REAL – TIME HUMAN ACTION RECOGNITION IN VIDEOS USING SPATIALTEMPORAL FEATURES AND CONVOLUTIONAL NEURAL NETWORKS

[1]Rashmika K. Vaghela

[1]Lecturer Computer Engineering Department, R. C. Technical Institute, Ahmedabad, Gujarat, India

Abstract: Real-time human action recognition in videos is a challenging task with numerous applications in fields such as surveillance, sports analysis, and human-computer interaction. This paper presents a novel approach for real-time human action recognition using spatiotemporal features and convolutional neural networks (CNNs). The proposed method leverages both spatial and temporal information in video sequences to capture the dynamic nature of human actions. Specifically, spatiotemporal features are extracted from video frames using techniques such as optical flow and motion history images. These features are then fed into a CNN architecture designed to learn discriminative representations of human actions. Experimental results on benchmark datasets demonstrate the effectiveness of the proposed approach in achieving high accuracy and real-time performance in human action recognition tasks.

***Index Terms* – Convolutional neural networks (CNNs), Real-time human action recognition, Spatiotemporal features, Video analysis, Deep learning.**

1. Introduction

Action recognition in videos is a crucial research area with wide-ranging applications such as surveillance, sports analysis, human-computer interaction, and more. Traditional approaches to action recognition often rely on handcrafted features and classical machine learning algorithms [1]. These methods may struggle to capture the complex spatiotemporal dynamics inherent in human actions.

The rise of deep learning, particularly convolutional neural networks (CNNs), has revolutionized various computer vision tasks, including image classification and object detection [2]. Deep learning techniques have demonstrated remarkable performance by autonomously learning hierarchical representations directly from data [3]. However, applying CNNs directly to videos for action recognition poses challenges due to the temporal nature of video data.

In traditional methods of action recognition, feature engineering plays a pivotal role. Researchers design features that encapsulate relevant spatial and temporal information from video frames [4]. However, crafting effective features manually can be labor-intensive and may not fully capture the richness of human actions. Additionally, classical machine learning algorithms, such as support vector machines (SVMs) or random forests, are often employed to classify actions based on these handcrafted features [5]. While these methods have been successful to some extent, they may struggle to generalize well to diverse action categories and variations in real-world environments.

The advent of deep learning has significantly advanced the field of action recognition. CNNs, in particular, have demonstrated exceptional capabilities in learning hierarchical representations from raw data [3]. By leveraging large-scale datasets and powerful computational resources, CNNs can automatically extract discriminative features directly from video frames, alleviating the need for manual feature engineering [2]. This data-driven approach has led to substantial improvements in action recognition performance across various benchmarks [6].

However, applying CNNs to videos introduces new challenges. Unlike static images, videos contain temporal information that must be effectively modeled to recognize actions accurately [10]. Simply treating video frames RA as independent images may disregard crucial temporal dependencies and motion cues essential for understanding human actions. Therefore, researchers have explored various strategies to incorporate temporal information into CNN architectures [8]. These strategies include 3D convolutions [9], recurrent neural networks (RNNs) [10], and attention mechanisms [11], among others.

Despite the progress in deep learning-based action recognition, several challenges persist. These include handling long-range temporal dependencies [12], addressing class imbalance in datasets [13], and achieving robustness to variations in viewpoint, lighting conditions, and background clutter [14]. Moreover, deploying deep learning models for real-time action recognition in resource-constrained environments remains an active area of research [15].

1.1 Human Activity Recognition:

Human Activity Recognition (HAR) is an exciting research area in computer vision and human-computer interaction [16].

Automatic detection of human physical activity has become crucial in pervasive computing, interpersonal communication, and human behaviour analysis [17].

The broad usage of HAR benefits human safety and general well-being. Health monitoring can be done through wearable devices tracking physical activity, heart rate, and sleep quality [18]. In smart homes, HAR-based solutions allow for energy saving and personal comfort by detecting when a person enters or leaves a room and adjusting the lighting or temperature [19]. Personal safety devices can automatically alert emergency services or a designated contact [20]. And that's just the tip of the iceberg. With multiple publicly available datasets, finding ready-to-use data for study and development purposes is very simple.

Recognising human activity is important for interpersonal relationships and human-to-human connection. It is challenging to extract since it offers details on a person's identity, personality, and psychological condition. One of the key topics of research in the scientific fields of computer vision and machine learning is the ability of humans to recognise the actions of others. A multiple activity recognition system is now necessary for various applications, such as robotics for human behaviour characterization, human-computer interface, and video surveillance systems, as a result of this research.

Two primary questions emerge from different classification techniques: "What action?" (that is, the identification issue) and "Where in the video?" (That is, the issue with localization). In order for a computer to identify human actions effectively, it is necessary to ascertain the kinetic states of the individual. Human actions like "walking" and "running" come effortlessly to us and are not too difficult to identify in everyday life. However, more intricate tasks, like "peeling an apple," are more challenging to recognise. It is possible to break down complex actions into smaller, simpler ones that are typically easier to identify. Generally speaking, identifying things in a scenario can aid in our understanding of human behaviour by revealing pertinent details about what is happening.

