Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33560

From Visual Perception to Multimodal Fusion: Application Analysis of Artificial Intelligence in Physical Retail Environment

Shuaichen Guo

Faculty of Innovation Engineering, Macau University of Science and Technology, Macau, China
1240006958@student.must.edu.mo

Abstract. Under the continuous impact of e-commerce, traditional physical retail needs to complete the digital transformation and upgrading to the "retail 4.0" mode. This paper systematically reviews the application of computer vision and multimodal fusion technology in the real retail scene. Firstly, the research sorted out the five stages of visual perception evolution from single mode to multimodal, and compared the core characteristics and development context of target detection and behavior recognition technology; Then, the mathematical model and collaboration mechanism of the fusion of vision and RFID, audio, radar and other modes are analyzed in depth, and an intelligent retail ecological framework covering bottom perception, middle fusion, and top-level applications is built, covering key application directions such as unmanned retail, digital twins, and supply chain optimization. The research confirms that single vision technology has the limitation of scene adaptation, and multimodal fusion is the necessary path to achieve high-precision customer insight and independent retail. Finally, this paper sorts out the problems of large-scale applications such as privacy protection, model robustness, edge computing resource constraints, and puts forward future research directions, such as lightweight end-to-side model development, interactive system construction of fusion generative large model, and exploration of human-computer deep collaboration mechanism.

Keywords: Computer vision, Multimodal fusion, Deep learning, Independent retail system

1. Introduction

With the rapid development of global e-commerce and profound changes in consumption patterns, the traditional physical retail industry is facing unprecedented transformation pressure. In order to reshape the core competitiveness in the fierce market competition, physical retail needs to break through the shackles of the traditional operation mode and accelerate the pace towards the "retail 4.0" mode. The core connotation of "retail 4.0" is to deeply rely on emerging technologies such as artificial intelligence (AI), Internet of things (IoT), and big data analysis to realize the comprehensive digital and intelligent transformation of physical shopping scenes, commodity management, and customer experience [1]. Its ultimate goal is to build a seamless Omni channel shopping experience, and achieve a leap forward improvement in operational efficiency through

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33560

accurate collection and in-depth mining of customer behavior data in the store. Early attempts to automate retail systems mainly relied on barcode scanning and basic radio frequency identification (RFID) technology. Although these technologies have played a fundamental role in inventory management and logistics tracking, they show obvious deficiencies in dealing with complex and dynamic retail scenarios. Especially in the face of high-precision application requirements such as customer fine-grained behavior understanding, contactless automatic checkout and scene semantic perception, the breadth and depth of their perception are difficult to support the goal of refined operation. In recent years, the growing maturity of deep learning technology and the significant enhancement of computing power have completely changed this situation and promoted the explosive application of computer vision technology in the retail field. By building a multi-layer neural network model, the machine can carry out multi-level abstract representation learning on massive visual data, so as to automatically extract the high-level semantic features in the image, which has made breakthroughs in key tasks such as commodity recognition, target detection and customer tracking [2]. In this context, the target detection algorithm has evolved from the traditional artificial feature extraction method to two-stage (such as Fast R-CNN) and single-stage (such as YOLO Series) detectors based on convolutional neural network (CNN). The introduction of these advanced algorithms has significantly improved the real-time and accuracy of commodity positioning and classification in the retail environment, and laid a solid perceptual foundation for the landing of unmanned retail. However, in the actual deployment process, pure visual perception technology still faces severe challenges. In a real retail environment, severe occlusion of the target, dramatic changes in lighting conditions, high similarity of commodity packaging, and complex customer interaction often lead to blind spots or recognition errors in a single visual mode. This inherent physical limitation has prompted academia and industry to gradually shift the focus of research to multimodal fusion technology. Multimodal fusion aims to simultaneously process and integrate information from multiple modes such as vision, audio, RF signals, weight sensors and radar by building a unified mathematical architecture [3]. The core advantage of this technology path is that it can make full use of the information complementarity and redundancy between different sensors, effectively overcome the perception bottleneck of single mode and improve the robustness of the system in extreme scenarios through cross modal collaborative verification and feature enhancement. For example, the "just walk out" technology adopted by Amazon go is a typical business example of the integration of multimodal perception and deep learning [4]. The system captures the visual trajectory through the top-level camera array, combines with the physical feedback of the shelf weight sensor, and accurately determines the shopping intention of customers through multi-source data fusion algorithm, so as to realize the real sense of non sensory payment experience. Although significant achievements have been made in related technologies, there are still some gaps in the existing literature in the systematic sorting of the evolution trajectory of multimodal technology, mathematical formal modeling and the construction of the top-level ecological framework. This paper aims to systematically fill this research gap and provide guiding technical reference and theoretical support for the intelligent upgrading of the retail industry. Specifically, this paper will carry out in-depth analysis from the following four dimensions: first, comb the core evolution process of single-mode visual perception technology, and focus on evaluating the performance of mainstream target detection algorithms in retail scenes; Secondly, it discusses the technology path from single visual perception to multimodal fusion, covering the fusion architecture and typical coordination mechanism of different dimensions such as feature level and decision-making level in detail; Then, based on the maturity of the underlying perception technology, an integrated intelligent retail ecological application framework is built; Finally, the

