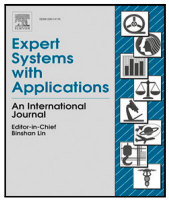
Expert Systems With Applications 255 (2024) 124635

Contents lists available at ScienceDirect

Expert Systems With Applications

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



Shelf Management: A deep learning-based system for shelf visual monitoring

Rocco Pietrini [a,*], Marina Paolanti [b], Adriano Mancini [a], Emanuele Frontoni [b], Primo Zingaretti [a]

[a] VRAI - Vision Robotics and Artificial Intelligence Lab, Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche, via Brecce Bianche 12, Ancona, 60131, Italy

[b] Department of Political Sciences, Communication and International Relations, University of Macerata, Via Don Minzoni 22A, Macerata, 62100, Italy

ARTICLE INFO

Dataset link: https://github.com/rokopi-byte/shelf_management

Keywords:
Shelf management
Retail
Shelf monitoring
SKU recognition
Planogram compliance
Planogram

ABSTRACT

Shelf monitoring plays a key role in optimizing retail shelf layout, enhancing the customer shopping experience and maximizing profit margins. The process of automating shelf audit involves the detection, localization and recognition of objects on store shelves, including diverse products with varying attributes in unconstrained environments. This facilitates the assessment of planogram compliance. Accurate product localization within shelves requires the identification of specific shelf rows. To address the current technological challenges, we introduce "Shelf Management", a deep learning-based system that is carefully tailored to redesign shelf monitoring practices. Our system can navigate the complexities of shelf monitoring by using advanced deep learning techniques and object detection and recognition models. In addition, a complex semantic module enhances the accuracy of detecting and assigning products to their designated shelf rows and locations. In particular, we recognize the lack of finely annotated datasets at the SKU level. As a contribution to the field, we provide annotations for two novel datasets: SHARD (SHelf mAnagement Row Dataset) and SHAPE (SHelf mAnagement Product dataset). These datasets not only provide valuable resources, but also serve as benchmarks for further research in the field of retail. A complete pipeline is designed using a RetinaNet architecture for object detection with 0.752 mAP, followed by a Deep Hough transform to detect shelf rows as semantic lines with an F1 score of 97%, and a product recognition step using a MobileNetV3 architecture trained with triplet loss and used as a feature extractor together with FAISS for fast image retrieval with an accuracy of 93% on top-1 recognition. Localization is achieved using a deterministic approach based on product detection and shelf row detection. Source code and datasets are available at https://github.com/rokopi-byte/shelf_management.

1. Introduction

Product placement within retail space is a strategic marketing practice that holds immense significance for brands. The location of a product on the shelf can have a profound impact on its visibility, accessibility, and ultimately, its sales performance (Mondal, Mittal, Saurabh, Chaudhary, & Reddy, 2023). Retail stores are dynamic environments where consumer behavior and decision-making are heavily influenced by the arrangement and presentation of products (Saqlain, Rubab, Khan, Ali, & Ali, 2022). By understanding the principles and strategies behind product placement, brands can optimize their visibility, attract customer attention, and increase the likelihood of purchase (Ediriginghe & Munson, 2023). Product placement can be considered as horizontal, regarding the shelf location on a retail store map, or vertical, regarding the location in terms of shelf row and the specific position on it.

Horizontal placement takes into consideration the layout of the store and is usually tackled by analyzing shopper trajectories coming either from camera systems or from active tracking systems (Gabellini, D'Aloisio, Fabiani, & Placidi, 2019; Paolanti, Liciotti, Pietrini, Mancini, & Frontoni, 2018; Rossi, Paolanti, Pierdicca, & Frontoni, 2021; Syaekhoni, Lee, & Kwon, 2018), while the vertical one focus on visual aspects, that will be deepened in this work. According to extensive marketing studies and research, the allocation of space within a supermarket has been consistently proven to have a significant and positive impact on several crucial aspects of product sale performance (Bianchi-Aguiar, Hübner, Carravilla, & Oliveira, 2021; Kan, Liu, Lichtenstein, & Janiszewski, 2023). These studies have provided compelling evidence that strategic product placement and allocation of shelf space can greatly enhance product visibility, increase consumer awareness, and stimulate demand. When products are strategically positioned in high-traffic areas, such as

* Corresponding author.
E-mail addresses: r.pietrini@staff.univpm.it (R. Pietrini), marina.paolanti@unimc.it (M. Paolanti), a.mancini@univpm.it (A. Mancini), emanuele.frontoni@unimc.it (E. Frontoni), p.zingaretti@univpm.it (P. Zingaretti).

https://doi.org/10.1016/j.eswa.2024.124635
Received 21 July 2023; Received in revised form 13 June 2024; Accepted 25 June 2024
Available online 29 June 2024

eye-level shelves or end caps, they are more likely to catch the attention of shoppers and draw their interest. Increased visibility plays a vital role in creating brand awareness and capturing the attention of potential customers. Products that are easily noticed and readily accessible are more likely to be considered and ultimately purchased by consumers. Moreover, studies have shown that well-planned product placement can influence consumer behavior and decision-making. When products are strategically placed within a supermarket, they tend to benefit from the mere exposure effect, where repeated exposure to a product increases familiarity and positively influences purchasing decisions (Hanaysha, Al Shaikh, & Alzoubi, 2021; Keh, Wang, & Yan, 2021).

1.1. Challenges

Shelf Space Allocation (SSA) is a complex task in retail management that involves efficiently organizing and distributing available shelf space to various products within a store. To tackle this challenge, a specific tool called *planogram* is commonly employed. A planogram is a visual representation or diagram that shows how products should be displayed on shelves or fixtures in a retail store. It is a tool used by retailers to ensure that merchandise is arranged in a way that maximizes sales and creates an organized and visually appealing shopping experience for customers. Planograms typically include information such as product placement, quantity, and pricing, and may be updated regularly to reflect changes in inventory, promotions, or seasonal offerings. These planograms are usually created by the retailer's headquarters and then distributed to every chain store so that store managers can update product layouts on shelves accordingly (Paolanti et al., 2019). It is essential to verify whether each store accurately adheres to the recommended planograms. This verification process is commonly known as planogram compliance checking (Liu & Tian, 2015). Planogram compliance is essential in retail as it ensures that products are displayed in a consistent and organized manner, which can improve the customer experience and increase sales. Non-compliance with planograms can result in out-of-stock items, misplaced products, and overall confusion for customers, which can negatively impact a retailer's reputation (Frontoni, Marinelli, Rosetti, & Zingaretti, 2017). In the fast-paced world of retail, keeping a vigilant eye on store shelves has evolved beyond mere planogram compliance. For brands and retailers, effectively monitoring shelves involves a holistic approach that encompasses various factors beyond planogram adherence. This comprehensive shelf monitoring strategy goes beyond ensuring products are correctly placed and explores aspects such as pricing, competitor analysis, share of shelf (SOS), and other crucial metrics (Düsterhöft & Hübner, 2023). The shelf monitoring strategy has been carried out manually for decades, brands and retailers employed qualified on-field workers to periodically visit the store and manually scan and measure products. This approach is labor-intensive, time-consuming, and prone to human errors. However, with advancements in technology, many retailers and brands are now actively seeking automated solutions to streamline this process.

Although computer vision has made significant progress in recent years, identifying retail products on a shelf is still considered a challenging task from a computer vision perspective. This is due to several factors, such as the vast number of product categories in a typical supermarket, which can number in the thousands, as noted by Goldman, Herzig, Eisenschtat, Goldberger, and Hassner (2019). New products are introduced everyday, or simply their package is revised, making them appear visually different. Additionally, similar-looking products, such as those from the same brand but with different sizes or flavors, can be difficult to distinguish and must be recognized as distinct products. Recently, by leveraging artificial intelligence (AI) and its subsets deep learning and machine learning algorithms, retailers aim to automate and optimize this crucial task (Wei, et al., 2020).

1.2. Nature and scope

We extend our previous preliminary works in this context (Pietrini, Galdelli, Mancini, & Zingaretti, 2023; Pietrini et al., 2022) by developing Shelf Management, an innovative expert system designed for automated visual shelf monitoring using either fixed or mobile cameras. These two different use cases serve different purposes. The mobile camera approach serves as a valuable tool for retailers and brand representatives during regular store visits, while the fixed camera setup enables real-time shelf monitoring, facilitating various analyses such as instant out-of-stock detection. The system proposes a complete pipeline to analyze a shelf image and identify each product and its position on the shelf for further analysis at a later stage. First, the system performs shelf row detection, using advanced deep learning techniques to identify the different rows present on store shelves. This step is crucial for subsequent analysis, as it provides a basis for further processing. Secondly, the system focuses on product detection, using state-of-the-art object detection algorithms to locate and outline individual products within each row of shelves. This step is critical in isolating and extracting the necessary information for subsequent analysis. The system then moves on to product identification, using powerful deep learning models to recognize and classify the detected products. By comparing the identified products to a reference gallery, the system can accurately determine their specific attributes and characteristics. Overall, this novel expert system offers a comprehensive solution for automating the shelf audit process. Its ability to perform shelf row detection, product detection, and product identification streamlines the process, improves efficiency, and provides valuable insights for retail businesses. In addition to its advanced functionalities, Shelf Management is designed to be user-friendly and flexible. The system can be accessed through a user-friendly mobile application, as shown in Fig. 1, making it convenient for retail employees and managers to monitor planogram compliance on the go. The mobile app provides real-time updates, allowing users to view the detected products, compliance assessment results, and detailed feedback directly on their mobile devices. Moreover, Shelf Management can rely on fixed cameras installed within the retail environment. These fixed cameras capture high-resolution images of the store shelves, providing a continuous stream of data for planogram compliance evaluation and out-of-stock detection. The use of fixed cameras ensures consistent and reliable monitoring of shelf organization, without the need for manual intervention or reliance on mobile devices. By combining the mobile app and fixed camera integration, Shelf Management offers a comprehensive solution for planogram compliance monitoring, catering to the diverse needs of retail businesses.