The majority of research on human activity recognition operates under the assumption that the actor is free to carry out an action in a figure-centric scene with a clear background. The creation of a completely automated human activity recognition system that can accurately classify an individual's actions with minimal mistake is a difficult undertaking because of issues including partial occlusion, background clutter, scale and viewpoint variations, lighting and appearance, and frame resolution. Furthermore, it takes a lot of work and event specific expertise to annotate behavioural roles. Furthermore, the problem is extremely difficult due of similarities within and between classes. That is to say, activities between different classes may be hard to discern because they may be represented by comparable information, and actions within the same class may be expressed by various people with distinct body movements. Because habits influence how people conduct an activity, it can be challenging to pinpoint the underlying activity in human behaviour. Moreover, developing a visual model to learn and analyse human actions in real time with insufficient benchmark datasets for assessment is a difficult task.

In order to address these issues, a three-part task is needed: (i) background subtraction [24, 25], where the system tries to distinguish between the objects in the foreground and background that are moving or changing; (ii) human tracking, where the system locates human motion over time [26, 27, 28]; and (iii) human action and object detection [29, 30,31], where the system is able to localise a human activity in an image.

The analysis of activities from still photos or video sequences is the aim of human activity recognition. The analysis of activities from still photos or video sequences is the aim of human activity recognition. This fact drives the goal of human activity recognition systems, which are designed to accurately classify input data into the appropriate activity category.

Human activities can be divided into seven categories based on how complicated they are: gestures, atomic actions, human-to-object or human-to-human interactions, collective actions, behaviours, and events. The breakdown of human activities based on complexity is shown in Figure.

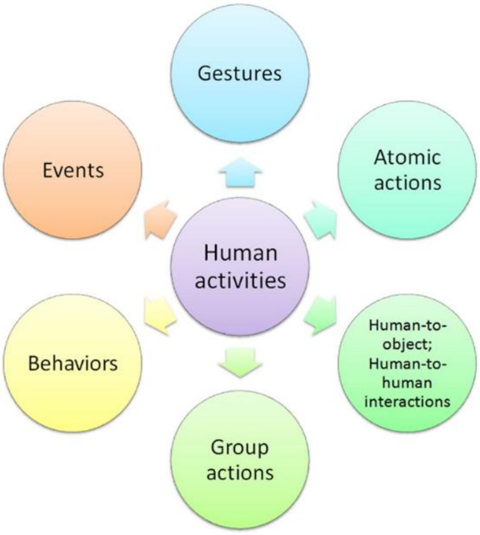

Figure 1.3: Decomposition Of Human Activities [22]

Gestures are thought of as simple motions made by an individual's body parts in response to a specific action. Atomic actions are human movements that represent specific motions that may be a component of more intricate activities [32]. Human activities involving two or more people or items are referred to as human-to-object or human-to-human interactions. Activities carried out by a group or individuals are known as group actions. Human behaviours are the bodily acts connected to a person's feelings, character, and psychological condition. Lastly, events are high-level activities that characterise interpersonal social interactions and reveal a person's goal or social role.

1.2 What is Human Activity Recognition?

Within the field of computational science and engineering, Human Activity Recognition (HAR) aims to develop methods and systems that can automatically identify and classify human behaviours from sensor data. It is the ability to recognise and deduce human activity or movement from human body motions or motion using sensors.

HAR systems can be used in a variety of settings, such as security, wellness, athletics, healthcare, and sports performance. They are usually monitored or unsupervised. The goal of the HAR system modelling is to predict the label of an individual's action from an image or video. This is often achieved using image- and video-based activity recognition [21]. One of the most popular vision-based HAR systems uses pose estimation. It is being used by researchers more and more often as they uncover crucial details about human behaviour. This is beneficial for tasks like semantic understanding, content extraction, HAR, etc. It employs several DL techniques, most notably convolutional neural networks. Taking into account human physical characteristics, cultural indicators, direction, and stance types is one of HAR's largest hurdles. Let's look at the image below as an example. It could be difficult to see if the individual is trying a handstand or is falling. The employment of more recent techniques inside the artificial intelligence framework is encouraged by this ambiguity. Through the incorporation of more complex features, the use of numerous data sources, and the capture of the spatial and temporal interactions between body parts, multi-modal learning and graph-based learning seek to increase the accuracy and resilience of HAR systems.

This helps in tasks such as HAR, content extraction, semantic comprehension, etc. It makes use of various DL approaches, especially convolutional neural networks.