DOI: 10.54254/2755-2721/2026.GU33560

paper prospectively discuss the core challenges faced by current technology applications, such as privacy protection and model generalization, and point out the future research direction.

2. Single mode visual perception technology and core evolution

2.1. Development track of computer vision in offline retail

The development of intelligent retail vision technology can be divided into five key stages (as shown in Figure 1). The first stage is the traditional perception stage (before 2012), which mainly relies on barcode scanning and manually designed image features (such as sift, hog, etc.). The system can only complete simple commodity code recognition, lacking a true understanding of scene semantics, and is difficult to cope with the changing conditions in the real retail environment. The second stage is the deep learning driven single-mode vision stage (2012-2018). AlexNet's breakthrough performance in the ImageNet competition opened the era of deep learning [5]. Fast r-cnn and YOLO further promoted the breakthrough development of target detection technology, making a qualitative leap in the accuracy and speed of commodity identification in the retail scene [6,7]. The third stage is the embryonic stage of visual maturity and multimodality (2019-2021). Visual technology is gradually maturing in the field of commodity recognition and multi-target tracking. Researchers began to pay attention to the limitations of single visual technology in the scene of occlusion, light changes and so on. At the same time, early multimodal attempts to combine vision with weight sensors and audio have appeared, which preliminarily verified the feasibility and effectiveness of cross modal information fusion [8,9]. The fourth stage is the multimodal collaboration and application deepening stage (2022-2023). The deep integration of vision and radar is widely used in scenes such as smart shopping carts and complex behavior analysis, and the multi-sensor collaborative sensing architecture is gradually engineering [10]. The fifth stage is the construction stage of intelligent retail ecosystem (2024-2025). The emergence of unmanned retail system marks the initial formation of intelligent retail ecosystem, and the system level perception, decision-making and service capabilities begin to be verified in large-scale business scenarios [11,12].

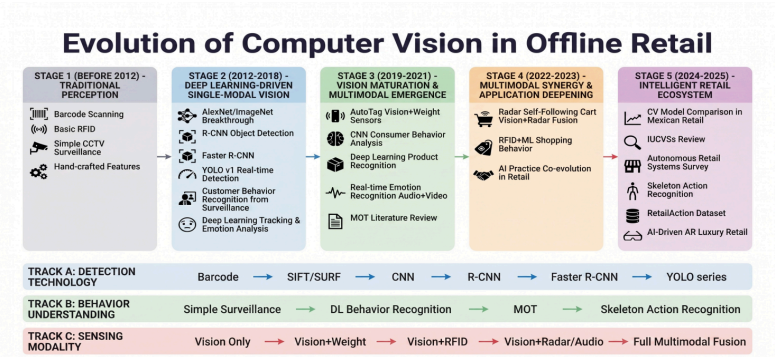


Figure 1. Development and evolution timeline of computer vision in offline retail environment (picture credit: original)

2.2. Evolution and mathematical evaluation of target detection algorithm

Target detection is the core task of retail visual perception, and the evolution of its algorithm architecture directly determines the practicability of the system. The two-stage detector (such as Fast R-CNN) uses the two-step process of "forming candidate regions and reclassifying and returning",

DOI: 10.54254/2755-2721/2026.GU33560

generates high-quality candidate boxes with the help of the region proposal network (RPN), and then carries out fine classification and boundary box regression on the candidate boxes through the classifier. The accuracy is high but the reasoning speed is relatively slow, which is suitable for commodity inventory and shelf analysis scenarios that do not require high real-time performance [6]. The single-stage detector (such as the YOLO series) unifies the target positioning and classification as a single regression problem, and simultaneously predicts the location and category of all targets in a single forward propagation. It has fast reasoning speed and is suitable for real-time customer behavior tracking and checkout scenarios [7]. The two types of architectures have their own emphasis in retail scenarios, and the actual system is often selected according to the trade-off between accuracy and speed.