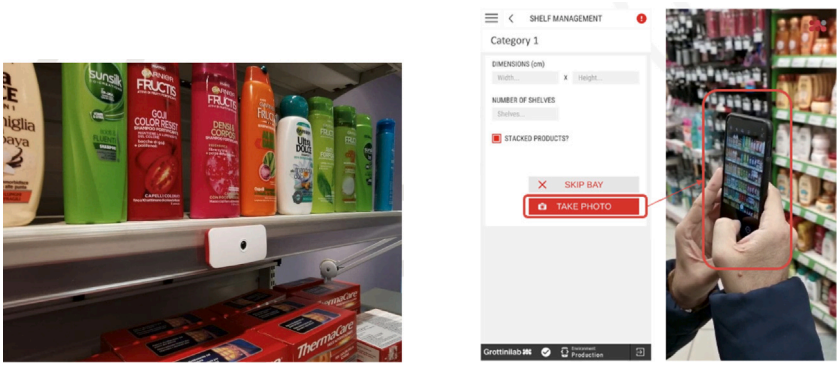
1.3. Contributions

The main contributions of this paper, in comparison to state-of-the-art approaches, are as follows:

(i) *Automation and Efficiency*: The proposed Shelf Management system offers a fully automated solution for the shelf audit. Unlike traditional manual methods or existing approaches, which often require extensive human intervention, the system adopts deep learning techniques to streamline the process. This automation significantly reduces the time and effort required for the shelf audit, enabling retailers and brands to conduct more frequent and comprehensive assessments.

(ii) *Comprehensive Analysis*: The system combines multiple stages, including shelf row detection, product detection, and product identification, into a unified framework. By considering the entire shelf layout and identifying individual products, the system provides a comprehensive assessment that goes beyond basic compliance checks.

(iii) *Accurate Product Identification*: With the integration of powerful deep learning models, Shelf Management achieves robust and accurate product identification. This is crucial for recognition-related tasks such as planogram compliance evaluation and share of shelf calculation, as



(a) Fixed Camera (b) Mobile Camera

Fig. 1. Shelf Management use cases.

it enables precise matching of detected products with their expected attributes and characteristics. By leveraging advanced recognition techniques, the system reduces false positives and enhances the accuracy of compliance assessments.

(iv) *Real-time Monitoring*: The Shelf Management system supports the use of both fixed and mobile cameras, enabling real-time monitoring of planogram compliance and out-of-stocks. This real-time capability allows retailers to proactively address compliance issues, promptly make necessary adjustments, and maintain planogram integrity. It provides timely feedback and ensures that product placements align with desired planogram specifications at all times. Overall, the main contributions of this paper lie in the automation, comprehensive analysis, accurate product identification, and real-time monitoring capabilities of the Shelf Management system.

(v) We introduce two specific datasets tailored for the domain of the shelf audit process. In fact, to address the lack of fine-grained, Stock Keeping Unit (SKU)-level annotated datasets in this field, two novel datasets are collected and released: SHelf mAnagement Row Dataset (SHARD) and SHelf Product datasEt (SHAPE). SHAPE is a dataset manually collected and annotated, providing a diverse collection of images encompassing various products commonly found on retail shelves. These datasets serve as a valuable resource for training and evaluating deep learning models for accurate product identification and shelf row detection. SHARD focuses on the detection and localization of shelf rows within retail environments. It contains annotated images capturing different shelf layouts, including various shelf row designs, positions, and display hooks. The dataset enables the development and evaluation of specialized algorithms for precise shelf row detection, a critical step in the shelf audit. By releasing these datasets to the research community, this paper contributes to the advancement of shelf audit methodologies. SHARD and SHAPE serve as benchmark resources, facilitating further research, algorithm development, and comparison of results in the domain of retail shelf management.

The use of computational techniques, in particular deep learning, in the automation of retail shelf audits is in line with the objective of applying advanced technologies to practical scenarios. Moreover, the integration of these techniques into a coherent system is in itself a significant contribution. The complexity comes from the need to deal with diverse products and unconstrained environments. The Shelf Management System has practical implications for retailers and brands. It enables them to make data-driven decisions, optimize shelf layouts, improve product availability and enhance the overall customer experience.

1.4. Paper outline

The paper is organized as follows: Section 2 provides a comprehensive review of the current literature in the field of visual shelf monitoring. It begins by defining the scope of visual shelf monitoring, followed by a discussion of the different approaches and techniques that have been explored in previous studies. The section also highlights the strengths and limitations of existing methods, providing a critical analysis of the current state of research. Section 3 presents our proposed method for visual shelf monitoring. The section begins with a conceptual overview of the approach, detailing the theoretical foundations and motivations behind its design. We then describe the datasets collected and used in this research, including their sources, the data collection process and the features of the data. This part is crucial as it not only presents the method, but also justifies the choices made during the research process. In Section 4, the experimental results obtained from applying the proposed method are discussed and analyzed. Section 5 summarizes the key findings of the study and highlights its contributions to the field of visual shelf monitoring. It discusses the practical implications of the research and its potential benefits. Finally, the section suggests directions for future research, identifying gaps in the current literature that could be addressed in subsequent studies, and suggesting ways to extend and improve upon the work presented in this paper.

2. Related works

In this section, we provide an overview of the existing literature and research related to shelf monitoring. We explore various approaches and techniques that have been proposed to address the challenges associated with traditional shelf monitoring practices.

2.1. Computer vision-based approaches for shelf monitoring

Traditional methods for shelf monitoring primarily rely on manual inspections conducted by human auditors. These methods are labor-intensive, time-consuming, and prone to human errors. While they have been widely used in the past, their limitations have led to the exploration of automated solutions. Computer vision techniques have been extensively studied for shelf monitoring tasks. These approaches typically involve the detection and recognition of products using image processing algorithms, but also some approaches for shelf row detection. Feature extraction, template matching, and machine learning-based classifiers are commonly employed to identify products on store shelves. Earlier studies tackled the problem using classical computer vision algorithms both for product detection and recognition, such as SVM classifier for detection and SIFT keypoints matching for recognition (Pietrini et al., 2019; Vaira et al., 2019). Ray, Kumar, Shaw, and Mukherjee (2018) presented an end-to-end solution for recognizing merchandise displayed in the shelves of a supermarket. Given images of individual products, which are taken under ideal illumination for product marketing, the challenge is to find these products automatically in the images of the shelves.

Shelf row detection have been investigated by Jubair and Banik (2013) proposing a method to identify books from images of library shelves, including a step for bookshelf row detection. This strategy employs the canny edge detector. This same methodology has been used in a retail context by Geng, Wang, Weng, Huang, and Zhu (2019). Alternatively, some, such as Saran, Hassan, and Maurya (2015), Sun, Hanata, Sato, Tsuchitani, and Akashi (2019), Varol, Kuzu, and Akgiil (2014), Yilmazer and Birant (2021), applied a blend of the Sobel filter and Hough transform, or Agnihotram et al. (2017), who combined color information and line detection. A significant work specifically in the field of shelf row detection was done by Fan and Zhang (2014), which tackled the problem with a multi-step approach. First, they identified the vanishing point and the corresponding shelf line segments, then divided the image into equal-angle wedges centered at the vanishing point, and projected these line segments into the wedges. Finally, the shelves were identified by analyzing these projections.

Conventional computer vision methods seem ill-suited to retail shelves, given that products can have various form factors and be stacked, which calls for further investigation.

2.2. Deep learning-based approaches for shelf monitoring

Deep learning has emerged as a powerful tool for shelf monitoring and planogram compliance. Convolutional Neural Networks (CNN) have shown remarkable performance in object detection and recognition tasks. Researchers have proposed deep learning architectures tailored for shelf monitoring, which utilize CNNs for accurate product detection and attribute recognition. These models leverage large-scale annotated datasets and achieve superior performance compared to traditional computer vision approaches (O'Mahony et al., 2020), in contrast to simpler neural networks that struggle to extract such intricate features (Wei, et al., 2020).

To determine the number of products present on store shelves, Higa and Iwamoto (2018) used surveillance cameras to capture videos of the shelves. The study employed background subtraction to track changed regions, removed moving objects, and employed a CNN based on CaffeNet for the classification of these regions. This approach achieved a success rate of 89.6%. Building upon this work, Higa and Iwamoto (2019) expanded the methodology by monitoring product availability using images from surveillance cameras. The Hungarian method was used to distinguish the foreground from successive images, and two deep networks, CIFAR-10 and CaffeNet, were employed for the classification of detected changed regions. Additionally, this methodology facilitated the identification of commonly accessed shelves. In cases where limited training data is available, another paper proposed a fast detection and recognition method based on fine-grained categories of products, achieving a mean Average Precision (mAP) of 52.16% for each product category (Karlinsky, Shtok, Tzur, & Tzadok, 2017). Sun, Zhang, and Akashi (2020) proposed a template-free, zero-shot product detection system that avoids using templates and instead detects products by segmenting shelves horizontally into layers and vertically into individual products. Horizontal layer classification was performed using GoogLeNet, while vertical division was achieved using another trained GoogLeNet. This approach demonstrated improved performance compared to existing methods, although it was negatively influenced by empty regions between products, making it less robust. Yilmazer and Birant (2021) proposed a semi-supervised deep learning-based image classification approach. The study combined the concepts of "semi-supervised" and "on-shelf availability (SOSA)". By leveraging both labeled and unlabeled data, semi-supervised learning was applied. The deep learning architecture YOLOv4 (Bochkovskiy, Wang, & Liao, 2020) was employed. Numerous researchers have contributed to the application of CNNs for product detection in retail settings. For instance, Jund, Abdo, Eitel, and Burgard (2016) employed CNNs to tackle the challenge of in-store product recognition, achieving an accuracy of 78.9%. Goldman and Goldberger (2020) addressed the task of large-scale fine-grained structure classification by leveraging contextual information in combination with deep networks. The research by Crăciunescu, Baicu, Mocanu, and Dobre (2021) presented a method for calculating shelf occupancy using a fully convolutional neural network. This network discerned the shelves and the background information from RGB-D images, thereby requiring a depth sensor.

