CHAPTER

19

Vision-based action recognition for the human-machine interaction

Ankush Verma[a], *Vandana Singh*[b], *Amit Pratap Singh Chouhan*[a], *Abhishek*[c], *and Anjali Rawat*[d]

[a]Department of Radiology, Sharda School of Allied Health Sciences, Sharda University, Greater Noida, Uttar Pradesh, India [b]Department of Microbiology, Sharda School of Allied Health Sciences, Sharda University, Greater Noida, Uttar Pradesh, India [c]Department of Radiology, Kalaish Institute of nursing and Paramedical sciences. Greater Noida, Uttar Pradesh, India [d]Department of Optometery, Sharda School of Allied Health Sciences, Sharda University, Greater Noida, Uttar Pradesh, India

19.1 Introduction

Vision-based action recognition plays a significant role in increasing human-machine interaction (HMI) by permitting contemporary machines to interpret and react to human gestures and movements [1]. This vision-based action recognition technology comprises the use of computer vision algorithms and mathematical approaches to evaluate and to interpret visual data, such as video feeds or picture sequences, to distinguish particular actions done by humans [2]. The identification process often starts with the collection of visual information, which may be in the form of video footage, gestures acquired by cameras or other imaging devices. The next phase of vision-based recognition entails preprocessing the visual input to increase its quality and extract essential elements [3]. Feature extraction by machines is an important aspect of action recognition, as it helps to identify key patterns and characteristics associated with different actions. Following preprocessing and feature extraction from the visual data, machine learning models—which are frequently trained on labeled datasets—come into play. Based on the traits that have been retrieved, these models are intended to categorize and identify different human acts [4]. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid models that combine CNN and RNN methods are popular machine learning techniques for action recognition. Vision-based action recognition has a wider range of natural applications in HMI [5]. Vision-based action recognition may be used in healthcare applications to track, examine, and document patient movements in order to aid in their rehabilitation [6]. As well as in the realm of augmented reality (AR) and virtual

Artificial Intelligence and Multimodal Signal Processing in Human-Machine Interaction
https://doi.org/10.1016/B978-0-443-29150-0.00011-1

reality (VR), this technology may improve immersive experiences by letting computers to react dynamically and correctly to human activities. Because of the enormous potential of HMI, research is still ongoing to develop vision-based action recognition systems for HMI. These advancements in systems must overcome obstacles like real-time processing, managing complex and dynamic environments, and guaranteeing robustness to changes in lighting and background [7]. As new technology continues to progress, this sector provides the potential for generating more seamless, accurate, and engaging interfaces between people and machines. Vision-based action recognition systems incorporate a number of sophisticated and current approaches to increase accuracy and performance. One of the most popular strategies in HMI is to mix geographical and temporal information. Spatial information in machine interaction entails examining the static aspects of a picture, such as the stance or arrangement of body parts, whereas temporal information focuses on the dynamics of the activity across time [8]. Making optimal use of 3D models to imitate human motions is another essential component with the use of depth data, which may be gathered via sophisticated technology such as depth-sensing cameras or stereo vision; these models can more accurately portray object in three dimensions. The system's ability to accurately recognize complex gestures and movements is improved by the recording of 3D natural activity. For vision-based action recognition to be more accurate and efficient, attention processes must be included [9]. Vision-based action recognition in HMI has a wide range of growing real-world applications. This technology aids in the development of gesture-controlled interfaces that are integrated into automobiles, improving driver convenience and safety [10]. In the retail sector, it might be used to improve the shopping experience, optimize store layouts, and analyze customer behavior. Even with all of this development, research is still being done to address issues like biases in the datasets, making sure that models are resilient across a range of demographics, and enhancing the interpretability of the decision-making processes in these systems. The area is developing, and as advanced technologies are incorporated into daily life, vision-based action recognition has the potential to completely transform HMIs in a variety of contexts [11].

19.2 Fundamentals of vision-based action recognition

Vision-based action recognition is an expanding field within computer vision that aims to translate and grasp human actions or activities through the analysis of visual data captured by cameras or other imaging devices inbuilt in the machine [12]. This process involves the acquisition of visual information, typically in the form of images or videos, and further followed by crucial steps in feature extraction, temporal modeling, and action classification. The first stage consists of acquiring visual data via cameras, with successive preprocessing processes geared at refining and boosting the input quality. This preprocessing method assures that the future analysis is based on clean and useful visual information. The process of discovering significant and intricate patterns within the visual data begins with the preprocessing of the data and continues with the feature extraction phase. For the extraction phase, methods such as optical flow analysis, CNNs, and histogram of oriented gradients (HOG) are often used. This is a critical stage because it establishes the framework for capturing discriminative attributes that aid in distinguishing between different movements or activities [13]. A fundamental component of vision-based action recognition is temporal modeling, which understands that actions take place throughout time. To show the usage or advantages of this temporal dimension, techniques like RNNs, long short-term memory (LSTM) networks, or 3D CNNs are typically included. The recognition system is made more resilient by this temporal modeling, which offers a more accurate and effective representation of dynamic activities [14]. Action classification is the last stage of vision-based recognition, in which the observed event is given a label or category based on the recognized characteristics and temporal information. This classification phase helps provide a significant analysis of the visual data, converting unprocessed visual data into actionable insights on human behavior [15]. To put it briefly, the first phase involves employing cameras or other imaging equipment to collect visual data. The quality of this raw data significantly influences the subsequent stages of the process. Preprocessing steps, such as noise reduction, image stabilization, and normalization, are crucial for refining and clarity of the input data [16]. By taking these steps, you can guarantee that the analysis that comes after is founded on precise and consistent visual data, which paves the way for trustworthy action recognition [17].