In the model evaluation, the intersection union ratio (IoU) is the core index, which is used to measure the coincidence degree between the prediction box $B_p$and the real box B_{gt} :

$$IoU = \frac{AreaofOverlap}{AreaofUnion} = \frac{B_p \cap B_{gt}}{B_p \cup B_{gt}} \tag{1}$$

In the YOLO algorithm, in order to ensure the reliability of the detection results, the core concept of confidence score is introduced. The confidence degree is used to measure the probability that the bounding box contains the target and the coincidence degree between the prediction box and the real box, and its mathematical expression is:

$$C = P_r(object) \cdot IoU(B_p, B^{gt}) \tag{2}$$

Where $\mathrm{P_r(object)}$ represents the probability that there is a target in the grid cell, and IoU ($\mathrm{B_p}$, $\mathrm{B^{GT}}$) represents the intersection ratio of the predicted bounding box and the real bounding box.

In order to eliminate overlapping redundant detection frames, the detection system usually adopts non maximum suppression (NMS) algorithm [6]. The core idea of the algorithm is: in all candidate detection boxes, first select the prediction box with the highest confidence as the reference box, and then calculate the intersection and union ratio (IoU) of the remaining candidate boxes and the reference box in turn, and suppress the candidate boxes whose IoU exceeds the preset threshold as redundant boxes, so as to ensure that each commodity target is detected only once in the final detection result, effectively avoiding the problem of repeated counting. In the retail scene, goods on shelves are densely arranged, and the threshold setting of NMS has a significant impact on the final detection accuracy. A high threshold may lead to missed detection, and a low threshold may lead to the false deletion of the detection box of adjacent goods.

The cross entropy loss function is usually used in model training. However, in the retail scene, the number of image background areas (negative samples) is often far more than the commodity target areas (positive samples). This extremely unbalanced sample distribution will lead to the model being dominated by a large number of easily classified background samples in the training process, thus weakening the learning ability of difficult to classify commodity targets. To solve this problem, focal loss loss function is often introduced:

$$L_{FL} = -\sum_{c=1}^{M} \alpha_c (1 - p_c)^{\gamma} y_c log(p_c) \tag{3}$$

Among them, γ is the focusing parameter, which is used to dynamically reduce the loss weight of easy to classify samples, so that the training focus of the model is focused on the commodity targets that are difficult to distinguish; α_c is used to balance the sample proportion between different

DOI: 10.54254/2755-2721/2026.GU33560

categories and further alleviate the problem of category imbalance. Focal loss has been proved to significantly improve the recognition ability of the model for small targets and occluded goods in the task of dense commodity detection.

3. Evolution from single visual modality to multimodal fusion

3.1. Architecture and mathematical formalization of multimodal fusion

In the retail scene, multimodal fusion aims to integrate the semantic information of vision and the physical attributes of other sensors, and realize the collaborative expression and joint decision-making of cross modal information by constructing a unified mathematical framework. There are three fusion strategies (as shown in Figure 2):

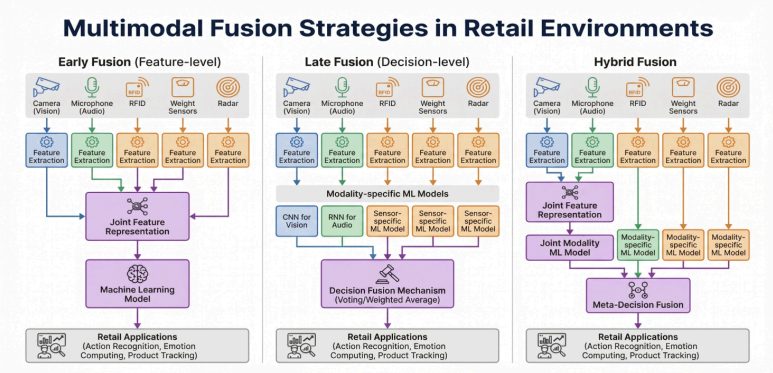


Figure 2. Multimodal fusion strategy architecture in retail environment (picture credit: original)

Suppose there are N modes, the original data of mode i is X_i, and the feature is expressed as $h_i = f_i(x_i)$.Late fusion (decision level fusion): after independent prediction of each mode, the final decision is obtained by weighted fusion.

$$\hat{y} = \sum_{i=1}^{N} w_i \, g_i(h_i) s.t. \sum_{i=1}^{N} w_i = 1 \tag{4}$$

Where w_i is the weight of dynamic learning through the attention mechanism:

$$w_i = \frac{exp(v^\top tanh(Wh_i + b))}{\sum_{j=1}^{N} exp(v^\top tanh(Wh_j + b))} \tag{5}$$

This method is highly flexible, convenient for modular design and independent optimization, and when a modal sensor fails, the system can still rely on other modes to maintain basic functions, with strong fault tolerance. It is the most widely used fusion strategy in the current intelligent retail system [13].