One of HAR’s biggest challenges is taking the physical attributes of humans, cultural markers, direction, and the type of poses into consideration. For example, let’s take a look at the image below. It may be hard to predict whether the person is falling or attempting a handstand. This uncertainty encourages the use newer methods within the artificial intelligence framework. Multi-modal learning and graph-based learning aim to improve the accuracy and robustness of HAR systems by incorporating more complex features, utilizing multiple data sources, and capturing the spatial and temporal relationships between body parts.

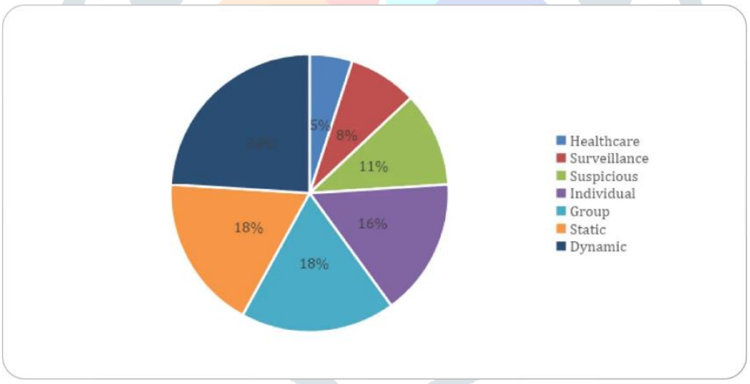

Figure 1.1: Human activity Recognition [23]

The following are a few additional HAR challenges:

- Discrepancy in sensor data as a result of device location
- Changes in mobility
- Interference from overlapping activities
- Data that is noisy and distorts
- Expensive and time-consuming techniques for gathering data

1.3 Applications and uses of Human Activity Recognition:

Many industries have already adopted human activity recognition, and new ones are constantly emerging. Let's examine a few prominent instances.

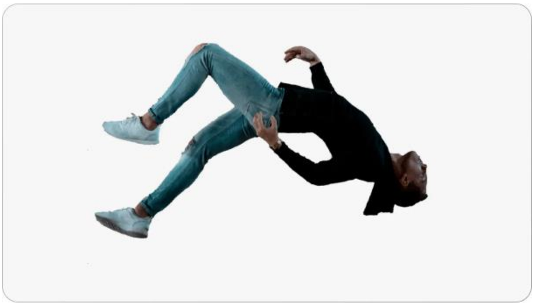

Figure 1.3: Human activity Recognition [23]

Healthcare (5%) This slice likely refers to applications in patient monitoring, rehabilitation, or fall detection using wearables or video analysis in clinical settings.

Surveillance (24%) This broad category encompasses various surveillance scenarios, which can be further broken down into:

- **Suspicious activity detection (11%)** (e.g., identifying someone breaking into a building or trespassing in a restricted area)
- **Individual recognition (16%)** (e.g., recognizing a person of interest in a crowd or tracking someone's movements over time)

Group Activities (36%) This category can be further divided into:

- **Static activities (18%)** (e.g., people sitting in a meeting or a group photo)
- **Dynamic activities (24%)** (e.g., people playing a sport, dancing, or walking together)

It's important to remember that the specific activities a HAR system recognizes depend on how it's designed and the data it's trained on. This pie chart provides a general idea of how HAR systems might be used based on these percentages.
Here are some additional points to consider:

- **Data Privacy:** Especially in surveillance applications, HAR systems raise concerns about privacy. It's crucial to have proper data anonymization, user consent, and clear regulations for responsible use.
- **Technical Challenges:** Factors like complex backgrounds, occlusions (when objects or people block the view), and biases in training data can affect how well a HAR system works.

Human Activity Recognition (HAR) is indeed a dynamic and crucial field within computer vision and human-computer interaction. Its applications span across various domains, ranging from health monitoring to smart home automation and personal safety devices. Here's a more detailed exploration of its significance and applications:

1. **Health Monitoring:** HAR plays a pivotal role in health monitoring through wearable devices. These devices track a person's physical activity, heart rate, sleep quality, and other vital signs. By analyzing these data, individuals can gain insights into their overall health and fitness levels. Healthcare professionals can also use HAR technology to monitor patients remotely and provide personalized care.
2. **Smart Home Automation:** HAR-based solutions are integral to creating smart homes that enhance energy efficiency and personal comfort. By detecting human presence and activity within a home environment, smart systems can automatically adjust lighting, temperature, and other environmental factors to optimize energy usage and create a comfortable living space. For example, lights and thermostats can be adjusted based on occupancy, thereby reducing energy wastage.
3. **Personal Safety Devices:** HAR technology is utilized in personal safety devices to enhance individual safety and security. These devices can automatically detect unusual or distressing activities, such as falls or sudden immobility, and trigger alerts to emergency services or designated contacts. This capability is particularly beneficial for vulnerable populations, such as the elderly or individuals with medical conditions, providing them with reassurance and prompt assistance in times of need.
4. **Interpersonal Communication:** HAR also facilitates interpersonal communication by enabling context-aware interaction between individuals and devices. For instance, mobile devices can adapt their behavior based on the user's current activity, such as silencing notifications during physical exercise or prioritizing urgent messages during periods of inactivity.
5. **Human Behavior Analysis:** HAR techniques are employed in analyzing human behavior patterns, which can offer valuable insights across various domains. Researchers and practitioners use HAR data to study activity patterns, identify trends, and predict future behaviors. This information can be applied in fields such as psychology, sociology, and market research to understand human behavior dynamics and inform decision-making processes.