In the combined sphere of detection and recognition, De Feyter and Goedemé (2023) proposed several training methods to explore the disparity between training with fully-annotated data and task-specific data. Their findings suggest that training on tightly cropped product images prevents the recognition part of the joint architecture from learning to handle context information available in feature maps during inference. This limitation restricts the model's performance, preventing it from achieving results comparable to those obtained from a model trained on fully-annotated data. However, this conclusion may not hold water in real-world scenarios where the range of products is vast and continually changing. Retraining a network in such dynamic environments is not a practical option, casting doubt on the viability of their findings in real-world applications. A fundamental work is the one by Goldman et al. (2019), where the authors proposed a new method for object detection in densely packed scenes, specifically targeting SKUs on retail shelves. The novelty was represented by the use, on top of a retinanet, of a Soft-IoU layer for estimating the overlap between predicted and (unknown) ground truth boxes. Then an EM-based (Expectation Maximization) unit was used for resolving bounding box overlap ambiguities. Chen et al. (2022) proposed a large-scale benchmark of basic visual tasks on products that challenge algorithms for detecting, reading, and matching in retail.

While both the approaches presented by Goldman et al. (2019), and Chen et al. (2022) make significant contributions to the field of object detection and recognition in densely packed scenes, such as retail environments, they only address parts of the overall problem. The method by Goldman et al. (2019) is instrumental in accurately identifying individual items and dealing with bounding box overlaps, it does not appear to go beyond that initial step of detection or pure classification. Likewise, Chen et al. (2022) benchmark provides an essential tool for assessing the performance of algorithms on fundamental visual tasks, but it also does not extend into aspects like product localization within a specific environment. What is even more crucial to note is that the retail landscape is highly dynamic, with new products emerging daily. Simultaneously, a single product may be available on the shelf in various packaging formats, such as for promotional campaigns, event collaborations, or rebranding initiatives. Consequently, employing a conventional classification approach with a predefined number of categories, or frequent model retraining, becomes impractical. Therefore, alternative approaches for recognition must be explored. Therefore, what seems to be missing from the current literature is a comprehensive pipeline that not only effectively detects objects in densely packed scenes but also recognizes the specific product and precisely localizes it on a specific shelf row in a retail context. Such a pipeline would need to integrate advanced object detection, recognition, and localization strategies into a cohesive system capable of dealing with the complexities of a real-world retail environment.

Lee, Kim, Lee, and Kim (2017) proposed the concept of a semantic line that separates different semantic regions in a scene. They also introduced a new method to detect these lines using a CNN with multi-task learning, treating line detection as a hybrid of classification and regression tasks. This method has been effectively employed for horizon estimation, composition enhancement, and image simplification. Although these techniques have adapted existing object detectors for line detection, the distinctive features of lines are not fully taken into account, resulting in sub-optimal performance. Lines, possessing simpler geometric properties than complex objects, can be represented more succinctly with a few parameters. Jin, Lee, and Kim (2020) built a detection network with mirror attention (D-Net) and comparative ranking and matching networks (RNet and M-Net) for semantic line

detection. In contrast, Zhao, Han, Zhang, Xu, and Cheng (2022) integrated the powerful learning capability of CNNs with the classic Hough transform, creating what they dubbed the 'Deep Hough Transform'.

We propose that shelf rows can be seen as a specific instance of semantic lines, serving to separate different semantic areas of a shelf. Accordingly, we suggest utilizing the Deep Hough Transform, fine-tuned on a freshly gathered dataset, to detect these lines (shelf rows). Such a comprehensive approach would be immensely beneficial in creating automated systems for retail management, enabling precise inventory tracking, automated restocking, and more efficient store layout planning. Current research, while foundational and instrumental, has yet to fully tackle this multidimensional problem. The task of pinpointing a product's location on store shelves involves taking the shelf's physical structure into account, such as its rows.

2.3. Datasets for shelf monitoring

Recognition algorithms for retail products have been investigated by researchers for several decades, even prior to the prevalence of deep learning in computer vision tasks. Various retail product datasets have been proposed to facilitate such research, including SOIL-47 by Koubaroulis, Matas, Kittler, and CMP (2002), Grozi-120 by Merler, Galleguillos, and Belongie (2007), and the Supermarket Produce (SP) dataset by Rocha, Hauagge, Wainer, and Goldenstein (2010). However, these early datasets typically have few images and products. For example, the RPC (Wei, Cui, Yang, Wang, & Liu, 2019) dataset is a large-scale dataset proposed for automatic checkout that contains 200 categories and 83,739 images. However, these images were captured under controlled lighting and a clean background, which may not accurately reflect real-world scenarios of product recognition on shelves. The TGFS dataset (Hao, Fu, & Jiang, 2019), on the other hand, uses images collected from self-service vending machines, making the image distribution closer to that of the checkout system in a natural environment. Nonetheless, it only contains 30 K images from 24 fine-grained and 3 coarse classes, and the resolution of each image is limited. In the review paper of Santra and Mukherjee (2019) many other retail related dataset are analyzed such as the Grocery Product Dataset (GPD) (George & Floerkemeier, 2014), Grocery Dataset (GD) (Jund et al., 2016) and Freiburg Groceries Dataset (FGD) (Varol & Kuzu, 2015) but all of them have less than 20,000 images and may not be suitable for today's data-demanding deep learning model. The SKU-110K dataset by Goldman et al. (2019) is currently the largest retail image dataset in terms of number of images, containing over 1M images from 11,762 store shelves. However, the dataset only provides bounding boxes of each object in the scene, without further annotating the category of the bounding boxes, which makes it unsuitable for object recognition purposes. The MVTec D2S dataset by Follmann, Bottger, Hartinger, Konig, and Ulrich (2018) is an instance-aware semantic segmentation dataset for retail products, providing 21,000 images of 60 object categories with pixel-wise labels. Although it may serve as an additional grocery-relevant component to other semantic segmentation datasets, the dataset was also captured in a laboratory environment with controlled camera settings and may not be ideal for object recognition in a real store. The RP2K dataset by Peng, Xiao, and Li (2020) was the first large scale product dataset annotated at the SKU level. It contains two components: the original shelf images and the individual object images cropped from the shelf images. The shelf images are labeled with the shelf type, store ID, and a list of bounding boxes of objects of interest. For each image cropped from its bounding box, rich annotations are provided including the SKU ID, product name, brand, product type, shape, size, flavor/scent and the bounding box reference to its corresponding shelf image. Meta-category labels are also provided for each object image in two different ways. One is categorized by its product type, which reflects the placement of the products, i.e., products with the same type usually placed on the same or nearby shelf. Included are 7 meta categories by product types: dairies, liquors, beers, cosmetics, non-alcoholic drinks, tobacco and seasonings. Another categorization method is by its product shape, with 7 shapes: bottle, can, box, bag, jar, handled bottle and pack, which covers all possible shapes that appeared in the dataset. Dataset comprises 10,385 high-resolution shelf images in total, with, on average, 37.1 objects in each image. The dataset contains in total 384,311 images of individual objects. Each individual object image represents a product from the 2388 SKUs.

Table 1
Datasets for Shelf monitoring.

Dataset	Categories	SKUs	Images
SOIL-47	NA	47	1974
Grozi-120	NA	120	11 870
SP	15	NA	2633
RPC	NA	200	83 739
TGFS	3	24	38 000
GPD	27	3235	3235
GD	10	NA	13 000
FGD	25	NA	4947
SKU-110K	NA	NA	1M
MVTec D2S	60	NA	21 000
RP2K	7	2388	384 311
Product-6K	NA	6348	12 917
Product-10K	NA	10K	190K
AliProducts	NA	50K	2.5M
Unitail-OCR	NA	1454	1454

Georgiadis et al. (2021) proposed Products-6K large-scale product dataset with nearly 6000 different SKUs and images captured both in real and studio setup and Bai, Chen, Yu, Wang, and Zhang (2020) Products-10K, which contains 10000 fine-grained SKU-level products frequently bought by online customers. AliProduct dataset proposed by Cheng et al. (2020) is crawled from web sources by searching 50K product names, consequently containing 2.5 million noisy images without human annotations. In Chen et al. (2022), the authors recently created Unitail-OCR dataset to sustain retail product recognition through product matching via robust reading. In the gallery of products to be recognized, there are 1454 fine-grained products with frontal photos. Among these products, there are 10,709 labeled text regions located and 7565 legible word transcriptions. Although the numbers of some datasets are relevant, they are far from representative of the huge number of categories present in a supermarket. Dataset are summarized in Table 1 along with the number of product categories regardless how the categorization was made, the number of SKUs and the total number of images. Regarding the shelf row detection problem, to the best of our knowledge, there are not any available datasets.

By addressing these gaps in the literature, this paper intends to provide a novel and comprehensive solution for efficient and accurate shelf monitoring in the retail industry.

3. Method

In this section, we introduce the Shelf Management system as well as the datasets used for evaluation. The framework is depicted in Fig. 2. In particular, the proposed method comprises several key steps that enable efficient and accurate evaluation: *Shelf Row Detection*, *Product Detection*, *Product Recognition*, *Product Localization*.

- *Shelf row detection*: the system uses advanced deep learning techniques to identify the different rows present on store shelves. This step is crucial as it provides a foundation for further analyses.
- *Product detection*: leveraging state-of-the-art object detection algorithms, the system locates and outlines individual products within each shelf row. This step isolates and extracts the necessary information for subsequent analysis.

