Engineering Applications of Artificial Intelligence 133 (2024) 108452

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Datasets and methods of product recognition on grocery shelf images using computer vision and machine learning approaches: An exhaustive literature review

Ceren Gülra Melek [a,*], Elena Battini Sönmez [b], Songül Varlı [c]

[a] İstanbul Arel University, Büyükçekmece, İstanbul, 34537, Turkey
[b] İstanbul Bilgi University, Eyüpsultan, İstanbul, 34060, Turkey
[c] Yıldız Technical University, Esenler, İstanbul, 34220, Turkey

ARTICLE INFO

Keywords:
Product recognition
Planogram compliance
Computer vision
Machine learning
Grocery shelf images

ABSTRACT

A product recognition system recognizes all products on the shelf images and determines their positions. A business equipped with an automatic product recognition system has a convenient follow-up of many human-powered activities while increasing customer satisfaction. That is, product recognition stands out with its benefits such as tracking shelf layouts and stocking their status, and improving the shopping experience for customers, especially the visually impaired ones. However, product recognition is a challenging problem of computer vision in terms of the difficulty of obtaining and updating datasets and the breadth of the product scale. On the other hand, the number of studies on product recognition is constantly increasing by using various computer vision and machine learning methods, and effective solutions are offered to this problem. This paper provide a comprehensive review in the field of researches of product recognition on grocery store shelves. In this article, data sets and approaches used in the literature for the development of an automatic product recognition system are examined and compared, and their benefits and limitations are commented. Finally, a guideline is provided for future researchers and new perspectives for future studies are presented.

1. Introduction

Technological developments and industrial transformation have caused people's expectations to rise even in their ordinary affairs activities in daily life. In retailing, it has become desirable to carry out merchandising activities, which require manpower and waste time, by automatic systems in order to respond quickly and easily to the expectations of customers (Sabolcik, 2021). That is, the trend towards the development and use of automatic product recognition systems in grocery stores is increasing day by day due to the several benefits it provides.

1. Efficient inventory management: By automatically recognizing the products on grocery shelf images, the system can track which products are in stock, which products are running low, and which products are out of stock. This allows the store to efficiently manage its inventory, restock products when needed, and reduce waste.
2. Improved customer experience: With automatic product recognition on grocery shelf images, customers can easily find the products they are looking for and identify the products that meet their specific needs. This can lead to a more positive shopping experience and increase customer satisfaction.
3. More accurate pricing: Automatic product recognition can ensure that the correct price is charged for each item, reducing the risk of errors and the need for price checks. This can reduce mistakes and disputes at the checkout, improving the customer experience.
4. Personalized marketing: With automatic product recognition on grocery shelf images, the system can suggest complementary products or offer personalized discounts based on the customer's previous purchases. This can help increase sales and improve customer loyalty.

That is, grocery product recognition has become an important area of research, both academically and industrially, due to the various benefits it provides. Many approaches such as barcode (Fernandcz et al., 2018),

* Corresponding author.
E-mail addresses: cerenmelek@arel.edu.tr (C.G. Melek), elena.sonmez@bilgi.edu.tr (E. Battini Sönmez), svarli@yildiz.edu.tr (S. Varlı).

https://doi.org/10.1016/j.engappai.2024.108452
Received 19 December 2023; Received in revised form 25 March 2024; Accepted 12 April 2024
Available online 2 May 2024

(Kulyukin and Kutiyanawala, 2010), radio-frequency identification (RFID) (Wolbitsch et al., 2019), (Busu et al., 2011), computer vision, machine learning have been used for product recognition from past to present. However, using barcodes requires a specialized scanner or a smartphone camera to read the product information and identify the item. RFID can identify products by using tags that contain radio frequency information to transmit data between devices, allowing for the identification of products. Overall, barcode and RFID technologies have proven to be effective for product recognition in many applications, but they have some limitations. Barcodes need line-of-sight requirement to be scanned and lead to slow down the scanning process. Besides, barcodes are causing delays or errors in product recognition with vulnerability to physical damage. On the other hand, some security concerns and limited range problems may occur when using RFID for product recognition. In parallel, the development of advanced technologies in the field of machine learning (ML), and computer vision has made it possible for automatic product recognition systems to become more accurate and reliable. Machine learning algorithms can be trained to recognize new items and adapt to changing environments, making them a more flexible option for product recognition than traditional technologies such as barcodes or RFID. Computer vision algorithms can identify products with high accuracy by analyzing features such as shape, size, and color. This can help to reduce errors and improve efficiency in product recognition. Furthermore, computer vision and machine learning technologies can be integrated with other systems such as inventory management, supply chain, and logistics, making them a more seamless and efficient option for item recognition and tracking. Also, reducing the cost of hardware and software structures required for acquiring and processing data and being more accessible facilitated the application of computer vision and machine learning methods for product recognition in retail stores. These advantages have further fueled the trend towards the adoption of computer vision and machine learning based systems in the retail industry and took the attention of researchers.

This paper aims to provide a comprehensive review in the field of researches of product recognition on grocery store shelves; it presents the challenges and benefits of this problem, it discusses the using methods in detail, it compares the results, and provides a guideline for future researchers. Several reviews about product recognition on grocery shelf images are available in the literature. Study (Melek and Sonmez, 2017) presents an overview of the methods used in product recognition but it is limited to few papers. Survey (Wei et al., 2020) is restricted to previous researches the studies using deep learning techniques. The most comprehensive survey (Santra and Mukherjee, 2019) presents detailed information of the methods and results on product recognition but it covers only the papers published until 2018. The contributions of this work are three-fold: (1) it examines the publicly available datasets mostly used for the product recognition issue, (2) it lists the most recent articles on automatic product recognition on grocery shelf images using computer vision and machine learning approaches, and (3) it compares and classifies all studies among each other's.

The rest of the article consists of the following sections. Section 2 introduces the most common datasets used in the papers related to product recognition. Section 3 introduces briefly the studies related to product recognition. Section 4 details the classification of grocery product recognition studies according to the handled problems and using methods. Section 5 compares the existing studies according to their benefits and limitations. Finally, Section 6 gives the concluding remarks and suggestions for future works.

2. Grocery product recognition datasets

The item identification problem on grocery shelves aims to obtain the location and class information of the products in a shelf image. The first challenge researches face when start working on this topic is the knowledge of all available datasets that they can use to train and test the product recognition algorithms they will develop. Challenging datasets enable the development of more accurate and robust models by providing a diverse range of images of grocery items under various lighting conditions, backgrounds, and viewpoints. Also, the use of publicly available datasets helps to standardize the labelling and annotation of grocery products placed in shelf images, making it easier for researchers to compare and evaluate different models.

In the literature, several collections are available with different item groups, features, limitations and complexities. In this review, we focus on datasets for product recognition from shelf images ((Zhang et al., 2007)- (Chen et al., 2022)); there are also grocery product datasets that address different problems such as recognition of checkout product (Retail Product Checkout (RPC) dataset (Wei et al., 2019), Densely Segmented Supermarket (D2S) dataset (Follmann et al., 2018)) or recognition of grabbed a product with hand from the shelf (Take Goods from Shelves (TGFS) dataset (Hao, 2019), Products-6K (Georgiadis et al., 2021)); which are not considered in this survey. Table 1 lists the publicly available datasets for product recognition from shelf images; it gives information about product variety, number of images in train and test set, contents of annotation file and which products are annotated. More in details:

WebMarket (Zhang et al., 2007): The WebMarket dataset[1] is one of the first published datasets on grocery products. It consists of 3153 shelf images. With Jpeg format of size 2272×1704 or 2592×1944; pictures were recorded from 10 shelves of the grocery, which consists of 18 pieces 30-m-long shelves. Shelf images are captured in a sequence starting from one end of the first shelf. Adjacent images overlap slightly to ensure full coverage of each shelf. Each image covers three or four shelf levels; it is about 1.5 m in height and 2 m in width. There were no restrictions on the viewpoint or distance during photography, although most of the images are frontal views, and no special illumination was used. Besides the shelf images, template images consist of 102 individual products each of which has 2 or 3 instances from different viewpoints or distances. In total 300 template images were recorded in front of a plain background. Examples of templates and shelf images are given in Table 2. The ground truth information of the dataset contains only shelf image information of each product, the location information of the products on the shelf image is not available.

Grozi-120 (Merler et al., 2007): The Grozi-120 dataset[2] consists of in vitro (template images for training) and in situ (shelf images for testing) images. It is the first dataset prepared for the problem of object recognition and localization. There are a total of 676 product images, called in vitro, collected from the web, belonging to 120 different product classes. Each product class has a minimum of 2 and a maximum of 14 samples represented in the images. Also, in situ images consist of 11,194 shelf images captured in every 5 frames of a 30 min length video of 29 distinct grocery shelves. Every class has a minimum of 14 and a maximum of 814 samples. The collection of template images was captured from various vendors or photo galleries, with a wide range of illuminations, sizes, and poses. On the other hand, the shelf images were taken from videos filmed in grocery stores, and showcase and present variations in illumination, scale, reflectance, pose, color, and rotation at the store level. Sample images of templates and shelves are given in Table 2. The ground truth information of shelf images include bounding box annotation with the spatial coordinates of the upper-left (x1, y1) and the lower-right (x2, y2) corners.

Grocery Products (George and Floerkemeier, 2014): Grocery Products[3] is different from the other datasets due to its fine-grained structure. It includes hierarchical categorical information about each

[1] http://yuhang.rsise.anu.edu.au accessed as on 5th August 2023.
[2] http://grozi.calit2.net/grozi.html accessed as on 5th August 2023.
[3] https://drive.google.com/file/d/1vvB1hvKhr4pE8zUpPImlogA6kofkLVN/view accessed as on 5th August 2023.

Table 1
Detailed information of publicly available dataset for grocery product recognition.

Dataset	Product Variety	Train Set	Test Set	Annotation File	Annotated Products
WebMarket (Zhang et al., 2007)	**102**	**300 TI (2 or 3 image per product)**	**3153 SI**	**PC**	**aSP**
Grozi-120 (Merler et al., 2007)	**120**	**676 TI (2 to 14 image per product)**	**11,194 SI**	**BBs + PC**	**aSP**
Grocery Products (George and Floerkemeier, 2014)	**27 upper classes**	**3235 TI (average of 112 PI belongs to each category - 25 to 415)**	**680 SI**	**BBm + PC**	**aSP**
	3235 sub-classes	**3235 TI (1 image per product)**	**680 SI**	**BBm + PC**	**aSP**
Grocery Dataset (Varol and Kuzu, 2015)	**10**	**3701 TI (274 to 598 per product)**	**354 SI**	**BBs + PC**	**aAP**
Freiburg Groceries (Jund et al., 2016)	**25**	**4947 PI**	**74 SI**	**PC**	**aSP**
CAPG-GP (Geng et al., 2018)	**102**	**177 PI (1 to 4 per product)**	**234 SI**	**BBs + PC**	**aSP**
SKU-110K (Goldman et al., 2019)	**1**	**8233 SI for training 588 SI for validation**	**2941 SI**	**BBs**	**aAP**
Locount (Cai et al., 2021)	**140**	**34,022 SI**	**16,372 SI**	**BBm + PC + NoI**	**aAP**
Unitail-Det (Chen et al., 2022)	**1**	**8216 SI for training 588 SI for validation**	**2940 SI for origin-domain 500 SI for cross-domain**	**BBs**	**aAP**

*aAP: annotated for All Products.
*aSP: annotated for Selected Products.
*BBm: single Bounding Box represent Multiple products.
*BBs: single Bounding Box represent Single product.
*NoI: Number of Instances
*PC: Product Categories.
*PI: Product Images.
*SI: Shelf Images.
*TI: Template Images.

product. According to this information, each product has one or more upper category. That is, the Grocery Products dataset has 80 broad product categories with 8350 sub-categories. However, only 27 of those 80 product categories have the ground truth information. Hence, only those 27 categories are taken into account in all studies working with this database. As result, the training set of the Grocery Products collection consists of 27 upper product categories under which 3235 fine-grained product templates were collected from the web. Each product template is represented by a single image taken with white background in studio-like conditions. Each fine-grained category has between 25 and 415 training photos, with an average of 112 different retail products. The test set consists of 680 shelf images taken by a cell phone in real grocery stores with various lighting circumstances, viewing perspectives, and zoom levels. The number of products in each test image varies from 6 to 30, depending on the picture. Table 2 shows examples of training and test images of the Grocery Product collection. Each test image's ground truth information is specified with bounding boxes and category. However, a single bounding box is used to delimit multiple instances of the same product; it does not isolate single items.

Grocery Dataset (Varol and Kuzu, 2015): The Grocery Dataset[4] is a more specific collection restricted to cigarette packages. The dataset consists of product images from 10 different categories and 354 shelf images with annotated ground truths. For each class, the 5 product packages were photographed from various groceries on a white background using 4 different cameras. Totally 3701 product images were captured, with each product presents between 274 and 598 times. Shelf images were captured from 40 different groceries with 4 different cameras to provide diversity in the dataset. Each image has 2 to 7 numbers of shelves and 2 to 137 number of products. The shelves are photographed by moving the frame one shelf down every time. Overall, there are 354 shelf images, and around 13,000 products in total. Typically, there are 200 items for each class on all shelf images. Additionally, 10,000 products were not classified that are categorized negatively. Sample images of products and shelves are given in Table 2. The ground truth information of shelf images given with the dataset, were obtained by drawing bounding boxes around each product package using the tool of Google Image Clipper.[5] Each bounding box have the spatial coordinates of the upper-left (x1, y1) corner and the length of width and height of product.