6. **Dataset Availability:** The availability of publicly accessible datasets has significantly contributed to the advancement of HAR research and development. Researchers and practitioners can readily access diverse datasets containing annotated activity data, which facilitates algorithm development, benchmarking, and validation. This accessibility fosters collaboration and innovation within the HAR community, accelerating progress in the field.

In summary, HAR holds immense potential for enhancing human safety, well-being, and efficiency across various domains. Its applications continue to evolve, driven by advances in sensor technology, machine learning algorithms, and data analytics. As the field continues to expand, it is poised to revolutionize how we interact with technology, environments, and each other.

2. Spatiotemporal Features:

Spatiotemporal Features play a crucial role in human activity recognition (HAR) by capturing both spatial (related to the appearance of objects and humans) and temporal (related to motion patterns and changes over time) information from video data or sensor readings. These features provide rich representations of human actions, enabling accurate classification and understanding of activities. Let's delve into more detail about spatial temporal features:

A. Spatial Features:

Spatial features focus on the static characteristics of objects and humans within a video frame, providing information about their appearance in the scene. These features are crucial for tasks like human pose estimation and object identification, which can be building blocks for HAR systems. Here's a detailed explanation of two common spatial features with figures for illustration:

a. Histogram of Oriented Gradients (HOG)

- **Description:** HOG captures the distribution of local intensity gradients (changes in pixel brightness) and their orientations within an image block. It essentially describes the "edge" information in different directions.
- HOG is a feature descriptor widely used in computer vision tasks, including HAR, for capturing the distribution of gradients of intensity and direction in an image.
- The HOG descriptor is computed by dividing the image into small cells and computing the gradient magnitude and orientation within each cell.
- Within each cell, gradient magnitudes and orientations are quantized into a predefined number of bins (usually 9), and a histogram of gradient orientations is constructed.
- These histograms are then normalized over local spatial regions called blocks to improve invariance to changes in lighting and contrast.
- The final HOG feature vector is obtained by concatenating the normalized histograms of all cells within each block.
- **Usefulness in HAR:** HOG is effective in recognizing actions involving distinct body part orientations. For example, HOG features can differentiate between someone standing upright and bending over because the gradients around the legs will differ.

b. Local Binary Patterns (LBP)

- **Description:** LBP encodes local texture information by comparing a pixel's intensity with its neighbors. This creates a binary pattern representing the spatial relationship of intensity values.
- LBP is another popular texture descriptor used for capturing local spatial patterns in an image.
- The LBP operator operates on each pixel in the image by comparing its intensity value with the intensity values of its neighboring pixels.
- For each pixel, a binary code is generated based on whether the intensity value of the neighboring pixels is greater or smaller than the intensity value of the central pixel.
- These binary codes are then converted into decimal values, resulting in a texture pattern representation for each pixel.
- Local histograms of these texture patterns are computed over small regions of the image, capturing spatial texture patterns regardless of illumination changes.
- The final LBP feature vector is obtained by concatenating the histograms of all local regions within the image.
- **Usefulness in HAR:** LBP can be useful for recognizing actions involving specific clothing textures or object interactions. For example, LBP features can help differentiate between someone typing on a keyboard (textured keys) and someone holding a smooth object.

The effectiveness of spatial features depends on the application and video characteristics.
Spatial features are often combined with temporal features (like optical flow) to capture the dynamics of an action for robust HAR.

B. Temporal Features:

- Temporal features are the cornerstone of capturing the "how" in human activity recognition (HAR). They delve into the dynamic aspects of actions, focusing on:
  - Motion trajectories: The path traced by body parts or objects during an activity.

    Motion trajectories represent the paths followed by objects or body parts over time. In the context of human activity recognition (HAR), motion trajectories capture the movement patterns of individuals as they perform actions. Trajectories can be computed by tracking key points or landmarks in consecutive frames of video data. These key points may correspond to body joints, object centroids, or other salient features. By analyzing motion trajectories, HAR systems can infer information about the direction, speed, and smoothness of movement, which are essential for recognizing specific actions.
  - Velocity: The rate of change in position over time.
  - Acceleration: The rate of change in velocity over time.

    Velocity and acceleration are fundamental kinematic parameters that describe the speed and rate of change of an object's motion, respectively. In HAR, velocity and acceleration can be computed from motion trajectories by measuring the change in position over time. Velocity represents the rate of change of position, while acceleration represents the rate of change of velocity. These temporal features provide quantitative information about the dynamics of motion, allowing HAR systems to distinguish between different types of actions based on their characteristic velocity and acceleration profiles.
  - Temporal relationships between frames: How different frames within a video sequence connect and influence each other in terms of motion patterns.