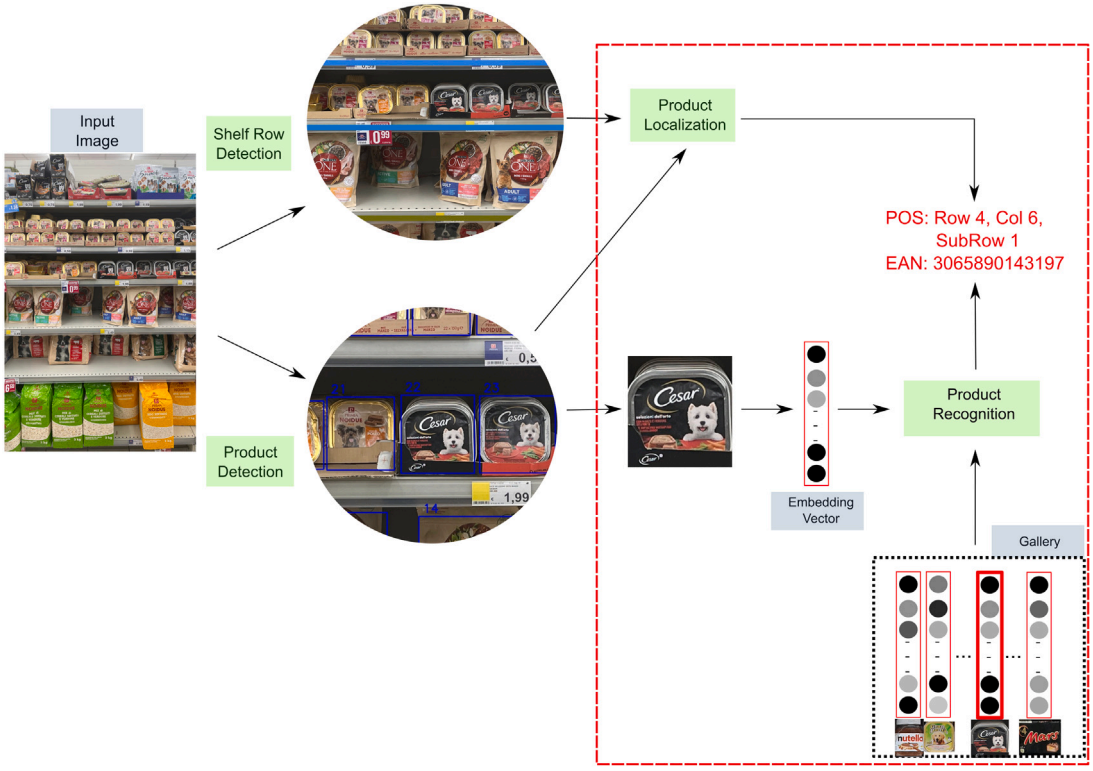


Fig. 2. Shelf Management system and its key components. The system encompasses several essential steps, including Shelf Row Detection, Product Detection, Product Recognition, and Product Localization. Primary steps are highlighted in green, essential data components are depicted in gray. The red segment signifies individual product processing, with this process occurring concurrently for each identified product.

- *Product recognition*: powerful deep learning models are employed to recognize and classify the detected products. By comparing the identified products against a reference gallery, the system accurately determines their specific attributes and characteristics.
- *Product localization*: assignment of position in terms of shelf row, column (horizontal relative position within the shelf row) and subrow (vertical relative position within the column and the row), all based on bounding box coordinate with respect to the shelf rows coordinate.

Steps are highlighted in algorithm 1 and detailed in this section. The input image, whether from a fixed camera or a mobile application, is independently processed by two modules: one for the detection of shelf rows and another for product detection (lines 1-2). The first module returns a list of detected shelf rows, assigning each row an incremental ID (starting from 0 for the bottom shelf row) and providing its vertical (y) coordinate. The second module returns a list of detected products in the entire image, assigning each product a bounding box (x1, x2, y1, y2) and an incremental ID (starting from 0 for the bottom-left product). Subsequently, the products are evaluated one by one in a loop (line 4). First, a check is performed (lines 5-10) to determine if there is a consistent overlap with any of the previously detected shelf rows. In the event of such an overlap, the product is removed from the list as it is considered a false positive (e.g., a price tag or promotional material). Next, the image is cropped based on the product's bounding box to isolate the specific product (line 11), and an embedding vector is computed for this product (line 12). This embedding vector, along with the gallery vectors, is used in the recognition step (line 13), which returns an ordered list of the top-N (where N is configurable) closest matches. We select the top-1 match, and if it exceeds a certain threshold, the product's information is stored in a list (lines 15-20). The 'assignLocation' procedure computes a row, column and subrow by comparing the product's centroid with the detected shelf rows and with other products in the same row, respectively (see 3.3 for details). Finally, the list will only contain the detected product, along with their location and identification. This innovative expert system provides a comprehensive solution for automating shelf audits. Its ability to perform row detection, product detection, product identification and localization streamlines the process, improves efficiency and provides valuable insights for retailers. With Shelf Management, retail businesses can ensure accurate and efficient shelf monitoring, leading to optimized shelf organization and enhanced customer experiences. The details of the system are given in the following Subsections. Shelf Management is comprehensively evaluated on "SHARD (SHelf mAnagement Row Dataset)" and "SHAPE (SHelf mAnagement Product datasEt)", two publicly available datasets specifically collected and manually labeled for this work (Section 3.5).

3.1. Shelf row detection

The distribution of products plays a fundamental role in the shelf audit process. Knowing exactly where a product is located in terms of shelves rows enables several shelf analysis, such as a planogram compliance check. In addition some important Key Performance Indicators (KPI) take into consideration the physical structure of the shelf, for example the SOS. This refers to the percentage or proportion of physical shelf space within a retail store that a particular product or brand occupies. A higher SOS indicates greater visibility for the product, which can lead to increased sales and market share. The only reliable way to calculate SOS for a brand is to detect its products and associate them to a particular shelf row. Detecting shelf rows is hence fundamental in this task to overcome limitations that may arise only using product detection, such as stacked items with the same SKU, which is often the case (only the lower item in the stack should be considered for the planogram analysis). Stacked items can also show not contiguous bounding boxes due to the camera perspective, making the correct association product-shelf row nearly impossible. In addition,

Algorithm 1: Pseudo-code for the proposed pipeline

```
shelf_rows ← detectShelfRows(input_image);
bounding_boxes ← DetectProducts(input_image);
products ← None;
foreach b_box in bounding_boxes do
    foreach shelf_row in shelf_rows do
        if b_box.area ∩ shelf_row.area > shelf_threshold then
            delete b_box;
            continue
        end
    end
    cropped ← image[b_box];
    embeddings ← extractEmbeddings(cropped);
    EANs_list, reliability_list ←
     search(embeddings,gallery);
    if reliability_list[0] > recognize_threshold then
        product.b_box = b_box;
        product.EAN = EANs_list[0];
        product.reliability = reliability_list[0];
        product.shelf_row, product.column =
          assignLocation(b_box);
        products.append(product);
    end
end
```

the identification of shelf rows facilitates object segmentation and viewpoint correction, which are crucial for object recognition. Considering that a shelf row serves as a separation between distinct semantic areas, we approached the detection of shelf rows as a specific case of semantic lines (Zhao et al., 2022). First a pixel-wise representation with a CNN-based encoder is extracted (ResNet50-FPN), and then the deep Hough transform (DHT) converts representations from feature space into parametric space. The global line detection problem is then converted in detecting peak response in the transformed features, making the problem simpler. Finally, a reverse Hough transform (RHT) converts the detected lines back to image space.

3.2. Product detection

The approach proposed by Goldman et al. (2019) was used for the detection task. A trained model on the SKU-110K dataset, with a RetinaNet as the backbone, has been integrated into the pipeline. The model's base is RetinaNet, a one-stage object detector, where focal loss is employed to allocate lower loss values to "easy" negative samples, directing the model's attention and resources towards more challenging samples. This approach enhances prediction accuracy. RetinaNet utilizes ResNet50-FPN as its backbone for feature extraction and incorporates two specialized subnetworks for classification and bounding box regression, collectively forming the RetinaNet architecture. This architecture attains state-of-the-art performance, surpassing Faster R-CNN, a renowned two-stage object detection method. However, given the challenges posed by the retail environment, authors proposed two additional module on top of the RetinaNet, a Soft-IoU layer estimates the Jaccard index between detected boxes and ground truth boxes, while an EM-Merger unit transforms detections and Soft-IoU scores into a Mixture of Gaussians (MoG) and resolves overlapping detections in dense scenes. Occasionally, this network produces false positive detections corresponding to price tags or promotional material. Both can be heterogeneous between different stores and need to be eliminated. The shelf rows detected in the previous step were also used for this purpose. Any bounding box that overlaps a shelf row by a threshold is discarded. While this phase of the pipeline involves the use of a reference model without significant modification, an experimental phase was undertaken to assess the optimal parameters using the SHARD dataset, as detailed in the results section.