Freiburg Groceries (Jund et al., 2016): The training images (D_1) of Freiburg Groceries dataset[6] were taken with four distinct cameras from different offices, apartments, and grocery stores. Each of the 25 product categories has instances between 97 and 370 per class; the total number of picture is 4,947, of size 256x256. In addition, the testing set (D2) of Freiburg Groceries was captured using a Kinect v2[7] camera with cluttered scenes, messy backgrounds and different lighting conditions in lab environment. The obtained dataset includes a point cloud of the scene that corresponds to each rack image with a resolution of 1920x1080, an RGB image, and a depth image. Each test images contains various products belonging to multiple classes. Examples of training and test images are given in Table 2. The annotation file of the dataset consists of only labeled product information of each test image, the bounding box information of the products is not available.

CAPG-GP (Geng et al., 2018): CAPG-GP dataset[8] is a fine-grained grocery product dataset. The training set is made of 177 stock-keeping unit (SKU) photos that were gathered from e-commerce sites. Training images consist of three different product types: tube packed items with 18 classes, bag packaged products with 15 classes, and box-like packaged products with 69 classes. Totally, there are 102 grocery products from 5 different brands represented with a minimum of 1 and a maximum of 4 samples. The test set consists of 234 shelf images photographed from 2 stores by mobile phone cameras. Each shelf image

[4] https://github.com/gulvarol/grocerydataset accessed as on 5th August 2023.

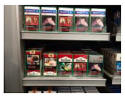
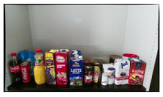
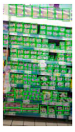
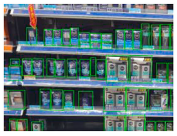
[5] https://code.google.com/archive/p/imageclipper/accessed as on 5th August 2023.

[6] https://github.com/PhilJd/freiburg_groceries_dataset accessed as on 5th August 2023.

[7] https://github.com/code-iai/iai_kinect2 accessed as on 5th August 2023.

[8] http://zju-capg.org/capg-gp.html accessed as on 5th August 2023.

Table 2
The sample images of grocery product recognition datasets.

Dataset	Examples of Training Images		Examples of Test Images	
WebMarket (Zhang et al., 2007)				
Grozi-120 (Merler et al., 2007)				
Grocery Products (George and Floerkemeier, 2014)				
Grocery Dataset (Varol and Kuzu, 2015)				
Freiburg Groceries (Jund et al., 2016)				
CAPG-GP (Geng et al., 2018)				
SKU-110K (Goldman et al., 2019)				
Locount (Cai et al., 2021)				
Unitail-Det (Chen et al., 2022)				

contains 20 object in average. Table 2 shows samples of training and test images. The annotation file of shelf images includes bounding box information of only 177 products belonging to certain categories. Furthermore, the annotation file of the training set gives information of which grocery product stores each SKU images.

SKU-110K (Goldman et al., 2019): SKU-110K dataset[9] is a very large-scale dataset that consists of shelf images both in the training set and the test set. Furthermore, SKU-110K have a variation under favor of densely structure with the number and variety of products on shelves. The average of product area in shelf images is approximately 0.27%. The images were captured by mobile phones from thousands of supermarkets in the United States, Europe and East Asia. Shelf images have different scales, viewing angles, lighting conditions, noise level, and other causes of unpredictability because there was not any regulation about photograph quality and view settings when capturing. The data set consists of a total of 11,762 images, of which 70% (8233 images - 1, 210,431 bounding boxes) is training, 5% (588 images - 90,968 bounding boxes) is validation, and 25% (2941 images - 432,312 bounding boxes) is test images. The split of dataset was realized with random selection in order to prevent the same shelf display from the same store from appearing in more than one of these subsets. Each product on the shelf in the dataset has been labeled as a single class with bounding box but product types were not specified. Sample images of SKU-110K dataset are given in Table 2.

Locount (Cai et al., 2021): It consists of 50,394 shelf images photographed from 28 different stores and apartments with different viewing angles and lighting conditions. The training set has 34,022 photos with 1,437,166 instances, and the testing set contains 16,372 images with 468,151 instances. Sample images of Locount dataset are given in Table 2. Locount dataset has a fine-grained structure with 9 broad different object categories that are Baby Stuffs, Drinks, Foodstuff, Daily Chemicals, Clothing, Electrical Appliances, Storage Appliances, Kitchen Utensils, and Stationery and Sporting Goods. Also, these 9 broad categories have an overall of 140 sub-categories containing products in stores. Locount collection differentiate from the other datasets due to the presence of an annotation file, which has been created using the

[9] https://github.com/eg4000/SKU110K_CVPR19 accessed as on 5th August 2023.

Colabeler tool[10]. The annotation file of the training and test set consist of bounding boxes of each object categories on shelf images and the information of the number of instances enclosed in the bounding box. The dataset contains a variety of object at different scales, and the distribution of object scales is separated into three subsets, small scale, medium scale, and large scale subset, based on the number of pixels in the bounding boxes. While 52.5% of the dataset consists of medium-sized products, large-scale products are present by 29.5% and small-scale products 18%. Furthermore, according to the statistics of dataset, the instance numbers of the annotated bounding boxes consist mostly in one-instance, some of them contain between 2 and 10 instances and few variants have more than 10 instances.

Unitail-Det (Chen et al., 2022): It is one of the United Retail (Unitail) Datasets[11] specialized for product detection; it has been created from two sources: origin domain and cross-domain. In the origin domain, training and testing photos are meant to be obtained from comparable angles in the same stores by the same sensors. 11,744 photos from SKU-110k were selected in order to create the origin domain. In the cross domain, 500 photos, with size 3024x4032, were taken from unused categories in various stores and using a variety of sensors and camera perspectives. Pictures depict a wide range of product categories, from delicatessen items to household goods, for a total of 1454 product categories. This dataset stands out with its quadrilateral annotation, which represents the frontal face of the bounding boxes for products. The bounding box consists of 4 coordinates ptl, ptr, pbr, pbl with 8 degrees of freedom (xtl, ytl, xtr, ytr, xbr, ybr, xbl, ybl). Furthermore, cuboid and cylindrical products are represented with the top-left corner (xtl,ytl) of the frontal faces, and the remaining points stand in for the remaining three corners, in clockwise order. The training set for the origin-domain has 8216 images annotated with 1,215,013 instances, whereas the validation set contains 588 images with 92,128 instances. The test set for the origin-domain has 2940 images annotated with 432,896. On the other hand, the test set has 500 images with 37,071 instances in cross-domain. Table 2 shows samples of training and test images in both the domains of Unitail-Det.

Additionally, Table 3 associates every database with the list of papers using it. Accordingly, the most utilized datasets are Grocery Products and Grozi-120, and the least used collections are Locount and Unitail-Det, which was expected since they are more recently published ones. These collections have many common and different features, as detailed in the previous section. Different data sets can be used to find solutions to the problems related to the product recognition task using alternative methods, depending to their features.

3. Grocery product recognition studies

With the need for new approaches and the increase in the ease of data collection, along with decreasing costs of data storage and processing, the utilization of computer vision and machine learning methods for solving product recognition issues on market shelves has increased. Machine learning algorithms offer more flexible solutions by enabling training to recognize new items and adapt to changing environments. Additionally, computer vision algorithms can analyze various features of images such as color, structure, morphology, and frequency to identify products with high accuracy. Consequently, the number of studies utilizing computer vision and machine learning methods for product recognition in market shelf images is on the rise.

In a comprehensive product recognition problem on market shelves, input is taken from a shelf image, and localization and classification results of all pre-defined products are obtained. Upon examining product recognition studies on market shelves, there are studies that only identify products locations, only classify products, or addressing both aspects together. Fig. 1 shows the differentiation between product classification, localization and detection on related datasets. As listed in Table 4, there are studies that address problems of classification and localization separately, as well as studies that address them together with detection.

Table 3
Grocery product recognition datasets and studies.

Datasets	Studies
WebMarket	(Zhang et al., 2007) (Zhang et al., 2009) (Ray et al., 2018) (Varadarajan et al., 2020) (Santra et al., 2020a) (Santra et al., 2021) (Kant, 2020) (Yilmazer and Birant, 2021) (Santra et al., 2022) (Cho et al., 2022) (Selvam and Koilraj, 2022)
Grozi-120	(Merler et al., 2007) (George and Floerkemeier, 2014) (Winlock et al., 2010) (Franco et al., 2017) (Karlinsky et al., 2017) (Geng et al., 2018) (Varadarajan and Srivastava, 2018) (Santra et al., 2020b) (Srivastava, 2020) (Santra et al., 2020a) (Santra et al., 2021) (Santra et al., 2022) (Selvam and Koilraj, 2022) (Jonathan and Kusuma, 2022) (Srivastava, 2022) (Wang et al., 2022) (Ghosh, 2023)
Grocery Products	(George and Floerkemeier, 2014) (Ray et al., 2018) (George et al., 2015) (Yörük et al., 2016) (Tonioni and Di Stefano, 2017) (Karlinsky et al., 2017) (Tonioni et al., 2018) (Geng et al., 2018) (Varadarajan and Srivastava, 2018) (Santra et al., 2020b) (Varadarajan et al., 2020) (Osokin et al., 2020) (Santra et al., 2020a) (Santra et al., 2021) (Santra et al., 2022) (Selvam and Koilraj, 2022) (Ghosh, 2023) (Yücel and Ünsalan, 2024) (Sinha et al., 2021) (Melek et al., 2023) (De Feyter, 2023)
Grocery Dataset	(Varol and Kuzu, 2015) (Varol, 2014) (Gokdag, 2019) (Gökdag, 2016) (Varadarajan and Srivastava, 2018) (Varadarajan et al., 2020) (Kant, 2020) (Ghosh, 2023) (Hurtik and Vlasanek, 2020)
Freiburg Groceries	(Jund et al., 2016) (Farren, 2017) (Ciocca et al., 2016) (Selvam and Koilraj, 2022)
CAPG-GP	(Geng et al., 2018) (Varadarajan et al., 2020) (Kant, 2020) (Srivastava, 2020) (Cho et al., 2022) (Srivastava, 2022)
SKU-110K	(Goldman et al., 2019) (Ye et al., 2016) (Pan et al., 2020) (Varadarajan et al., 2020) (Kant, 2020) (Rong et al., 2021) (Deng and Yang, 2021) (Cho et al., 2022) (Wang et al., 2020) (Moon et al., 2022) (Ravi et al., 2022) (Xu et al., 2022) (Strohmayer and Kampel, 2023) (Vasanthi and Mohan, 2024)
Locount	Cai et al. (2021)
Unitail-Det	Chen et al. (2022)

Product classification is the process of identifying specific products within an image or video frame. This involves categorizing the product into predefined classes, which could be brands, type of products or different categories of products. For example, “L’oreal” and “Head&Shoulders” are the brand classes, “shampoo” is the class of product type and “personal care” is the product category. Product classification is essential in cases where it is only important to know whether the product is on the shelf or not, but its location is not important. The datasets of WebMarket (Zhang et al., 2007) and Freiburg Groceries (Jund et al., 2016) are specially created for product classification.

The study (Zhang et al., 2007) that the WebMarket dataset was created, aims to find shelf images containing a given product template. The Harris-Affine interest region detector (Mikolajczyk and Schmid, 2004) was employed to identify regions of interest in shelf images. Subsequently, a vocabulary of visual words was created using hierarchical k-means clustering from Scale-Invariant Feature Transform (SIFT) (Lowe, 2004) features, and representations of product templates were formed using histograms of visual word occurrences. The results were compared using four different similarity measures to match shelf images with target product images. However, none of the obtained performances are sufficient for a real-world application. The another study (Zhang et al., 2009) using the WebMarket dataset (Zhang et al., 2007), propose a method that use weighting visual words obtained from Hessian-Affine features to deal with the difficulty in recognizing products caused by scale differences in images. Additionally, a combination of matching scores was used to match shelf and product images. The achieved performance of (Zhang et al., 2009) is better than (Zhang et al.,

[10] http://www.colabeler.com/accessed as on 5th August 2023.
[11] https://unitedretail.github.io/accessed as on 5th August 2023.

Fig. 1. Categorization of the problem.

Table 4
The studies handled product classification, product localization and product detection.

Categorization of the Problem	Studies
Product Classification	(Zhang et al., 2007) (Jund et al., 2016) (Zhang et al., 2009) (Ray et al., 2018) (Santra et al., 2020b) (Srivastava, 2020) (Jonathan and Kusuma, 2022) (Srivastava, 2022) (Wang et al., 2022) (Ghosh, 2023) (George et al., 2015) (Farren, 2017) (Ciocca et al., 2016)
Product Localization	(Goldman et al., 2019) (Varadarajan et al., 2020) (Kant, 2020) (Cho et al., 2022) (Varadarajan and Srivastava, 2018) (Yücel and Ünsalan, 2024) (Ye et al., 2016) (Pan et al., 2020) (Rong et al., 2021) (Deng and Yang, 2021) (Wang et al., 2020) (Moon et al., 2022) (Ravi et al., 2022) (Xu et al., 2022) (Vasanthi and Mohan, 2024)
Product Detection	(Merler et al., 2007) (George and Floerkemeier, 2014) (Varol and Kuzu, 2015) (Geng et al., 2018) (Cai et al., 2021) (Chen et al., 2022) (Santra et al., 2020a) (Santra et al., 2021) (Yilmazer and Birant, 2021) (Santra et al., 2022) (Selvam and Koilraj, 2022) (Winlock et al., 2010) (Franco et al., 2017) (Karlinsky et al., 2017) (Yörük et al., 2016) (Tonioni and Di Stefano, 2017) (Tonioni et al., 2018) (Osokin et al., 2020) (Sinha et al., 2021) (Melek et al., 2023) (De Feyter, 2023) (Varol, 2014) (Gokdag, 2019) (Gökdag, 2016) (Hurtik and Vlasanek, 2020) (Strohmayer and Kampel, 2023)

2007).