- **Feature extraction:** Feature extraction is an essential step where the system/machine identifies and captures relevant patterns within the visual data [18]. Techniques like CNNs' excel software help in capturing hierarchical and spatial features, while optical flow analysis is effective in understanding the motion dynamics within a sequence. The choice of feature extraction method is contingent upon the specific requirements of the action recognition task.
- **Temporal modeling:** Understanding the temporal progress of actions is critical for accurate and efficient vision-based recognition. While 3D CNNs bring classic 2D convolutional procedures into the temporal dimension, RNNs, with their recurrent connections, are adept at capturing temporal relationships [19]. The capacity of the system to handle dynamic and complicated scenarios is enhanced by temporal modeling, which guarantees that the system can discern between activities that unfold over time.
- **Action classification:** The final stage of vision-based recognition involves assigning a label or category to the recognized action. This multilayered approach enables computers to interpret and understand human actions, paving the way for applications in fields like surveillance, human-computer interaction, gesture analysis, and sports analysis [19].

19.3 Human-machine interaction

HMI is a dynamic, growing, and diverse area focusing on the seamless integration of people and machines, seeking to improve the overall user experience and aid in bridging the gap between human cognitive capacities and machine functionalities [20]. HMI attempts to build spontaneous, accurate, and efficient interfaces that facilitate effective communication between humans and machines. This connection of human-machine goes beyond standard input/output systems to embrace a variety of modalities, including touchscreens, speech recognition, gesture control, and even brain-computer interfaces. The objective of HMI is to build interfaces that not only recognize human intents and actions but also adapt to user preferences and change with the user's cognitive processes over time. The capacity of computers to perceive and respond to human directions and gestures is made possible by the crucial roles played by machine learning, computer vision, and natural language processing in HMI systems [21]. The developments in haptic feedback technologies lead to providing more immersive and tactile experiences, boosting the user's sensation of engagement with the machine. This immersive sort of engagement has great promise across several sectors, including gaming, education, vehicle, healthcare, and training simulations. In healthcare, for instance, HMI helps the construction of surgical training simulators, while in education, vision-based recognition HMI provides interactive learning experiences. A significant field of research called HMI influences how humans connect with technology as well as how robots are absorbed into society at large and how we live our daily lives. The future of HMI will be characterized by how well-integrated and user-friendly HMI systems are as technology develops [22]. This will be a major aspect in choosing the acceptance and success of new technologies.

19.4 Role of vision in HMI

In HMI, vision plays a crucial role as a basis for designing intuitive and effective interfaces that allow seamless interaction or communication between people and machines [23]. For computers to detect and react to visual inputs from humans, vision plays a significant part in input modalities (Table 19.1). This is the integration of computer vision technologies with humans that enable computers to grasp gestures, facial emotions, and even eye movements. For instance, by recognizing people and customizing user interfaces to their preferences, facial recognition technology provides personalized user experiences. The development of augmented reality (AR) and virtual reality (VR) applications relies on computer vision algorithms, as the merging of the digital and physical worlds requires a promising knowledge of the visual environment with humans [24]. Through the use of vision-based technology, augmented reality (AR) boosts situational awareness and gives contextually relevant information by superimposing digital data over the real-world perspective. In VR, vision is vital for constructing realistic environments by detecting head and body movements, letting users to interact with digital regions as if they were present in the human body. The role of vision in HMI becomes increasingly significant in applications connected to autonomous systems. In self-driving cars, vision sensors, such as cameras and LiDAR, enable the vehicle to interpret information about its surroundings, identify obstructions, and make real-time judgments for safe navigation. Vision-based systems help with tasks such as object recognition, navigation, and manipulation in robotics, enabling robots to interact with their surroundings intelligently [25]. As HMI continues to evolve, the integration of visual technologies is going beyond standard computer displays, embracing more natural, effective, accurate, and immersive modes of interaction. The collaborative coupling of computer vision with HMI not only increases user experiences but also brings up new chances

TABLE 19.1 Showing different feature and description of human-machine interaction.

Feature	Description
Object recognition	Employing computer vision techniques to recognize and classify things in the surrounding environment
Gesture recognition	Examining and interpreting bodily and hand gestures to determine the intents of the user
Facial expression analysis	Identifying and comprehending facial expressions in order to engage users and recognize emotions
Hand-object interaction	Ability to recognize and decipher interactions between the user's hands and scene elements
Visual feedback	Provide people with visual feedback in real time based on their activities or system reactions
Environmental awareness	Raising awareness of the external environment in order to modify system behavior

for applications across numerous disciplines, from healthcare and education to entertainment and industrial automation. The expanding significance of vision in HMI underscores its relevance in shaping the future of HMI and cooperation and the way we connect with technology.