3.3. Product localization

In a computer vision system designed to automatically analyze a planogram from a picture of a store shelf, the localization component is a crucial step that follows product detection and shelf row detection. In many retail environments, certain types of products are often stacked vertically within the same shelf row. For instance, canned or packaged goods might be placed in stacks. The localization process must account for these stacking arrangements to avoid misidentifying product quantities and placements. Shelf analysis like SOS usually takes into consideration only the first layer of stacked products. This localization process aims to precisely place each detected product within its corresponding shelf row and also within a specific column (position) within the shelf row. Additionally, the system defines a subrow for stacked products, which is the row relative to each column. Each product has a subrow value of 1 unless stacked, in which case the subrow value increments going up. Columns increase from left to right and are relative to each row, as each row can have a varying number of columns. The first step involves assigning a row to each detected product box, considering the detected shelf rows. After filtering out bounding boxes that overlap with shelf rows (such as price tags or promotional tags), the Y-coordinate of the center of each bounding box is compared with the Y-coordinates of the shelf rows to determine the appropriate shelf row for each product. This process is illustrated in Procedure AssignRow. In the second step, for each shelf row, the bounding boxes assigned to it are sorted according to their center X-coordinates from left to right. Each product is then assigned the column based on its relative position in the row. This process is illustrated in Procedure AssignColumn. In the final step, each bounding box is initially assigned a subrow value of 1. For each shelf row, each bounding box is compared with all the others in the same shelf row. If the center X-coordinate of a bounding box is less than the X2-coordinate (right edge) of another bounding box and the center Y-coordinate is less than the Y1-coordinate (upper edge) of the same bounding box, the subrow is incremented. This basically means a product is stacked on another one. This process is illustrated in Procedure AssignSubRow. In Fig. 3 a sample shelf is depicted to illustrate the logic.

Procedure AssignRow(b_box)

```
row ← 1;
foreach shelf_row in shelf_rows do
    if b_box.center_y > shelf_row then
        row ← shelf_row;
        break;
    end
end
```

Procedure AssignColumn(b_box)

```
foreach shelf_row in shelf_rows do
    sorted_list ← sort(bounding_boxes where
      b_box.row = shelf_row);
    i = 0;
    foreach b_box in sorted_list do
        b_box.column ← i;
        i ← i + 1;
    end
end
```

3.4. Product recognition

Every retail product across the globe is uniquely denoted by a specific code, such as the European Article Number (EAN) in Europe, the Universal Product Code (UPC) in the UK, Japan, and Australia among others. Given the huge number of products in a supermarket and the high temporal variability, it was necessary to use an approach

Procedure AssignSubRow(b_box)

```
foreach shelf_row in shelf_rows do
    sorted_list ← sort(bounding_boxes where
      b_box.row = shelf_row);
    foreach b_box in sorted_list do
        b_box.subrow = 1;
        foreach b_box_2 in sorted_list do
            if b_box_2.x2 > b_box.center_x and
              b_box_2.y1 > b_box.center_y then
                b_box.subrow ← b_box.subrow + 1;
            end
        end
    end
end
```

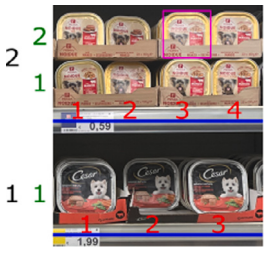


Fig. 3. Localization schema for a sample shelf with 2 shelf rows (in blue). In black rows number, in red columns number and in green subrows number. In this example the sample product bounded in pink is assigned row = 2, column = 3, subrow = 2.

with an independent number of classes. Recognizing a product from its image can be seen as an image retrieval problem. First, we need a gallery of known products, where each product is identified by a unique code and has one or more images. Then, a query image of an unknown product can be compared with all images in the gallery to find the most similar one according to a defined metric. Image retrieval can be viewed as a vector similarity problem in a high-dimensional image feature space. Classification networks, such as CNNs trained for image classification tasks, can be used as feature extractors. This is because, during training, these networks learn to represent the data in a way that facilitates the classification task by mapping the input images to vectors of real numbers, called embedding vectors. These vectors aim to capture the semantic or syntactic meaning of the objects, with similar objects having similar vectors. We built a gallery on the SHAPE dataset, generating an embedding vector for each image. In this way, searching for a product simply means generating the embedding vector for the query image and finding the closest embedding vector in the gallery. The approach used to generate the embedding vectors was inspired by the work of Schroff, Kalenichenko, and Philbin (2015), moving the application domain, from face recognition to the identification of retail products. Part of the SHAPE dataset, has been held out as a test set, with a particular strategy: for SKUs that featured multiple images in the gallery, one image was chosen to be part of the test set. This selection ensured that the system would encounter these particular images for the first time during testing, having never been exposed to them during training, but also guaranteed to have valid query images that the model should recognize. In this way, we can distinguish whether a low similarity score comes from an unknown product or poor recognition. This test dataset comprises 813 images, evenly distributed across 62 product categories. The remaining images were allocated for model training, with 80% of them used for training and the remaining 20% for validation purposes. We then used different state-of-the-art classification CNNs, pre-trained for classification task on ImageNet (Deng et al., 2009) from the Keras Applications repository. From these models we removed the last classification layer, adding in place a dropout layer and a final dense layer of size 256. Therefore, the output of the models resulted in an vector of size 256 on which the L2 normalization has also been applied. L2-normalizing the vectors moves the value in the range 0-1 which is more convenient for matching purposes. The loss used for the training is the Triplet Hard Loss (Schroff et al., 2015) with a margin of 1.0 (soft margin). Triplet loss is a loss function particularly effective for learning useful representations of data by distance comparison. The idea is to take three images (anchor, positive, negative) — a "triplet" — where two of the images are more similar to each other than they are to the third. The goal of the triplet loss is to make sure that the anchor and the positive image (which are of the same EAN in this case) are closer together in the learned feature space than the anchor and the negative image (different EAN). The triplet loss function tries to achieve this by minimizing the distance between the anchor and the positive and maximizing the distance between the anchor and the negative. The term "hard" in "triplet hard loss" refers to the strategy of selecting the triplets (Hermans, Beyer, & Leibe, 2017). Hard triplets are those where the negative is closer to the anchor than the positive in the feature space. These are the most informative triplets to train on because they are the ones that the model is currently getting most wrong. By focusing on hard triplets, the model learns more effectively to distinguish between similar-looking but different classes, leading to better performance. This strategy is known as hard negative mining. We decided to mine online triplets using the approach of Hermans et al. (2017). In fact, they show how the use of triplets mined online can greatly increase the accuracy of the model and reduce training times. After training the model a test phase have been conducted in order to assess the quality of the learned feature in the recognition process. The test set have been used for this, image-by-image we extracted the embedding vector and compared against the gallery. The cosine similarity was the selected criteria for similarity evaluation between embedding vectors of the query images and the gallery. The cosine similarity between two generic numerical vectors A and B is represented by the following equation and a score between -1 e $+1$, where values close to $+1$ represent a greater similarity ($+1$ corresponds to same orientation):

$$cosine\ similarity = \frac{A \cdot B}{\|A\| \, \|B\|} \tag{1}$$

Cosine similarity quantifies the degree of similarity between two vectors in a space with an inner product. It does this by calculating the cosine of the angle between these two vectors, thereby indicating if the vectors are approximately oriented in the same direction. The element in the gallery with the highest cosine similarity with the query element is selected as the matching candidate.

3.5. Shelf management datasets

As already stated, this paper introduces two specific datasets that have been meticulously collected and manually labeled for the purpose of the work. In order to address the scarcity of fine-grained, SKU-level annotated datasets in the field of planogram compliance checking, we have created two novel datasets: SHARD and SHAPE. The motivation behind introducing these datasets is rooted in the need for high-quality resources tailored specifically for planogram compliance checking. In the field of planogram analysis, there is a lack of datasets with fine-grained annotations at the SKU-level, which hinders the development and evaluation of accurate and reliable algorithms. Existing datasets often lack the detailed product attribute annotations and precise shelf rows localization necessary for comprehensive planogram compliance evaluation. To bridge this gap, we have carefully curated and labeled SHARD and SHAPE datasets and the details are given in the following Sections 3.5.1 and 3.5.2.

Fig. 4. Sample shelf picture from the SHARD dataset belonging to the pet food category.

3.5.1. SHARD dataset

SHARD (SHelf mAnagement Row Dataset) which contains ca. 22K images was collected using smartphones and resized, with a maximum height of 1000 pixels and a resolution of 96 dpi, to reduce storage requirements while still providing sufficient detail for the detection of shelf rows. These images were gathered over a span of two months across 2000 different supermarkets in Italy, representing various retail chains. As a result, the dataset encompasses a wide range of visual characteristics, including shelf row materials, colors, price tags, shapes, sizes of displayed goods, the number of rows, and the presence of hooks, among others. Each image displays a single shelf, ensuring that all categories are represented. These images are in JPG format to further optimize storage, and the annotation data is provided in a CSV file, which includes the indication of the vertical coordinate of each shelf row. The annotation file is structured as follow:

$$image\_name, p_1, p_2, \ldots, p_n \tag{2}$$

where p_i is the Y-coordinate shelf row expressed as the percentage of the picture's height in pixels, starting from the top, and n, where $n >= 1$, is the number of shelf rows present in the image. An example of an image of the dataset is shown in Fig. 4, while its relative annotation regarding key information is shown below:

$$00a03bd5 - a6fa - 471d - b4e3 - 6d5cf8aece27.jpg;$$
$$0.13, 0.23, 0.31, 0.44, 0.59, 0.77, 0.98$$

3.5.2. SHAPE dataset

SHAPE (SHelf mAnagement Product datasEt) contains 50K images of 17K different SKU belonging to 62 different categories, fine-grained labeled with their European Article Number (EAN). Images have been collected in 2000 different supermarkets in Italy, using smartphones of different models with a minimum resolution of 8Mpx. Each SKU have a variable number of images, where each image is cropped exactly on the package boundaries. In addition the dataset is structured according to the GS1[1] hierarchical classification, this system is called Global Product Classification (GPC). GPC classifies products by grouping them into categories based on their essential properties as well as their relationships to other products. This specific classification is structured into five tiers, with the fifth tier offering the highest level of detail. To provide a snapshot of these divisions, the first tier could include categories like groceries and beverages, while the second could comprise of subcategories such as bakery items, cereals, and water. Further, the third tier could involve more specific items like wafers and sparkling water, and the fourth could delve into even finer classifications like vanilla flavor or quantities of 0-50 centilitres. Lastly, the fifth and final tier could feature granular distinctions like portioned items and plastic packaging. We chose to operate up to the second tier as we found this level of detail to be sufficiently granular. An example of an image of the dataset is shown in Fig. 5 with 6 product images of different categories. The dataset is structured as follows: first-level folders represent categories as previously defined combining the first and the second tier of the GPC (see GS1 hierarchical classification), while second-level folders represents SKUs. Both categories and EANs are anonymized using progressive numbers (1, 2, 3) but could be released under a commercial agreement.