3.2. Typical multimodal coordination mechanism

In the process of multi-modal collaborative development of intelligent retail, the fusion application of vision technology and a variety of different types of sensors and signals has effectively made up for the limitations of a single vision technology and promoted the deepening and upgrading of technology applications. The representative fusion forms mainly include three types. In the fusion application of visual and physical sensors (Vision+RFID/Weight), the AutoTag system successfully

DOI: 10.54254/2755-2721/2026.GU33560

solves the key pain point of high labeling cost in the smart retail scene by combining the shelf weight sensor with the camera and using the data collaboration of the two to realize automatic domain adaptation [8]. The core innovation of this scheme is that when there is ambiguity in visual recognition, the weight change signal can be used as a strong constraint to assist in determining the commodity category, so as to realize the model migration and deployment across stores without a large number of manual annotation. The integration of visual and auditory signals (vision+Audio) further expands the perception dimension. The AudioSense system indirectly infers customers' shopping behavior by accurately analyzing various audio signals such as commodity collision sound and packaging friction sound, effectively making up for the lack of perception of a single visual technology when there is visual occlusion in the scene [9]. Audio mode has a unique perception advantage for the key action of customers' taking goods. Especially in the scene of more shelves and limited camera perspective, it can significantly improve the recall rate of behavior recognition. The deep fusion of vision and radar (vision+radar) improves the accuracy and reliability of scene perception. Millimeter wave radar can provide accurate range and speed information, which is not affected by lighting conditions. By organically integrating it with advanced vision algorithms, an intelligent shopping cart with high-precision target tracking and autonomous obstacle avoidance capabilities has been successfully constructed [10]. The precise depth information provided by radar can effectively compensate the inherent defects of traditional 2D vision technology in distance perception, and provide strong support for the stable operation of intelligent shopping cart in crowded shopping mall environment.

4. Integrated application framework: building an intelligent retail ecosystem

The maturity of the underlying perception technology promotes the systematization of top-level applications. With the gradual maturity of visual perception and multimodal integration technology, the intelligent retail system began to evolve from a single functional module to an integrated ecological framework. Figure 3 shows the conceptual framework of the smart retail ecosystem, which covers three levels: the bottom layer of perception acquisition, the middle layer of AI processing and multimodal integration, and the top layer of application and ecosystem.

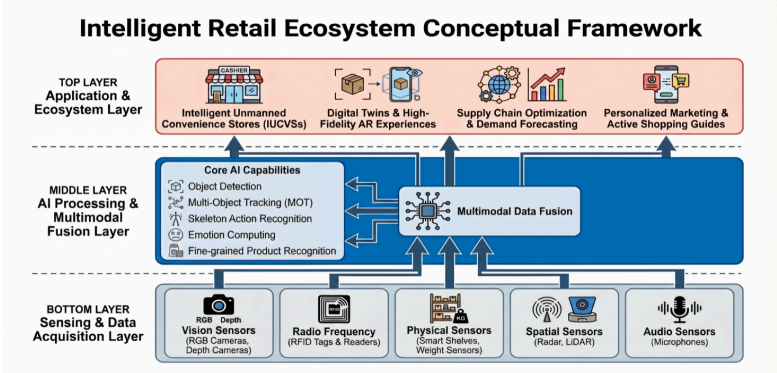


Figure 3. Conceptual framework of smart retail ecosystem (picture credit: original)

The maturity of the underlying perception technology has promoted the systematic development of top-level applications, and gradually formed an integrated application framework, thus supporting the construction of the intelligent retail ecosystem. Figure 3 shows the conceptual framework of the intelligent retail ecosystem. As a typical application carrier of multimodal sensing technology, intelligent unmanned convenience stores (iucvss) constitute the core implementation form of