19.5 Action recognition techniques

A broad array of ways are employed in action recognition techniques to evaluate and grasp human behaviors from visual input, such as movies or image sequences. CNNs are especially excellent at extracting hierarchical features, which allows them to discern complicated patterns within individual frames—a vital capacity for action recognition [26]. Another essential way is to design LSTM or RNNs, which are suited for temporal modeling. Due to the sequential pattern of activities, temporal dependencies are well recorded by RNNs and LSTMs, enabling the system to grasp how actions dynamically change over time. 3D CNNs expand the classic 2D convolutional processes into the temporal domain, improving the efficacy of data by integrating both spatial and temporal information for a more thorough comprehension of activities. The employment of feature-based methodologies, such as optical flow analysis and HOG, provides replacement ways for obtaining essential data from video material. HOG focuses on spatial gradient information data, whereas optical flow examines the motion patterns between successive frames (Fig. 19.1). A complete approach to action recognition

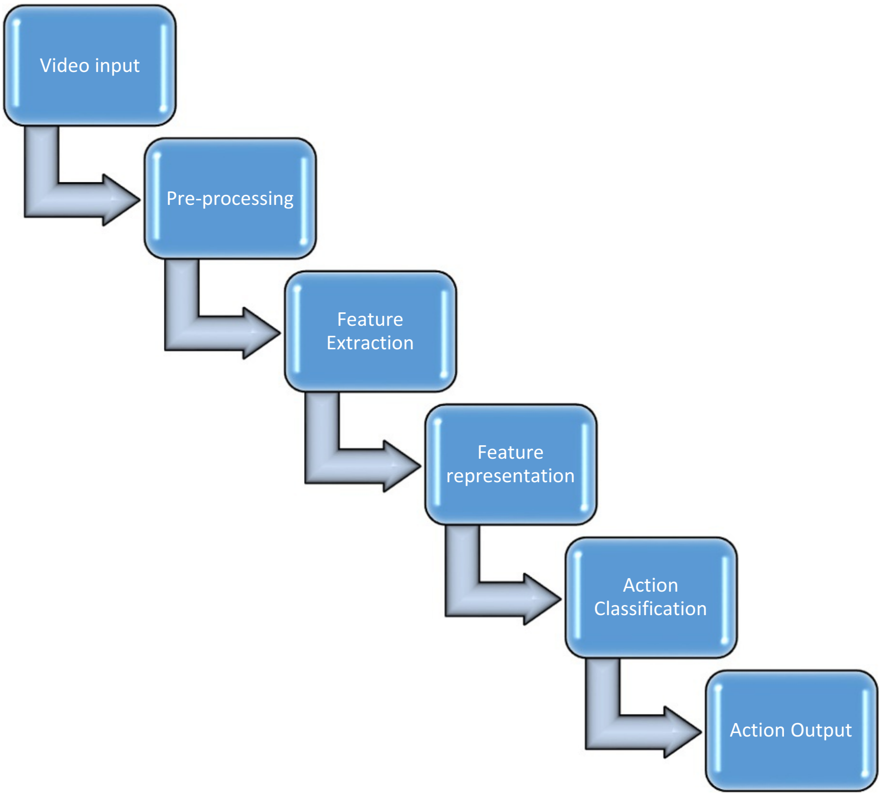

FIG. 19.1 Showing action recognition technique.

is made feasible by the integration of these numerous approaches, which integrate temporal and geographical information to produce trustworthy and exact conclusions. The fusion of these techniques and the investigation of multimodal approaches, incorporating information from different sensor modalities, can further enhance the skills of action recognition systems across various modern and advanced applications, from video surveillance and human-computer interaction to sports analysis and healthcare monitoring [27].

19.5.1 Classical approaches

Classical approaches in action recognition techniques mainly involve handcrafted feature extraction and traditional machine learning algorithms, offering a distinct paradigm when compared to the deep learning methods that have gained popularity in recent years. One classical technique uses the spatiotemporal features derived from object motion information within video sequences. Despite the rise of deep learning, classical techniques remain relevant, and the fusion of classical approaches with modern approaches is an active area of research, aiming to combine the interpretability, accuracy, and efficiency of classical methods with the realistic power of deep neural networks for improved action recognition across diverse and modern applications.

- **Rule-based systems:** Rule-based systems play a key role in action recognition techniques because they offer a structured and interpretable framework for modeling and analyzing human behaviors from visual input. Rule-based systems in action recognition may manage linguistic representations, utilizing natural language rules to define action recognition patterns. One issue with rule-based systems resides in their scalability and flexibility to handle various, vast datasets and with sophisticated behaviors. Rule-based systems, despite these problems, offer interpretability and transparency, which could be crucial in instances where grasping the decision-making process is vital. Rule-based systems may be used in tandem with other methods, such as contemporary machine learning classifiers, to develop hybrid models that incorporate the benefits of both approaches for action recognition [28].
- **Template matching:** Template matching is a traditional and inbuilt approach applied in action recognition to recognize and discover predetermined patterns or templates inside video or picture sequences. Template matching is theoretically uncomplicated and quick to implement; however, its efficacy relies on the quality and representativeness of the templates, making it suited for cases where actions are unique in visual patterns. Template matching can be used for both spatial dimensions and temporal dimensions. While temporal template matching expands the method to take into account the sequential evolution of patterns over numerous frames, reflecting the dynamic character of activities, spatial template matching focuses on finding patterns inside individual frames [29]. Template matching is still a useful approach, particularly in hybrid models when it is supplemented with other techniques. To improve resilience and flexibility to various contexts, it might be included, for instance, into a more comprehensive action recognition pipeline that incorporates feature extraction or machine learning components [30].