4. Results and discussions

In this section the experimental part is discussed.

4.1. Product detection results

The model described in 3.2, pre-trained on the SKU-110K dataset was tested on 1000 randomly chosen images from the SHARD dataset manually annotated for product detection. Experiments in this step were devoted to assess the best parameters for the model, such as find optimal value for the score threshold and the hard score rate used to filter detections, produced by the model in terms of bounding boxes, soft and hard scores. The two scores can be seen as distinct yet complementary approaches for assessing the quality of localization: the hard score assesses the extent to which the patch within the bounding box resembles an object, while the soft score gauges the degree of overlap between the bounding box and the underlying object. Soft and hard scores are averaged in a single confidence score depending on the hard score rate that we decided to set equal to 0.5 to give them the same importance. The score threshold represent a minimum threshold that the detection confidence score must satisfy, we set it to 0.35 after an experimental phase devoted to maximize the Average Precision metrics. This model has been used without significant modification since its initial development for the specific task of detection in densely populated retail environments. We use COCO (Lin et al., 2014) evaluation metrics, including the average precision (AP) at IoU = .50:.05:.95, AP at IoU = .5 (denoted as AP^{50}), and AP at IoU = .75 (AP^{75}). To assess the practical feasibility of our system, we also investigate counting metrics. Let $\{P'_i\}_{i=1}^n$ represent the predicted number of products in each test image, where $i \in [1, n]$, and $\{t_i\}_{i=1}^n$ denote the actual number of products per image. The Mean Absolute Error (MAE) is calculated as $\frac{1}{n}\sum_{i=1}^n |P'_i - t_i|$, while the Root Mean Squared Error (RMSE) is computed as $\sqrt{\frac{1}{n}\sum_{i=1}^n (P'_i - t_i)^2}$. Full experiments are reported in Table 2 where different backbones have been compared for the RetinaNet. While all ResNet backbones gives similar results, the best result is achieved by ResNet-152, with an average precision of 0.772, all exceeding the performance achieved by the authors in their original paper on the SKU-110K dataset. This improvement can be attributed to the characteristics of the images in SHARD, which are relatively simpler as they represent single cropped shelves. However, this closely reflects

[1] https://www.gs1.org/standards/gpc

(a) EAN: 8000500179864

(b) EAN: 8000580203329

(c) EAN: 5000159384186

(d) EAN: 8000965232609

(e) EAN: 615357118874

(f) EAN: 8009470041461

Fig. 5. Sample images from the SHAPE dataset. 6 product images of different categories are displayed within their EAN.

Fig. 6. Qualitative results for product detection.

Table 2
Product detection. In bold the selected configuration for the system.

Backbone	AP	AP^{50}	AP^{75}	MAE	RMSE
ResNet-50	**0.752**	**0.834**	**0.787**	**6.587**	**11.745**
ResNet-101	0.762	0.842	0.797	5.487	10.702
ResNet-152	0.772	0.851	0.802	5.354	9.170

real-world scenarios where shelf analysis is typically performed on a per-shelf basis. The decision to use ResNet-50 as the backbone was influenced by computational requirements and the feasibility of the approach without relying on a GPU. ResNet-50, the smallest variant tested, has 25 million parameters, compared to the 44 million and 60 million parameters of the larger versions. Qualitative results are depicted in Fig. 6.

4.2. Shelf row detection results

The system's shelf row detection algorithm successfully identifies different rows present on store shelves, providing a solid foundation for subsequent analyses. Model was pre-trained on the NLK dataset (Zhao et al., 2022) and then fine-tuned on the Shelf Row Dataset (SHARD) (detailed in Section 3.5.1) for 30 epochs (with early stopping), using a learning rate of 0.0002 with Adam optimizer, with a batch size of 16. A common split ratio of 80:20 has been used for training and validation. A separate set of 5000 images never seen by the model have been used in testing to assess the performance of such model. In Zhao et al. (2022), the authors proposed a new metric, named EA-score, that considers both Euclidean and Angular distance between a pair of lines. Let l_i, l_j be a pair of lines to be measured, the angular distance S is defined as:

$$S_\theta = 1 - \frac{\theta(l_i, l_j)}{\pi/2} \tag{3}$$

where $\theta(l_i, l_j)$ is the angle between l_i and l_j. The Euclidean distance is defined as:

$$S_d = 1 - D(l_i, l_j) \tag{4}$$

where $D(l_i, l_j)$ is the Euclidean distance between midpoints of l_i and l_j. Note that the image is normalized into a unit square before calculating the Euclidean distance. Finally, the EA-score, is calculated as follows:

$$S = (S_\theta \times S_d)^2 \tag{5}$$

Fig. 7. Qualitative results for shelf row detection.

The quality of line detection is evaluated by measuring Precision, Recall, and F-measure. To achieve this, the predicted lines in set P and the ground-truth lines in set G are matched using bipartite matching. The goal is to find a matching that ensures each ground-truth line g_i corresponds to at most one detected line p_j and vice versa. The maximum matching of a bipartite graph can be solved efficiently using the classical Hungarian method, which has a polynomial time complexity. Predicted lines that are matched with ground-truth lines are classified as true positives, while detected lines that are not matched with any ground-truth line are considered false positives. Ground-truth lines without a corresponding predicted line are false negatives. To further assess the detection performance also Precision (Prec) (Eq. (6)), Recall (Rec) (Eq. (7)) and F-measure (Eq. (8)) were computed as secondary metrics,

$$Prec = \frac{TP}{TP + FP} \tag{6}$$

$$Rec = \frac{TP}{TP + FN} \tag{7}$$

$$F = \frac{2 \times Prec \times Rec}{Prec + Rec} \tag{8}$$

where TP, FP, FN are the correct detections, the wrong detections and the missed detection, respectively. A series of thresholds $\tau = 0.01, 0.02, \ldots, 0.99$ to prediction and ground-truth pairs, therefore a series of Precision, Recall, and F-measure scores are computed. Finally, the performance is evaluated in terms of average precision, recall, and F-measure. Different experiments using the SHARD dataset have been conducted to choose the best Hough quantization level. The results can be seen in Table 3 and show how feasible the approach is in terms of both accuracy and inference time. After analyzing the table, it is clear that there is little discernible difference between the different quantization levels. Consequently, we have chosen to set the value to 100 for our system in order to minimize the inference time. In Fig. 7, qualitative results of the approach are presented, wherein the model correctly identifies all the shelf rows, even in the presence of a challenging situation. In this challenging scenario, products are stacked or enclosed in a cardboard divider that closely resembles a line.

4.3. Product localization results

Product localization is achieved through the application of geometrical if-then rules described in Section 3.3. These rules are applied to the results obtained from product detection and shelf row detection, ensuring precise placement of each product within its designated shelf row, column and subrow. Any errors in localization are propagated from the preceding steps of product detection and shelf row detection, as the localization process itself is deterministic and relies on the accuracy of these earlier stages. Consequently, no experiments are conducted for this part, as the process strictly relies on the accuracy of these earlier stages.

Table 3
Quantitative performance comparison of shelf row detection at different quantization levels in parameter space as in Zhao et al. (2022). In bold the selected configuration for the system.

Quantization levels	Precision	Recall	F-1	Inference time
100	**0.9753**	**0.9671**	**0.9663**	**0.0235**
120	0.9704	0.9712	0.9691	0.0250
130	0.9730	0.9827	0.9744	0.0279
150	0.9529	0.9814	0.9721	0.0329

Table 4
Comparison between CNNs used as backbone in terms of accuracy. In bold the selected configuration for the system.

Network	Top-1	Top-5	Top-10
MobileNetV3-Large (Howard et al., 2019)	**0.93**	**0.96**	**0.97**
MobileNetV3-Small (Howard et al., 2019)	0.86	0.95	0.97
EfficientNetV2-B0 (Tan & Le, 2021)	0.91	0.96	0.97
Inception-V3 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016)	0.91	0.95	0.96

4.4. Product recognition results

The SHAPE dataset was utilized to evaluate the recognition performance, with the strategy highlighted in Section 3.4. The first experiment aimed to assess the recognition accuracy and it is showed in Table 4. We compared 4 lightweight (MobileNetV3 was used in its 2 variants) state-of-art CNNs in terms of recognition accuracy. As we can see from the table MobileNetV3-Large (Howard et al., 2019) achieved the best performance in all the metrics, although all the other networks are close, demonstrating the feasibility of our approach for product recognition. To gain a better understanding of the model's performance Top-5 and Top-10 accuracy have been also investigated, because even if in this specific domain the exact recognition is fundamental (SKU-level recognition), still there can be ambiguities that could be solved later by a fine-grained approach, as will be described in Section 5. The deep learning models employed for product recognition exhibit high accuracy in recognizing and classifying the detected products.

Inference time for the feature extraction step has been evaluated to assess the computational requirement and the feasibility of the approach without relying on a GPU. MobileNetV3-Small outperformed the other networks even though all of them are really close, the approach is

Table 5
Comparison between the networks used as backbone in terms of inference time both on GPU (Nvidia K80) and CPU (Intel i7-5500U) in s. In bold the selected configuration for the system.

Network	GPU	CPU
MobileNetV3-Large (Howard et al., 2019)	**0.002**	**0.026**
MobileNetV3-Small (Howard et al., 2019)	0.001	0.012
EfficientNetV2-B0 (Tan & Le, 2021)	0.002	0.046
Inception-V3 (Szegedy et al., 2016)	0.002	0.095

Table 6
Comparison between search strategies for image retrieval.

