Chapter 6
Spatiotemporal Object Detection and Activity Recognition

Vimal Kumar, Shobhit Jain, and David Lillis

Abstract Spatiotemporal object detection and activity recognition are essential components in the advancement of computer vision, with broad applications spanning surveillance, autonomous driving, and smart stores. This chapter offers a comprehensive overview of the techniques and applications associated with these concepts. Beginning with an introduction to the fundamental principles of object detection and activity recognition, we discuss the challenges and limitations posed by existing methods. The chapter progresses to explore spatiotemporal object detection and activity recognition, which entails capturing spatial and temporal information of moving objects in video data. A hierarchical model for spatiotemporal object detection and activity recognition is proposed, designed to maintain spatial and temporal connectivity across frames. Additionally, the chapter outlines various metrics for evaluating the performance of object detection and activity recognition models, ensuring their accuracy and effectiveness in real-world applications. Finally, we underscore the significance of spatiotemporal object detection and activity recognition in diverse fields such as surveillance, autonomous driving, and smart stores, emphasizing the potential for further research and development in these areas. In summary, this chapter provides a thorough examination of spatiotemporal object detection and activity recognition, from the foundational concepts to the latest techniques and applications. By presenting a hierarchical model and performance evaluation metrics, the chapter serves as a valuable resource for researchers and practitioners seeking to harness the power of computer vision in a variety of domains.

Keywords Spatiotemporal data · Object detection · Activity recognition
Challenges and limitations

V. Kumar (✉) · D. Lillis
University College Dublin, Dublin, Ireland
e-mail: vimal.kumar@ucd.ie; david.lillis@ucd.ie

S. Jain
The University of Texas at Dallas, Richardson, TX, USA

J. A et al. (eds.), *Spatiotemporal Data Analytics and Modeling*, Big Data Management, https://doi.org/10.1007/978-981-99-9651-3_6

6.1 Introduction

Spatiotemporal data analytics is an interdisciplinary field that uses computational methods and algorithms to examine data that changes over space and time. With advancements in sensing technologies and a rapid increase in data volume, the significance of spatiotemporal analytics has grown exponentially, finding extensive applications in fields such as environmental monitoring, urban planning, transportation, health surveillance, and increasingly in intelligent systems like autonomous vehicles and smart stores [1]. Central to this interdisciplinary domain is spatiotemporal data that captures information across both spatial and temporal dimensions. Its importance has amplified with technological advancements and the widespread use of surveillance systems, wearable devices, and video platforms. One pivotal application of spatiotemporal data is in object detection and activity recognition in videos, which involves identifying, localizing, and classifying activities performed by specific objects in video sequences [2].

Object detection, an integral component of any video understanding system, seeks to locate and classify objects of interest within video frames. Despite challenges posed by variations in object appearance, occlusions, and changes in camera viewpoint, this process has witnessed significant improvements due to the advent of deep learning techniques [3–5]. After object detection, activity recognition comes into play, focusing on identifying the actions or behaviors exhibited by the detected objects over a sequence of frames. Activity recognition finds applications across diverse contexts, including sports, surveillance, health care, and wildlife tracking, despite the inherent ambiguity and variability in human actions [2, 4, 6, 7].

This chapter seeks to provide a comprehensive overview of the techniques for spatiotemporal object detection and activity recognition. We will discuss the evolution of these methodologies, current challenges and limitations, and state-of-the-art techniques. Moreover, the chapter will outline various performance evaluation metrics crucial for ensuring model accuracy and effectiveness. It will also explore the application of spatiotemporal object detection and activity recognition in diverse domains, including surveillance, autonomous driving, and smart stores, thereby showcasing current research and the potential for future advancements.

Following this introduction, Sect. 6.2 will delve into the history and basic techniques of object detection and activity recognition. Section 6.3 will address the challenges and limitations in this field, illustrated with examples. Section 6.4 will focus on the relevance of spatiotemporal data to object detection and activity recognition. Section 6.5 will offer a detailed discussion of the hierarchical model, while Sect. 6.6 will cover various performance evaluation metrics with comparison across different models. Lastly, Sect. 6.7 will discuss applications and potential future developments in the field.

In conclusion, this chapter aims to serve as a comprehensive guide to spatiotemporal object detection and activity recognition. It intends to offer a broad understanding of the concepts and techniques, proving invaluable for both novice readers and experienced researchers keen on exploring the dynamic world of video-based understanding systems. By summarizing key points, exploring potential impact on computer vision, and potential for future research, it aims to foster further exploration and innovation in this exciting field.

6.2 Fundamentals of Object Detection and Activity Recognition

To comprehend the complexity and breadth of object detection and activity recognition, we begin by exploring their historical progression, proceed to deconstruct the underlying algorithms, and finally delve into their importance and diverse applications across various industries.

Object detection, a cornerstone of computer vision systems, can trace its evolution back to the incipient stages of image processing. In the beginning, the goal was relatively straightforward: to identify and categorize objects within static images [8]. However, with the advent and proliferation of video technologies, the need to decipher moving visual content became paramount, propelling object detection to a whole new realm of challenges and possibilities. Early object detection methods primarily focused on discerning simple, distinct features such as edges or corners, that were considered as characteristic signatures of objects within an image. The emergence of machine learning and, more specifically, convolutional neural networks (CNNs) marked a significant breakthrough, empowering modern techniques to accurately identify and classify multiple objects within a single frame [9].

