), on the image classification and object detection tasks using the generated image datasets from the vending machine setup.

4.1 Image Classification

First, we address the image classification problem. Concretely, we test the potential of the three CNNs to correctly classify labelled image patches, i.e. identify which one of the in total N=39 products from the vending machine sample is displayed in an image snippet.

RQ1. Can current CNNs yield a sufficiently high accuracy for image-based product classification in a realistic retail environment?

To address RQ1, we used classes with a sufficiently large k of labelled image patches and selected 100 (training) + 20 (test) instances per each of the N=39 applicable product classes as the training dataset. As depicted in Figure 4, when using the entire training dataset, we observe relatively high accuracy across the three CNNs: Resnet achieved a 95% accuracy, Mobilenet reached a 97.1% accuracy and Inception even demonstrated an accuracy rate of 97.7%. Given that CNN-based accuracy rates in image classification problems in recent studies range above 84% [42], we can confirm RQ1 as it seems very well possible to reach accuracies above 95% percent with all network architectures as Figure 4 shows.

RQ2. How many instances of product images are required to achieve 90% (95%) accuracy in image-based product classification in a realistic retail environment?

As publicly available labelled datasets for packaged products are lacking, we addressed the research question on how many images are needed at least to achieve a sufficient accuracy in order to support reliable image classification. Especially since most retailers or brands to not share labelled images from realistic retail environment, the development of computer vision-based identification on packaged products is hindered. To address RQ2, we varied the number of *k* image patches that were used for training of the CNNs.

As depicted in Figure 4, we can report that as little as only six instances of a product are enough to create a classifier with the Inception CNN that is able to classify products with an accuracy of 90% and 26 instances for an accuracy of 95%. The Inception architecture with its relatively high complexity is the CNN with the lowest *k* necessary to achieve high accuracy rates. But the other architectures follow swiftly with Resnet requiring 10 images for 90% accuracy and 35 for 95%, as well as Mobilenet needing 20 instances for 90% and 51 images for 95% accuracy. To answer RQ2, we conclude that at least six images can be sufficient for less critical applications, where false positives or false negatives are not harmful, and where 90% accuracy is sufficient. For more robust image classification, at least 26 images or more seem necessary. While for lower *k* there are large differences between the different networks, for higher *k* the architectures converge and achieve similar accuracy rates.

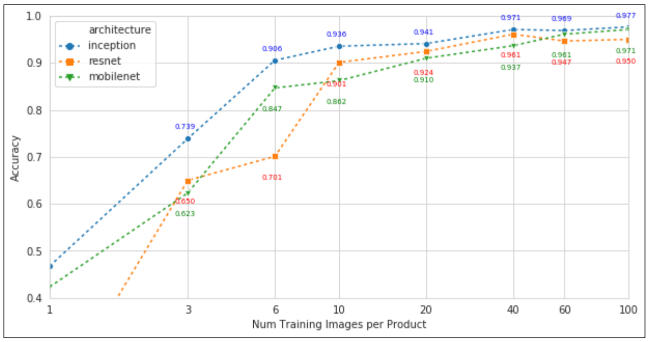

Figure 4: Product classification accuracy per number of training instances per product.

4.2 Object Detection

Next, we assessed the potential of object detection networks (ODN) to support the product detection within images of the retail environment. Again, the study context is the vending machine to represent a realistic retail environment. The object detection task includes the identification of image patches that contain products and the subsequent correct classification of the detected image patches (Figure 2). This means, in order to achieve a high accuracy, the ODN have to achieve both, i) detecting the position of objects in the vending machine assortment and ii) correctly classify the detected objects against the labelled ground truth. Finally, the achieved performance is assessed by calculating the mean average precision (mAP) [25] for an intersect over union (IoU) [25] of 0.5, as recommended by similar studies in other fields.

RQ3. Can current ODNs yield a sufficiently high enough mean average precision for image-based product detection?