19.5.2 Machine learning-based approaches

Machine learning-based approaches have emerged as modern, accurate, powerful, and versatile techniques in action recognition, revolutionizing the field by enabling automated learning of discriminative features and patterns directly from data [31]. One prevalent method of machine learning is the application of deep learning architectures, particularly CNNs and RNNs and hybrid models. These techniques can be used to continue to evolve; they hold immense potential for applications ranging from surveillance and human-computer interaction to healthcare and sports analytics, contributing to the development of more intelligent and context-aware systems [32].

- **Supervised learning:** Supervised learning is a fundamental model in action recognition techniques, playing an important role in training machine models to recognize and classify human actions based on labeled training data [33]. Supervised learning involves combining the algorithm with a dataset consisting of video sequences or image frames where each instance is interpreted with the equivalent action label by the machine. This algorithm knows how to associate the visual features extracted from the input data with the specific action categories during the training phase. Supervised models may lack functions when challenged with variations in viewpoint, lighting conditions, or complex interactions that were not adequately represented in the training data [34]. Supervised learning remains a basic in action recognition, providing a crucial and interpretable framework for training models that can accurately and efficiently identify and classify human actions across various applications, from video surveillance and human-computer interaction to sports analytics and healthcare monitoring.
- **Unsupervised learning:** Unsupervised learning has greatly come under the headings of action recognition techniques as an alternative pattern to supervised learning, offering valuable insights into extracting patterns and structures from unlabeled or partially labeled data. Unsupervised learning has the main objective of uncovering inherent structures and representations within video data without relying on preexisting action labels. Unsupervised learning in action recognition is particularly advantageous in scenarios where labeled training data is rarely analyzed and difficult to obtain. The effectiveness of unsupervised learning in vision-based action recognition continues to be a subject of ongoing research, and the integration of unsupervised techniques with supervised learning or semisupervised approaches holds promise for developing more full-bodied and versatile action recognition systems capable of handling real-world complexities [35].

19.5.3 Deep learning approaches

Deep learning approaches have revolutionized the field of action recognition, offering modern and advance methods for automatically learning higher representations and discerning complex patterns from raw visual data. These techniques have been successfully applied across diverse domains, from surveillance and human-computer interaction to sports analytics and healthcare monitoring to know more about their depth and strength [36]. The capabilities of deep learning-based action recognition systems should continue to grow as deep learning research advances due to advancements in multimodal integration, architecture design, and model interpretability. These developments will help create applications that are more intelligent and context wise.

- **Convolutional neural networks:** CNNs have emerged as a cornerstone in vision-based action identification approaches, giving a strong and effective mechanism to automatically extract discriminative spatial characteristics from visual input. The use of CNNs' ability to learn spatial hierarchies accounts for much of their efficacy in action recognition. CNNs automatically recognize complicated visual information via the use of numerous convolutional layers, ranging from basic edges and textures to more abstract and sophisticated representations connected to specific actions. 3D CNNs are perfectly suited for action recognition in video sequences as they are a temporal extension of conventional CNNs. Through the inclusion of filters in both spatial and temporal dimensions, 3D CNNs can capture spatiotemporal elements that are critical for interpreting dynamic activities. CNNs have been demonstrated to be incredibly successful in several action detection applications, such as sports analytics, video surveillance, and human-computer interaction.

- **Recurrent neural networks:** In action recognition algorithms, RNNs are crucial because they imitate the temporal dynamics present in sequential data, including video sequences. RNNs add a cyclical structure, which distinguishes them from normal feed-forward neural networks and allows them to keep hidden states and record associations across time steps [37]. A special sort of RNN called LSTM networks is commonly applied in action recognition because it may decrease the issues caused by disappearing and inflating gradients, making it simpler to explain long-range temporal correlations in video data. When employing RNNs for action recognition, feature sequences obtained from video frames or optical flow data that illustrates motion patterns are typically employed as inputs. When it comes to action recognition approaches, RNNs are a helpful tool since they can replicate the temporal evolution of actions and serve as a base for more sophisticated, context-aware systems that can accurately recognize and classify human activities in changing contexts.
- **Hybrid models:** To improve performance and flexibility, hybrid models in action recognition methods are a synergistic mixing of different neural network architectures or machine learning algorithms. The merging of RNNs with Convolutional Neural Networks (CNNs) is one prominent hybrid technique [38]. Although RNNs are more proficient at modeling temporal connections in sequential data, CNNs are excellent at extracting spatial attributes from image or video frames. Combining these architectures enables hybrid models to extract discriminative spatial patterns while also representing the dynamic nature of actions across time, resulting in more robust and accurate action recognition systems. CNNs are often utilized in these hybrid architectures to extract spatial features, which collect fine visual information from each frame. Hybrid models in action recognition methods give the flexibility and adaptability essential to manage the nuances of real-world occurrences [39].

