Vision Systems for Retail Shelf Management

Rajnish Kumar
Computer Science and Engineering,
Lovely Professional University
Email:-rajnishk71249@gmail.com

Sagar Kumar
Computer Science and Engineering,
Lovely Professional University
.Email:-sagarre5661@gmail.com

Rameshwar Mishra
Computer Science and Engineering,
Lovely Professional University
Email:-rameshwarmishra411@gmail.com

Jobanpreet Singh
Assistant Professor,
Lovely Professional University
Email:-jobanpreet.32357@lpu.in

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 managements Row Dataset) and SHAPE (Shelf managements 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

I. 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 (Edirisinghe & 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 et al., 2019, Paolanti et al., 2018, Rossi et al., 2021, Syaekhoni et al., 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 et al., 2021, Kan et al., 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 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

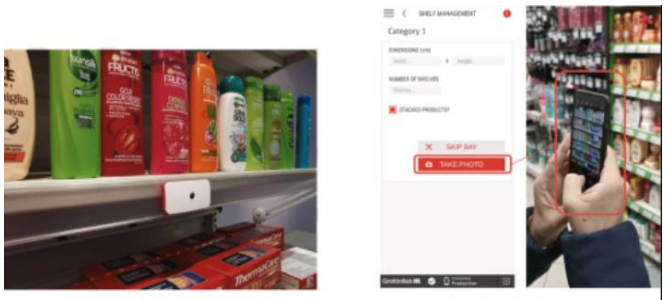
product increases familiarity and positively influences purchasing decisions (Hanaysha et al., 2021, Keh et al., 2021)..

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. 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 (Frontenis). 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 (Dusterhoft). 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 labour-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.

1.2. Nature and scope

We extend our previous preliminary works in this context (Pietrini et al., 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. Collaborative Filtering Techniques:



(a) Fixed Camera (b) Mobile Camera

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 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.. 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 managements 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.

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.

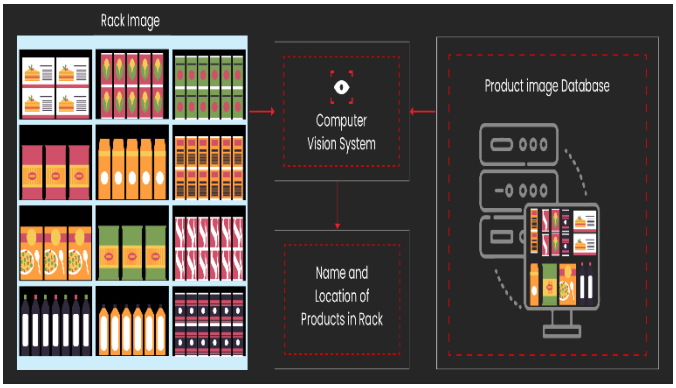
II. Related Works

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.Data Collection Module: This module gathers data on user interactions, including song plays, skips, ratings, and playlist additions. It also collects demographic information (e.g., age, location) and song attributes (e.g., genre, artist) to enhance the dataset. The data is periodically updated to track users' evolving preferences. Interactions are organized into a user-song interaction matrix, forming the foundation for collaborative filtering algorithms, particularly in the context of K-Means clustering.

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.

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 regionsThis approach demonstrated improved performance compared to existing methods,. 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.

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.

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) 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, 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.

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.

Table 1. Datasets for Shelf monitoring.

Dataset	Categories	SKUs	Images
SOIL-47	NA	47	1974
Grozi-120	NA	120	11870
SP	15	NA	2633
RPC	NA	200	83739
TGFS	3	24	38000
GPD	27	3235	3235
GD	10	NA	13000
FGD	25	NA	4947
SKU-110K	NA	NA	1M
MVTec D2S	60	NA	21000
RP2K	7	2388	384311
Product-6K	NA	6348	12917
Product-10K	NA	10K	190K
liProducts	NA	50K	2.5M
Unitail-OCR	NA	1454	1454

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.

•*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). 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).

Algorithm 1: Pseudo-code for the proposed pipeline

```
1  shelf_rows ← detectShelfRows(input_image);
2  bounding_boxes ← DetectProducts(input_image);
3  products ← None;
4  foreach b_box in bounding_boxes do
5      foreach shelf_row in shelf_rows do
6          if b_box.area ∩ shelf_row.area > shelf_threshold then
7              delete b_box;
8              continue
9          end
10     end
11     cropped ← image[b_box];
12     embeddings ← extractEmbeddings(cropped);
13     EANs_list, reliability_list ← search(embeddings, gallery);
14     if reliability_list[0] > recognize_threshold then
15         product.b_box = b_box;
16         product.EAN = EANs_list[0];
17         product.reliability = reliability_list[0];
18         product.shelf_row, product.column = assignLocation(b_box);
19         products.append(product);
20     end
21 end
```

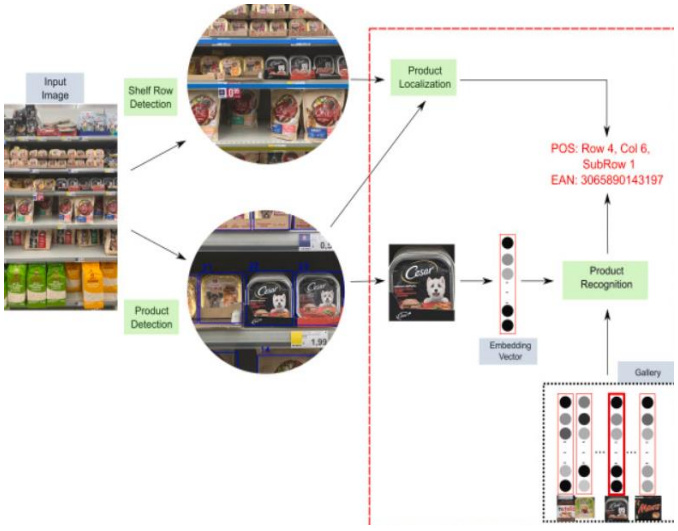


Fig. 2. Shelf Management system and its key components. The system encompasses several essential steps, including Shelf Row Detection, Product Detection, Product Recognition, and Product Localization. Primary steps are highlighted in green, essential data components are depicted in gray. The red segment signifies individual product processing, with this process occurring concurrently for each identified product.gure 4: Trade-off between accuracy and diversity across recommendation methods.

