ST-ANet: Enhanced Human Activity Recognition Framework with WiFi CSI

by

Shengxiong Xiao

A thesis submitted to the
Department of Computer Science
in conformity with the requirements for
the degree of Master of Science

Bishop's University
Canada
April 2024

Copyright © Shengxiong Xiao, 2024
released under a CC BY-SA 4.0 License

Abstract

This thesis introduces ST-ANet, an innovative framework that uses WiFi channel state information (CSI) to improve human activity recognition. ST-ANet combines advanced signal processing techniques and machine learning to extract valuable insights from WiFi CSI data. The thesis explains how WiFi CSI data can be collected without intruding on individuals' privacy during the data collection process. It also provides a detailed explanation of the feature extraction method and the dual-channel spatiotemporal framework used in ST-ANet. Furthermore, it explores deep learning algorithms for activity recognition, showcasing their effectiveness in handling complex CSI matrices. The main findings of the study highlight the significant potential of ST-ANet. By utilizing WiFi CSI, this framework achieves higher levels of accuracy and detail when identifying human activities compared to traditional deep learning methods. It accomplishes this while ensuring the protection of individuals' privacy. The system accurately identifies various activities through the extracted BVP files. In summary, ST-ANet not only enhances human activity recognition but also finds practical applications in healthcare, smart homes, and security systems. This framework demonstrates how WiFi CSI data can be harnessed to improve our understanding of human behavior in various environments.

Keywords: Human Activity Recognition, Deep Learning Algorithms, Artificial Intelligence, Internet of Things, Slow-Fast, Self-attention Mechanism.

Acknowledgements

I wish to extend my heartfelt gratitude to all those who provided unwavering support and guidance throughout my research journey and the successful completion of this thesis. I am deeply indebted to my thesis advisor, Professor Madjid Allili, for his invaluable mentorship, constant insights, and unwavering commitment to my academic development. The guidance from Madjid played a pivotal role in shaping the direction of this research. I sincerely appreciate the esteemed faculty members of the Department of Computer Science at Bishop's University for their intellectual contributions and constant encouragement. Their expertise and constructive feedback significantly enriched this study. My family deserves special mention for their unwavering support, understanding, and encouragement throughout my graduate studies. Their love and trust have always been a wellspring of motivation for me. Furthermore, I am grateful to the university staff, technical teams, and administrative personnel for their assistance, generous funding, and provision of essential resources. Their contributions have greatly facilitated the research process.

In conclusion, the successful culmination of this thesis is the collective result of the collaborative efforts and support of numerous individuals, and I extend my profound gratitude to every one of them.

Sincerely,

Shengxiong Xiao

Contents

1 Introduction **3**
- 1.1 The IoT Infrastructure: Building the Foundation for HAR Research . 3
  - 1.1.1 Unlocking the Benefits of HAR 5
- 1.2 Main approaches to HAR and Ways of Data Collection 6
  - 1.2.1 Challenges in Conventional Human Activity Detection 7
- 1.3 An Alternative: Wi-Fi CSI 9
  - 1.3.1 Widar System Overview 11
- 1.4 Exploration of HAR Through Wi-Fi CSI 14

2 Related Works **16**
- 2.1 Deep Learning Recognition on WiFi-based HAR 16
  - 2.1.1 Two-pathway Action Recognition Model: SlowFast 19
- 2.2 Related Studies Using SlowFast 20
- 2.3 Enhance Learning of Temporal Features Using Attention Mechanism 23
  - 2.3.1 Self-attention Modules 24
  - 2.3.2 SlowFast HAR Models Utilizing Attention Mechanism 25

3 Methodology and Evaluation **29**
- 3.1 Proposed ST-ANet: Architecture and Multipath Components 29
  - 3.1.1 Fast Pathway: Capturing Temporal Information 31
  - 3.1.2 Slow Pathway: Capturing Spatial Information 33
  - 3.1.3 Feature Fusion Using Attention Mechanisms 35
- 3.2 Evaluation Metrics and Experimental Setup 36
- 3.3 Widar Dataset 37

4 Experiment and Results **40**
- 4.1 Analysis and Ablation Study 40
  - 4.1.1 Performance Comparison with Existing Models 42
  - 4.1.2 Ablation Study 43