Activity recognition, an advancement of object detection, seeks to comprehend the dynamic behavior of detected objects within video sequences. This process of deciphering the semantics of motion necessitates not only recognizing the individual actions performed by objects but also understanding the context and temporal sequencing of these actions [10, 11].

At a rudimentary level, object detection and activity recognition operate on a sequential pipeline comprising preprocessing of data, feature extraction, and finally, classification or recognition (Fig. 6.1). The initial preprocessing stage involves

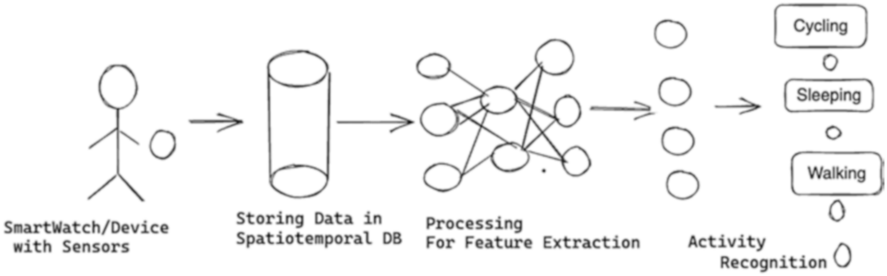
SmartWatch/Device with Sensors
Storing Data in Spatiotemporal DB
Processing For Feature Extraction
Activity Recognition
Cycling
Sleeping
Walking

Fig. 6.1 The pipeline of object detection and activity recognition

filtering noise and normalizing the input data to a standard format. Following this, the feature extraction phase seeks to distinguish key characteristics that help differentiate various objects or activities. In the context of object detection, these features could be geometric attributes such as shape, color intensity, or texture gradients. Conversely, for activity recognition, spatiotemporal features like motion trajectories, sequential object interactions, or action primitives form the crux [8]. The final stage of the pipeline involves classification or recognition, which employs machine learning algorithms to segregate the extracted features into distinct classes or activities.

Object detection and activity recognition techniques have demonstrated immense utility and versatility across numerous sectors. For instance, in surveillance systems, these techniques enable detection of anomalies or intrusions, effectively enhancing security measures [9]. In health care, object detection and activity recognition can contribute toward patient behavior monitoring and assist in intricate surgical procedures [6]. In the realm of autonomous vehicles, object detection facilitates the recognition of pedestrians, vehicles, and potential obstructions, whereas activity recognition aids in interpreting their movement patterns to predict future trajectories. Even in sports analytics, these techniques can augment player performance evaluation and strategizing by comprehending player movements and actions [2].

To summarize, the rapid evolution and convergence of object detection and activity recognition techniques have been instrumental in the development of sophisticated systems capable of understanding and interpreting dynamic visual content. The integration of these techniques with cutting-edge machine learning algorithms has not only expanded the boundaries of possibilities but also paved the way for novel challenges in the field of spatiotemporal data analytics (Table 6.1). In the ensuing section (Sect. 6.3), we shall delve deeper into the specific challenges and potential roadblocks encountered in contemporary techniques of object detection and activity recognition, paving the way for discussion on future research directions.

Table 6.1 A summary highlighting diverse applications of object detection and activity recognition across various sectors

Sector	Applications of object detection and activity recognition
Surveillance and security	(a) Enable detection of anomalies and intrusions (b) Monitor estates in real time (c) AI powered video analytics to find anomalies in a recorded video footage
Sports	(a) Assist referee in match decisions (b) Evaluate player's performance
Health care	(a) Patient behavior monitoring (b) Assist in critical surgeries
Retail	(a) Contactless checkout by creating a virtual shopping cart using AI (b) Inventory management to monitor stocks and its arrangements for easy access (c) Foot traffic analysis to perform crowd management (d) Informing employees for customer assistance when there is a need
Industries	(a) Monitoring the use of personal protective equipment (PPE) kits by workers (b) Product assembly to enable automated production in industries (c) Defect detection to maintain quality standards (d) Productivity improvement by monitoring workers activity
Transportation and automobile	(a) Driverless/autonomous vehicles (b) Collision detection (c) People counting for safety and planning (d) Parking occupancy monitoring (e) In-cabin monitoring for safety and medical emergencies (f) Road maintenance analysis

6.3 Challenges and Limitations in Object Detection and Activity Recognition

The promising developments in object detection and activity recognition have substantially aided our comprehension and interpretation of visual data. However, there exist numerous formidable challenges and limitations associated with these methodologies that need to be addressed.