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, 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.Potential Strategies for Addressing the Cold Start Problem:

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.. 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.

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 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. 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.

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

4. Results and discussions

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 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 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. 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

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

Backbone	AP	AP50	AP75	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

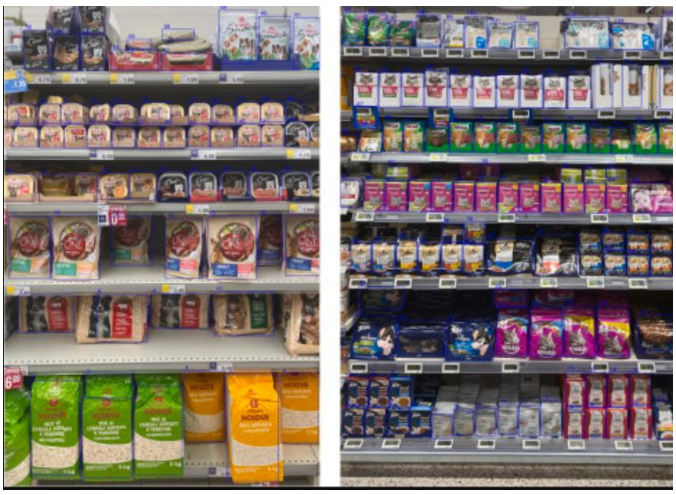
4.2. Shelf row detection results

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 li, lj be a pair of lines to be measured, the angular distance S is defined as:

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	0.966	0.0235
120	0.9704	0.9712	0.969	0.0250
130	0.9730	0.9827	0.974	0.0279
150	0.9529	0.9814	0.972	0.0329



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.

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 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.

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

Network	GPU	CPU
MobileNetV3-Large	0.002	0.026
MobileNetV3-Small	0.001	0.012
EfficientNetV2-B0	0.002	0.046
Inception-V3	0.002	0.095

III. Conclusion and Future Work

Summary:

This paper demonstrates the effectiveness of collaborative filtering techniques in generating personalized music playlists. By leveraging user-song interactions and applying similarity

metrics, the system successfully predicts songs that align with individual preferences, leading to improved recommendation accuracy, diversity, and overall user satisfaction. Among the analysed approaches, the **hybrid model**, which combines both user-based and item-based collaborative filtering, proved to be the most effective in balancing recommendation accuracy and music diversity. Evaluation results confirmed that collaborative filtering outperforms content-based filtering, particularly in terms of personalization. Despite these advantages, challenges such as the **cold start problem** and **popularity bias** persist. Strategies like hybridization with content-based filtering and leveraging **demographic data** offer promising solutions to mitigate these issues.

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 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 processThese decisions were supported by rigorous pre-testing to ensure that each component integrated into our system was optimized for performance in a retail environment. 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

IV. REFERENCES

[1] L. Suchman. Plans and Situated Actions, The Problem of Human-Machine Communication. Cambridge University Press (1987).

[2] S. Tamminen, A. Oulasvirta, K. Toiskallio and A. Kankainen. Understanding Mobile Contexts. Proc. Mobile HCI, 17-31 (2003). [3] M.P. Vossen. Local Search for Automatic Playlist Generation. Masters Thesis, Technische Universiteit Eindhoven (2006).

[4] N. Kravtsova, G. Hollemans, T.J.J. Denteneer, and J. Engel. Improvements in the Collaborative Filtering Algorithms for a Recommender System. Technical Note NL-TN-2001/542, Philips Research, Eindhoven, (2002).

[5] S. R. Covey. The 7 Habits of Highly Effective People. SimonSchuster UK LTD, (2004)

[6] A.G. Greenwald, T.C. Brock, T.M. Ostrum. Psychological Foundatins of Attidue. Academic Press, New York. (1973). [7] M. Tollos, R. Tato, T. Kemp. Mood-Basd Navigation Through Large Collection of Musical Data.. The 5th International Conference on Music Information Retrieval, Barcelona (Spain), ISMIR (2004). [8] D. Liu, L. Lu and H-J. Zhang. Automatic Mood Detection from Acoustic Data. John Hopkins University, (2003).

[9] L. Edwards and T. Torcellini. A Literature Review of the Effects of Natural Light on Building Occupants. Technical Note NREL/TP-550-30769, National Renewable Energy Laboratory, Colorado, (2002). [10] A.G. Barnston. The Effect of Weather on Mood, Productivity, and Frequency of Emotional Crisis in a Temperate Continental Climate. International Journal of Biometeorology. Vol. 32, No. 2, (1998). [11] AES, Demystifying Audio Metadata, Journal of the Audio Engineering Society, Vo. 51, No.7/8, 744-751 (2003).

[12] Orio. Music Retrieval: A Tutorial and Review. Foundations and Trends in Information Retrieval. Vol. 1, No. 1, 1-90 (2006).

[13] J-J. Arcourtier and F. Pachet. Music Similarity Measures: What's the Use? The 3rd International Conference on Music Information Retrieval, Paris (France), ISMIR (2002). [14] J.R. Brown. The Effects of Stressed Tempo Music on Performance Times of Track Atheletes. Florida State University. Florida, (2005).