To address RQ3, we used a sufficiently large dataset and therefore decided to leverage the entire 100 labelled training images of the vending machine assortment with its subsequent labelled image patches for the N=39 product classes. As depicted in Figure 5, only the Resnet/Retinanet architecture can achieve a mAP of over 0.9 when using the 100 images and patches of the vending machine assortment (93.4%). The Resnet/Retinanet can even reach an accuracy of 98.6%, when used with all available data (i.e. classes have varing numbers of training images, from 100 to 1000 images, with a mean around 250. Due to the uneven distribution, the 98.6% are not depicted in Figure 5). Inception was able to break 90% when the entire data set was used with 94.5% mAP and the Mobilenet architecture was not able to achieve a mAP of over 90%. Therefore, we confirm RQ3, that object detection can be achieved with the right architecture choice of combination of CNN and ODN.

RQ4. How many instances of product images are required to achieve 90% (95%) mean average precision (mAP) in image-based product detection in a fast-paced realistic retail environment?

As depicted in Figure 5, during the object detection task, only the Resnet/Retinanet architecture reached a mean average precision (mAP) [25] of 0.9 for an intersect over union (IoU) [25] of 0.5 for a *k* lower than 100, requiring at least 42 images. None of the neural networks can achieve an mAP of over 0.95 with less than 100 instances and only Resnet/Retinanet with the full training dataset. Therefore, for RQ4 we can say the k and thus the effort for object detection are far larger and we cannot even calculate a definite number for 95% mAP. The architecture seems to play a much larger role as the focal pyramid network architecture of the Retinanet was far superior when used to detect objects. However an explanation might be just the perfect amount of regularization through downsized images of 320 x 320 pixels.

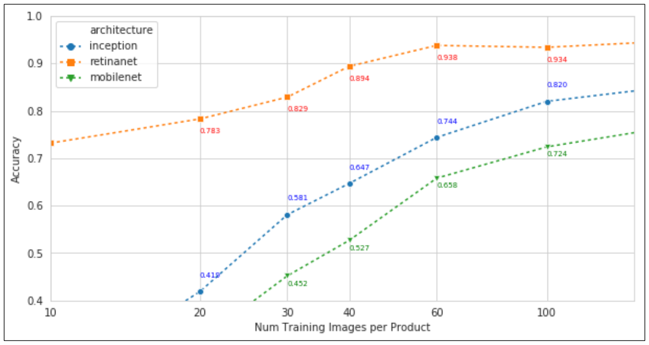

Figure 5: Product detection mean average precision (mAP) per number of training instances (bounding box) per product.

In realistic circumstances within retail environment, where there is most likely a video stream, the error rates can theoretically be reduced further by using multiple frames within a second of a video to detect the object, using a mean confidence pooling approach. For a network with already high accuracy such as 93.8% for Retinanet, thus predicting the right object is far more likely than predicting the same wrong object multiple times, and a possible frame rate of 6 per second as we have evaluated for a detection through gRPC on the cloud (in line with [14]), the accuracy could reach nearly 100% if we use the following formula to calculate the chance of not finding the object with those frames $error = (1 - mAP)^{fps}$. In this case not only the accuracy and the architecture are important but as well the latency and the hardware, as for example Mobilenet can run at 12 fps on a mobile device (tested with authors mobile phone). Those considerations will be very important for real use-cases in industry.