DOI: 10.54254/2755-2721/2026.GU33560

unmanned retail system. By integrating the overhead camera array and the shelf weight sensor, this kind of system can realize the whole process of automatic settlement from consumers' entry to departure [4]. The realization of this technology path depends on multi-level perception cooperation: the rgb-d vision sensor at the bottom is responsible for capturing the customer's spatial position and hand movements, and the shelf weight sensor provides the physical confirmation signal of goods' taking and placing. The two conduct joint reasoning through the multi-modal fusion model at the middle layer, and finally output the accurate shopping list at the top application layer. In addition, the collaborative evolution of retail practice is also an important part of the construction of intelligent retail ecosystem. The theory of practical co evolution points out that the introduction of artificial intelligence technology is not only a deployment at the technical level, but also a process of collaborative participation of multiple subjects [14]. Artificial intelligence has a reshaping effect on retail work practice, and it is better to show a synergistic relationship with human labor, rather than a simple replacement. At the top application level, the combination of digital twin technology and high fidelity ar experience provides retailers with real-time store operation visualization and simulation optimization capabilities; The supply chain optimization and demand forecasting module realizes the intellectualization and refinement of inventory management by integrating multimodal perception data and historical sales records; The personalized marketing and active shopping guidance functions rely on the customer behavior analysis model to provide consumers with real-time commodity recommendation and shopping guidance services, which further improves the personalized degree of shopping experience.

5. Discussion and future direction

5.1. Main challenges

At present, intelligent retail vision technology still faces many challenges in the process of large-scale application, mainly in three aspects. In terms of privacy protection and data security, the deployment of large-scale video surveillance has triggered widespread privacy concerns and ethical disputes. There are significant differences in consumers' acceptance of facial recognition and behavior tracking technology. Federated learning, as a distributed machine learning paradigm, allows collaborative training of models without sharing original data, so as to achieve cross store model optimization while protecting user privacy, providing a feasible path for solving this problem that takes into account both performance and compliance [15]. In terms of model robustness and generalization ability, the lighting conditions, shelf layout and commodity display mode of real stores are highly variable. The study found that the performance of the model can be reduced by 10% -30% in actual deployment, which shows that the model trained in the controlled laboratory environment is often difficult to migrate directly to complex real scenes, so the research of domain adaptation technology is particularly urgent. In terms of edge computing resource constraints, running complex multimodal fusion models requires high hardware computing power, while the computing resources and power budget of retail terminal devices are often very limited [16]. Model compression technologies such as knowledge distillation are the key technology path for deploying high-performance models on resource constrained devices [17]. How to maintain the model accuracy while significantly reducing the computational overhead is the core problem of the current engineering implementation.

DOI: 10.54254/2755-2721/2026.GU33560

5.2. Future research directions

In response to the above challenges, future research can focus on three core directions. The first is the lightweight end-to-side model. Through network pruning, quantitative perception training and neural architecture search (NAS) technologies, the paper develops an efficient multimodal architecture that can run in real time on edge devices, and realize the perception accuracy close to cloud reasoning under the constraint of limited computing resources, so as to reduce the hardware deployment threshold and operation cost of the intelligent retail system. Second, combined with the interactive system of generative large model, the large language model (LLM) has strong cross domain reasoning and natural language interaction ability [18]. The visual language model (VLM) is introduced into the retail field, which can realize the paradigm shift from passive video surveillance to active voice shopping guide, so that the system can understand customers' intentions and provide personalized real-time shopping suggestions. The third is human-computer deep collaboration, which focuses on exploring how to realize the real empowerment of AI technology in the retail front line by optimizing the human-computer interaction interface and workflow, and promote retail employees to free themselves from repetitive work and focus on high-value service links that require human judgment and emotional connection, so as to ultimately maximize the effectiveness of human-computer collaboration [14].

6. Conclusion

This paper systematically summarizes the application status and development trend of computer vision and multimodal fusion technology in the real retail scene. The research shows that with the continuous improvement of the complexity of retail scene, the single visual perception technology has been unable to meet the needs of high-precision customer behavior understanding and independent retail system, and the evolution to multimodal fusion (vision+sensor/audio/radar) is the only way to achieve accurate customer insight and unmanned retail. This paper embodies three major academic contributions. It firstly systematically combs the five evolutionary stages of intelligent retail visual technology from traditional perception to multimodal ecology and builds a clear analytical framework for historical cognition in this field. It secondly constructs a multimodal mathematical formal model containing early fusion, late fusion and hybrid fusion, realizes adaptive learning of dynamic weights by introducing the attention mechanism, and offers theoretical support for subsequent algorithm design. It finally puts forward an intelligent retail ecological framework covering underlying perception acquisition, middle-level AI processing and top-level application ecology, which provides systematic path guidance for academic research and enterprise digital transformation. With the continuous improvement of algorithm efficiency and the continuous reduction of software and hardware costs, multimodal AI technology will continue to reshape the business model and service form of retail industry in the future. However, challenges such as the compliance of privacy protection, the robustness of models across scenarios, and the efficient use of edge computing resources still need continuous attention. Future research should explore in-depth in the direction of lightweight deployment, generative large-scale model interaction, and human-computer collaboration, so as to promote the implementation of intelligent retail technology from laboratory to large-scale commerce.