The study (Jund et al., 2016) that introduced the product recognition dataset Freiburg Groceries, product images obtained by manually cropping shelf images were classified using a convolutional neural network (CNN) model. The CaffeNet architecture (Jia et al., 2014), an adaptation of AlexNet (Krizhevsky and Hinton), was retrained as the CNN model. The classification performance obtained by averaging five different training-test partitions is 78.9%. The study serves as a foundation for both the provided dataset and product classification using deep networks. In the study (Ciocca et al., 2016), features obtained from the last average pooling layer before the network division in the DenseNet-169 architecture were classified using cubic Support Vector Machine (SVM). Although a higher accuracy was achieved, it is not comparable with (Jund et al., 2016) due to using the different classification strategy.

Pre-trained Alexnet was used for product classification in the study (Santra et al., 2020b). Random forests (Breiman, 2001) have been used to deterministically identify and terminate unimportant connections in the network. According to the results obtained on the Grozi-120 (Merler et al., 2007) and Grocery Products (George and Floerkemeier, 2014) datasets, an improvement in classification accuracy was achieved.

In the study (Srivastava, 2020), a pre-trained convolutional network, ResNext-INet, on Instagram, was fine-tuned using a novel neural network layer called Local Concepts-Accumulation (LCA) and maximum entropy loss to classify grocery products. According to the obtained results on Grozi-120 (Merler et al., 2007) and CAPG-GP (Geng et al., 2018), the proposed method achieved higher accuracy compared to image matching based on keypoint detection and ResNext-INet without fine-tuning. In this case, having multiple training images for each class is crucial for the training of the network. Additionally, the classification accuracy of (Srivastava, 2020) on Grozi-120 (Merler et al., 2007) dataset is higher than (Jonathan and Kusuma, 2022) which is used a modified Visual Geometry Group (VGG) −16 (Simonyan and Zisserman, 2015). In another product classification study (Srivastava, 2022), representations of images taken from a pre-trained model are used instead of fine-tuning throughout the entire backbone in (Srivastava, 2020). This approach leads to achieving accuracy similar to or better than the (Srivastava, 2020). The study (Wang et al., 2022), a developed Siamese Neural Network (SNN) (Koch, 2011) was utilized for product classification. The proposed algorithm surpassing traditional techniques and has overcome the significant issue encountered in the training phase, which is the insufficient data problem in the retail product recognition domain as seen in (Srivastava, 2020).

In the study (George et al., 2015), both textual and visual features were utilized to recognize products. For textual features, words on product packages were detected using Optical Character Recognition (OCR) (Smith, 2005), while distinctive patches were employed for visual features. The obtained features were classified using SVM. However, the proposed method can only predict product categories and cannot predict subcategories.

In study (Ghosh, 2023), product recognition was conducted through text recognition. Faster R–CNN (Ren et al., 2017) was utilized for detecting the areas containing text, while Long-Short Term Memory (LSTM) (Ghosh et al., 2019) and Bidirectional Long-Short Term Memory (BLSTM) (Ghosh et al., 2019) models of Recurrent Neural Network (RNN) classifier were used for recognizing the letters within text blocks. The usage of Faster R–CNN (Ren et al., 2017) yielded faster and more accurate results compared to other methods. In the other study (Farren, 2017), guided prunning was performed on the CaffeNet (Jia et al., 2014) model used for product recognition, resulting in a simpler and more accurate algorithm.

In study (Ray et al., 2018), a two-layer model was presented, comprising hypothesis generation and verification. The initial layer predicts potential merchandise at designated shelf locations through correlation-based and NeoSURF-based image matching between shelf and product images. In the subsequent layer, a novel graph-theoretic method verifies these hypotheses. The proposed approach outperforms competing methods in (Zhang et al., 2007) and (George and Floerkemeier, 2014) on the WebMarket (Zhang et al., 2007) and Grocery Products (George and Floerkemeier, 2014) datasets in terms of performance.

Product localization involves to determine the precise items' location within the image. This typically means providing the coordinates or bounding boxes around each recognized product in the image. Product localization is an important task in scenarios where it's necessary to know not just which products are present but also where

they are situated in the picture. The SKU-110K (Goldman et al., 2019) and Unitail-Det (Chen et al., 2022) datasets are specially created for product localization.

The studies (Goldman et al., 2019), (Varadarajan et al., 2020), (Kant, 2020), (Pan et al., 2020), (Xu et al., 2022) and (Vasanthi and Mohan, 2024) used SKU-110K (Goldman et al., 2019) dataset to localize all products on shelf images. In the study (Goldman et al., 2019), the authors propose a framework combining RetinaNet (Lin et al., 2020), developed for dense object detection where many products are present, with EM-Merger. Improved performance is achieved compared to existing algorithms presented in (Ren et al., 2017), (Lin et al., 2020) and (Redmon and Farhadi, 2017). Furthermore, study (Kant, 2020) proposes two different models, Gaussian Decoder Network (GDN), an extended version of RetinaNet (Lin et al., 2020), and Gaussian Layer Network (GLN), which has fewer parameters and higher accuracy compared to GDN. Both GDN and GLN outperform the method proposed in study (Goldman et al., 2019). Additionally, the highest performance is attained with the Dynamic Refinement Network (DRN) consisting of two modules (feature selection and dynamic refinement) presented in study (Pan et al., 2020). Another approach proposed in study (Xu et al., 2022), comprises Cascade R–CNN (Cai and Vasconcelos, 2018), Residual Network (ResNet) (He et al., 2016), Feature Pyramid Network (FPN), and balanced L1 loss. Low-quality region proposals resulting from dense scenes are improved with Cascade R–CNN (Cai and Vasconcelos, 2018), and the position loss from the loss function is balanced and constrained using balanced L1 loss. Moreover, FPN leads to an increase in performance for detecting small-scale objects. The study concludes that the proposed method achieves higher detection accuracy compared to well-known object detection methods like Faster R–CNN (Ren et al., 2017), RetinaNet (Lin et al., 2020) and You Only Look One Version 3 (YOLOv3) (Redmon and Farhadi, 2017). However, due to the cascaded structure used, the proposed method's speed is slower, which does not meet the real-time requirements for detecting products on market shelves. The paper (Vasanthi and Mohan, 2024) proposes the utilization of a Ghost Convolution Block (GCB) within a nested-transformer encoder block in the feature refinement network to address these complexities. The proposed method has demonstrated higher performance compared to state-of-the-art object detection methods and (Goldman et al., 2019).

In addition to object localization, studies (Ye et al., 2016), (Deng and Yang, 2021) and (Moon et al., 2022) focused on different problems such as counting objects and training on unlabeled data. ResNet (He et al., 2016) integrated with FPN used as feature extractor and Faster R–CNN (Ren et al., 2017) used as the detector head in (Ye et al., 2016). The difference from the other studies is realized the training process of network with unlabeled data. In the study (Deng and Yang, 2021), Faster R–CNN (Ren et al., 2017) and multiple-step sampling were used for localize and count the products on SKU-110K dataset (Goldman et al., 2019). In the other object localization and counting study (Moon et al., 2022), the proposed method Distance Between Circles - Non-Maximum Suppression (DBC-NMS) achieved better performance from (Goldman et al., 2019).

The other study (Wang et al., 2020) focused on improving speed performance of Non-maximum suppression (NMS) (Neubeck, 2006). For this purpose, hashing-based non-maximum suppression (HNMS) was proposed and applied with Faster R–CNN (Ren et al., 2017) and RetinaNet (Lin et al., 2020). Results on different datasets show the significant speed improvement with comparable accuracy. In the study (Ravi et al., 2022), a new loss function balanced-Intersection over Union (BIoU) was proposed and the use of this loss function in Mask R–CNN (He et al., 2020) and YOLOv5 (Jocher et al., 2020) models increased the performance.

In the study (Varadarajan et al., 2020), retrained Faster R–CNN (Ren et al., 2017) with NMS (Neubeck, 2006) was used to locate the positions of products. According to results, while significant performance were not obtained on the SKU-110K (Goldman et al., 2019) dataset, results close to previous studies were achieved in the Grocery Products (George and Floerkemeier, 2014) and CAPG-GP (Geng et al., 2018) datasets. In the other study (Cho et al., 2022), a ResNet (He et al., 2016) based FPN was employed to locate the positions of products. Additionally, advanced weighted Hausdorff distance (AWHD) and hard negative-aware anchor (HNAA) were utilized in the process. The obtained results on the datasets of WebMarket (Zhang et al., 2007), CAPG-GP (Geng et al., 2018) and SKU-110K (Goldman et al., 2019) outperformed according to state-of-the-art methods. In (Varadarajan and Srivastava, 2018), fully convolutional network with VGG framework (Simonyan and Zisserman, 2015). Convolutional Encoder-Decoder (ConvAE) and Refine-Net module were used. Although the proposed method demonstrates good performance on the applied datasets, there is a need to enhance the method's performance at higher thresholds.

The study (Yücel and Ünsalan, 2024) used different local feature extraction methods: SIFT (Lowe, 2004), Speeded Up Robust Features (SURF) (Bay et al., 2006), Oriented Features from Accelerated Segment Test (FAST), Rotated Binary Robust Independent Elementary Features (BRIEF) (ORB) (Rublee et al., 2011), Accelerated-KAZE (AKAZE) (Ghosh et al., 2019), and Binary Robust Invariant Scalable Keypoints (BRISK) (Leutenegger et al., 2011). Brute force search and implicit shape model used for detecting object center points in a given shelf image. The modified Needleman-Wunsch algorithm used for planogram compliance control. Then, iterative search step used for improving the performance of the algorithm. When looking at the results obtained with the proposed method, a lower number of false negatives is obtained, while a higher number of false positives is obtained.

In addition, there are also problems that address classification and localization together, this task is called product detection. It is used in problems such as planogram matching and stock tracking, where product class and product location information is required simultaneously. The datasets of Grozi-120 (Merler et al., 2007), Grocery Products (George and Floerkemeier, 2014), Grocery Dataset (Varol and Kuzu, 2015), CAPG-GP (Geng et al., 2018) and Locount (Cai et al., 2021) have the class and location information of product. The datasets of Grozi-120 (Merler et al., 2007), SKU-110K (Goldman et al., 2019) and Locount (Cai et al., 2021) have annotation with spatial coordinates of the upper-left and the lower-right corners of the product. Grocery Products (George and Floerkemeier, 2014) has also the spatial coordinates but in different order. The Grocery Dataset (Varol and Kuzu, 2015) and CAPG-GP (Geng et al., 2018) have coordinates of the upper-left corner; width and height of the product. The bounding box of Unitail-Det (Chen et al., 2022) consists of 4 coordinates ptl (point of top left), ptr (point of top right), pbr (point of bottom right), pbl (point of bottom left) with 8 degrees of freedom. Fig. 2 shows examples of different annotation format on shelf images in distinct datasets.

In the studies (Varol and Kuzu, 2015), (Varol, 2014)- (Gökdag, 2016), methods consisting of three different stages, namely shelf detection, product localization, and product classification, were tested on the Grocery Dataset (Varol and Kuzu, 2015). Detecting shelf lines in shelf images has a performance-enhancing effect in terms of searching for products in a more limited area and eliminating products detected in wrong areas. In the study (Varol, 2014), Hough transform (HT) (Duda and Hart, 1972) was used to detect shelf lines, achieving an accuracy of 83.4% on 229 shelf images. Since HT (Duda and Hart, 1972) was found not to be resilient to changes in shelf design and computationally more complex, in the study (Varol and Kuzu, 2015), shelf lines were detected by calculating the histogram of the projection of products on the y-axis. In the study (Gökdag, 2016), a Gaussian mixture model was used, achieving an accuracy of 99.03% on 350 shelf images. The authors of studies (Varol and Kuzu, 2015), (Varol, 2014)- (Gökdag, 2016) addressed the problem of product detection using different methods and tested their performance on various combinations of the Grocery Dataset (Varol and Kuzu, 2015). In these studies, Histogram of Oriented Gradients (HoG) (Dalal and Triggs) features for product detection were classified using a cascaded object detection algorithm (Viola and Jones,

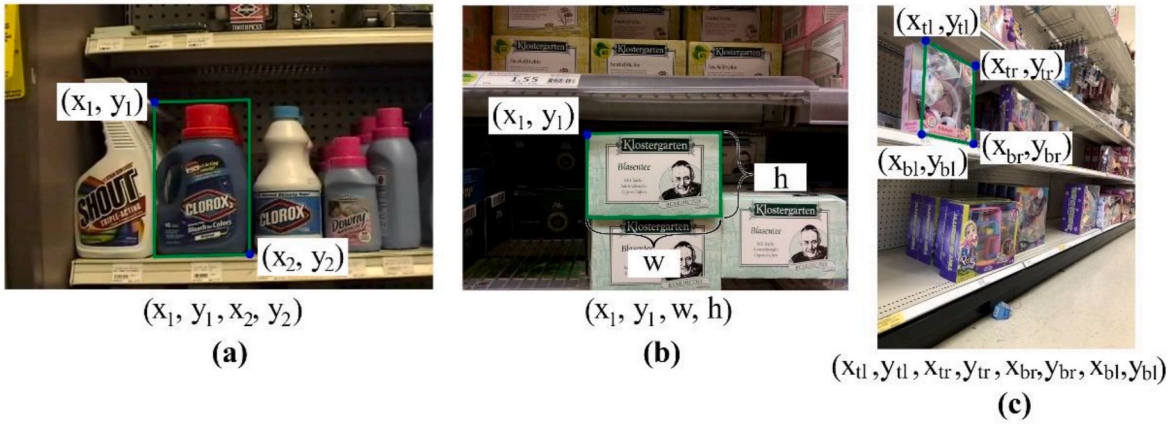


Fig. 2. Distinct types of bounding box information on different datasets: (a) Bounding box example from Grozi-120 (Merler et al., 2007), (b) Bounding box example from Grocery Products (George and Floerkemeier, 2014), (c) Bounding box example from Unitail-Det (Chen et al., 2022).