One of the foremost challenges in object detection arises from substantial variations in scale, pose, and illumination of objects that are frequently encountered in real-world scenarios [12]. Objects of different sizes, orientations, and lighting conditions demand different levels of discernment, often putting a strain on

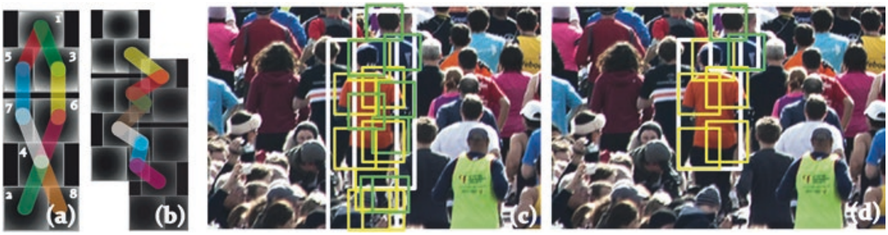

Fig. 6.2 Instance of occlusion in a crowded scene [14]

conventional detection techniques. Furthermore, the instances of overlapping objects and occlusion also compound the problem of accurate object detection. Occlusion, where objects are partially or completely hidden behind other objects, often leads to detection failures [13]. Figure 6.2 can depict an instance of occlusion in a crowded scene, demonstrating the complexity of accurate detection.

Another recurring issue in object detection is the detection of small, thin, or tiny objects [1]. The effective detection and classification of such objects remain challenging due to their poor representation in the feature maps. Traditional object detectors often fail to capture the nuanced details of such objects, leading to reduced detection performance.

Alongside these, the presence of background clutter and highly similar objects also pose significant difficulties [5]. Distinguishing objects from their surroundings or differentiating between similar-looking objects requires the implementation of highly sophisticated feature extraction and classification mechanisms.

Turning to activity recognition, the challenges grow multifold due to the added temporal dimension. Activities are inherently temporal in nature and can exhibit significant variability in terms of duration, speed, and sequence [13]. The need to comprehend the context of activities, including the recognition of interactions between multiple entities or the interpretation of multi-activity scenes, poses a significant challenge [7]. Understanding these interactions is critical in many application scenarios, such as surveillance and elderly care, but it is still a challenging task for current systems.

Moreover, intra-class variability, where the same activity is performed differently by different individuals or by the same individual at different times, is another prominent issue in activity recognition [3]. Different individuals have unique styles or habits, and these variations can drastically affect the recognition results. A notable case that brings these challenges to the forefront is in the field of sports video behavior recognition [2]. Identifying individual players' actions, interpreting team strategies, or recognizing significant events such as goals in soccer or aces in tennis necessitates sophisticated object detection and activity recognition methodologies. These tasks become exceedingly complex due to fast player movements, occlusion scenarios, and constant changes in the field of view.

In conclusion, while object detection and activity recognition technologies have made considerable strides, significant challenges and limitations persist. Further

Table 6.2 Encapsulates the main challenges and limitations in object detection and activity recognition

Serial number	Current challenges and limitations	Description
1	Interactions between multiple entities simultaneously	It will be more complex when a scene contains multiple objects and their activities during interactions.
2	Multi-activity scenarios	When there are two objects performing different actions, the complexity of identifying it through different frames becomes challenging.
3	Intra-class variability and identification	In a people monitoring application, there are different styles of walking by each individual and it creates ambiguity in identification of actions.
4	Sports video behavior recognition	For detecting fouls in a sport, sports video will be really helpful, but, since players are already in motion, it will be a major challenge to recognize.

research and technological advancements are required to address these issues and enhance the robustness and adaptability of these systems to effectively handle real-world variability and complexity (Table 6.2).

6.4 Spatiotemporal Object Detection and Activity Recognition

Spatiotemporal object detection and activity recognition concern the understanding of both spatial and temporal dynamics within video data, crucial for a comprehensive representation of events and activities in the real world. Here, the spatial aspect refers to the recognition of objects in the video frames, while the temporal aspect refers to the understanding of these objects' behaviors and movements over time.

The intrinsic value of incorporating spatiotemporal information within video data analysis lies in its potential to capture more nuanced and sophisticated understandings of complex events. For example, in sports analytics, not only the recognition of individual players and their actions (spatial) but also the pattern and sequence of these actions (temporal) are of paramount importance for strategy analysis [2]. Figure 6.3 can demonstrate the process of spatiotemporal object detection and activity recognition within a sports game scenario.

In another context, an application of spatiotemporal recognition is visible in surgical procedures. Here, the recognition of surgical instruments (spatial) and their usage over time (temporal) can aid in monitoring and evaluating surgical procedures [6]. Despite its significant potential, spatiotemporal object detection and activity