DOI: 10.54254/2755-2721/2026.GU33560

References

[1] Har, L. L., Rashid, U. K., Chuan, L. T., Sen, S. C., & Xia, L. Y. (2022). Revolution of retail industry: From perspective of retail 1.0 to 4.0. Procedia Computer Science, 200, 1615–1625. https: //doi.org/10.1016/j.procs.2022.01.362

[2] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https: //doi.org/10.1038/nature14539

[3] Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443. https: //doi.org/10.1109/TPAMI.2018.2798607

[4] Szabó-Szentgróti, E., Rámháp, S., & Kézai, P. K. (2023). Systematic review of cashierless stores (just walk out stores) revolutionizing the retail. Management & Marketing, 18(s1), 427–448. https: //doi.org/10.2478/mmcks-2023-0023

[5] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (Vol. 25). Curran Associates, Inc. https: //proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[6] Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 1137–1149. https: //doi.org/10.1109/TPAMI.2016.2577031

[7] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 779–788). https: //www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html

[8] Ruiz, C., Falcao, J., & Zhang, P. (2019). AutoTag: Visual domain adaptation for autonomous retail stores through multi-modal sensing. In Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 518–523). ACM. https: //doi.org/10.1145/3341162.3345604

[9] Sharma, A., & Lee, Y. (2017). AudioSense: Sound-based shopper behavior analysis system. In Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 488–493). ACM. https: //doi.org/10.1145/3123024.3124440

[10] Yao, L., Li, T., Tong, R., Wang, K., & Zhang, L. (2023). Radar self-following shopping cart based on multi-sensor fusion. IEEE Access, 11, 77055–77072. https: //doi.org/10.1109/ACCESS.2023.3297889

[11] Kohli, T. P., Malhotra, S., & Pingali, V. (2025). Revolutionising retail: A review of intelligent unmanned convenience stores and their market implications. Technological Forecasting and Social Change, 219, 124273. https: //doi.org/10.1016/j.techfore.2025.124273

[12] Rukundo, S., et al. (2025). A survey of challenges and sensing technologies in autonomous retail systems. arXiv preprint arXiv: 2503.07997. https: //arxiv.org/abs/2503.07997

[13] Kalisetty, S., & Lakkarasu, P. (2025). Deep learning frameworks for multi-modal data fusion in retail supply chains: Enhancing forecast accuracy and agility. International Journal of Intelligent Systems and Applications in Engineering, 2(1).

[14] Bonetti, F., Montecchi, M., Plangger, K., & Schau, H. J. (2023). Practice co-evolution: Collaboratively embedding artificial intelligence in retail practices. Journal of the Academy of Marketing Science, 51(4), 867–888. https: //doi.org/10.1007/s11747-022-00896-1

[15] Truong, N., Sun, K., Wang, S., Guitton, F., & Guo, Y. (2021). Privacy preservation in federated learning: An insightful survey from the GDPR perspective. Computers & Security, 110, 102402. https: //doi.org/10.1016/j.cose.2021.102402

[16] Fernández Del Carpio, A. (2024). Analyzing computer vision models for detecting customers: A practical experience in a Mexican retail. International Journal of Advances in Intelligent Informatics, 10(1), 131. https: //doi.org/10.26555/ijain.v10i1.1112

[17] Moslemi, A., Briskina, A., Dang, Z., & Li, J. (2024). A survey on knowledge distillation: Recent advancements. Machine Learning with Applications, 18, 100605. https: //doi.org/10.1016/j.mlwa.2024.100605

[18] Raza, M., Jahangir, Z., Riaz, M. B., Saeed, M. J., & Sattar, M. A. (2025). Industrial applications of large language models. Scientific Reports, 15(1), 13755. https: //doi.org/10.1038/s41598-025-98483-1