2001). Additionally, in the study (Gokdag, 2019) and (Gökdag, 2016), the obtained incorrect results were rectified using a Gaussian mixture model. Furthermore, average and median filters were used to create boxes for undetected products. Various solutions were proposed for the product classification stage in studies (Varol and Kuzu, 2015), (Varol, 2014) and (Gökdag, 2016). For recognizing brands in the areas obtained from product detection, SIFT (Lowe, 2004) with HSV color features were utilized in study (Varol and Kuzu, 2015), while HoG (Dalal and Triggs) and HSV color features with SVM were employed in study (Varol, 2014). In contrast, in study (Gökdag, 2016), the dense SIFT (Lowe, 2004) feature vectors were classified using Extreme Learning Machines (ELM) (Ã et al., 2006), resulting in an accuracy of 99.21%.

In the study (Merler et al., 2007), comparative results of end-to-end product recognition processes using three different methods with five different scales on window patches extracted from shelf images are provided. According to the precision and recall values obtained, the best performance was achieved with SIFT (Lowe, 2004), followed by color histogram matching (CHM) and finally by features similar to Boosted Haar-like (BHaar) (Viola and Jones, 2001). Although this study does not yield successful results, it lays the groundwork for solving the product recognition problem.

In the study (George and Floerkemeier, 2014), the proposed method consists of three main steps. In the first two steps, where product detection and recognition stages are addressed together, a test image is filtered through two consecutive ranking procedures to identify the possible categories to which it belongs, and then matched with all images in the filtered categories. For this purpose, a region-constrained linear coding model using dense SIFT features (Lowe, 2004) was trained with a random forest algorithm (Breiman, 2001). Subsequently, fast dense pixel matching was performed through deformable spatial pyramid matching. In the final step, an energy function was globally minimized to obtain the ultimate list of recognized products along with their inferred locations. In this study, a large-scale dataset was created for fine-grained product recognition, and the proposed method laid the foundation for it.

In the study (Winlock et al., 2010), the authors propose an application aimed at helping visually impaired users locate items listed in shopping lists from video footage of retail shelves. SURF descriptors (Bay et al., 2006) and a multi-class Naive Bayes classifier (Barrington et al., 2008) are used to recognize products from regions obtained through an optical flow-based method (Lucas, 1981). The system's performance is measured by detecting 10 products from 25 different shopping lists in 10 different video frames. According to the research results, as the threshold for obtaining the correct product increases, the number of missed products also increases. Additionally, as the threshold value decreases, the number of incorrectly collected products also increases. Therefore, the study does not recommend a suitable threshold value for users to complete their shopping lists.

In another study (Karlinsky et al., 2017), products are reclassified using a pre-trained CNN model (VGG-f network (Chatfield et al., 2014)) after predicting the shortlist of possible categories based on SIFT features (Lowe, 2004) using a probabilistic inference model. The proposed method achieves a higher performance compared to previous studies (George and Floerkemeier, 2014) (Melek and Sonmez, 2017) and (Ren et al., 2017) on different datasets Grozi-120 (Merler et al., 2007) and Grocery Products (George and Floerkemeier, 2014). The proposed model is useful in cases where there is insufficient data collection for training deep detection models. However, if more training data is available, deep models perform better ("Detecting retail products in situ).

In the study (Geng et al., 2018), SIFT (Lowe, 2004), SURF (Bay et al., 2006) and BRISK (Leutenegger et al., 2011) features were used to create attention map. VGG-16 (Simonyan and Zisserman, 2015) was used to classify features. Proposed approach offers scalability with the identification of product instances does not necessitate training and provides a rough indication of the class label with high recall. However, the localization of product instances relies on recurring patterns, which are susceptible to variations such as occlusions, highlights, blurriness is the limitation of the study (Geng et al., 2018).

In another study (De Feyter, 2023), different procedures were proposed to demonstrate the performance of task-specific training of a shared product detection model. Firstly, this shared model was trained on a labeled dataset containing all products. Then, the training process was modified to complete training on task-specific datasets. Through the use of various procedures in the study, it was shown that separating the training into a detection and classification stage did not lead to a decrease in performance compared to training the model as a conventional multi-class detector.

In study (Yörük et al., 2016), a Hough voting scheme is employed to predict the poses of products in the shelf image based on the matching SURF key points (Bay et al., 2006) of the shelf and product images in the Hough space. Subsequently, fine-grained classification is performed using a pose-class histogram in the Hough space. While the proposed method achieves high success in categorization prediction, it does not demonstrate the same success in predicting sub-product categories.

Study (Tonioni and Di Stefano, 2017) consists of three stages: unconstrained product recognition, graph-based consistency checking, and product verification. In the unconstrained product recognition stage, feature matching and generalized (HT) (Duda and Hart, 1972) are employed. In the graph-based consistency checking part, the alignment between planned product positions on the shelves (planogram) and the positions obtained in the previous stage is verified. Finally, various

methods are used for product verification, among which the most successful method is BRISK (Leutenegger et al., 2011). The utilization of planogram information in addition to training and test datasets has resulted in better performance in terms of determining product locations and capturing missed product.

In another product detection study (Franco et al., 2017), the proposed method consists of pre-selection, fine selection, and post-processing steps. In the fine selection stage, using SURF features (Bay et al., 2006) from a simpler dataset yields more accurate results, while deep neural network features perform better on a more complex dataset. The proposed method is quite complex due to the requirement of generating numerous separate candidate windows for each product in the first stage and the increased processing time required for feature extraction in the second stage.

In another study (Tonioni et al., 2018), a customized YOLOv2 (Redmon and Farhadi, 2017) for the product detection step, descriptors learned with VGG-16 (Simonyan and Zisserman, 2015) for recognition, and finally, some optimization strategies were employed. According to the results obtained in study (Tonioni et al., 2018), higher recognition performance was achieved compared to (Tonioni and Di Stefano, 2017) and (Ren et al., 2017). However, training YOLOv2 (Redmon and Farhadi, 2017) requires an additional dataset. Additionally, market product recognition datasets have fewer images belonging to a larger number of classes compared to commonly used object recognition datasets.

Other end-to-end studies (Selvam and Koilraj, 2022) and (Chen et al., 2022) have attempted to solve the problem of product recognition by recognizing text on products. The aim of these studies is to improve the shopping experience, especially for visually impaired individuals. The stages of study (Selvam and Koilraj, 2022) are preprocessing, product detection, and product recognition (including text detection and text recognition) addressed sequentially. For detecting products, YOLOv5 (Jocher et al., 2020) is used; for text detection, a ResNet50 (He et al., 2016) based FPN with a new post-processing technique is employed, and finally, a Selective Context Attentional Text Recognizer (SCATTER) (Litman et al., 2020) is utilized to recognize the text information of the products. The proposed method enhances the effectiveness of existing techniques.

In another study (Chen et al., 2022), consisting of a deep learning-based product detection stage and a visual feature matching-based product classification stage, the FPN with Dense-Box style architecture (Huang et al., 2015) is adopted for product detection. In the product classification stage, feature vectors obtained through spatial encoding based on text location are matched with the Hungarian algorithm (Robinson and Assignment, 1950). The dataset created in this study differs from other datasets as it contains coordinates of the four corners of the product, overlapping more with the product itself. Thus, it serves as a reference for text-based product recognition studies with a well-aligned, textually enhanced large-scale dataset.

The study (Cai et al., 2021) proposed Cascaded Localization and Counting Network (CLCNet) with ResNet-50 (He et al., 2016) based the feature pyramid architecture. The result on Locount dataset that was introduced in (Cai et al., 2021), show the effectiveness of the proposed method.

The study (Santra et al., 2020a) started with the process of using Regions with CNN Features (R–CNN) for generating region proposals around products on a rack, followed by classification using a convolutional neural network. Greedy-NMS is then applied to resolve overlapping proposals. Extensive experiments show that this approach outperforms existing methods by up to 7% across WebMarket (Zhang et al., 2007), Grozi-120 (Merler et al., 2007) and Grocery Products (George and Floerkemeier, 2014). Another study (Santra et al., 2021) using the datasets WebMarket (Zhang et al., 2007), Grozi-120 (Merler et al., 2007) and Grocery Products (George and Floerkemeier, 2014), consist of three stage: exemplar-driven region proposal, classification followed by non-maximal suppression of the region proposals. In exemplar-driven region proposal stage, BRISK descriptors were clustering with DBSCAN clustering. Classification stage was realized by ResNet-101 based CNN model. Finally, greedy-NMS was used for refinement of results. The suggested method surpasses alternative approaches, leading to an enhancement of detection accuracy by approximately 4%.

Study (Yilmazer and Birant, 2021) proposed a semi-supervised learning on on-shelf availability with the different backbone RetinaNet (Lin et al., 2020), YOLOv3 (Redmon and Farhadi, 2017) and YOLOv4 (Wang and Liao). Using dataset WebMarket (Zhang et al., 2007) was labeled again for this study based on five classes. According to the obtained results, using the combination of three different CNN improves accuracy.

The authors of (Santra et al., 2022) proposed a Reconstruction-Classification Network (RC-Net) for object-level classification. Additionally BRISK was used for part-level features. Discriminative features were identified through unsupervised search and encoded using convolutional Long Short-Term Memory (LSTM). The classification process was performed by R–CNN based on these features. In benchmark datasets, (Santra et al., 2022) achieved close or superior values when compared to existing methods.

The main difference of study (Osokin et al., 2020) from other studies is that the classes of objects used for training and testing do not overlap. For object detection, ResNet (He et al., 2016), correlation matching, and TransformNet were sequentially used. This method proves useful in situations where obtaining a sufficient amount of labeled data is challenging.

Study (Sinha et al., 2021) was proposed Faster R–CNN based object localizer and ResNet-18 based image encoder to detect products on shelf images. According to obtained results on Grocery Products (George and Floerkemeier, 2014) achieved better performance from (George and Floerkemeier, 2014), (Yörük et al., 2016) and (Tonioni et al., 2018). The other study (Melek et al., 2023) was proposed a DNN based product localization and a hybrid usage of different feature extraction methods (SURF (Bay et al., 2006), BRISK (Leutenegger et al., 2011) and ORB (Rublee et al., 2011)) for product classification, than refine the results with neighborhood clustering algorithm. Detecting product locations independent of product types provides flexibility, while the hybrid use of features has also improved performance.

Study (Hurtik and Vlasanek, 2020) proposes You Only Look Once And See Contours (YOLO-ASC), which performs quadrangular detection instead of rectangular-based product localization, ignoring areas outside the products and improving product recognition result. Another study (Strohmayer and Kampel, 2023) focused on real-time product recognition using YOLOv7tiny (Wang et al., 2023), MobileNetV3s (Howard et al., 2019) as embedder, Principal Component Analysis (PCA) to reduce the embedding dimensionality and K-Nearest Neighbors (KNN) to classify products.

4. Classification of grocery product recognition studies

This section provides a systematic review of computer vision and machine learning methods used in the studies on the publicly available data sets, which are described in the previous paragraphs. More in details, it introduces the studies on that data sets under two different subtitles: firstly, it examines how these works addressed the product recognition issue; secondly, it details which methods were used in those papers.

4.1. Classification of the studies according to the handled problems

In addition to studies that find solutions to product classification, localization and both, there are also studies that address more specific problems. That is, object retrieval, shopping assistive systems, fine-grained classification, planogram compliance, real-time product recognition, one-shot object detection, stock-tracking and dense product detection are sub-categories of the product recognition task. Table 5

Table 5
Classification of the studies according to the handled problems.

Sub-category of the Problem	Studies
Object Retrieval	(Zhang et al., 2007) (Zhang et al., 2009) (Ray et al., 2018) (Santra et al., 2021) (Franco et al., 2017) (Sinha et al., 2021) (Melek et al., 2023)
Shopping Assistive Systems	(Merler et al., 2007) (George and Floerkemeier, 2014) (Jund et al., 2016) (Winlock et al., 2010) (Jonathan and Kusuma, 2022) (George et al., 2015) (Yörük et al., 2016) (Tonioni et al., 2018) (Farren, 2017)
Fine-grained Classification	(George and Floerkemeier, 2014) (Geng et al., 2018) (Santra et al., 2022) (Cho et al., 2022) (Karlinsky et al., 2017) (George et al., 2015) (Tonioni et al., 2018) (Ciocca et al., 2016)
Planogram Compliance	(Varol and Kuzu, 2015) (Ray et al., 2018) (Tonioni and Di Stefano, 2017) (Yücel and Ünsalan, 2024) (Varol, 2014) (Gokdag, 2019) (Gökdag, 2016)
Real-time Product Recognition	(Moon et al., 2022) (Strohmayer and Kampel, 2023)
One-shot Object Detection	(Wang et al., 2022) (Osokin et al., 2020) (De Feyter, 2023)
Stock-tracking	(Cai et al., 2021) (Yilmazer and Birant, 2021) (Deng and Yang, 2021) (Moon et al., 2022)
Dense Product Detection	(Goldman et al., 2019) (Varadarajan et al., 2020) (Kant, 2020) (Pan et al., 2020) (Rong et al., 2021) (Deng and Yang, 2021) (Wang et al., 2020) (Ravi et al., 2022) (Xu et al., 2022)

associates every sub-category with the list of papers that address it. As can be seen, a study can address more than one problem at the same time.