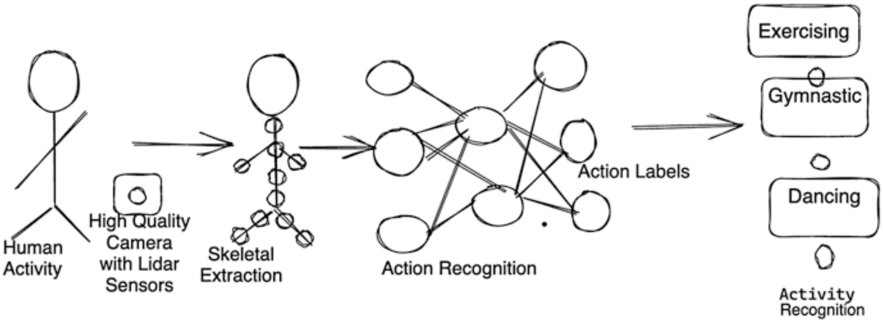

Fig. 6.3 Spatiotemporal object detection and activity recognition in a sports game

recognition present complex challenges. One of these challenges involves the extraction of useful spatiotemporal features from raw video data, which is a computationally expensive process requiring efficient and effective algorithms [11]. Moreover, the successful segmentation of foreground objects, which is crucial for activity semantics understanding, remains a difficult task due to varying lighting conditions, occlusions, and complex background activities [15]. Another key challenge involves the accurate detection and recognition of activities in real time. Many applications, such as surveillance systems and elderly care service robots, require instant responses, making the temporal aspect of activity detection significantly important [3].

Recognizing activities that involve multiple interacting objects is also a formidable challenge. Multiple objects can perform interdependent activities, resulting in a complex scenario that demands not only the recognition of individual activities but also an understanding of the interactions between them [7]. Recent advancements have addressed some of these challenges using sophisticated methodologies. For instance, using spatiotemporal depth cuboid similarity features extracted from depth camera data has shown promising results in activity recognition [16]. Further, employing the idea of spatiotemporal attention-based LSTM networks has achieved a significant breakthrough in 3D action recognition and detection [17].

In summary, spatiotemporal object detection and activity recognition is a burgeoning area of research with the potential to drastically improve our understanding and interpretation of complex real-world events. Yet, it also comes with its own set of challenges that necessitate further technological advancements and innovative research.

6.5 Model for Spatiotemporal Object Detection and Activity Recognition

This section focuses on the hierarchical model for spatiotemporal object detection and activity recognition. The model maintains both spatial and temporal connectivity across frames, a crucial feature for robust and efficient video understanding.

6.5.1 Hierarchical Model

The hierarchical model in this research is inspired by the concept of maintaining spatial and temporal connectivity across frames in videos. It adopts a twofold approach for activity recognition, which involves spatial object detection followed by temporal activity recognition. The object detection stage involves identifying key objects in each frame, and then the activity recognition stage analyzes the temporal evolution of these objects over several frames. The entire process can be viewed as a dynamic pipeline (as shown in Fig. 6.4), with object detection feeding into activity recognition.

Object detection relies on a robust foreground object segmentation mechanism, where key objects of interest are identified and segmented from the background [5, 15]. This approach is often employed in sports video behavior recognition, where key objects like players and the ball need to be identified for effective activity analysis [2, 18].

The model incorporates a multiscale spatiotemporal context approach, similar to those found in related work [11, 17]. It maintains the spatiotemporal continuity across frames by leveraging advanced computer vision techniques like Hough forests [19], LSTM networks with spatiotemporal attention [17], and depth cuboid similarity features [16]. Activity recognition, on the other hand, utilizes a combination of advanced temporal augmentation strategies and machine learning techniques [6]. This stage of the model emphasizes on the temporal evolution of key objects, recognizing patterns that signify specific activities. This aspect is important in applications like surgical activity recognition, where it is crucial to understand the temporal evolution of activities from video data [6].

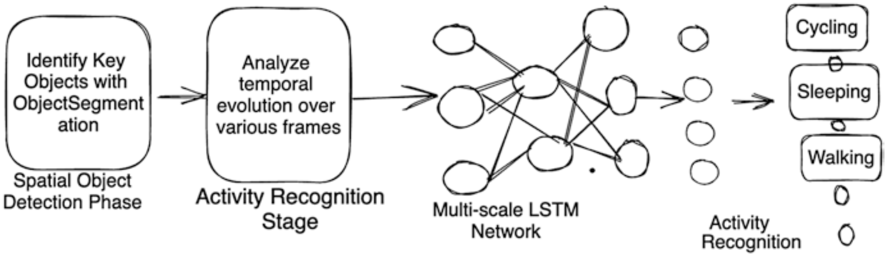
Identify Key Objects with ObjectSegment ation
Spatial Object Detection Phase
Analyze temporal evolution over various frames
Activity Recognition Stage
Multi-scale LSTM Network
Cycling
Sleeping
Walking
Activity Recognition

Fig. 6.4 The Hierarchical model for spatiotemporal object detection and activity recognition

Table 6.3 Comparison of different techniques for maintaining spatiotemporal connectivity

S. No.	Techniques	Merits	Demerits
1	Hierarchical framework [10]	Encodes contextual information at different levels—Point, intra-trajectory, and intertrajectory.	Relative motion between objects can reduce the performance of action recognition.
2	Human-object interaction model [20]	(i) Human action is predicted using human-object interaction as spatiotemporal features. (ii) Computationally efficient and can be used in real-time processing.	Convolution networks will not be able to encode semantic information into the spatiotemporal features.
3	Self-supervised framework [12]	(i) To reduce computational complexity, it uses patches extracted from video frames divided. (ii) To reduce memory constraints, this approach uses a permutation strategy instead of a random strategy.	No significant demerits identified.
4	Semantic representation technique [21]	(i) A graph driven approach to detect visual patterns. (ii) It has faster search time.	Event patterns at different hierarchies are not handled.
5	Cascaded region proposal and detection (CRPAD) framework [9]	This approach identifies activities at a fine-grain level.	Temporal modeling is not used in this approach.