Search strategy	Index	nlist	nprobe	Time (ms)	Accuracy
Brute Force	–	–	–	181	0.93
FAISS (Johnson et al., 2019)	FlatIP	-	-	**0.41**	**0.93**
FAISS (Johnson et al., 2019)	IVFFLAT	4096	4	0.09	0.81
FAISS (Johnson et al., 2019)	IVFFLAT	16 384	32	0.2	0.9

indeed feasible on a CPU with an average inference time of 0.01 s per image. The complete benchmark is highlighted in Table 5. The ideal trade-off, and consequently the network selected as feature extractor is MobileNetV3-Large.

The search strategy was also studied, since comparing each product image on a shelf with tens of thousands of images in the gallery can take a considerable amount of time, as the gallery grows over time. The main requirement in this step is to maintain recognition accuracy, which is fundamental in this specific domain. The basic approach is brute force, which simply compares a query image with all the items in the gallery one by one, maintaining accuracy as it is an exhaustive search. A smarter strategy is to further speed up the similarity search by using an off-the-shelf Approximate Nearest neighbor (ANN) indexing library. These ANN indexing schemes convert the traditional $\mathcal{O}(n \log 2n)$ sort algorithm into a $\mathcal{O}(\log 2n)$ serial time algorithm by efficiently parallelizing it using multi-threading, BLAS and machine SIMD vectorization. We chose to use FAISS (Johnson, Douze, & Jégou, 2019) an open-source library developed by Facebook's AI Research (FAIR) team, mainly because it provides an option to perform batch queries on the index, therefore latency is much lower. It is designed for efficient similarity search and clustering of large-scale datasets. FAISS supports different indexing structures, including the popular index structures like IVF (Inverted File), which enable fast retrieval of nearest neighbors. FlatIP index implement an exhaustive search strategy and we can see from Table 6 that is 400x faster than the basic brute force without sacrify the accuracy. We then tested a cell-probe method called IVFFLAT with different configurations. The feature space is partitioned into *nlist* cells. The gallery vectors are assigned to one of these cells thanks to a quantization function (K-means is used with the assignment to the centroid closest to the query), and stored in an inverted file structure formed of *nlist* inverted lists. At query time, a set of *nprobe* inverted lists is selected. The query is compared with each of the gallery vectors assigned to these lists. This search strategy can speed up the search by a factor of 2 while losing only 1% of the accuracy. Since the main objective was to maintain accuracy in a domain where exact recognition is required, we chose the FlatIP index of FAISS to be part of the pipeline.

5. Conclusions and future works

Visual shelf monitoring plays a vital role in the success of retailers and brands. With the retail industry becoming increasingly competitive, ensuring product availability, correct placement and adherence to planogram specifications is critical to maximizing sales and improving the customer experience. Visual shelf monitoring allows retailers and brands to accurately assess the presentation and organization of products on store shelves, ensuring that they are visually appealing and easily accessible to customers. By using automated systems such as Shelf Management, retailers can streamline the monitoring and evaluation process, reduce human error and gain detailed insight into planogram compliance. This in turn enables them to make data-driven decisions on inventory management, product placement and overall store layout, ultimately leading to improved customer satisfaction and increased profitability. This paper presents Shelf Management, an innovative expert system designed to automate shelf audits using fixed or mobile cameras. The system demonstrated its effectiveness in accurately detecting, recognizing and evaluating products on store shelves. Using advanced deep learning techniques, Shelf Management successfully performed key steps including shelf row detection, product detection, product identification and localization. The results obtained from applying Shelf Management to the two datasets, SHelf mAnagement Row Dataset (SHARD) and Shelf MProduct Dataset (SHAPE), demonstrated its ability to streamline the evaluation process, improve efficiency and provide valuable insights for retailers. The unique challenges of the retail sector, including the need for high accuracy in product detection and the diverse range of products and shelf layouts, have guided our decisions throughout the development process. Our choice of pipeline components, from advanced object detection algorithms to advanced product recognition models, was driven by their proven effectiveness in handling the complexities of the retail environment. The demands of the retail environment for efficiency, accuracy and scalability required a careful approach to selecting the most appropriate components for our pipeline. These decisions were supported by rigorous pre-testing to ensure that each component integrated into our system was optimized for performance in a retail environment. Our decisions were influenced not only by the empirical results, but also by the strategic importance of aligning with the evolving needs and challenges of the retail domain. The paper also highlighted the strengths and limitations of Shelf Management, emphasizing its ability to automate shelf checking, reduce human error and deliver consistent results. However, challenges such as occlusions, variations in product shape (non-rigid packaging) and complex shelf layouts can impact system performance in certain scenarios. In addition, as the approach is not end-to-end, errors tend to propagate along the pipeline. These limitations provide opportunities for future research and improvement. Our future work will focus on a comprehensive evaluation of the product recognition capabilities by annotating and testing a larger set of images from the SHARD dataset to confirm the robustness of the model in different retail scenarios. In addition, we plan to improve the system's dynamic adaptability and real-time analysis capabilities, test its scalability in different retail environments, and develop more accurate evaluation metrics to capture the complexity of its performance. This comprehensive approach will significantly advance the theoretical and practical contributions of our work in retail shelf management, and strengthen the system's relevance and effectiveness in the ever-evolving retail sector.

CRediT authorship contribution statement

Rocco Pietrini: Data curation, Methodology, Software, Resources, Validation. **Marina Paolanti:** Writing – review & editing, Visualization. **Adriano Mancini:** Writing – review & editing, Visualization,

Supervision. **Emanuele Frontoni:** Conceptualization, Writing – review & editing, Visualization, Supervision. **Primo Zingaretti:** Supervision, Project administration.

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

Data and codebase are available at https://github.com/rokopi-byte/shelf_management.

Acknowledgment

This work is funded by Egocentric and exocentric views for an object-level human behavior analysis and understanding through tracking in complex spaces (EXTRA EYE) project, Piano Nazionale di Ripresa e Resilienza Missione 4 - Componente 2 – Investimento 1.1 "Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)", Codice CUP D53D23008900001. The authors would like to express their grateful to Grottini lab http://www.grottinilab.com.

References

Agnihotram, G., Vepakomma, N., Trivedi, S., Laha, S., Isaacs, N., Khatravath, S., et al. (2017). Combination of advanced robotics and computer vision for shelf analytics in a retail store. In *2017 international conference on information technology* (pp. 119–124). http://dx.doi.org/10.1109/ICIT.2017.13.

Bai, Y., Chen, Y., Yu, W., Wang, L., & Zhang, W. (2020). Products-10k: A large-scale product recognition dataset. arXiv preprint arXiv:2008.10545.

Bianchi-Aguiar, T., Hübner, A., Carravilla, M. A., & Oliveira, J. F. (2021). Retail shelf space planning problems: A comprehensive review and classification framework. *European Journal of Operational Research*, *289*(1), 1–16.

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv:2004.10934.

Chen, F., Zhang, H., Li, Z., Dou, J., Mo, S., Chen, H., et al. (2022). Unitail: Detecting, reading, and matching in retail scene. In *Computer vision–ECCV 2022: 17th European conference, tel aviv, Israel, October 23–27, 2022, proceedings, part VII* (pp. 705–722). Springer.

Cheng, L., Zhou, X., Zhao, L., Li, D., Shang, H., Zheng, Y., et al. (2020). Weakly supervised learning with side information for noisy labeled images. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XXX 16* (pp. 306–321). Springer.

Crăciunescu, M., Baicu, D., Mocanu, S., & Dobre, C. (2021). Determining on-shelf availability based on RGB and ToF depth cameras. In *2021 23rd International conference on control systems and computer science* (pp. 243–248). http://dx.doi.org/10.1109/CSCS52396.2021.00047.

De Feyter, F., & Goedemé, T. (2023). Joint training of product detection and recognition using task-specific datasets. In *Proceedings of the 18th international joint conference on computer vision, imaging and computer graphics theory and applications-volume 5: VISAPP* (pp. 715–722). SciTePress.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Ieee.

Düsterhöft, T., & Hübner, A. (2023). Problems and opportunities of applied optimization models in retail space planning. In *Retail space analytics* (pp. 161–181). Springer.

Edirisinghe, G. S., & Munson, C. L. (2023). Strategic rearrangement of retail shelf space allocations: Using data insights to encourage impulse buying. *Expert Systems with Applications*, *216*, Article 119442.

Fan, J., & Zhang, T. (2014). Shelf detection via vanishing point and radial projection. In *2014 IEEE international conference on image processing* (pp. 1575–1578). http://dx.doi.org/10.1109/ICIP.2014.7025315.

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* (pp. 569–585).

Frontoni, E., Marinelli, F., Rosetti, R., & Zingaretti, P. (2017). Shelf space re-allocation for out of stock reduction. *Computers & Industrial Engineering*, *106*, 32–40.

Gabellini, P., D'Aloisio, M., Fabiani, M., & Placidi, V. (2019). A large scale trajectory dataset for shopper behaviour understanding. In *New trends in image analysis and processing–ICIAP 2019: iCIAP international workshops, bioFor, patReCH, e-BADLE, deepRetail, and industrial session, Trento, Italy, September 9–10, 2019, revised selected papers 20* (pp. 285–295). Springer.

Geng, Z., Wang, Z., Weng, T., Huang, Y., & Zhu, Y. (2019). Shelf product detection based on deep neural network. In *2019 12th international congress on image and signal processing, bioMedical engineering and informatics* (pp. 1–6). http://dx.doi.org/10.1109/CISP-BMEI48845.2019.8965694.

George, M., & Floerkemeier, C. (2014). Recognizing products: A per-exemplar multi-label image classification approach. In *Computer vision–ECCV 2014: 13th European conference, Zurich, Switzerland, September 6-12, 2014, proceedings, part II 13* (pp. 440–455). Springer.