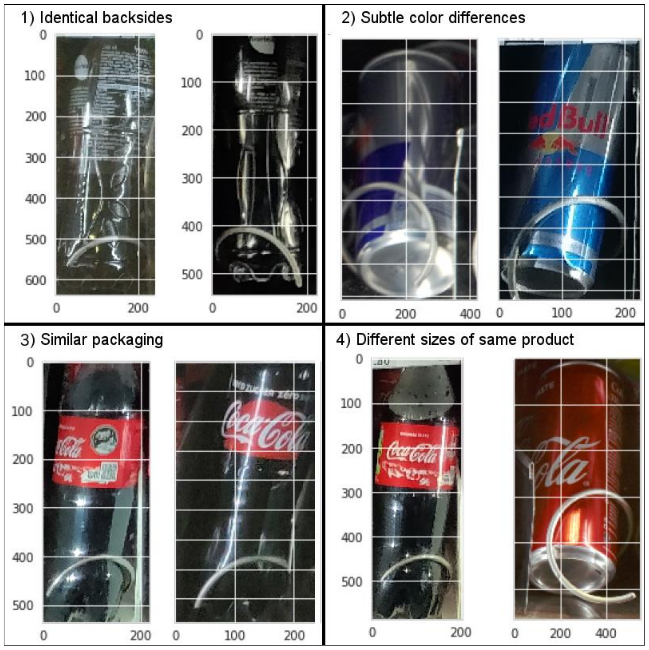
1) Identical backsides
2) Subtle color differences
3) Similar packaging
4) Different sizes of same product
0
100
200
300
400
500
600

Figure 6: Edge cases that were challenging for product classification via computer vision in retail environments.

4.3 Edge Cases

Besides addressing the research questions, the study also allows for discussion of challenges in computer vision-based product identification as the edge cases where predictions went wrong can be retrieved (Figure 6). In most retail-related use-cases, false positives or false negatives might only affect the user experience, since a consumer might see a wrong advertisement in mixed reality, or an employee might have to scan a barcode if an item was not detected correctly via the headset through computer vision. Thus, compared to other computer vision applications such as autonomous cars or tumor predictions, the cost of misclassification is mostly of low severity user experience or low financial impact.

The most common misclassifications resulted from the following four edge cases. Products might have similar backsides such as Fuze IceTea Lemon (1.1) and Fuze IceTea Peach (1.2) which mainly differ in the color of their lids that were hidden by the shelf. Some products also feature only subtle color nuances, such as Red Bull (2.1) and Red Bull sugarfree (2.1) which both feature varying shades of blue. Some producers use similar packaging among different products, such Coca Cola (3.1) and Coca Cola Zero (3.2) which have very similar fonts, colors and designs. Finally, a product can come in different sizes, such as Coca Cola 0.5L PET (4.1) and Coca Cola 0.33L can (4.2), thereby impairing accuracy of object detection within retail environments.

5 Discussion

In this study, we addressed the potential of image classification and object detection to identify packaged products within a typical retail environment. Computer vision-based identification of packaged products is still a nascent field and is lacking publicly available datasets and published research. Still, image classification and object detection are promising approaches towards visual product identification as they allow for several advantages over conventional marker-based or barcode-based identification, such as passively triggered detection (instead of active scanning), no need for store-installed hardware or planogram interfaces, identification of multiple objects simultaneously, retrieval of product dimensions and relative user positioning via spatial computing.

Regarding the image classification (IC) as the correct identification of a product within an image patch, our study demonstrated feasibility and observed accuracy rates of 95%-97.7% (RQ1) with 100 images per class through transfer learning. Further, the assessment of the minimum number of required images to support IC (RQ2) concluded that at least six (for 90% accuracy) to 26 images (for 95% accuracy) are required for training of relevant models. This study therefore confirms the feasibility of IC on packaged products, and also demonstrated that the integration of IC within a research or real-world implementation is already feasible with limited amount of investment and effort as current labeling services indicate a pricing of 35$ per 1000 images, leading to labeling costs of just 0.91 USD per product to support IC with an accuracy of 95%.

Object detection (OD) as extension to IC involves identifying the position and the class of one or multiple objects within an overall image. For RQ3, we observed that OD is possible, albeit heavily depends on the architecture choice, as out of three architectures only the Resnet/Retinanet architecture succeeded in achieving at reaching an mAP of over 90% (i.e. 93.4%) after having been trained on 100 images. Addressing RQ4 revealed, that we can confirm that it is feasible to achieve an mAP of 90% with 42 instances per class, also using the Resnet/Retinanet architecture. An mAP of over 95% could only also only be observed with the Resnet/Retinanet architecture, but only with far more than 100 images per product. Theoretically, the accuracy for already performant networks can be improved by using video data and factoring latency and infrastructure choices. This study therefore also confirm feasibility of OD on packaged products, albeit demonstrating that the effort for object detection is far higher, as it not only requires more data, but labeling this data is as well more expansive at 69$ per 1000 results, leading to cost for 2.90 USD per product class for an mAP of 90%.