6.5.2 *Spatial and Temporal Connectivity*

This model's unique feature is its emphasis on maintaining spatial and temporal connectivity across frames, thereby creating a holistic understanding of the scene's dynamic evolution. This connectivity is maintained via an efficient spatial-temporal graph, similar to what was proposed by Sun et al. [10]. This graph-based approach enables the model to establish object-object and object-scene relations, which contribute to a deeper understanding of the activity being performed [20]. This concept of connectivity is particularly significant in complex video environments, such as those found in surveillance or sports videos, where the context and multiscale motion awareness play a significant role in activity recognition [7, 9]. Additionally, the model incorporates advanced techniques for maintaining spatiotemporal connectivity. These include knowledge graphs for spatiotemporal event pattern matching [21], optimization models for human activity recognition inspired by information on human-object interaction [20], and the application of unsupervised learning of spatiotemporal context for video action recognition [12] (Table 6.3).

6.6 Performance Analysis of Object Detection and Activity Detection

In this section, we present and discuss various metrics used to evaluate the performance of object detection and activity recognition models. These metrics are paramount for ensuring model accuracy and effectiveness. In complex video understanding tasks, the effectiveness of a model can be evaluated by how well it performs in spatial object detection and temporal activity recognition, quantified using specific metrics [2, 15].

6.6.1 Analysis for Object Detection

Precision (P): This metric measures the proportion of correctly detected objects among all detections made by the model. It is calculated as:

$$P = \frac{TP}{TP + FP}$$

where TP is the number of true positives, and FP is the number of false positives [4].

Recall (R): This metric measures the proportion of correctly detected objects among all the ground truth objects. It is calculated as:

$$R = \frac{TP}{TP + FN}$$

where FN is the number of false negatives [7].

F1 score: This is the harmonic mean of precision and recall, providing a single metric that balances both. It is calculated as

$$F1 = 2 * \frac{P * R}{P + R}$$

It is particularly useful when the cost of false positives and false negatives are significantly different [18].

Intersection over union (IoU): This metric measures the overlap between the predicted bounding box and the ground truth bounding box. A high IoU score means the prediction closely aligns with the ground truth. It is calculated as

$$\mathrm{IoU} = \frac{\text{Area of overlap}}{\text{Area of union}}$$

IoU is an essential metric for object detection tasks as it quantifies the goodness of fit of the predicted bounding box [3].

6.6.2 Analysis for Activity Recognition

For activity recognition, the following metrics are commonly used:

Classification accuracy: This is the proportion of correctly classified activities among all activities. It is calculated as

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

where TN is the number of true negatives [20].

Mean average precision (mAP): This is the average precision at different recall levels. It calculates the area under the precision-recall curve and provides a single measure that balances precision and recall over all classes [6].

Confusion matrix: Although not a single metric, a confusion matrix gives a comprehensive view of how well the model performs across all classes of activities. It allows researchers to see the number of false positives and false negatives for each class, providing more insight into the model's performance [8] (Table 6.4).

As an example, consider the use case of a model detecting and recognizing activities in sports videos, such as tennis [18]. Here, the precise detection of players (object detection) and the accurate recognition of their activities (activity recognition) such as "serving," "forehand," "backhand," etc. are crucial. These evaluation metrics can provide a comprehensive performance assessment.

By carefully considering these metrics during the development and evaluation of models, researchers can ensure a thorough and rigorous performance evaluation. A high-performing model should demonstrate satisfactory performance in these metrics across multiple diverse datasets.

Table 6.4 Example of a confusion matrix for activity recognition [22]

True activity (down)/ predicted activity (right)	Boxing	HandClapping	HandWaving	Jogging	Running	Walking
Boxing	0.98	0.02	0	0	0	0
HandClapping	0.09	0.91	0	0	0	0
HandWaving	0.03	0.09	0.88	0	0	0
Jogging	0	0	0	0.83	0.10	0.07
Running	0	0	0	0.08	0.88	0.04
Walking	0	0	0	0.02	0	0.98

6.7 Applications of Spatiotemporal Object Detection and Activity Recognition

The fusion of spatiotemporal object detection and activity recognition carries substantial implications for a multitude of research fields, given its capacity to automate the understanding of object interactions and activities within a spatial and temporal context [13, 17]. The following section discusses these implications in the context of surveillance systems, autonomous vehicles, smart retail, sports analysis, and elderly care, demonstrating the breadth of applicability and potential avenues for future research [23].

6.7.1 Surveillance in Urban Environments

Spatiotemporal activity recognition enhances surveillance systems by enabling predictive and anomaly detection capabilities [10]. Techniques that incorporate context and multiscale motion awareness can facilitate advanced behavioral analysis, an area of research with promising implications for smart city management and public safety [7, 21, 24, 25]. Future research might consider how to fine-tune these systems for specific urban environments or unique public safety challenges (Fig. 6.5).