Georgiadis, K., Kordopatis-Zilos, G., Kalaganis, F., Migkotzidis, P., Chatzilari, E., Panakidou, V., et al. (2021). Products-6K: a large-scale groceries product recognition dataset. In *The 14th PErvasive technologies related to assistive environments conference* (pp. 1–7).

Goldman, E., & Goldberger, J. (2020). CRF with deep class embedding for large scale classification. *Computer Vision and Image Understanding*, *191*, Article 102865.

Goldman, E., Herzig, R., Eisenschtat, A., Goldberger, J., & Hassner, T. (2019). Precise detection in densely packed scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5227–5236).

Hanaysha, J. R., Al Shaikh, M. E., & Alzoubi, H. M. (2021). Importance of marketing mix elements in determining consumer purchase decision in the retail market. *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET)*, *12*(6), 56–72.

Hao, Y., Fu, Y., & Jiang, Y.-G. (2019). Take goods from shelves: A dataset for class-incremental object detection. In *Proceedings of the 2019 on international conference on multimedia retrieval* (pp. 271–278).

Hermans, A., Beyer, L., & Leibe, B. (2017). In defense of the triplet loss for person re-identification. arXiv:arXiv:1703.07737.

Higa, K., & Iwamoto, K. (2018). Robust estimation of product amount on store shelves from a surveillance camera for improving on-shelf availability. In *2018 IEEE international conference on imaging systems and techniques* (pp. 1–6). IEEE.

Higa, K., & Iwamoto, K. (2019). Robust shelf monitoring using supervised learning for improving on-shelf availability in retail stores. *Sensors*, *19*(12), 2722.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., et al. (2019). Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1314–1324).

Jin, D., Lee, J.-T., & Kim, C.-S. (2020). Semantic line detection using mirror attention and comparative ranking and matching. In *Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part XX 16* (pp. 119–135). Springer.

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535–547.

Jubair, M. I., & Banik, P. (2013). A technique to detect books from library bookshelf image. In *2013 IEEE 9th international conference on computational cybernetics* (pp. 359–363). http://dx.doi.org/10.1109/ICCCyb.2013.6617619.

Jund, P., Abdo, N., Eitel, A., & Burgard, W. (2016). The freiburg groceries dataset. arXiv preprint arXiv:1611.05799.

Kan, C., Liu, Y., Lichtenstein, D. R., & Janiszewski, C. (2023). EXPRESS: The negative and positive consequences of placing products next to promoted products. *Journal of Marketing*, Article 00222429231172111.

Karlinsky, L., Shtok, J., Tzur, Y., & Tzadok, A. (2017). Fine-grained recognition of thousands of object categories with single-example training. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4113–4122).

Keh, H. T., Wang, D., & Yan, L. (2021). Gimmicky or effective? The effects of imaginative displays on customers' purchase behavior. *Journal of Marketing*, *85*(5), 109–127.

Koubaroulis, D., Matas, J., Kittler, J., & CMP, C. (2002). Evaluating colour-based object recognition algorithms using the soil-47 database. In *Asian conference on computer vision, vol. 2* (pp. 840–845).

Lee, J.-T., Kim, H.-U., Lee, C., & Kim, C.-S. (2017). Semantic line detection and its applications. In *2017 IEEE international conference on computer vision* (pp. 3249–3257). http://dx.doi.org/10.1109/ICCV.2017.350.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13* (pp. 740–755). Springer.

Liu, S., & Tian, H. (2015). Planogram compliance checking using recurring patterns. In *2015 IEEE international symposium on multimedia* (pp. 27–32). http://dx.doi.org/10.1109/ISM.2015.72.

Merler, M., Galleguillos, C., & Belongie, S. (2007). Recognizing groceries in situ using in vitro training data. In *2007 IEEE conference on computer vision and pattern recognition* (pp. 1–8). IEEE.

Mondal, A., Mittal, R., Saurabh, S., Chaudhary, P., & Reddy, P. K. (2023). An inventory-aware and revenue-based itemset placement framework for retail stores. *Expert Systems with Applications*, [ISSN: 0957-4174] *216*, Article 119404. http://dx.doi.org/10.1016/j.eswa.2022.119404, URL https://www.sciencedirect.com/science/article/pii/S095741742202423X.

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., et al. (2020). Deep learning vs. traditional computer vision. In *Advances in computer vision: proceedings of the 2019 computer vision conference (CVC), volume 1 1* (pp. 128–144). Springer.

Paolanti, M., Liciotti, D., Pietrini, R., Mancini, A., & Frontoni, E. (2018). Modelling and forecasting customer navigation in intelligent retail environments. *Journal of Intelligent and Robotic Systems*, *91*, 165–180.

Paolanti, M., Romeo, L., Martini, M., Mancini, A., Frontoni, E., & Zingaretti, P. (2019). Robotic retail surveying by deep learning visual and textual data. *Robotics and Autonomous Systems*, *118*, 179–188.

Peng, J., Xiao, C., & Li, Y. (2020). RP2K: A large-scale retail product dataset for fine-grained image classification. arXiv preprint arXiv:2006.12634.

Pietrini, R., Galdelli, A., Mancini, A., & Zingaretti, P. (2023). Embedded vision system for real-time shelves rows detection for planogram compliance check. In *International design engineering technical conferences and computers and information in engineering conference, vol. 87356*. American Society of Mechanical Engineers, Article V007T07A003.

Pietrini, R., Manco, D., Paolanti, M., Placidi, V., Frontoni, E., & Zingaretti, P. (2019). An IOT edge-fog-cloud architecture for vision based planogram integrity. In *Proceedings of the 13th international conference on distributed smart cameras* (pp. 1–5).

Pietrini, R., Rossi, L., Mancini, A., Zingaretti, P., Frontoni, E., & Paolanti, M. (2022). A deep learning-based system for product recognition in intelligent retail environment. In *Image analysis and processing–ICIAP 2022: 21st international conference, lecce, Italy, May 23–27, 2022, proceedings, part II* (pp. 371–382). Springer International Publishing Cham.

Ray, A., Kumar, N., Shaw, A., & Mukherjee, D. P. (2018). U-pc: Unsupervised planogram compliance. In *Proceedings of the European conference on computer vision* (pp. 586–600).

Rocha, A., Hauagge, D. C., Wainer, J., & Goldenstein, S. (2010). Automatic fruit and vegetable classification from images. *Computers and Electronics in Agriculture*, *70*(1), 96–104.

Rossi, L., Paolanti, M., Pierdicca, R., & Frontoni, E. (2021). Human trajectory prediction and generation using LSTM models and GANs. *Pattern Recognition*, *120*, Article 108136.

Santra, B., & Mukherjee, D. P. (2019). A comprehensive survey on computer vision based approaches for automatic identification of products in retail store. *Image and Vision Computing*, *86*, 45–63.

Saqlain, M., Rubab, S., Khan, M. M., Ali, N., & Ali, S. (2022). Hybrid approach for shelf monitoring and planogram compliance (hyb-smpc) in retails using deep learning and computer vision. *Mathematical Problems in Engineering*, *2022*, 1–18.

Saran, A., Hassan, E., & Maurya, A. K. (2015). Robust visual analysis for planogram compliance problem. In *2015 14th IAPR international conference on machine vision applications* (pp. 576–579). http://dx.doi.org/10.1109/MVA.2015.7153257.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE conference on computer vision and pattern recognition* (pp. 815–823). http://dx.doi.org/10.1109/CVPR.2015.7298682.

Sun, H., Hanata, K., Sato, H., Tsuchitani, I., & Akashi, T. (2019). Segmentation based non-learning product detection for product recognition on store shelves. In *2019 nicograph international (nicoInt)* (pp. 9–16). http://dx.doi.org/10.1109/NICOInt.2019.00009.

Sun, H., Zhang, J., & Akashi, T. (2020). TemplateFree: product detection on retail store shelves. *IEEJ Transactions on Electrical and Electronic Engineering*, *15*(2), 242–251.

Syaekhoni, M. A., Lee, C., & Kwon, Y. S. (2018). Analyzing customer behavior from shopping path data using operation edit distance. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *48*, 1912–1932.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

Tan, M., & Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International conference on machine learning* (pp. 10096–10106). PMLR.

Vaira, R., Pietrini, R., Pierdicca, R., Zingaretti, P., Mancini, A., & Frontoni, E. (2019). An IOT edge-fog-cloud architecture for vision based pallet integrity. In *New trends in image analysis and processing–ICIAP 2019: iCIAP international workshops, bioFor, patReCH, e-BADLE, deepRetail, and industrial session, trento, Italy, September 9–10, 2019, revised selected papers 20* (pp. 296–306). Springer International Publishing.

Varol, G., & Kuzu, R. S. (2015). Toward retail product recognition on grocery shelves. In *Sixth international conference on graphic and image processing, vol. 9443* (pp. 46–52). SPIE.

Varol, G., Kuzu, R. S., & Akgiil, Y. S. (2014). Product placement detection based on image processing. In *2014 22nd Signal processing and communications applications conference* (pp. 1031–1034). http://dx.doi.org/10.1109/SIU.2014.6830408.

Wei, X.-S., Cui, Q., Yang, L., Wang, P., & Liu, L. (2019). RPC: A large-scale retail product checkout dataset. arXiv preprint arXiv:1901.07249.

Wei, Y., Tran, S., Xu, S., Kang, B., Springer, M., et al. (2020). Deep learning for retail product recognition: Challenges and techniques. *Computational Intelligence and Neuroscience*, *2020*.

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), 327.

Zhao, K., Han, Q., Zhang, C.-B., Xu, J., & Cheng, M.-M. (2022). Deep hough transform for semantic line detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 4793–4806. http://dx.doi.org/10.1109/TPAMI.2021.3077129.