It is very hard to solve computer vision problems for all products, viewing angles, shelf situations. Some objects where it is even impossible for humans to differentiate the products from the backside without reading the detailed print, such as the edge cases mentioned in this study. Still, the overall accuracy of around 95% should not be a critical problem for future human-product interaction use-cases such as advertisement, displaying nutrients or recipes or services. However, when 100% accuracy is necessary the barcode should still be used. For now, computer vision cannot provide such reliability, but their other advantages make them a promising and convenient, alternative identification technology.

6 Conclusion

Given the nascent state of computer vision on packaged retail products, this paper contributes to the development in this field by demonstrating feasibility of object detection and image classification under realistic circumstances within an in-the-wild retail environment. Contributions to research include the adaption to a new domain, with results for realistic circumstances. Further we show promising object detection results for multi object detection without detection groups. All of which provides the viability of the technology for future human interaction research and use-cases. Managerial implications include the dataset requirements for enabling current retail environment to support computer-vision based product identification, combined with the call for public labeled dataset needed for the field, Finally, this study contributes to the development by open sourcing the dataset of labelled images to advance the research in this area. We believe that the community should engage in this process by extending public databases such as 'Verified by GS1' or Openfoodfacts with image collection, labeling and segmentation, to support the basis for similar research at larger, potentially one day global scale.

For example, the display of nutritional information in mixed reality displayed as digital overlay on the vending machine surface detected through spatial computing demonstrates a potential use-case of computer vision-based product identification that has previously not been possible with barcode scanning (Figure 7). Such product-consumer interactions could be supported by GS1's Digital Link standard [5], that aims to enable digital interactions between consumers and products.

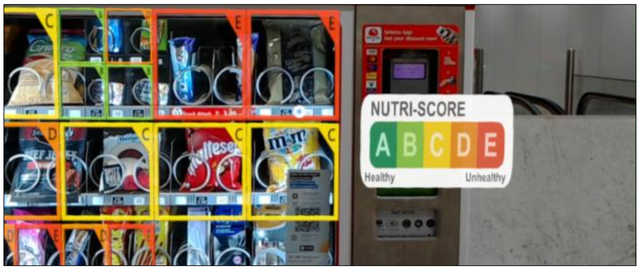

Figure 7: Computer vision on packaged products allow novel human-object interactions, e.g. passively triggered diet interventions in absence of markers or barcodes (Screenshot of Microsoft HoloLens viewing the study's vending machine).

The findings of this paper shall be considered under certain limitations. Currently, there is no feedback loop to gather for image interpretation, since the research questions evolved around initial feasibility of IC and OD. which can then use a feedback loop to gather more data. Future progress in object detection and zero-shot learning can reduce the number of required items, as well as improvements in hardware such as specialized inference chips such as the mobile TPU presented recently by Google. Finally, the number of classes is far from the millions of products that exist in the world or the thousands for a single shop. Increasing the number of classes drastically always creates a new challenge to work with. Thus, we will continue to explore product identification from visual representation and try to establish an approach with zero-shot learning capabilities. For the future, interesting areas of research are how to reduce the required *k* to detect images to possibly even one, through the use of ensemble methods, such as detection of text, product group detection and positional data of products in shelves and vending machines, through IoT data or by using digital anchors in a version of SLAM where products are simultaneously detected and mapped. All of this data combined in a global knowledge graph which allows to perfectly identify any product would allow for shopping experiences with completely new, more natural and immersive human-product interactions (for example via mixed-reality headsets as suggested in Figure 1 and Figure 7). Finally, the release of the 300 images of vending machines assortments with 15k labeled instances of 90 products is planned to stimulate research on computer vision on packaged retail products.