In the **object retrieval** issue, the query image of a product is given and the challenge is to find the same products from a collection of images (Zhang et al., 2007). **Shopping assistive systems** aim to improve the customer's shopping experience. Customers can find items in the shopping list quickly, compare prices of products, obtain information whether the product he/she is looking for is on the shelf or not, where it is located and if it does not exist, he/she receives recommendation of existing alternative items with their location info (Yörük et al., 2016). In addition, shopping assistive systems offer the convenience of shopping on your own (Winlock et al., 2010). **Fine-grained classification** refers to a level of classification that goes beyond broad product categories and focuses on distinguishing between subtle differences within a particular category (George and Floerkemeier, 2014). The well-known fine grained dataset is the Grocery Products (George and Floerkemeier, 2014), which contains 3235 sub-classes connected to 27 upper classes. For example, the class of "cheese" is connected to "dairy" and class of dairy is connected to the "food" category. Fine-grained classification is effective in accurately distinguishing between products within similar categories, especially when subtle differences are key to identification. This process typically involves moving from broader product categories to more specific classes, which can lead to faster and more reliable results as the recognition process refines the classification step by step. **Planogram compliance** is a term commonly used in the retail industry to describe the extent to which a retail reflects the planograms set by the retailer or brand. A planogram is a visual representation or diagram that specifies how products should be arranged and displayed on store shelves (Tonioni and Di Stefano, 2017). **Real-time product recognition** is necessary for situations where recognizing the product quickly is as important as recognizing it correctly. Since product recognition requires high computational cost processing in multi-class data sets, it is challenging to achieve this in real time. **One-shot object detection** refers to models trained to recognize a new object or class from just one or a few examples without requiring much training data (Osokin et al., 2020). **Stock-tracking** contains the applications of counting and on-shelf availability. The aim of product counting is to estimate how many items in each category will be present in a given shelf image (Cai et al., 2021). On-shelf availability measures the extent to which products are in stock and available for purchase on store shelves at the time when customers wish to buy them (Yilmazer and Birant, 2021). **Dense product detection** is used when shelf images have a large number of products closely spaced and/or frequently similar or identical-looking objects (Goldman et al., 2019).

4.2. Classification of the studies according to using methods

In this section, previous papers are examined according to the methods used for extraction and classification of features, detection of products and refinement of localization and classification results.

4.2.1. Feature extraction methods

A basic product recognition algorithm finds potential item locations in the shelf image, and determines the class of the product in this location by matching this potential area with the product template images. These processes are addressed together or in separate stages. In both cases, the various features of images were used to localize and recognize grocery products. Although most of the studies are taken advantage of the visual features of the item, there are also studies (Chen et al., 2022), (Selvam and Koilraj, 2022), (Ghosh, 2023) and (George et al., 2015) that use the textual attributes of the goods. As listed in Table 6, both visual and textual features were extracted with traditional attributes extraction methods or deep learning based architectures. To achieve better recognition performance, automatic product recognition systems often use several feature extraction methods separately or together for different categories of product recognition. The choice of the attributes and their combination depends on the specific requirements and challenges of the product recognition task. That is, systems may use different combinations of features based on factors like the availability of data, the nature of the products, the quality of the images and text, and the desired level of recognition accuracy. That is, using combination of attributes allows the system to leverage the strengths of different feature extraction methods.

Table 6
Classification of the studies according to using features in different stage.

Main Category of the Problem	Traditional feature extraction	Deep learning based feature extraction
Product Classification	(Zhang et al., 2007) (Zhang et al., 2009) (Ray et al., 2018) (Franco et al., 2017) (Ghosh, 2023) (George et al., 2015) (Melek et al., 2023)	(Jund et al., 2016) (Chen et al., 2022) (Santra et al., 2021) (Franco et al., 2017) (Santra et al., 2020b) (Srivastava, 2020) (Jonathan and Kusuma, 2022) (Srivastava, 2022) (Wang et al., 2022) (Farren, 2017) (Ciocca et al., 2016)
Product Localization	(Santra et al., 2021) (Selvam and Koilraj, 2022) (Franco et al., 2017) (Yücel and Ünsalan, 2024) (Melek et al., 2023)	(Goldman et al., 2019) (Chen et al., 2022) (Varadarajan et al., 2020) (Selvam and Koilraj, 2022) (Varadarajan and Srivastava, 2018) (Melek et al., 2023) (Ye et al., 2016) (Pan et al., 2020) (Rong et al., 2021) (Deng and Yang, 2021) (Wang et al., 2020) (Moon et al., 2022) (Ravi et al., 2022) (Xu et al., 2022)
Product Detection	(Merler et al., 2007) (George and Floerkemeier, 2014) (Varol and Kuzu, 2015) (Geng et al., 2018) (Winlock et al., 2010) (Karlinsky et al., 2017) (Yörük et al., 2016) (Tonioni and Di Stefano, 2017) (Varol, 2014) (Gokdag, 2019) (Gökdag, 2016)	(Cai et al., 2021) (Santra et al., 2020a) (Kant, 2020) (Yilmazer and Birant, 2021) (Santra et al., 2022) (Cho et al., 2022) (Karlinsky et al., 2017) (Tonioni et al., 2018) (Osokin et al., 2020) (Sinha et al., 2021) (De Feyter, 2023) (Hurtik and Vlasanek, 2020) (Strohmayer and Kampel, 2023)

4.2.1.1. Traditional feature extraction methods. Traditional feature extraction methods provide a comprehensive understanding of the products from attributes such as "color", "local", and "shape and contour" features, as listed in Table 7. **Color features capture the color distribution in product images, including color histograms**. A color histogram counts the number of pixels that fall into different color bins. CHM involves comparing the color histograms of two images and adjusting one of the histograms to make it more similar to the other (Lovak and Shafer). Different color spaces were preferred in product recognition studies. Lab color space in (Merler et al., 2007) and (Winlock et al., 2010), Hue, Saturation, Value (HSV) color space in (Varol and Kuzu, 2015) and (Varol, 2014) were used. In Lab color space, "L" represents the lightness of the color ranges from 0 (black) to 100 (white). "a" represents the position of the color on the green to red axis. Negative values represent green, and positive values represent red. "b" represents the position of the color on the blue to yellow axis. Negative values represent blue, and positive values represent yellow. In HSV color space, hue represents the type of color, such as red, green, or blue. It is measured in degrees on the color wheel, ranging from 0 to 360. In the HSV color space, red is typically around 0°, green is around 120°, and blue is around 240°. Saturation measures the intensity or purity of the color. A saturation value of 0 represents a shade of gray (no color), and a saturation value of 100 represents the full intensity of the color. Value represents the brightness of the color and ranges of color is same with "L" of Lab color space. According to the comparison of color spaces YCbCr, HSV and Lab in (Merler et al., 2007), Lab gives the best result.

Local features are key point descriptors used to detect and match distinctive local features. The commonly used local features extractors are SIFT (Lowe, 2004), SURF (Bay et al., 2006), BRISK (Leutenegger et al., 2011), ORB (Rublee et al., 2011), AKAZE (Ren et al., 2017) and Hessian-Affine (Mikolajczyk and Schmid, 2004). SIFT is a method for detecting and describing local features in images. It is widely used for object recognition and matching, and it's invariant to scale, rotation, and translation (Lowe, 2004). SURF is similar to SIFT but designed to be computationally more efficient. It uses a box filter approximation and integral images to accelerate key point detection (Bay et al., 2006). BRISK is a feature detector and descriptor that is designed to be both efficient and robust. It operates on the binary strings of intensity comparisons to describe key points (Leutenegger et al., 2011). ORB is a fast and efficient alternative to SIFT and SURF. It uses a combination of the FAST key point detector and the BRIEF descriptor, making it suitable for real-time applications (Rublee et al., 2011). AKAZE is an accelerated version of KAZE, which is a feature detector and descriptor. It is designed to be both fast and robust and is particularly suitable for image matching and object recognition (Ren et al., 2017). Hessian-Affine is a method for detecting affine-invariant features. It uses the Hessian matrix to analyze the local structure of an image and identify distinctive and repeatable features (Mikolajczyk and Schmid, 2004). The number of features, computational cost, and robustness to scale, rotation and affine transformation are key factors to select more appropriate local feature descriptors for product recognition. As it can be seen in Table 7, more than one method was used in some studies. One reason for this is to compare the results obtained from different methods (Geng et al., 2018), (Yücel and Ünsalan, 2024), and the other reason is to use more than one method as a hybrid to increase system success (Tonioni and Di Stefano, 2017), (Melek et al., 2023). While different methods provide best result for different problems on different datasets in (Yücel and Ünsalan, 2024), SIFT provide best result in (Geng et al., 2018). On the other hand, while in (Tonioni and Di Stefano, 2017), the combined use of SURF and BRISK did not provide a significant contribution, in (Melek et al., 2023), the hybrid usage of SURF, BRISK and ORB increased the performance.

Shape and contour features include information about object shapes and boundaries extracted from images. The used shape and contour features in product recognition studies are Harris-Affine (Mikolajczyk and Schmid, 2004), BHaar (Viola and Jones, 2001), HoG (Dalal and Triggs) and HT (Duda and Hart, 1972). Harris-Affine is an extension of the Harris corner detector, designed to be invariant to affine transformations that identifies keypoints that are robust to changes in scale, rotation, and skew (Mikolajczyk and Schmid, 2004). BHaar is likely a variant or extension of Haar-like features, which are used in object detection, particularly in the Viola-Jones face detection algorithm (Viola and Jones, 2001). HoG is a feature descriptor that captures information about the distribution of gradient orientations in an image (Dalal and Triggs). The HT is a technique used for detecting simple geometric shapes (such as lines or circles) in images. It is often used as a feature extraction method in image analysis (Duda and Hart, 1972). Shape and contour features can be computationally expensive, have limitation on rotation invariance, scale and sensitivity to noise. Therefore, shape and contour features were used less than other features in the product recognition task. Additionally, it has been used with some color or local features to increase system performance. For instance, Harris-Affine was used with SIFT in (Zhang et al., 2007), BHaar was used with CHM in (Merler et al., 2007), HoG was used with CHM in (Varol and Kuzu, 2015), HT was used with SURF in (Yörük et al., 2016).

Table 7
Using methods of traditional feature extraction in different stage.

Type of Traditional Features	Methods of Traditional Feature Extraction	Studies
Color features	**CHM** (Lovak and Shafer)	(Merler et al., 2007) (Varol and Kuzu, 2015) (Winlock et al., 2010) (Varol, 2014)
Local Features	**SIFT** (Lowe, 2004)	(Zhang et al., 2007) (Merler et al., 2007) (George and Floerkemeier, 2014) (Varol and Kuzu, 2015) (Geng et al., 2018) (Tonioni and Di Stefano, 2017) (Yücel and Ünsalan, 2024) (Gökdag, 2016)
	SURF (Bay et al., 2006)	(Geng et al., 2018) (Ray et al., 2018) (Winlock et al., 2010) (George et al., 2015) (Yörük et al., 2016) (Tonioni and Di Stefano, 2017) (Yücel and Ünsalan, 2024) (Melek et al., 2023)
	BRISK (Leutenegger et al., 2011)	(Geng et al., 2018) (Santra et al., 2021) (Santra et al., 2022) (Tonioni and Di Stefano, 2017) (Yücel and Ünsalan, 2024) (Melek et al., 2023)
	ORB (Rublee et al., 2011)	(Tonioni and Di Stefano, 2017) (Yücel and Ünsalan, 2024) (Melek et al., 2023)
	AKAZE (Alcantarilla and Bartoli)	(Tonioni and Di Stefano, 2017) (Yücel and Ünsalan, 2024)
	Hessian-Affine (Mikolajczyk and Schmid, 2004)	Zhang et al. (2009)
Shape and Contour Features	**Harris-Affine** (Mikolajczyk and Schmid, 2004)	Zhang et al. (2007)
	BHaar (Viola and Jones, 2001)	Merler et al. (2007)
	HoG (Dalal and Triggs)	Varol (2014)
	HT (Duda and Hart, 1972)	(Yörük et al., 2016) (Varol, 2014)

4.2.1.2. Deep learning based Feature Extraction Algorithms. **Deep learning features are extracted from deep** CNN that have been trained on large datasets of product images. According to the previous studies on product recognition, CaffeNet (Jia et al., 2014), ResNet (He et al., 2016), AlexNet (Krizhevsky and Hinton), Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), VGG (Simonyan and Zisserman, 2015), SuperPoint (Detone and Rabinovich), ResNext (Xie et al., 2017), SNN (Koch, 2011), DenseNet-169 (Huang and Weinberger) and MobileNet (Howard et al., 2019) were used for deep learning based feature extraction. Table 8 associates to every deep

Table 8
The Studies Using Methods of Deep Learning based Feature Extraction.