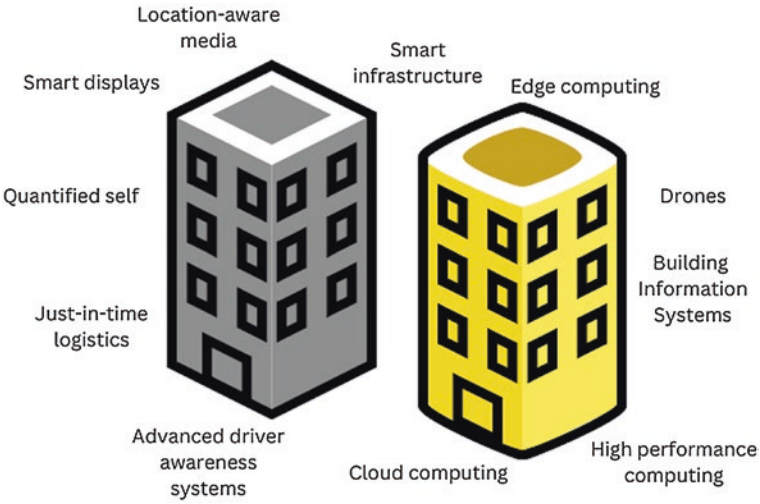

Fig. 6.5 Diagram depicting application of spatiotemporal object detection and activity recognition in urban surveillance

6.7.2 Autonomous Driving

Within the realm of autonomous driving, understanding dynamic traffic scenarios is critical, and spatiotemporal object detection offers a solution [10]. Research should continue to refine these models, ensuring they accurately recognize vehicles, pedestrians, and evolving traffic conditions. The potential for developing even more sophisticated models, capable of predicting and responding to unpredictable road scenarios, remains a compelling direction for further study (Fig. 6.6).

6.7.3 Smart Stores

The retail sector can utilize these technologies to gain detailed insights into customer behaviors and interactions [15, 20]. By analyzing customer-object interactions, retailers can optimize store layout and sales strategies. This opens a new area of research focused on refining and expanding the capabilities of these systems to provide more nuanced consumer behavior insights (Fig. 6.7).

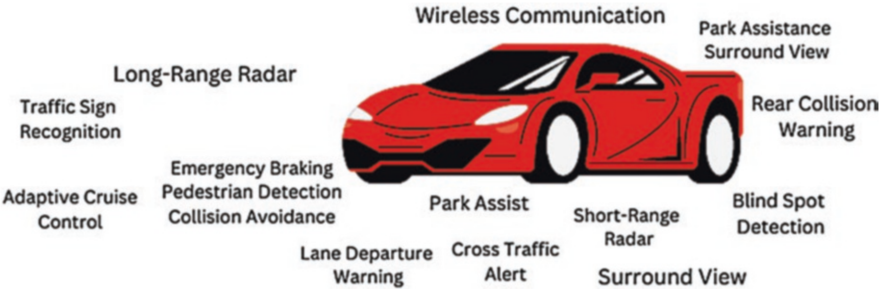

Fig. 6.6 Diagram showing the use of these technologies in autonomous vehicles

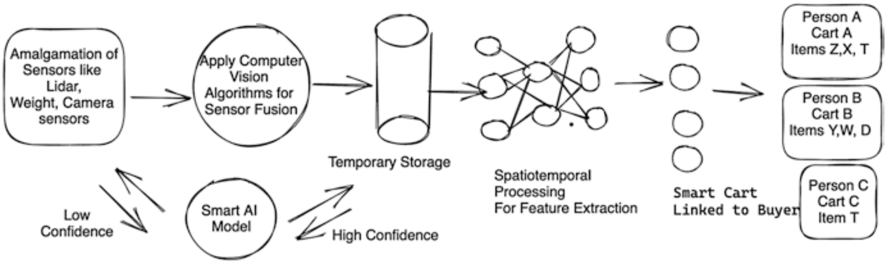
Amalgamation of Sensors like Lidar, Weight, Camera sensors
Apply Computer Vision Algorithms for Sensor Fusion
Temporary Storage
Spatiotemporal Processing For Feature Extraction
Smart Cart Linked to Buyer
Person A Cart A Items Z,X, T
Person B Cart B Items Y,W, D
Person C Cart C Item T
Low Confidence
Smart AI Model
High Confidence

Fig. 6.7 Illustration of the application of these techniques in smart retail

6.7.4 Sports Analysis

The sports industry increasingly employs spatiotemporal object detection and activity recognition for performance improvement and broadcasting [2, 18]. This application provides a rich area for exploration, particularly in the development of sport-specific algorithms or models capable of analyzing complex, rapid movements.

6.7.5 Elderly Care

These technologies offer innovative solutions for elderly care, as evidenced by Wang et al.'s [3] temporal action detection model. Future research could investigate how to further tailor these systems for individual care needs, exploring their potential in tracking health indicators, or predicting and preventing accidents.