REFERENCES

[1] AiFi emerges from stealth with its own take on cashier-free retail, similar to Amazon Go: 2018. *https://techcrunch.com/2018/02/27/aifi-emerges-from-stealth-with-its-own-take-on-cashier-free-retail-similar-to-amazon-go/*. Accessed: 2019-06-15.

[2] Cadena, C. et al. 2016. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*. 32, 6 (2016), 1309–1332. DOI:https://doi.org/10.1109/TRO.2016.2624754.

[3] Csakvary, B. 2017. Msc Thesis: Promoting healthier food choices with the application of Augmented Reality. *Wageningen University - Department of Social Sciences Marketing and Consumer Behaviour Group (MCB)*. (2017).

[4] Dai, J. et al. 2016. R-FCN: Object Detection via Region-based Fully Convolutional Networks. (2016). DOI:https://doi.org/10.1016/j.jpowsour.2007.02.075.

[5] Digital Link: 2019. *https://www.gs1.org/standards/gs1-digital-link*. Accessed: 2019-08-04.

[6] Du, C.J. and Cheng, Q. 2014. Computer vision. *Food Engineering Series*.

[7] ElSayed, N.A.M. et al. 2016. Situated Analytics: Demonstrating immersive analytical tools with Augmented Reality. *Journal of Visual Languages and Computing*. (2016). DOI:https://doi.org/10.1016/j.jvlc.2016.07.006.

[8] Geng, W. et al. 2018. Fine-Grained Grocery Product Recognition by One-Shot Learning. 2, (2018), 1706–1714.

[9] George, M. et al. 2015. Fine-Grained Product Class Recognition for Assisted Shopping. *Proceedings of the IEEE International Conference on Computer Vision* (2015).

[10] Global Trade Item Number (GTIN) | GS1: 2015. .

[11] He, K. et al. 2015. Deep Residual Learning for Image Recognition. *Arxiv.Org*. (2015). DOI:https://doi.org/10.1016/0141-0229(95)00188-3.

[12] Howard, A.G. et al. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *ArXiv*. (2017). DOI:https://doi.org/arXiv:1704.04861.

[13] Huang, G. et al. 2017. Densely connected convolutional networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. 2017-Janua, (2017), 2261–2269. DOI:https://doi.org/10.1109/CVPR.2017.243.

[14] Huang, J. et al. 2017. Speed/accuracy trade-offs for modern convolutional object detectors. *arXiv*. (2017).

[15] Ilg, E. et al. FlowNet : Learning Optical Flow with Convolutional Networks.

[16] Jerauld, R. 2017. Warable Food Nutrition Feedback System. US009646511. 2017.

[17] Karlinsky, L. et al. 2017. Fine-grained recognition of thousands of object categories with single-example training. *CVPR*. (2017).

[18] Karpathy, A. and Leung, T. 2014. Large-scale Video Classification with Convolutional Neural Networks. *CVPR 2014* (2014).

[19] Kor, A.L. 2019. Qualitative spatial reasoning for orientation relations in a 3-D context. *Advances in Intelligent Systems and Computing* (2019).

[20] Kristan, M. et al. 2017. The Visual Object Tracking VOT2017 challenge results. *Proceedings of the 2017 International Conference of Computer Vision, Computer Vision Foundation*. (2017), 1949–1972.

[21] Krizhevsky, A. et al. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*. (2012), 1–9. DOI:https://doi.org/http://dx.doi.org/10.1016/j.protcy.2014.09.007.

[22] Kuznetsova, A. et al. 2019. The Open Images Dataset V4: Unified image classification , object detection , and visual relationship detection at scale. *arXiv*. (2019), 1–20.

[23] Lecun, Y. et al. 2015. Deep learning. *Nature*. 521, 7553 (2015), 436–444. DOI:https://doi.org/10.1038/nature14539.

[24] Li, B. et al. 2018. SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks. *arXiv*. (2018).