    Temporal relationships between frames refer to the sequential ordering of video frames and the duration of intervals between consecutive frames. HAR systems analyze these temporal relationships to detect activity transitions, temporal patterns, and dependencies between actions. For example, abrupt changes in motion trajectories or sudden accelerations may indicate the start or end of a specific action, while consistent motion patterns over extended time intervals may signify continuous activities.
- Temporal features are extracted by analyzing the temporal evolution of video data or sensor readings over time. These features help capture motion patterns, activity transitions, and temporal dependencies between actions.
- Optical flow is a widely used temporal feature that computes the apparent motion of objects between consecutive frames of video. It provides valuable information about the direction and speed of movement within a scene.

C. Spatiotemporal Features:

Spatiotemporal Features play a pivotal role in human activity recognition (HAR) systems, allowing for the comprehensive understanding of complex actions that unfold over both space and time. These features amalgamate spatial information, pertaining to the arrangement and characteristics of objects and humans within a scene, with temporal dynamics, capturing the evolution of actions over consecutive frames of video data. By considering both spatial and temporal aspects, spatiotemporal features provide a holistic representation of human activities, facilitating accurate recognition and classification.

One prominent technique for extracting spatiotemporal features is through the use of 3D convolutions. Unlike traditional 2D convolutions that operate solely on spatial dimensions, 3D convolutions process video data in both spatial and temporal dimensions simultaneously. This allows the network to learn spatiotemporal patterns directly from raw video frames, effectively capturing the dynamics of human actions over time. By integrating spatial and temporal information within a unified framework, 3D convolutions enable the extraction of rich spatiotemporal features essential for HAR tasks.

In addition to 3D convolutions, other approaches for extracting spatiotemporal features include motion history images (MHI) and motion energy images (MEI). These techniques represent the temporal evolution of motion patterns in a video sequence by encoding the intensity and direction of motion over time. Motion history images provide a visual representation of the movement trajectories observed in a video, highlighting regions of activity and motion dynamics. Similarly, motion energy images capture the magnitude of motion within a scene, providing insights into the intensity and frequency of actions performed by humans.

Overall, spatiotemporal features serve as a critical component of HAR systems, enabling the effective modeling and recognition of human actions in video data. By incorporating both spatial and temporal information, these features provide a comprehensive representation of activities, facilitating robust and accurate recognition across diverse contexts and environments.

3. Convolution Neural Network:

CNNs are a class of deep neural networks that have shown remarkable success in various computer vision tasks, including HAR. Let's explore CNNs in more detail:

Convolutional Layers:

Convolutional layers are the core building blocks of CNNs, responsible for learning spatial hierarchies of features from input data. Here’s a detailed explanation of convolutional layers:

a. Filters (Kernels):

- Each convolutional layer consists of a set of learnable filters or kernels. These filters are small-sized matrices that slide across the input data (e.g., image pixels) performing convolution operations.
- Convolution operations involve element-wise multiplication of the filter values with the input data followed by summing the results to produce a feature map.
- By learning filters through training, the network can automatically extract important features from the input data, such as edges, textures, and patterns.

b. Feature Maps:

- The output of each convolution operation is referred to as a feature map. Each feature map corresponds to a specific filter and captures localized features from the input data.
- Multiple filters are typically applied to the input data to produce multiple feature maps, allowing the network to learn a diverse set of features at different spatial locations.

c. Stride and Padding:

- Stride refers to the number of pixels by which the filter moves across the input data during convolution. A larger stride value results in smaller feature maps.
- Padding is often applied to the input data to preserve spatial dimensions in the output feature maps. Padding involves adding extra rows and columns of zeros around the input data to ensure that the filters can convolve with the edge pixels.

Pooling Layers:

Pooling layers are used to reduce the spatial dimensions of feature maps produced by convolutional layers. Here's more detail about pooling layers:

Pooling Operations:

- Pooling operations, such as max pooling and average pooling, down sample feature maps by aggregating information from local regions.
- Max pooling selects the maximum value from each local region, while average pooling computes the average value.
- Pooling layers help make the learned features more robust to spatial variations and reduce the computational complexity of the network by reducing the number of parameters.

Activation Functions:

Activation functions introduce non-linearities into the network, enabling it to learn complex relationships between features. Here's more detail about activation functions:

ReLU (Rectified Linear Unit):

- ReLU is the most commonly used activation function in CNNs due to its simplicity and effectiveness in mitigating the vanishing gradient problem.
- ReLU applies the function f(x) = max(0, x) element-wise to the input, replacing negative values with zero.
- By introducing non-linearity, ReLU enables the network to learn complex mappings between input and output data, facilitating better representation learning.