Case Study: Wildlife Monitoring
A case study by Clapham et al. [4] displays the potential of these technologies in wildlife monitoring. They developed a deep learning model capable of identifying individual brown bears, providing invaluable data for conservation efforts. This work mirrors advancements in human activity analysis [8, 26], indicating potential for future research in animal behavior studies and conservation efforts.

As research continues to push the boundaries of these technologies, novel approaches such as TRandAugment [6], and the usage of Hough forests for action recognition [19], illuminate potential paths of exploration. By considering these findings, researchers can identify areas for advancement, fostering innovation in the application of spatiotemporal object detection and activity recognition across various sectors.

6.8 Summary

In this chapter, we have explored the fascinating field of spatiotemporal object detection and activity recognition, investigating its concepts, techniques, applications, and the significance it holds for computer vision research. Let us summarize the key findings and highlight the potential for future research.

Beginning with the fundamentals in Sect. 6.2, we emphasized the importance of spatiotemporal analysis in understanding object interactions and activities within dynamic environments. We discussed the challenges posed by temporal variations and the need for effective models that maintain spatial and temporal connectivity across frames.

Moving forward, Sect. 6.3 focused on spatiotemporal object detection, where we examined various techniques, including region-based methods, motion cues, and two-stream networks. These approaches enable the detection of objects and their trajectories over time, contributing to improved scene understanding and object recognition accuracy.

Section 6.4 delved into activity recognition, where we explored the techniques used to recognize and categorize human activities in videos. We discussed the use of local spatiotemporal patterns, context-awareness, and multiscale motion awareness to enhance activity understanding. These approaches have shown promising results in accurately recognizing and interpreting human actions in dynamic scenes.

In Sect. 6.5, we introduced the concept of hierarchical models for spatiotemporal object detection and activity recognition. These models leverage multi-level representations and preserve contextual information across different scales, leading to improved accuracy and effectiveness in complex scenes.

Section 6.6 delved into performance evaluation metrics, highlighting their importance in assessing the performance of object detection and activity recognition models. We discussed various metrics used to evaluate the accuracy, precision, recall, and F1 score of these models, providing researchers with tools to compare and benchmark their approaches effectively.

Furthermore, in Sect. 6.7, we explored the diverse applications of spatiotemporal object detection and activity recognition, including surveillance systems, autonomous driving, and smart stores. These applications have witnessed significant advancements in situational awareness, decision-making, and safety, showcasing the transformative impact of spatiotemporal analysis in various domains.

Considering these key points, it is evident that spatiotemporal object detection and activity recognition play a pivotal role in computer vision research. They provide the means to comprehend complex scenes, identify objects, track their interactions, and infer activities in a spatial and temporal context.

Looking ahead, the potential for further advancements in this field is vast. Researchers can focus on developing more robust and efficient models that can handle real-world challenges such as occlusions, scale variations, and complex environmental conditions. Exploring novel applications, such as fine-grained activity detection or behavior prediction in specific domains, can lead to significant contributions. Moreover, there is a growing need to address interpretability and explainability in spatiotemporal models to ensure transparency and build trust with end-users.

In conclusion, spatiotemporal object detection and activity recognition have emerged as vital components of computer vision research, enabling us to extract meaningful information from dynamic scenes. Their impact extends beyond academia, reaching various industries and domains. By fostering collaboration, sharing knowledge, and embracing the challenges, we can drive innovation, refine techniques, and unlock the full potential of spatiotemporal analysis.