Methods of Deep Learning based Feature Extraction	Studies
CaffeNet (Jia et al., 2014)	(Jund et al., 2016) (Farren, 2017)
ResNet (He et al., 2016)	(Geng et al., 2018) (Cai et al., 2021) (Santra et al., 2020a) (Santra et al., 2021) (Kant, 2020) (Cho et al., 2022) (Selvam and Koilraj, 2022) (Osokin et al., 2020) (Sinha et al., 2021) (De Feyter, 2023) (Ye et al., 2016) (Xu et al., 2022)
AlexNet (Krizhevsky and Hinton)	(Franco et al., 2017) (Santra et al., 2020b)
BERT (Devlin et al., 2019)	Chen et al. (2022)
VGG (Simonyan and Zisserman, 2015)	(Chen et al., 2022) (Santra et al., 2022) (Karlinsky et al., 2017) (Varadarajan and Srivastava, 2018) (Jonathan and Kusuma, 2022) (Tonioni et al., 2018)
SuperPoint (Detone and Rabinovich)	Srivastava (2020)
ResNext (Xie et al., 2017)	(Srivastava, 2020) (Srivastava, 2022)
SNN (Koch, 2011)	Wang et al. (2022)
DenseNet-169 (Huang and Weinberger)	Ciocca et al. (2016)
MobileNet (Howard et al., 2019)	Strohmayer and Kampel (2023)

learning method the list of studies using it. CaffeNet is a deep CNN architecture based on the AlexNet architecture and was one of the early models used in the Caffe deep learning framework (Jia et al., 2014). ResNet introduced the concept of residual learning, where shortcut connections are used to skip one or more layers (He et al., 2016). Residual connections help address the vanishing gradient problem and facilitate the training of very deep neural networks. AlexNet (Krizhevsky and Hinton) consists of multiple convolutional and fully connected layers, incorporating techniques like ReLU activation and dropout (Krizhevsky and Hinton). BERT is a transformer-based model uses bidirectional context to understand the meaning of words in a sentence more effectively (Devlin et al., 2019). VGG is a CNN architecture known for its simplicity and effectiveness. It consists of multiple convolutional layers with small receptive fields, followed by max-pooling layers (Simonyan and Zisserman, 2015). SuperPoint is a neural network designed for keypoint detection in images.It is trained to predict keypoints and their descriptors in an unsupervised manner (Detone and Rabinovich). ResNext is an extension of the ResNet architecture, incorporating the concept of grouped convolutions. It introduces the idea of cardinality which represents the number of groups in grouped convolutions (Xie et al., 2017). SNN consist of two identical subnetworks with shared weights for tasks involving similarity or distance measurement with one-shot learning (Koch, 2011). DenseNet is a convolutional neural network architecture characterized by densely connected blocks, facilitating feature reuse and efficient learning having 169 layers for image classification tasks (Huang and Weinberger). MobileNet provide efficient feature extraction while minimizing computational requirements of CNN architecture designed for mobile and edge devices (Howard et al., 2019).

4.2.2. Methods for features classification

In order to recognize the product or find its location, the obtained features must be classified by matching them with the template images of the product. As listed in Table 9, different traditional and deep learning classification methods have been used for this: hierarchical k-means clustering (Stork, 2016), SVM, VGG-16 (Simonyan and Zisserman, 2015), RNN (Ghosh et al., 2019), ELM (Ä et al., 2006), Hungarian Algorithm (Robinson and Assignment, 1950), Density Based Spatial Clustering of Applications with Noise clustering (DBSCAN) (Ester et al., 1996), Naive Bayes classifier (Barrington et al., 2008), random forest (Breiman, 2001) and KNN classifier. Hierarchical k-means clustering (Stork, 2016) arranges data into a hierarchical structure resembling a

Table 9
Methods used for classification of features.

Methods	Studies
Hierarchical K-means Clustering (Stork, 2016)	Zhang et al. (2007)
SVM	(Varol and Kuzu, 2015) (George et al., 2015) (Varol, 2014) (Ciocca et al., 2016)
VGG-16 (Simonyan and Zisserman, 2015)	Geng et al. (2018)
RNN (Ghosh et al., 2019)	Ghosh (2023)
ELM (Ä et al., 2006)	Gökdag (2016)
Hungarian Algorithm (Robinson and Assignment, 1950)	Chen et al. (2022)
DBSCAN clustering (Ester et al., 1996)	Santra et al. (2021)
Naive Bayes Classifier (Barrington et al., 2008)	Winlock et al. (2010)
Random Forest (Breiman, 2001)	Santra et al. (2020b)
KNN Classifier	(Tonioni et al., 2018) (Strohmayer and Kampel, 2023)

tree based on the k-means algorithm. It clusters similar features together in images to classify features. SVM is a supervised machine learning algorithm that works by finding the hyperplane that best separates data points into different classes. VGG-16 (Simonyan and Zisserman, 2015) is a convolutional neural network architecture with 16 wt layers that has a straightforward structure with small receptive fields in the convolutional layers. RNNs are designed for sequential data processing and have connections that form directed cycles, allowing them to maintain a hidden state that captures information from previous time steps (Ghosh et al., 2019). ELM (Ä et al., 2006) is a type of neural network that consists of a single hidden layer of neurons and a simple linear output layer. The difference of ELM from traditional neural networks is the weights between the input layer and the hidden layer are randomly generated and remain fixed. The output weights are then computed analytically using a least squares approach. ELM provide fast training speed and good generalization performance, especially in scenarios with large datasets. Hungarian Algorithm (Robinson and Assignment, 1950) which is used for solving assignment problems, was adapted to match features or objects between two images. DBSCAN (Ester et al., 1996) is a clustering algorithm that groups together data points based on their density. It recognizes products from the created clusters that have varying shapes and sizes. Naive Bayes classifier (Barrington et al., 2008) is a probabilistic classification algorithm that assumes independence between features based on Bayes' theorem. Random forest (Breiman, 2001) is an ensemble learning method that builds multiple decision trees during training and outputs the mode of the classes as the prediction. KNN classifier is a simple and intuitive classification algorithm that assigns a data point to the majority class among its k-nearest neighbors. The choice of the method depends on factors such as the nature of the data, the complexity of the problem, and the available of the resources. Each of these method has its strengths and it is suitable for specific types of tasks of product recognition problem.

4.2.3. Algorithms for product detection

Some studies find product positions directly, regardless the product type, or find product positions with class information. As listed in Table 10, in addition to the common object detection algorithms Faster R–CNN (Ren et al., 2017), RetinaNet (Lin et al., 2020), Single Shot MultiBox Detector (SSD) (Liu et al., 2016), Cascade R–CNN (Cai and Vasconcelos, 2018), CenterNet (Zhou et al., 2019), Cascaded Object Detector (Viola and Jones, 2001), YOLOv2 (Redmon and Farhadi, 2017), YOLOv3 (Redmon and Farhadi, 2018), YOLOv4 (Wang and Liao), YOLOv5 (Jocher et al., 2020), YOLOv7tiny (Wang et al., 2023), Mask R–CNN (He et al., 2020), CenterNet2 (Zhou and Koltun, 2018), Refine-Net (Varadarajan and Srivastava, 2018), YOLO-ASC (Hurtik and Vlasanek, 2020) algorithms were developed specifically for product

Table 10
Methods used for product detection.

Methods	Studies
Faster R–CNN (Ren et al., 2017)	(Varadarajan et al., 2020) (Ghosh, 2023) (Sinha et al., 2021) (De Feyter, 2023) (Deng and Yang, 2021) (Ye et al., 2016) (Wang et al., 2020)
RetinaNet (Lin et al., 2020)	(Kant, 2020) (Yilmazer and Birant, 2021) (Wang et al., 2020)
Refine-Net (Varadarajan and Srivastava, 2018)	Varadarajan and Srivastava (2018)
SSD (Liu et al., 2016)	Melek et al. (2023)
Cascade R–CNN (Cai and Vasconcelos, 2018)	(Cai et al., 2021) (Rong et al., 2021) (Xu et al., 2022)
CenterNet (Zhou et al., 2019)	Pan et al. (2020)
Cascaded Object Detector (Viola and Jones, 2001)	(Varol, 2014) (Gokdag, 2019) (Gökdag, 2016)
YOLO-ASC (Hurtik and Vlasanek, 2020)	Hurtik and Vlasanek (2020)
YOLOv2 (Redmon and Farhadi, 2017)	Tonioni et al. (2018)
YOLOv3 (Redmon and Farhadi, 2018)	Yilmazer and Birant (2021)
YOLOv4 (Wang and Liao)	Yilmazer and Birant (2021)
YOLOv5 (Jocher et al., 2020)	(Selvam and Koilraj, 2022) (Ravi et al., 2022)
YOLOv7tiny (Wang et al., 2023)	Strohmayer and Kampel (2023)
Mask R–CNN (He et al., 2020)	Ravi et al. (2022)
CenterNet2 (Zhou and Koltun, 2018)	Moon et al. (2022)

detection.

Faster R–CNN (Ren et al., 2017) is a two-stage object detection model that introduces a Region Proposal Network (RPN) to generate region proposals for potential objects, which are then classified and refined. RetinaNet (Lin et al., 2020) is designed to address the issue of class imbalance in dense object detection. It introduces the Focal Loss to assign higher weights to hard-to-classify examples, improving the detection of rare objects. SSD (Liu et al., 2016) is a one-stage object detection model that simultaneously predicts multiple bounding boxes and their associated class scores. It operates on feature maps of multiple scales to capture objects of various sizes. Cascade R–CNN (Cai and Vasconcelos, 2018) is an extension of Faster R–CNN (Ren et al., 2017) that introduces a cascade of detectors, where each stage focuses on refining the proposals from the previous stage, gradually improving the accuracy of the detection. CenterNet (Zhou et al., 2019) directly predicts object centers and regresses bounding boxes from those centers. It achieves competitive accuracy with a simpler architecture compared to two-stage detectors. Cascaded Object Detector (Viola and Jones, 2001) is a method that applies a cascade of classifiers with increasing levels of complexity to refine the detection results, enhancing the precision of the object detection task. YOLOv2 (Robinson and Assignment, 1950) also known as YOLO9000, extends the YOLO architecture to handle object detection in multiple classes. It introduces anchor boxes and hierarchical classification to improve accuracy. YOLOv3 (Redmon and Farhadi, 2018) is an evolution of the YOLO architecture with improvements in accuracy and speed. It introduces additional features such as multi-scale detection, a darknet-53 backbone, and the ability to handle various object sizes. YOLOv4 (Wang and Liao) is a further enhancement to the YOLO series, introducing new features such as the CSPDarknet53 backbone, PANet, and Mish activation function, aiming for state-of-the-art performance in object detection. YOLOv5 (Jocher et al., 2020) is another iteration of the YOLO series, emphasizing simplicity and efficiency. It introduces a streamlined architecture and focuses on ease of use and deployment. YOLOv7tiny (Wang et al., 2023) is a lightweight version of the YOLO architecture, designed for real-time applications with reduced model size and complexity. Mask R–CNN (He et al., 2020) extends Faster R–CNN (Ren et al., 2017) to perform instance segmentation, providing pixel-level segmentation masks in addition to bounding boxes and class scores. CenterNet2 (Zhou and Koltun, 2018) is an improved version of CenterNet, incorporating advancements in object detection by focusing on keypoint-based center prediction. Refine-Net (Varadarajan and Srivastava, 2018) is designed for semantic segmentation, providing a multi-path refinement network to improve the segmentation accuracy by combining information from different scales and levels of abstraction. YOLO-ASC (Hurtik and Vlasanek, 2020) is an improvement to the YOLO architecture, introducing adaptive selection cascades to enhance detection accuracy. These methods represent various approaches to object detection, each with its unique characteristics and trade-offs. The choice of a specific method depends on factors such as task requirements, dataset characteristics, and computational constraints.

4.2.4. Algorithms used in the refinement stage

The refinement stage aims to get the final result by merging several overlapped bounding boxes corresponding to every item present in the shelf image into a single bounding box. As listed in Table 11, NMS (Neubeck, 2006), Greedy-NMS (Felzenszwalb et al., 2009), Hashing-based NMS (HNMS) (Wang et al., 2020), clustering algorithms, graph-based approaches and overlap filter (Gokdag, 2019) were used in the refinement stage of the product recognition task. NMS (Neubeck, 2006) is a post-processing technique commonly used in object detection tasks. It addresses the problem of multiple bounding box proposals for the same object. During NMS, bounding boxes are ranked based on their confidence scores, and overlapping boxes with lower scores are suppressed, keeping only the box with the highest confidence. Greedy-NMS (Felzenszwalb et al., 2009) is a variant of NMS that employs a greedy strategy to iteratively select the bounding box with the highest confidence and remove overlapping boxes with lower scores. It is a computationally efficient approach commonly used in real-time applications. HNMS (Wang et al., 2020) is an extension of NMS that incorporates hashing techniques. It efficiently handles large-scale object detection scenarios by utilizing hash functions to index bounding boxes, making the suppression process more scalable. Clustering helps to organize the bounding boxes into coherent groups, potentially improving the accuracy of localization. Graph-based approaches in the context of product recognition often involve representing relationships between detected objects as a graph and leveraging graph-based algorithms to refine recognition results. Overlap filter is a simple technique where bounding boxes with significant overlap are filtered to retain only the box with the highest confidence. It is a straightforward approach to reducing redundancy in detection results. These methods are commonly used to refine the output of product recognition and localization systems, especially in scenarios where multiple bounding box proposals are generated for the same object. Refinement techniques are crucial for improving the precision and reliability of the detected objects and ensuring that only the most confident and accurate predictions are retained in the final results. The choice of a specific refinement method depends on factors such as computational efficiency, scalability, and the characteristics of the detection task.

Table 11
Methods used in refinement stage.

Methods	Studies
NMS (Neubeck, 2006)	(Varadarajan et al., 2020) (Franco et al., 2017) (Moon et al., 2022)
Greedy-NMS (Felzenszwalb et al., 2009)	(Santra et al., 2020a) (Santra et al., 2021)
HNMS (Wang et al., 2020)	(Wang et al., 2020)
Clustering Algorithms	(Franco et al., 2017) (Melek et al., 2023)
Graph-based Approach	(Ray et al., 2018) (Santra et al., 2020a) (Tonioni and Di Stefano, 2017)
Overlap Filter	Gokdag (2019)

5. Benefits and limitations of grocery product recognition studies

A successful automatic product recognition system provides many benefits to manufacturers, suppliers, and consumers. Improving shopping experience of visually impaired people, instant price comparison, being able to find the product you are looking for faster and getting product recommendations are significant advantages provides to customers. On the other hand, tracking stock information and the placement of products on the shelves, reducing the amount of man-hours and human error are facilitated benefits to suppliers and manufacturers.

This survey allows to reveal some benefits and limitations of automatic product recognition systems, datasets and methods. The automatic grocery product recognition has many situations that increase the complexity of this task. First of all, there are many variety of products that need to be recognized and new ones can be added day by day. In addition, the large number of products having inter-class similarity increases the difficulty of the problem. Grocery stores have different designs, complexity (high density or visual clutter of shelves) and lighting conditions. This diversity in grocery stores effect directly to captured images from shelves cause to challenging viewing angle, increasing blurring and brightness in images and occlusions the items each other. Despite the fact that all of these make it challenging to create an automated system for product recognition, there are studies that overcome these limitations.