19.6 Applications of vision-based action recognition

Vision-based action recognition has found applications in many fields, exploiting the capacity to detect and understand human actions from visual input (Fig. 19.2). Some noteworthy uses are as follows:

- **Surveillance and security**: In video surveillance systems, identification of actions based on vision is crucial. It makes it feasible for suspicious activities or anomalous behavior in crowded environments to be automatically recognized and classed, increasing security monitoring in public locations, transportation hubs, and essential infrastructure [40].
- **Human-computer interaction (HCI):** Natural and intuitive interactions between people and computers are enabled by vision-based action recognition in HCI. In interactive systems and virtual

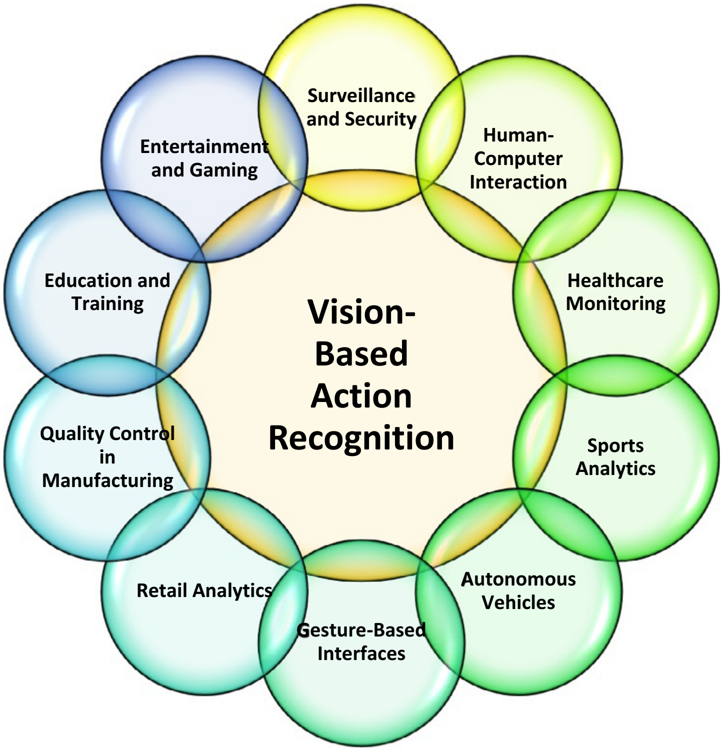

FIG. 19.2 Showing applications of vision-based action recognition.

environments, for example, gesture recognition enhances user experience by allowing users to manipulate devices or programs using hand gestures.

- **Healthcare monitoring**: Vision-based activity recognition assists in patient monitoring and rehabilitation in the medical area. To enhance patient care, systems may monitor vital signs, assess physical therapy exercises based on body movement analysis, and warn medical specialists of any odd or harmful behavior [41].
- **Sports analytics:** In sports analytics, vision-based action recognition is routinely applied to assess and grasp player performance, actions, and strategy. Coaches and analysts may gain important information by utilizing it to examine and assess a range of activities, including goal scoring, player positioning, and overall team dynamics.
- **Autonomous cars**: Vision-based action recognition is vital for interpreting the surrounding environment in the development of autonomous cars. It makes interaction possible to recognize bicycles, pedestrians, and other vehicles, which supports the automobile in making judgment calls for safe navigation and accident avoidance [42].
- **Gesture-based interfaces**: Motion-based user interfaces make use of gesture recognition, a subset of vision-based action recognition. Devices that enable users to interact and communicate with the interface via hand or body gestures, such as smart TVs and gaming consoles, exploit this technology.

- **Retail analytics**: Vision-based action recognition is utilized in retail contexts to assess customer behavior. It helps retailers to better evaluate client interactions with merchandise, quantify foot circulation inside shops, and improve store layouts for enhanced consumer engagement and enjoyment.
- **Quality control in manufacturing**: Vision-based action recognition helps to control quality in industrial contexts by examining activity on production lines. It may discover faults, monitor machine performance, and check that safety rules are obeyed, thereby enhancing overall efficiency and product quality.
- **Education and training**: Vision-based action recognition is utilized in educational settings to enhance interactive learning experiences and training simulations. It provides real-time feedback on physical activities such as good exercise forms or hands-on education in fields such as surgery or vocational skills.
- **Entertainment and gaming**: Vision-based action recognition is utilized in the entertainment industry to enhance gaming experiences by allowing players to move characters and interact with virtual environments using their bodies. It provides virtual reality applications and games with a more immersive quality [43]. Vision-based action recognition systems' versatility stimulates innovation in a broad variety of applications, delivering solutions that increase user engagement, safety, and efficiency in many disciplines.