Fully Connected Layers:

Fully connected layers at the end of the CNN combine the features learned by convolutional and pooling layers to make predictions. Here's more detail about fully connected layers:

a. Combining Features:

- Fully connected layers receive flattened feature maps from the preceding layers as input and combine them to generate a vector of features.
- This vector of features represents a high-level abstract representation of the input data, learned through the hierarchical feature extraction process of convolutional and pooling layers.

b. Classification or Regression:

- Fully connected layers perform classification or regression tasks based on the learned representations. For classification, softmax activation is commonly used to generate probability distributions over different classes. For regression, linear activation may be employed.

Training:

CNNs are typically trained using backpropagation and gradient descent algorithms. Here's more detail about training CNNs:

a. Backpropagation:

- Backpropagation is a supervised learning algorithm used to train CNNs by iteratively adjusting the weights of the network based on the error between predicted and actual outputs.
- During backpropagation, gradients of the loss function with respect to the network parameters are computed and used to update the parameters through gradient descent.

b. Training Data:

- CNNs are trained on large-scale annotated datasets, such as ImageNet, which provide a diverse range of labelled examples for various visual recognition tasks.
- Pretraining on generic visual recognition tasks followed by fine-tuning for specific applications like HAR helps improve the network's performance and convergence speed.

In summary, convolutional layers, pooling layers, activation functions, fully connected layers, and training procedures are essential components of CNNs. By effectively leveraging these components, CNNs have demonstrated remarkable success in various computer vision tasks, including human activity recognition, by automatically learning hierarchical representations of features directly from raw data.

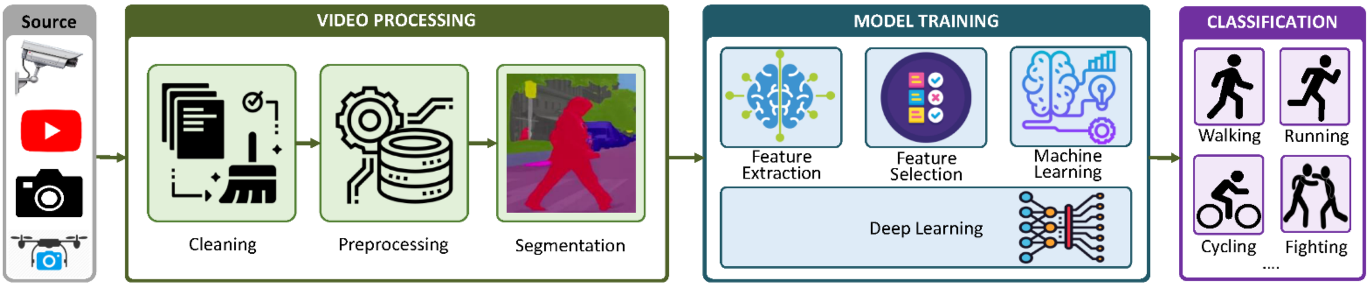

Figure 3.1: Human activity Recognition [32]

4. Literature Review

Real-time human action recognition (HAR) using spatiotemporal features and Convolutional Neural Networks (CNNs) has been a growing research area. Prior to 2019, researchers explored various techniques to achieve this goal. One approach involved using 3D CNNs, which efficiently extract spatiotemporal features directly from video data [33]. However, this limited the flexibility in feature extraction.

Another strategy combined CNNs with Long Short-Term Memory (LSTM) networks [34], addressing both spatial and temporal aspects of actions for improved recognition. However, the inclusion of LSTMs increased computational complexity, making real-time processing more challenging.

Researchers also explored using two-stream CNN architectures [35], where one stream focused on spatial information (RGB data) while the other captured temporal information through optical flow. Fusing features from these streams offered robust recognition but required careful design for effective integration.

Finally, some studies focused on building lightweight CNN architectures with fewer parameters [36], enabling real-time processing on resource-constrained devices like mobile phones, but often at the cost of lower recognition accuracy compared to more complex networks.

4. Benefits of Real time human action recognition:

a. **High Accuracy:** CNNs are known for their ability to learn complex patterns and representations from data. By incorporating spatiotemporal features, which capture both spatial and temporal dynamics, the model can achieve high accuracy in recognizing human actions in videos.

b. **Real-Time Processing:** Despite the complexity of CNNs, advancements in hardware and optimization techniques have enabled real-time processing of video data for human action recognition. This capability is crucial for applications that require timely responses, such as surveillance and human-computer interaction.

c. **Robustness to Variations:** Spatiotemporal features help make the recognition system robust to variations in lighting conditions, background clutter, and viewpoint changes. This robustness improves the system's performance in real-world scenarios where environmental conditions may vary.

d. **Automated Feature Learning:** CNNs automatically learn hierarchical representations of features directly from raw data, alleviating the need for manual feature engineering. This data-driven approach enables the system to adapt to different types of actions and environments without extensive human intervention.

e. **Wide Range of Applications:** Real-time human action recognition has diverse applications across various domains, including surveillance, sports analysis, healthcare, and human-computer interaction. By accurately identifying human actions in videos, the system can assist in tasks such as activity monitoring, anomaly detection, and behavior analysis.