As detailed in section 2, publicly available datasets of grocery products have different properties in term of variety of products, features, labelling, annotations and complexities. These datasets allow researchers to test their proposed methods and to get comparable results with other studies. The condition of retail store environment or methods of image acquisition are caused by poor image quality with the properties such as blurring, brightness, shadows, noise. Some pre-processing techniques may be required to overcome the difficulties as suggested in (Varol and Kuzu, 2015) and implemented in (Selvam and Koilraj, 2022) and (Franco et al., 2017). Grocery product datasets have not enough images for successful implementation of some methods. Synthetic images are used in (Merler et al., 2007), (Santra et al., 2020a), (Santra et al., 2021), (Santra et al., 2022), (Selvam and Koilraj, 2022), (Karlinsky et al., 2017), (Varadarajan and Srivastava, 2018), (Hurtik and Vlasanek, 2020), (Ravi et al., 2022), (Strohmayer and Kampel, 2023) and data augmentation is used in (Geng et al., 2018), (Chen et al., 2022), (Varadarajan et al., 2020)- (Santra et al., 2021), (Yilmazer and Birant, 2021), (Santra et al., 2022), (Selvam and Koilraj, 2022), (Karlinsky et al., 2017)- (Jonathan and Kusuma, 2022), (Wang et al., 2022), (Tonioni et al., 2018), (Osokin et al., 2020), (Ciocca et al., 2016)- (Pan et al., 2020), (Moon et al., 2022)- (Strohmayer and Kampel, 2023) to overcome this limitation of datasets. In addition, some pre-trained models are used in (Jund et al., 2016), (Geng et al., 2018), (Chen et al., 2022), (Santra et al., 2020a), (Santra et al., 2021), (Yilmazer and Birant, 2021), (Selvam and Koilraj, 2022), (Franco et al., 2017)- (Srivastava, 2022), (Tonioni et al., 2018), (Osokin et al., 2020), (Sinha et al., 2021), (De Feyter, 2023), (Hurtik and Vlasanek, 2020), (Ye et al., 2016), (Deng and Yang, 2021), (Moon et al., 2022)- (Xu et al., 2022) to increase performance of CNNs. In real world problems, new products and changing product packages must be regularly added to the system. In this situations, using in vitro images collected from internet that were captured under ideal imaging conditions is more useful and practical than acquiring new images every time (Merler et al., 2007). Also, to prevent retrained the system each time, incremental learning was used in (Cai et al., 2021), (Varol, 2014)- (Gökdag, 2016), (Rong et al., 2021), (Xu et al., 2022) or one-shot learning in (Wang et al., 2022), (Osokin et al., 2020), (De Feyter, 2023).

According to the comparison of these feature extraction types (Franco et al., 2017), deep learning based features have strong robustness and invariance and are, generally, more effective in complex scenarios. Traditional features are more effective when imposing global structural similarity in simpler scenarios. Traditional features can perform well with smaller datasets (i.e. WebMarket (Zhang et al., 2007)) consisting of a few number of product template images, where deep learning models might require large amounts of labeled data (i.e. SKU-110K (Goldman et al., 2019)) for effective training. There are studies (Varol and Kuzu, 2015), (Geng et al., 2018), (Santra et al., 2021), (Winlock et al., 2010), (George et al., 2015), (Tonioni and Di Stefano, 2017), (Melek et al., 2023), (Varol, 2014) in which system success is increased by using different feature extraction methods together. In this way, mismatching features are detected. Additionally, recognizing products from text features ((Chen et al., 2022), (Selvam and Koilraj, 2022), (Ghosh, 2023) and (George et al., 2015)) can be advantageous, especially for products with similar packaging. However, in real market shelf images, text may not be visible or may appear incomplete depending on the product's position or the angle of the photograph. Therefore, attempting to recognize products solely from text may not always be sufficient. The combined use of different visual or deep features can enhance performance, but some of the proposed methods for recognizing products from text may not be preferred due to computational costs.

At the product detection stage, two-stage models Faster R–CNN (Ren et al., 2017), RetinaNet (Lin et al., 2020), Cascade R–CNN (Cai and Vasconcelos, 2018), Mask R–CNN (He et al., 2020) and one-stage models YOLO-ASC (Hurtik and Vlasanek, 2020), YOLOv2 (Redmon and Farhadi, 2017), YOLOv3 (Redmon and Farhadi, 2018), YOLOv3 (Redmon and Farhadi, 2018), YOLOv4 (Wang and Liao), YOLOv5 (Jocher et al., 2020), YOLOv7tiny (Wang et al., 2023), CenterNet (Zhou et al., 2019), CenterNet2 (Zhou and Koltun, 2018) were generally used. Two-stage detectors involve two main stages. The first stage is responsible for generating region proposals, which are potential bounding box locations that might contain objects. The second stage performs object classification and refines the proposed bounding boxes. It classifies the objects within the proposed regions and adjusts the bounding box coordinates. One-stage detectors combine this two stage in a single stage. So, one-stage detectors are faster than two-stage detectors. One-stage detectors can perform better in scenarios where objects are of various scales and sizes, whereas two-stage detectors can excel in more challenging scenarios with smaller objects or dense scenes. As a result, one-stage detectors are preferable for real-time systems and two-stage detectors are more preferable for dense product detection.

Additionally, leveraging context and relationship information from graphical models to extract insights from relationships between products and shelf layouts is important for improving product recognition results. Graph-based approached in (Ray et al., 2018), (Santra et al., 2020a), (Tonioni and Di Stefano, 2017), confidence set based on the potential product arrangements on the shelf and the integrated visual hierarchy between brands in (Baz et al., 2019) and neighborhood relational clustering algorithm in (Melek et al., 2023) provide to increase system performance.

6. Conclusion

This paper presents an exhaustive literature review of product recognition on grocery shelf images. With this aim, the most common datasets used for product recognition are introduced. Previous studies are introduced and classified according to the used methods and handled problems. By examining the past researches, its benefits and limitations were revealed. The analysis of this paper is useful for further studies to increase automatic product recognition system performance.

The problem of product recognition cannot be addressed as a single problem, and achieving high success with a single method is quite challenging. In an end-to-end system where the location and class information of all products in a given shelf image are obtained, suitable detection algorithms and feature extraction methods can be used based on priorities such as speed and performance. Additionally, the specific characteristics of the problem being addressed also influence the

selection of methods. For example, algorithms capable of capturing small objects are suitable for dense product detection, while methods that can distinguish very similar features are more suitable for fine-grained classification.

The problem of recognizing products on retail shelves continues to be a significant challenge that is extensively researched both in academia and industry. Future work may include the use of hybrid models that utilize different features together, the integration of the visual features with other relevant sensor data. Despite the success of deep learning-based methods in many areas of object detection and classification, datasets of retail products are not sufficient to train these networks effectively. Working on larger datasets with deeper models is necessary. Therefore, there is a need to develop new approaches and methods to increase the size of retail product datasets and obtain labeled data. For future research, it is crucial to focus on making necessary improvements to ensure that the system operates faster and is more computationally efficient. Additionally, there is a need to explore new solutions that do not require frequent retraining of the system. Given the difficulty and cost associated with obtaining labeled data, developing unsupervised algorithms is also an important area of focus for future studies. The success of product recognition can be enhanced by leveraging context and relationship information from graphical models, which can extract important insights from relationships between products and shelf layouts. In addition to conventional methods, transformer-based approaches, which have gained popularity recently (DETR-DEtection Transformer, ViT-Vision Transformer, Swin Transformer, etc.) and demonstrated success in diverse domains, can also be employed for product recognition.

CRediT authorship contribution statement

Ceren Gülra Melek: Conceptualization, Formal analysis, Investigation, Resources, Validation, Visualization, Writing – original draft. **Elena Battini Sönmez:** Investigation, Methodology, Resources, Supervision, Validation, Writing – review & editing. **Songül Varlı:** Investigation, Methodology, Resources, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Ā, G.H., Zhu, Q., Siew, C., 2006. Extreme Learning Machine : Theory and Applications, vol. 70, pp. 489–501. https://doi.org/10.1016/j.neucom.2005.12.126.

P. F. Alcantarilla and A. Bartoli, “Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces”...

Barrington, L., Marks, T.K., Hsiao, J.H.W., Cottrell, G.W., 2008. Nimble: a kernel density model of saccade-based visual memory. J. Vis. 8 (14), 1–14. https://doi.org/10.1167/8.14.17.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. LNCS 3951 - SURF: Speeded up Robust Features.

Baz, I., Yoruk, E., Cetin, M., 2019. Context-aware confidence sets for fine-grained product recognition. IEEE Access 7, 76376–76393. https://doi.org/10.1109/ACCESS.2019.2921994.

Breiman, L.E.O., 2001. Random Forests, pp. 5–32.

Busu, M.F.M., Ismail, I., Saaid, M.F., Norzeli, S.M., 2011. Auto-checkout system for retails using Radio Frequency Identification (RFID) technology. Proc. - 2011 IEEE Control Syst. Grad. Res. Colloquium, ICSGRC 2011 193–196. https://doi.org/10.1109/ICSGRC.2011.5991855.

Cai, Y., Wen, L., Zhang, L., Du, D., Wang, W., 2021. Rethinking object detection in retail stores. 35th AAAI Conf. Artif. Intell. AAAI 2021 2A, 947–954. https://doi.org/10.1609/aaai.v35i2.16178.

Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Return of the devil in the details: delving deep into convolutional nets. BMVC 2014 - Proc. Br. Mach. Vis. Conf. 2014 1–11. https://doi.org/10.5244/c.28.6.

Chen, F., et al., 2022. Unitail: detecting, reading, and matching in retail scene. Lect. Notes Comput. Sci. 13667 (LNCS), 705–722. https://doi.org/10.1007/978-3-031-20071-7_41.

Cho, S., Paeng, J., Kwon, J., 2022. Densely-packed object detection via hard negative-aware anchor attention. In: Proc. - 2022 IEEE/CVF Winter Conf. Appl. Comput. Vision, WACV 2022, pp. 1401–1410. https://doi.org/10.1109/WACV51458.2022.00147.

Ciocca, G., Napoletano, P., Locatelli, S.G., 2016. Multi-task Learning for Supervised and Unsupervised Classification of Grocery Images, pp. 1–14.

Cai, Z., Vasconcelos, N., 2018. Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6154–6162. https://doi.org/10.48550/arXiv.1712.00726.

N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection”...

De Feyter, F., 2023. Joint Training of Product Detection and Recognition Using Task-Specific Datasets 5 (Visigrapp), 715–722. https://doi.org/0011725100003417.

Deng, Z., Yang, C., 2021. Multiple-step Sampling for Dense Object Detection and Counting, pp. 1036–1042.

D. Detone and A. Rabinovich, “SuperPoint: Self-Supervised Interest Point Detection and Description”...

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf. 1 (Mlm), 4171–4186.

Duda, R.O., Hart, P.E., 1972. Use of the Hough transformation to detect lines and curves in pictures, 15 (1). https://doi.org/10.1145/361237.361242.

Ester, M., Kriegel, H., Xu, X., Miinchen, D.-, 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.

Farren, D., 2017. Classifying Food Items by Image Using Convolutional Neural Networks. Stanford University, Stanford, CA, USA.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D., 2009. Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. 32 (9), 1627–1645.

Fernandcz, W.P., Xian, Y., Tian, Y., 2018. Image-based barcode detection and recognition to assist visually impaired Persons. In: 2017 IEEE 7th Annu. Int. Conf. CYBER Technol. Autom. Control. Intell. Syst. CYBER 2017, pp. 1241–1245. https://doi.org/10.1109/CYBER.2017.8446388.

Franco, A., Maltoni, D., Papi, S., 2017. Grocery product detection and recognition. Expert Syst. Appl. 81, 163–176. https://doi.org/10.1016/j.eswa.2017.02.050.

Geng, W., et al., 2018. Fine-grained Grocery Product Recognition by One-Shot Learning, vol. 2, pp. 1706–1714.

George, M., Floerkemeier, C., 2014. LNCS 8690 - Recognizing Products: A Per-Exemplar Multi-Label Image Classification Approach.

George, M., Mircic, D., Sörös, G., Floerkemeier, C., Mattern, F., 2015. Fine-grained product class recognition for assisted shopping. Proc. IEEE Int. Conf. Comput. Vis. 2015-Febru, 546–554. https://doi.org/10.1109/ICCVW.2015.77.

Georgiadis, K., et al., 2021. Products-6K : A Large-Scale Groceries Product Recognition Dataset Products-6K : A Large-Scale Groceries Product Recognition Dataset (July 2022). https://doi.org/10.1145/3453892.3453894.

Ghosh, R., 2023. Product Identification in Retail Stores by Combining Faster R-Cnn and Recurrent Neural Network. https://doi.org/10.1007/s11042-023-15633-1.

Ghosh, R., Vamshi, C., Kumar, P., 2019. RNN Based Online Handwritten Word Recognition in Devanagari and Bengali Scripts Using Horizontal Zoning, vol. 92, pp. 203–218. https://doi.org/10.1016/j.patcog.2019.03.030.

Gökdag, Ü., 2016. Planogram Matching Control in Grocery Products by Image Processing. https://doi.org/10.1109/SIU.2016.7495972.

Gokdag, U., 2019. Raf Görüntüleri Üzerinde Nesne Tanımaya Dayalı Planogram Eşleştirme no. October.

Goldman, E., Herzig, R., Eisenschtat, A., Goldberger, J., Hassner, T., 2019. Precise detection in densely packed scenes. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. 2019-June, 5222–5231. https://doi.org/10.1109/CVPR.2019.00537.