19.7 Challenges and future directions

Despite tremendous advancements, vision-based action recognition still confronts several obstacles that will continue to influence future developments in the area and research paths [44]. Some key challenges and potential directions for the future include:

- **Data variability**: Datasets used to train vision-based action recognition models may be insufficiently diverse in terms of surroundings, lighting conditions, and demography, resulting in models that fail to generalize to real-world scenarios. The future direction of data variability in producing more diverse and representative datasets may assist in designing models that are more robust and widely applicable. Examples of these datasets include information from a variety of cultural situations, camera viewpoints, and demographics.
- **Temporal variability:** Temporal variability is still tough to record and comprehend the temporal dynamics of action sequences, particularly when the actions are complicated and detailed. More complete temporal modeling is necessary to account for changes in the length, pace, and sequence of activities. The future route of models may be better able to detect temporal fluctuations in activities by investigating more intricate temporal modeling approaches, such as spatiotemporal graph networks, attention processes, and long-term dependency modeling [42].
- **Real-time processing:** Real-time processing faces certain challenges; computational hurdles arise when attempting real-time processing, particularly in applications such as live monitoring or autonomous cars. High-precision deep learning models might require a large computing budget. Future challenges of real-time processing will be essential to investigate model compression methods, create effective architectures, and make use of hardware acceleration (such as GPUs and TPUs) to provide real-time processing without sacrificing accuracy.
- **Depth and multimodal integration:** Challenges with fusion techniques and

multimodal data processing arise when depth information and other modalities (such as inertial sensors) are integrated into vision-based action recognition systems. For the future challenges of depth and multimodal integration to increase identification accuracy, it will be crucial to look into efficient fusion methods like early, late, or hybrid fusion and how depth information might support visual signals.

- **Fine-grained action recognition:** The challenges for vision-based models, distinguishing between subtle and fine-grained motions, are still a difficult issue, especially in crowded or complicated settings. The future challenges are comprehension and categorization of complex actions which can be improved by developing methods for fine-grained action recognition that use attention processes and high-resolution data.
- **Explainability and interpretability:** It is difficult to comprehend the logic behind the predictions made by deep learning models, especially sophisticated neural networks, because they are not interpretable [45]. The future challenges are the development of interpretability frameworks, visualization tools, and explainable AI approaches which will improve the transparency and reliability of vision-based action recognition systems, especially in crucial applications such as security or healthcare.
- **Limited labeled data:** The challenges of limited labeled data are labor-intensive process of labeling action sequences in big datasets though annotation restricts the availability of different and comprehensive datasets. In the future, to mitigate the problems caused by a lack of labeled data, one might look at semisupervised and unsupervised learning strategies as well as effective data augmentation methods.
- **Adaptability to new environments:** Vision-based action recognition models are less flexible when used in real-world deployments because they frequently have difficulty adjusting to novel and unfamiliar contexts, and the future directions. To ensure that models perform well in a variety of situations, research on domain adaptation approaches and transfer learning procedures can improve the models' capacity to adapt to new surroundings [46].
- **Ethical considerations and bias:** Due to the nature of the training data, vision-based action recognition systems may display biases that might give rise to unethical situations and unfair behaviors [47]. The future challenges in order to ensure responsible deployment, it will be essential to address biases through impartial and fair data gathering, model review, and the development process.
- **Human-computer interaction challenges:** The challenge in human-computer interaction is necessary to handle user privacy concerns, provide natural and intuitive interactions, and take accessibility for a variety of user groups into account when designing vision-based action recognition systems for seamless human-computer interaction. The future direction for vision-based action recognition in human-computer interaction will improve by concentrating on privacy-preserving methods, user-centered design, and inclusive strategies that accommodate users with varying skills [48].

19.8 Conclusion

Vision-based action recognition has undergone great development, pushed by breakthroughs in deep learning, computer vision, and the availability of large-scale datasets. From human-computer interaction and sports analytics to surveillance and healthcare, this technology offers immense potential for usage in a broad variety of disciplines. Despite the field's tremendous advances, several challenges still need to be handled, such as the demand for real-time processing, data variability, and the complexity of temporal modeling. For vision-based action recognition to continue evolving and being utilized, it will be vital to overcome these challenges and plan out future objectives, such as extending dataset variety, enhancing temporal modeling methodologies, and obtaining real-time efficiency. Researchers, practitioners, and policymakers working together will be important in determining how vision-based action recognition grows in the future and realizing its promise to reform sectors, promote safety, and enhance people's lives in general. The combination of cutting-edge technology and diversified cooperation gives the possibility to establish new ground in the continuing domain of vision-based action recognition. Decentralized and real-time action recognition systems with applications in smart cities, autonomous cars, and responsive environments may be ushered in by the convergence of vision-based recognition with edge computing, the Internet of Things (IoT), and 5G connection. It is critical to comprehend how models arrive at certain predictions, particularly in applications where clinical acceptability depends on interpretability, such as healthcare diagnostics. Interactive storytelling, training simulations, and gaming may all be redefined by immersive experiences that react to users' movements and actions. Globally adaptable and inclusive models will benefit from the inclusion of domain adaptation techniques, transfer learning, and data handling procedures from many populations and countries. In order to create intelligent systems that can see, comprehend, and react to human activities in a variety of ways and promote safer, more effective, and more engaging settings in both everyday life and industry, vision-based action recognition is poised to play a revolutionary role.

References

[1] L. Guo, Z. Lu, L. Yao, Human-machine interaction sensing technology based on hand gesture Recognition: A Review, IEEE Trans. Hum.-Mach. Syst. 51 (2021) 300–309, https://doi.org/10.1109/THMS.2021.3086003.

[2] D.R. Beddiar, B. Nini, M. Sabokrou, A. Hadid, Vision-based human activity recognition: A survey, Multimed. Tools Appl. 79 (41) (2020) 30509–30555.

[3] A. Wang, W. Zhang, X. Wei, A review on weed detection using ground-based machine vision and image processing techniques, 2019.