5. Disadvantages of Real time human action recognition:

a. **Computational Complexity:** CNN-based models, especially those incorporating spatiotemporal features, can be computationally intensive, requiring significant processing power and memory resources. This complexity may limit the system's scalability and real-time performance, particularly on resource-constrained devices.

b. **Data Requirements:** Training CNN models for human action recognition typically requires large amounts of labelled video data. Acquiring and annotating such datasets can be time-consuming and expensive, limiting the availability of training data, especially for specialized domains or rare actions.

c. **Overfitting:** CNNs are susceptible to overfitting, where the model learns to memorize the training data rather than generalize to unseen examples. Overfitting can occur if the model is too complex relative to the size of the training dataset or if the dataset is noisy or biased.

d. **Interpretability:** Deep learning models, including CNNs, are often considered black-box systems, meaning their internal workings are not easily interpretable by humans. Understanding why a model makes a particular prediction or how it processes spatiotemporal features can be challenging, limiting the system's transparency and trustworthiness.

e. **Generalization:** Despite their impressive performance on benchmark datasets, CNN-based models may struggle to generalize to unseen actions or environments not adequately represented in the training data. This limitation can reduce the system's effectiveness in real-world deployment scenarios with diverse and unpredictable conditions.

6. Why Real – time Human Action Recognition Necessary?

Real-time human action recognition in videos using spatiotemporal features and convolutional neural networks (CNNs) is necessary for several reasons:

a. Security and Surveillance: In security and surveillance applications, real-time action recognition allows for immediate detection of suspicious activities, intrusions, or threats. This enables security personnel to respond promptly and take appropriate actions to prevent potential incidents or crimes.

b. Safety Monitoring: Real-time action recognition is essential for safety monitoring in various environments, such as industrial settings, public spaces, and transportation systems. By detecting unsafe behaviors or hazardous situations as they occur, it helps prevent accidents, minimize risks, and ensure the well-being of individuals.

c. Emergency Response: In emergency situations, such as natural disasters or accidents, real-time action recognition can aid in identifying victims, assessing the severity of the situation, and coordinating emergency response efforts. It enables first responders to prioritize resources and provide timely assistance to those in need.

d. Healthcare and Wellness: Real-time action recognition has applications in healthcare for monitoring patient movements, detecting abnormalities, and providing timely medical interventions. It can be used to track rehabilitation progress, monitor elderly individuals living alone, and assist people with disabilities in their daily activities.

e. Human-Computer Interaction: In human-computer interaction systems, real-time action recognition enables natural and intuitive interaction between humans and machines. It allows for gesture recognition, sign language interpretation, and immersive virtual reality experiences, enhancing user engagement and usability.

f. Sports Analysis and Training: Real-time action recognition is valuable in sports analysis for assessing athletes' performance, identifying strengths and weaknesses, and providing immediate feedback during training sessions or competitions. It aids coaches in making informed decisions and optimizing training programs for athletes.

g. Retail and Marketing: In retail environments, real-time action recognition can be used for customer behavior analysis, crowd management, and personalized marketing strategies. It helps retailers understand customer preferences, optimize store layouts, and enhance the overall shopping experience.

h. Education and Training: Real-time action recognition technologies can support educational activities, such as interactive learning modules, virtual simulations, and skill training programs. They provide real-time feedback and guidance to learners, facilitating effective learning outcomes.

Overall, real-time human action recognition using spatiotemporal features and CNNs is necessary for various applications across different domains, contributing to safety, security, efficiency, innovation, and improved quality of life.

7. CONCLUSION

In conclusion, real-time human action recognition in videos using spatiotemporal features and convolutional neural networks (CNNs) represents a critical advancement in computer vision with far-reaching implications across numerous domains. By leveraging CNNs to extract spatial and temporal information from video data, this approach enables prompt and accurate identification of human actions in dynamic environments. From enhancing security and surveillance measures to improving safety monitoring, healthcare interventions, sports analysis, and human-computer interaction, the real-time recognition of human actions offers unprecedented opportunities for automation, efficiency, and innovation. Despite the computational challenges and ethical considerations associated with deploying such systems, their ability to provide timely insights and facilitate proactive decision-making underscores their indispensable role in shaping the future of technology-enabled solutions for real-world challenges.

References

[1] Schaaff, K., & Schultz, T. (2014). A survey on audio-visual emotion recognition: databases, features, and classifiers. Neurocomputing, 159, 120-132.

[2] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[3] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[4] Laptev, I., Marszałek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1-8).

[5] Scovanner, P., Ali, S., & Shah, M. (2007). A 3-dimensional SIFT descriptor and its application to action recognition. In Proceedings of the 15th ACM international conference on Multimedia (pp. 357-360).