[25] Lin, T.-Y. et al. 2015. Microsoft COCO: Common Objects in Context. *Arxiv.Org*. (2015). DOI:https://doi.org/10.1109/CVPR.2014.471.

[26] Lin, T.Y. et al. 2017. Focal Loss for Dense Object Detection. *Proceedings of the IEEE International Conference on Computer Vision*. 2017-Octob, (2017), 2999–3007. DOI:https://doi.org/10.1109/ICCV.2017.324.

[27] Liu, Y. and Huang, H. 2016. Car plate character recognition using a convolutional neural network with shared hidden layers. *Proceedings - 2015 Chinese Automation Congress, CAC 2015* (2016).

[28] Lucas, B.D. 1981. An Iterative Image Registration Technique with an Application to Stereo Vision. 130, (1981), 121–130.

[29] Montemerlo, M. et al. 2002. FastSLAM: A Factored Solution to the Simultaneous Localization and Mapping Problem Michael. *AAAI-02 Proceedings*. 35, 3–4 (2002), 221–232. DOI:https://doi.org/10.1080/02703149.2012.684583.

[30] Open Images V4 Challenge Results: *https://storage.googleapis.com/openimages/web/challenge.html*.

[31] Openfoodfacts Mobile Applications: 2019. .

[32] Pandey, P. et al. 2017. FoodNet: Recognizing Foods Using Ensemble of Deep Networks. *IEEE Signal Processing Letters*. (2017). DOI:https://doi.org/10.1109/LSP.2017.2758862.

[33] Phillips, P.J. et al. 2018. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*. (2018). DOI:https://doi.org/10.1073/pnas.1721355115.

[34] Radzi, S.A. and Khalil-Hani, M. 2011. Character recognition of license plate number using convolutional neural network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2011).

[35] Ren, S. et al. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Arxiv.Org*. (2016). DOI:https://doi.org/10.1016/j.biocon.2012.08.014.

[36] Russakovsky, O. et al. 2014. ImageNet Large Scale Visual Recognition Challenge. *arXiv Artificial Intelligence (cs.AI)*. (2014).

[37] Saeedi, S. et al. 2018. Navigating the Landscape for Real-Time Localization and Mapping for Robotics and Virtual and Augmented Reality. *Proceedings of the IEEE*. 106, 11 (2018), 2020–2039. DOI:https://doi.org/10.1109/JPROC.2018.2856739.

[38] Sandler, M. et al. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. (2018). DOI:https://doi.org/10.1134/S0001434607010294.

[39] Sandler, M. et al. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. (2018). DOI:https://doi.org/10.1134/S0001434607010294.

[40] Szegedy, C. et al. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. (2016). DOI:https://doi.org/10.1016/j.patrec.2014.01.008.

[41] Szegedy, C. et al. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. (2016). DOI:https://doi.org/10.1016/j.patrec.2014.01.008.

[42] Tan, M. and Le, Q. V 2019. EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks. (2019).

[43] Tonioni, A. and Stefano, L. Di 2019. A deep learning pipeline for product recognition on store shelves. *arXiv*. (2019).

[44] Tonioni, A. and Stefano, L. Di 2019. Domain invariant hierarchical embedding for grocery products recognition. *Computer Vision and Image Understanding*. January (2019). DOI:https://doi.org/10.1016/j.cviu.2019.03.005.

[45] Viscovery: Computer Vision based Self-checkout Terminals: 2019. *https://www.viscovery.com/*. Accessed: 2019-06-15.

[46] Welcome to Amazon Go: 2018. *https://www.amazon.com/b?ie=UTF8&node=16008589011*. Accessed: 2019-06-15.

[47] Zhang, S. et al. 2017. Single-Shot Refinement Neural Network for Object Detection. (2017), 4203–4212. DOI:https://doi.org/10.1109/CVPR.2018.00442.

[48] Zhao, Z.-Q. et al. 2018. Object Detection with Deep Learning: A Review. 14, 8 (2018). DOI:https://doi.org/10.1016/0272-7757(84)90036-0.