[4] C. Lea, M.D. Flynn, R. Vidal, A. Reiter, G.D. Hager, Temporal Convolutional Networks for Action Segmentation and Detection, 2017, pp. 156–165.

[5] H.B. Zhang, Y.X. Zhang, B. Zhong, Q. Lei, L. Yang, J.X. Du, D.S. Chen, A comprehensive survey of vision-based human action recognition methods, Sensors 19 (5) (2019) 1005.

[6] M.A.R. Ahad, A.D. Antar, O. Shahid, Vision-Based Action Understanding for Assistive Healthcare: A Short Review, CVPR Workshops, 2019, pp. 1–11.

[7] M. Ramanathan, W.-Y. Yau, E.K. Teoh, Human action Recognition with video data: research and evaluation Challenges, IEEE Trans. Hum.-Mach. Syst. 44 (2014) 650–663, https://doi.org/10.1109/THMS.2014.2325871.

[8] J.K. Aggarwal, S. Park, Human motion: modeling and recognition of actions and interactions, in: Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004, pp. 640–647, https://doi.org/10.1109/TDPVT.2004.1335299. 3DPVT 2004.

[9] V.I. Pavlovic, R. Sharma, T.S. Huang, Visual interpretation of hand gestures for human-computer interaction: a review, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1997) 677–695, https://doi.org/10.1109/34.598226.

[10] A.M.P. Faria, Automotive Gesture Interfaces based on Infrared Technology, Master Thesis, 2017.

[11] H. Gammulle, D. Ahmedt-Aristizabal, S. Denman, L. Tychsen-Smith, L. Petersson, C. Fookes, Continuous Human Action Recognition for Human-machine interaction: a review, ACM Comput. Surv. 55 (2023) 272. 1-272:38 https://doi.org/10.1145/3587931.

[12] I. Jegham, A. Ben Khalifa, I. Alouani, M.A. Mahjoub, Vision-based human action recognition: an overview and real world challenges, Forensic Sci. Int.: Digit. Investig. 32 (2020) 200901, https://doi.org/10.1016/j.fsidi.2020.200901.

[13] S. Mitra, T. Acharya, Gesture recognition: a survey, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 37 (2007) 311–324, https://doi.org/10.1109/TSMCC.2007.893280.

[14] H.F. Nweke, Y.W. Teh, M.A. Al-Garadi, U.R. Alo, Deep learning algorithms for human activity recognition

using mobile and wearable sensor networks: state of the art and research challenges, Expert Syst. Appl. 105 (2018) 233–261.

[15] D. Fisher, M. Meyer, Making Data Visual: A Practical Guide to Using Visualization for Insight, O'Reilly Media, Inc., 2017.

[16] K. Gibert, Sànchez–Marrè M, Izquierdo J., A survey on pre-processing techniques: relevant issues in the context of environmental data mining, AI Commun. 29 (2016) 627–663, https://doi.org/10.3233/AIC-160710.

[17] A. Ferrari, D. Micucci, M. Mobilio, P. Napoletano, On the personalization of classification models for human activity recognition, IEEE Access 8 (2020) 32066–32079, https://doi.org/10.1109/ACCESS.2020.2973425.

[18] T. Benbarrad, M. Salhaoui, S.B. Kenitar, M. Arioua, Intelligent machine vision model for defective product inspection based on machine learning, J. Sens. Actuator Netw. 10 (2021) 7, https://doi.org/10.3390/jsan10010007.

[19] P. Turaga, R. Chellappa, A. Veeraraghavan, Advances in video-based human activity analysis: challenges and approaches, in: M.V. Zelkowitz (Ed.), Advanced Computing, vol. 80, Elsevier, 2010, pp. 237–290, https://doi.org/10.1016/S0065-2458(10)80007-5.

[20] J. Cannan, H. Hu, Human-Machine Interaction (HMI): A Survey, 27, University of Essex, 2011, pp. 46–64.

[21] R. Yin, Wearable sensors-enabled human–machine interaction systems: from design to application, in: Advanced Functional Materials, Wiley Online Library, 2021. https://onlinelibrary.wiley.com/doi/abs/10.1002/adfm.202008936. (Accessed 17 February 2024).

[22] A. Henriksson, J. Rehnmark, Human-Machine Interaction within Teleoperation of Autonomous Vehicles : Designing Remote Operation from A User Experience Perspective, 2023.

[23] M. Nardo, D. Forino, T. Murino, The evolution of man–machine interaction: the role of human in industry 4.0 paradigm, Prod. Manuf. Res. 8 (2020) 20–34, https://doi.org/10.1080/21693277.2020.1737592.

[24] E. Olshannikova, A. Ometov, Y. Koucheryavy, T. Olsson, Visualizing big data with augmented and virtual reality: challenges and research agenda, J. Big Data 2 (2015) 22, https://doi.org/10.1186/s40537-015-0031-2.

[25] Y. Cong, R. Chen, B. Ma, H. Liu, D. Hou, C. Yang, A comprehensive study of 3-D vision-based robot manipulation, IEEE Trans. Cybern. 53 (3) (2021) 1682–1698.