Hao, Y., 2019. Take Goods from Shelves : A Dataset for Class-Incremental Object Detection, pp. 271–278. https://doi.org/10.1145/3323873.3325033.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2020. Mask R-CNN. IEEE Trans. Pattern Anal. Mach. Intell. 42 (2), 386–397. https://doi.org/10.1109/TPAMI.2018.2844175.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H., 2019. Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 1314–1324.

G. Huang and K. Q. Weinberger, “Densely Connected Convolutional Networks”...

Huang, L., Yang, Y., Deng, Y., Yu, Y., 2015. Dens: Unifying Landmark Localization with End to End Object Detection 1–13 [Online]. Available: http://arxiv.org/abs/1509.04874.

Hurtik, P., Vlasanek, P., 2020. YOLO-ASC : You Only Look once and See Contours.

G. Jocher et al., “ultralytics/yolov5: Initial Release.” Zenodo, 2020. doi: 10.5281/zenodo.3908560..

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia, pp. 675–678.

Jonathan, Kusuma, G.P., 2022. Retail product classification on distinct distribution of training and evaluation data. Pattern Recogn. Image Anal. 32 (1), 142–152. https://doi.org/10.1134/S105466182104012X.
Jund, P., Abdo, N., Eitel, A., Burgard, W., 2016. The Freiburg Groceries Dataset [Online]. Available: http://arxiv.org/abs/1611.05799.
Kant, S., 2020. Learning Gaussian Maps for Dense Object Detection 1–13 [Online]. Available: http://arxiv.org/abs/2004.11855.
Karlinsky, L., Shtok, J., Tzur, Y., Tzadok, A., 2017. Fine-grained recognition of thousands of object categories with single-example training. Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017 2017-Janua, 965–974. https://doi.org/10.1109/CVPR.2017.109.
Koch, G., 2011. Siamese Neural Networks for One-Shot Image Recognition.
A. Krizhevsky and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," pp. 1–9...
Kulyukin, V., Kutiyanawala, A., 2010. From ShopTalk to ShopMobile: vision-based barcode scanning with mobile phones for independent blind grocery shopping. Proc. 2010 Rehabil. Eng. Assist. Technol. Soc. North Am. Conf. (RESNA 2010), Las Vegas, NV 703, 1–5 [Online]. Available: http://digital.cs.usu.edu/~vkulyukin/vkweb/pubs/RESNA2010_VKulyukin1.pdf.
Leutenegger, S., Chli, M., Siegwart, R.Y., 2011. BRISK: binary Robust invariant scalable keypoints. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2548–2555. https://doi.org/10.1109/ICCV.2011.6126542.
Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P., 2020. Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. 42 (2), 318–327. https://doi.org/10.1109/TPAMI.2018.2858826.
Litman, R., Anschel, O., Tsiper, S., Litman, R., Mazor, S., Manmatha, R., 2020. Scatter: Selective context attentional scene text recognizer. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 11959–11969. https://doi.org/10.1109/CVPR42600.2020.01198.
Liu, W., et al., 2016. SSD: single shot multibox detector. Lect. Notes Comput. Sci. 9905 (LNCS), 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.
Carol N. Lovak and Steven A. Shafer, "Anatomy of Color Histogram."..
D. G. Lowe, "Distinctive image features from scale-invariant keypoints," vol. 60, no. 2, pp. 91–110, 2004..
Lucas, B.D., 1981. An Iterative Image Registration Technique with an Application to Stereo Vision.
Melek, C.G., Sonmez, E.B., 2017. A Survey of Product Recognition in Shelf Images, pp. 13–18.
Melek, C.G., Sonmez, E.B., Ayral, H., Varli, S., 2023. Development of a Hybrid Method for Multi-Stage End-To-End Recognition of Grocery Products in Shelf Images, pp. 1–23.
Merler, M., Galleguillos, C., Belongie, S., 2007. Recognizing groceries in situ using in vitro training data. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. https://doi.org/10.1109/CVPR.2007.383486.
Mikolajczyk, K., Schmid, C., 2004. Scale & affine invariant interest point detectors, 60 (1), 63–86.
Moon, J., Lim, S., Lee, H., Yu, S., 2022. Smart Count System Based on Object Detection Using Deep Learning.
Neubeck, A., 2006. Efficient Non-maximum Suppression.
Osokin, A., Sumin, D., Lomakin, V., 2020. OS2D: one-stage one-shot object detection by matching anchor features. Lect. Notes Comput. Sci. 12360 (LNCS), 635–652. https://doi.org/10.1007/978-3-030-58555-6_38.
Pan, X., et al., 2020. Dynamic refinement network for oriented and densely packed object detection. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 11204–11213. https://doi.org/10.1109/CVPR42600.2020.01122.
Ravi, N., Naqvi, S., El-sharkawy, M., 2022. BIoU : an Improved Bounding Box Regression for Object Detection.
Ray, A., Kumar, N., Shaw, A., Mukherjee, D.P., 2018. U-PC: unsupervised planogram compliance. Lect. Notes Comput. Sci. 11214 (LNCS), 598–613. https://doi.org/10.1007/978-3-030-01249-6_36.
Wang, J., Yin, X., Wang, L., Zhang, L., 2020. Hashing-based non-maximum suppression for crowded object detection. https://doi.org/10.48550/arXiv.2005.11426 arXiv preprint arXiv:2005.11426.
Redmon, J., Farhadi, A., 2017. YOLO9000: better, faster, stronger. Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017 2017-Janua, 6517–6525. https://doi.org/10.1109/CVPR.2017.690.
Redmon, J., Farhadi, A., 2018. YOLOv3: an incremental improvement [Online]. Available: http://arxiv.org/abs/1804.02767.
Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39 (6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031.
Robinson, J., Assignment, S., 1950. The Hungarian method for the assignment problem. Nav. Res. Logist. Q. 2, 14–16 [Online]. Available: https://onlinelibrary.wiley.com/doi/epdf/10.1002/nav.3800020109.
Rong, T., et al., 2021. A SOLUTION TO PRODUCT DETECTION IN DENSELY, pp. 1–6.
Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. ORB: an efficient alternative to SIFT or SURF. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2564–2571. https://doi.org/10.1109/ICCV.2011.6126544.
Sabolcik, R., 2021. Five Smart retail technology trends in store for 2021. Silicon Labs 8.
Santra, B., Mukherjee, D.P., 2019. A comprehensive survey on computer vision based approaches for automatic identification of products in retail store ☪. Image Vis Comput. 86, 45–63. https://doi.org/10.1016/j.imavis.2019.03.005.
Santra, B., Shaw, A.K., Mukherjee, D.P., 2020a. Graph-based Non-maximal Suppression for Detecting Products on the Rack R, vol. 140, pp. 73–80. https://doi.org/10.1016/j.patrec.2020.09.023.
Santra, B., Paul, A., Mukherjee, D.P., 2020b. Deterministic Dropout for Deep Neural Networks Using Composite Random Forest, vol. 131, pp. 205–212. https://doi.org/10.1016/j.patrec.2019.12.023.
Santra, B., Shaw, A.K., Mukherjee, D.P., 2021. An end-to-end annotation-free machine vision system for detection of products on the rack. Mach. Vis. Appl. 32 (3), 1–13. https://doi.org/10.1007/s00138-021-01186-6.
Santra, B., Shaw, A.K., Mukherjee, D.P., 2022. Part-based annotation-free fine-grained classification of images of retail products. Pattern Recogn. 121, 108257 https://doi.org/10.1016/j.patcog.2021.108257.
Selvam, P., Koilraj, J.A.S., 2022. A deep learning framework for grocery product detection and recognition. Food Anal. Methods 15 (12), 3498–3522. https://doi.org/10.1007/s12161-022-02384-2.
Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp. 1–14.
Sinha, A., Banerjee, S., Chattopadhyay, P., 2021. An Improved Deep Learning Approach for Product Recognition on Racks in Retail Stores, pp. 1–13.
Smith, R., 2005. An Overview of the Tesseract OCR Engine.
Srivastava, M.M., 2020. Bag of tricks for retail product image classification. Lect. Notes Comput. Sci. 12131 (LNCS), 71–82. https://doi.org/10.1007/978-3-030-50347-5_8.
Srivastava, M., 2022. Using Contrastive Learning and Pseudolabels to Learn Representations for Retail Product Image Classification, pp. 659–663. https://doi.org/10.5220/0010911000003124.
Stork, D.G., 2016. Pattern Classification.
Strohmayer, J., Kampel, M., 2023. REAL-TIME supermarket product recognition on mobile devices using scalable pipelines TU wien , computer vision lab. In: 2023 IEEE Int. Conf. Image Process, pp. 420–424. https://doi.org/10.1109/ICIP49359.2023.10223137.
Tonioni, A., Di Stefano, L., 2017. Product recognition in store shelves as a sub-graph isomorphism problem. Lect. Notes Comput. Sci. 10484 (LNCS), 682–693. https://doi.org/10.1007/978-3-319-68560-1_61.
Tonioni, A., Serra, E., Di Stefano, L., 2018. A deep learning pipeline for product recognition on store shelves. In: IEEE 3rd Int. Conf. Image Process. Appl. Syst. IPAS 2018, pp. 25–31. https://doi.org/10.1109/IPAS.2018.8708890.
Follmann, P., Bottger, T., Hartinger, P., Konig, R., Ulrich, M., 2018. MVTec D2S: densely segmented supermarket dataset. In: Proceedings of the European conference on computer vision (ECCV), pp. 569–585.
Varadarajan, S., Srivastava, M.M., 2018. Weakly Supervised Object Localization on grocery shelves using simple FCN and Synthetic Dataset (Mil).
Varadarajan, S., Kant, S., Srivastava, M.M., 2020. Benchmark for generic product detection: a low data baseline for dense object detection. Lect. Notes Comput. Sci. 12131 (LNCS), 30–41. https://doi.org/10.1007/978-3-030-50347-5_3.
Varol, G., 2014. Product Placement Detection Based on Image Processing. https://doi.org/10.1109/SIU.2014.6830408.
Varol, G., Kuzu, R.S., 2015. Toward retail product recognition on grocery shelves (March). https://doi.org/10.1117/12.2179127.
Vasanthi, P., Mohan, L., 2024. Biomedical Signal Processing and Control Ensemble of ghost convolution block with nested transformer encoder for dense object recognition. Biomed. Signal Process Control 88 (PB), 105645. https://doi.org/10.1016/j.bspc.2023.105645.
Viola, P., Jones, M., 2001. Rapid Object Detection Using a Boosted Cascade of Simple Features, pp. 1–8.
C. Wang and H. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection"...
Wang, C., Huang, C., Zhu, X., Zhao, L., 2022. One-shot retail product identification based on improved siamese neural networks. Circ. Syst. Signal Process. 41 (11), 6098–6112. https://doi.org/10.1007/s00034-022-02062-y.
Wei, X.S., Cui, Q., Yang, L., Wang, P., Liu, L.. RPC: A large-scale retail product checkout dataset. https://doi.org/10.48550/arXiv.1901.07249.
Wei, Y., Tran, S., Xu, S., Kang, B., Springer, M., 2020. Deep learning for retail product recognition: challenges and techniques. Comput. Intell. Neurosci. 2020 https://doi.org/10.1155/2020/8875910.
Winlock, T., Christiansen, E., Belongie, S., 2010. Toward real-time grocery detection for the visually impaired. In: 2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Work. CVPRW 2010, pp. 49–56. https://doi.org/10.1109/CVPRW.2010.5543576.
Wolbitsch, M., Hasler, T., Goller, M., Gutl, C., Walk, S., Helic, D., 2019. RFID in the wild - analyzing stocktake data to determine detection probabilities of products. 2019 6th Int. Conf. Internet Things Syst. Manag. Secur. IOTSMS 2019, 251–258. https://doi.org/10.1109/IOTSMS48152.2019.8939247.
Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M., 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7464–7475.
Xu, C., Zheng, Y., Zhang, Y., Li, G., Wang, Y., 2022. A Method for Detecting Objects in Dense Scenes, pp. 75–82.
Ye, C., Zhang, H., Xu, X., Cai, W., Qin, J., Choi, K., 2016. Object Detection in Densely Packed Scenes via Semi-supervised Learning with Dual Consistency, pp. 1245–1251.
Yilmazer, R., Birant, D., 2021. Shelf auditing based on image classification using semi-supervised deep learning to increase on-shelf availability in grocery stores. Sensors 21 (2), 1–26. https://doi.org/10.3390/s21020327.
Yörük, E., Öner, K.T., Akgül, C.B., 2016. An efficient Hough transform for multi-instance object recognition and pose estimation. Proc. - Int. Conf. Pattern Recognit. 0, 1352–1357. https://doi.org/10.1109/ICPR.2016.7899825.
Yücel, M.E., Ünsalan, C., 2024. Planogram compliance control via object detection, sequence alignment, and focused iterative search. Multimed. Tool. Appl. 83 (8), 24815–24839.

Zhang, Y., Wang, L., Hartley, R., Li, H., 2007. Where's the weet-bix? Lect. Notes Comput. Sci. 4843 (PART 1), 800–810. https://doi.org/10.1007/978-3-540-76386-4_76.
Zhang, Y., Wang, L., Hartley, R., Li, H., 2009. Handling significant scale difference for object retrieval in a supermarket. In: DICTA 2009 - Digit. Image Comput. Tech. Appl., pp. 468–475. https://doi.org/10.1109/DICTA.2009.79. June 2014.
Zhou, X., Koltun, V., 2018. Probabilistic Two-Stage Detection.
Zhou, X., Wang, D., Krähenbühl, P., 2019. Objects as points [Online]. Available: http://arxiv.org/abs/1904.07850.
Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492–1500.