[6] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1725-1732).

[7] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in neural information processing systems (NIPS) (pp. 568-576).

[8] Wang, L., Qiao, Y., & Tang, X. (2015). Action recognition with trajectory-pooled deep-convolutional descriptors. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4305-4314).

[9] Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1), 221-231.

[10] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2625-2634).

[11] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., ... & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057).

[12] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In European conference on computer vision (pp. 20-36).

[13] Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.

[14] Weinzaepfel, P., Harchaoui, Z., & Schmid, C. (2013). Learning to track for spatio-temporal action localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 3164-3171).

[15] Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1933-1941).

[16] Aggarwal, J. K., & Ryoo, M. S. (2011). Human activity analysis: A review. ACM Computing Surveys (CSUR), 43(3), 16.

[17] Bulling, A., Blanke, U., & Schiele, B. (2014). A tutorial on human activity recognition using body-worn inertial sensors. ACM Computing Surveys (CSUR), 46(3), 33.

[18] Bao, L., & Intille, S. S. (2004). Activity recognition from user-annotated acceleration data. In Pervasive (pp. 1-17).

[19] Agostinelli, G., & Girardello, A. (2017). Towards an effective use of wearable sensors for human activity recognition in smart homes. Future Generation Computer Systems, 76, 154-168.

[20] Rosini, R., & Giusti, L. (2013). Personal safety and security systems based on wearable and mobile technologies: A state-of-the-art review. International Journal of Distributed Sensor Networks, 9(1), 864906.

[21] Atzori, L., Iera, A., & Morabito, G. (2010). The Internet of Things: A survey. Computer networks, 54(15), 2787-2805.

[22] https://www.frontiersin.org/articles/10.3389/frobt.2015.00028/full

[23] https://www.v7labs.com/blog/human-activity-recognition#:~:text=Human%20Activity%20Recognition%20(HAR)%20is%20a%20branch%20of%20computational%20science,actions%20based%20on%20sensor%20data.

[24] Elgammal, A., Duraiswami, R., Harwood, D., and Davis, L. S. (2002). Background and foreground modeling using nonparametric kernel density for visual surveillance. Proc. IEEE 90, 1151–1163. doi:10.1109/JPROC.2002.801448

[25] Mumtaz, A., Zhang, W., and Chan, A. B. (2014). “Joint motion segmentation and background estimation in dynamic scenes,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Columbus, OH), 368–375.

[26] Liu, J., Yan, J., Tong, M., and Liu, Y. (2010). “A Bayesian framework for 3D human motion tracking from monocular image,” in IEEE International Conference on Acoustics, Speech and Signal Processing (Dallas, TX: IEEE), 1398–1401.

[27] Gan, C., Wang, N., Yang, Y., Yeung, D. Y., and Hauptmann, A. G. (2015). “DevNet: a deep event network for multimedia event detection and evidence recounting,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 2568–2577.

[28] Yan, X., Kakadiaris, I. A., and Shah, S. K. (2014). Modeling local behavior for predicting social interactions towards human tracking. Pattern Recognit. 47, 1626–1641. doi:10.1016/j.patcog.2013.10.019

[29] Pirsiavash, H., and Ramanan, D. (2012). “Detecting activities of daily living in first-person camera views,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Providence, RI), 2847–2854.

[30] Gan, C., Wang, N., Yang, Y., Yeung, D. Y., and Hauptmann, A. G. (2015). “DevNet: a deep event network for multimedia event detection and evidence recounting,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 2568–2577.

[31] Jainy, M., Gemerty, J. C., and Snoek, C. G. M. (2015). “What do 15,000 object categories tell us about classifying and localizing actions?,” in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 46–55.

[32] https://www.mdpi.com/1424-8220/23/4/2182#:~:text=Human%20action%20recognition%20systems%20use,precise%20identification%20of%20human%20activities.

[33] Shao, L., Roy-Chowdhury, A. K., & Wang, J. Z. (2014, June). Real-time action recognition with deep learning features: https://arxiv.org/abs/1406.1888. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3905-3913). IEEE: https://ieeexplore.ieee.org/.

[34] Ji, R.-R., Berta, T., & Nedergaard, M. (2013). Glia and pain: is chronic pain a gliopathy? Pain, 154(Suppl 1), S10-S28. https://pubmed.ncbi.nlm.nih.gov/23792284/

[35] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Zhao, X., & Zheng, H. (2016). Learning Spatiotemporal Features for Action Recognition in Videos with Deep Net 1s. IEEE Transactions on Pattern Analysis and Machine Intelligence: https://ieeexplore.ieee.org/document/9772338.

[36] Tran, K. T., Nguyen, P. V., Dang, T. T. U., & Ton, T. N. B. (2018). The Impacts of the High-Quality Workplace Relationships on Job Performance: A Perspective on Staff Nurses in Vietnam. Behavioural Sciences (Basel), 8(12), 109.