[26] L. Chen, N. Ma, P. Wang, J. Li, P. Wang, G. Pang, et al., Survey of pedestrian action recognition techniques for autonomous driving, Tsinghua Sci. Technol. 25 (2020) 458–470, https://doi.org/10.26599/TST.2020.9010018.

[27] M. Hoai, A. Zisserman, Improving human action recognition using score distribution and ranking, Computer Vision—ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1–5, 2014, Revised Selected Papers, Part V 12, Springer International Publishing, 2015, pp. 3–20.

[28] A.A. Yilmaz, M.S. Guzel, E. Bostanci, I. Askerzade, A novel action recognition framework based on deep-learning and genetic algorithms, IEEE Access 8 (2020) 100631–100644.

[29] I. Tyukin, T. Tyukina, C. van Leeuwen, Invariant template matching in systems with spatiotemporal coding: A matter of instability, Neural Netw. 22 (2009) 425–449, https://doi.org/10.1016/j.neunet.2009.01.014.

[30] M. Talele, R. Jain, Complex facial emotion recognition - a systematic literature review, in: 2023 Third IEEE Sponsored International Conference on Advances in Electrical, Computing, Communications and Sustainable Technologies, ICAECT, 2023, pp. 1–8, https://doi.org/10.1109/ICAECT57570.2023.10117836.

[31] A. Malekloo, E. Ozer, M. AlHamaydeh, M. Girolami, Machine Learning and Structural Health Monitoring Overview with Emerging Technology and High-Dimensional Data Source Highlights, 2022. https://journals.sagepub.com/doi/full/10.1177/14759217211036880. (Accessed 20 February 2024).

[32] S.K. Baduge, S. Thilakarathna, J.S. Perera, M. Arashpour, P. Sharafi, B. Teodosio, et al., Artificial intelligence and smart vision for building and construction 4.0: machine and deep learning methods and applications, Autom. Constr. 141 (2022) 104440, https://doi.org/10.1016/j.autcon.2022.104440.

[33] K.K. Verma, B.M. Singh, A. Dixit, A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system, Int. J. Inf. Technol. 14 (1) (2022) 397–410.

[34] A. Prest, C. Schmid, V. Ferrari, Weakly supervised learning of interactions between humans and objects, IEEE Trans. Pattern Anal. Mach. Intell. 34 (3) (2011) 601–614.

[35] G. James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor, Unsupervised learning, in: An Introduction to Statistical Learning, Springer International Publishing, Cham, 2023, pp. 503–556, https://doi.org/10.1007/978-3-031-38747-0_12.

[36] S. Mittal, Y. Hasija, Applications of deep learning in healthcare and biomedicine, Deep Learning Techniques for Biomedical and Health Informatics, 2020, pp. 57–77.

[37] A.K. Tyagi, A. Abraham, Recurrent Neural Networks: Concepts and Applications, CRC Press, 2022.

[38] M.A. Khan, HCRNNIDS: hybrid convolutional recurrent neural network-based network intrusion detection system, Processes 9 (5) (2021) 834.

[39] M. Lv, W. Xu, T. Chen, A hybrid deep convolutional and recurrent neural network for complex activity

recognition using multimodal sensors, Neurocomputing 362 (2020) 33–40, https://doi.org/10.1016/j.neucom.2020.06.051.

[40] A.G. D'Sa, B.G. Prasad, A survey on vision based activity recognition, its applications and challenges, 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), IEEE, 2019, pp. 1–8.

[41] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, Z. Li, A review on human activity recognition using vision-based method, J. Healthcare Eng. 2017 (1) (2017) 3090343.

[42] F. Abdul Manaf, S. Singh, Computer vision-based survey on human activity recognition system, challenges and applications, in: 2021 3rd International Conference on Signal Processing and Communication, ICPSC, 2021, pp. 110–114, https://doi.org/10.1109/ICSPC51351.2021.9451736.

[43] K. Yun, Vision-based garbage dumping action detection for real-world surveillance platform, ETRI J. (2020). Wiley Online Library https://onlinelibrary.wiley.com/doi/full/10.4218/etrij.2018-0520. (Accessed 20 February 2024).

[44] M. Al-Faris, J. Chiverton, D. Ndzi, A.I. Ahmed, A Review on computer vision-based methods for human action recognition, J. Imaging 6 (2020) 46, https://doi.org/10.3390/jimaging6060046.

[45] É. Zablocki, H. Ben-Younes, P. Pérez, M. Cord, Explainability of deep vision-based autonomous driving systems: review and challenges, Int. J. Comput. Vis. 130 (10) (2022) 2425–2452.

[46] S. Kalimuthu, T. Perumal, Human Activity Recognition Based on Smart Home Environment and Their Applications, Challenges, IEEE Conference Publication | IEEE Xplore, 2021. https://ieeexplore.ieee.org/abstract/document/9404753. (Accessed 20 February 2024).

[47] K. Muhammad, A. Ullah, J. Lloret, J. Del Ser, V.H.C. de Albuquerque, Deep learning for safe autonomous driving: current challenges and future directions, IEEE Trans. Intell. Transp. Syst. 22 (7) (2020) 4316–4336.

[48] N. Robinson, B. Tidd, D. Campbell, D. Kulić, P. Corke, Robotic vision for human-robot interaction and collaboration: A survey and systematic review, ACM Trans. Hum.-Rob. Interact. 12 (1) (2023) 1–66.