References

1. Pei, T., Huang, Q., Wang, X., Chen, X., Liu, Y., Song, C., … Zhou, C. (2021). Big geodata aggregation: Connotation, classification, and framework. National Remote Sensing Bulletin, 25(11), 2153–2162. doi: https://doi.org/10.11834/jrs.20210480
2. Liu, Y., & Jing, H. (2022). A Sports Video Behavior Recognition Using Local Spatiotemporal Patterns. Mobile Information Systems, 2022. doi: https://doi.org/10.1155/2022/4805993
3. Wang, K., Li, X., Yang, J., Wu, J., & Li, R. (2021). Temporal action detection based on two-stream You Only Look Once network for elderly care service robot. International Journal of Advanced Robotic Systems, 18(4). doi: https://doi.org/10.1177/17298814211038342
4. Clapham, M., Miller, E., Nguyen, M., & Darimont, C. T. (2020). Automated facial recognition for wildlife that lack unique markings: A deep learning approach for brown bears. Ecology and Evolution, 10(23), 12883–12892. doi: https://doi.org/10.1002/ece3.6840
5. Akilan, T. (2018). Video foreground localization from traditional methods to deep learning (Doctoral dissertation, University of Windsor (Canada)).
6. Ramesh, S., Dall'Alba, D., Gonzalez, C., Yu, T., Mascagni, P., Mutter, D., Padoy, N. (2023). TRandAugment: temporal random augmentation strategy for surgical activity recognition from videos. International Journal of Computer Assisted Radiology and Surgery. doi: https://doi.org/10.1007/s11548-023-02864-8
7. Cardoso, D. B., Campos, L. C. B., & Nascimento, E. R. (2022). An Action Recognition Approach with Context and Multiscale Motion Awareness. In Proceedings - 2022 35th Conference on Graphics, Patterns, and Images, SIBGRAPI 2022 (pp. 73–78). Institute of Electrical and Electronics Engineers Inc. doi: https://doi.org/10.1109/SIBGRAPI55357.2022.9991807
8. SankaranNampoothiri, S., & Anoop BK (2014). Review on Vision based Human Activity Analysis. International Journal of Computer Applications, 99(2), 9–14. doi: https://doi.org/10.5120/17343-6240
9. Aakur, S., Sawyer, D., Balazia, M., & Sarkar, S. (2020). An examination of proposal-based approaches to fine-grained activity detection in untrimmed surveillance videos. In 2018 TREC Video Retrieval Evaluation, TRECVID 2018. National Institute of Standards and Technology (NIST).
10. Sun, J., Wu, X., Yan, S., Cheong, L. F., Chua, T. S., & Li, J. (2009). Hierarchical spatio-temporal context modeling for action recognition. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009 (pp. 2004–2011). IEEE Computer Society. doi: https://doi.org/10.1109/CVPRW.2009.5206721
11. Wang, J., Chen, Z., & Wu, Y. (2011, June). Action recognition with multiscale spatio-temporal contexts. In CVPR 2011 (pp. 3185-3192). IEEE.
12. Ahsan, U., Madhok, R., & Essa, I. (2019, January). Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 179-189). IEEE.
13. Liu, L., Shao, L., Li, X., & Lu, K. (2016). Learning spatio-temporal representations for action recognition: A genetic programming approach. IEEE Transactions on Cybernetics, 46(1), 158–170. doi: https://doi.org/10.1109/TCYB.2015.2399172.
14. Haroon Idrees, Khurram Soomro and Mubarak Shah, Detecting Humans in Dense Crowds using Locally-Consistent Scale Prior and Global Occlusion Reasoning, Pattern Analysis and Machine Intelligence (PAMI), IEEE Transactions, 2015.
15. Yu, T. W., Sarwar, M. A., Daraghmi, Y. A., Cheng, S. H., Ik, T. U., & Li, Y. L. (2022). Spatiotemporal Activity Semantics Understanding Based on Foreground Object Segmentation: iCounter Scenario. IEEE Access, 10, 57748–57758. doi: https://doi.org/10.1109/ACCESS.2022.3178609

16. Xia, L., & Aggarwal, J. K. (2013). Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2834-2841).
17. Song, S., Lan, C., Xing, J., Zeng, W., & Liu, J. (2018). Spatio-Temporal Attention-Based LSTM Networks for 3D Action Recognition and Detection. IEEE Transactions on Image Processing, 27(7), 3459–3471. doi: https://doi.org/10.1109/TIP.2018.2818328
18. Kanimozhi, S., Mala, T., Kaviya, A., Pavithra, M., & Vishali, P. (2022). Key Object Classification for Action Recognition in Tennis Using Cognitive Mask RCNN. In Lecture Notes in Networks and Systems (Vol. 287, pp. 121–128). Springer Science and Business Media Deutschland GmbH. doi: https://doi.org/10.1007/978-981-16-5348-3_9
19. Gall, J., Yao, A., Razavi, N., Van Gool, L., & Lempitsky, V. (2011). Hough forests for object detection, tracking, and action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(11), 2188–2202. doi: https://doi.org/10.1109/TPAMI.2011.70
20. Liu, X., You, T., Ma, X., & Kuang, H. (2018). An optimization model for human activity recognition inspired by information on human-object interaction. In Proceedings - 10th International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2018 (Vol. 2018-January, pp. 519–523). Institute of Electrical and Electronics Engineers Inc. doi: https://doi.org/10.1109/ICMTMA.2018.00131
21. Yadav, P., Salwala, D., Das, D. P., & Curry, E. (2020). Knowledge graph driven approach to represent video streams for spatiotemporal event pattern matching in complex event processing. International Journal of Semantic Computing, 14(3), 423–455. doi: https://doi.org/10.1142/S1793351X20500051
22. M. H. Rahman and N. Bouguila, "Efficient Feature Mapping in Classifying Proportional Data," in IEEE Access, vol. 9, pp. 3712-3724, 2021, doi: https://doi.org/10.1109/ACCESS.2020.3047536.
23. Liz Oz, Always AI (2022), 17 interesting applications of Object Detection for businesses https://alwaysai.co/blog/object-detection-for-businesses June 4, 12.30PM PST
24. Abdellah Chehri, Hussein T. Mouftah, Autonomous vehicles in the sustainable cities, the beginning of a green adventure, Sustainable Cities and Society, Vol 51, 2019, 101751, ISSN 2210-6707, doi: https://doi.org/10.1016/j.scs.2019.101751.
25. Torrens PM. Smart and Sentient Retail High Streets. Smart Cities. 2022; 5(4):1670-1720. doi: https://doi.org/10.3390/smartcities5040085.
26. Lee J, Ahn B. Real-Time Human Action Recognition with a Low-Cost RGB Camera and Mobile Robot Platform. Sensors. 2020; 20(10):2886. doi: https://doi.org/10.3390/s20102886