5 Conclusion **45**

Bibliography **47**

List of Figures

1.1 Data transactions in IoT [6]. 4
1.2 Distinguished by different methods of data collection [13]. 7
1.3 CSI Phase of Different Activities [22] 10
1.4 Common Gestures Utilized in Human-Computer Interaction [23]. . . 11
1.5 CSI Phase of Different Activities [22] 12
1.6 Structure of deep learning model for activity recognition [13]. 13

2.1 Structure of deep learning model for activity recognition [27]. 17
2.2 Dual-path networks: SlowFast [34]. 19
2.3 Add a audio path to SlowFast [36]. 21
2.4 Learnable positional embedding in Transformer architecture [33]. . . 24
2.5 Overall architecture of The Spatio-Temporal SlowFast Self-Attention Network [26]. 25
2.6 The pipeline of Evo-ViT: token selection and SlowFast token updating [50]. 27

3.1 Overall architecture of the ST-ANet. 29
3.2 The fast pathway operates at higher frame rates. 31
3.3 The content of a residual block. 32
3.4 The slow pathway to capture static spatial information. 33
3.5 Overall architecture of the ST-ANet. 35

4.1 Training Performance of ST-ANet. 41

List of Tables

4.1 Widar Dataset - Model Performance 43
4.2 Comparison of Individual Components' Performance 44

Chapter 1

Introduction

1.1 The IoT Infrastructure: Building the Foundation for HAR Research

HAR aims to equip machines with the ability to analyze and interpret human movements. Using algorithms and computer systems can identify and categorize various human activities such as walking, running, and jumping. However, it is essential to note that individual activities can have significant differences. For example, the patterns present in human activities can reveal specific ideas, habits, and cultures. We believe that learning these patterns can help machines better understand human behavior.

In the meantime, the emergence of the IoT has enabled billions of electronic devices and appliances to be equipped with advanced sensors and wireless networks. Combined with appropriate embedded or cloud computing, these devices have become smarter and more convenient [1]. This thesis seeks to establish an effective HAR system within the IoT environment. Furthermore, it aims to integrate this HAR system with a multi-channel Convolutional Neural Network (CNN) to attain high precision, efficiency, and low energy consumption in activity recognition.

IoT has spawned numerous wearable intelligent devices and smart home applications. In addition, adopting 5G communication standards has impacted all aspects of the world [2]. This technology enables different physical devices to connect to the Internet and exchange data constantly, and the cost of this data exchange is gradually decreasing. As a result, connected devices, including mobile phones, social networks, and electronic communications collect and transmit real-time information [3]. The data generated by these devices contain a large amount of information and knowledge and have become a valuable resource, and mastering its use will drive the world toward increased intelligence [4]. The objective of HAR-related tasks lies in automated comprehension of human activities through data analysis. This has become an increasingly relevant tool in smart surveillance,

offering potential improvements in safeguarding public areas and critical infrastructure.

The vast amount of data generated by the IoT and new communication technologies have evolved into a new field known as Big Data. Big data encompasses structured data, such as organizational databases, and unstructured data, such as images, videos, and audio [5]. As the computational power required by data analysis exceeds the limits of humans, computers that excel in performing repetitive tasks at high speed have become crucial. This shift has driven the field of computer science towards data-driven discoveries and made deep learning one of the hottest trends in the rapidly evolving digital world.

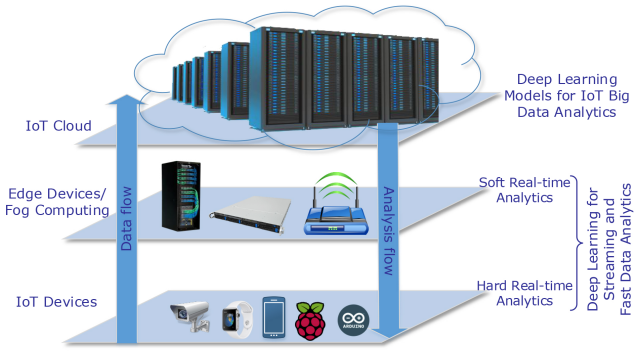


Figure 1.1: Data transactions in IoT [6].

Inspired by biological observations on human brain mechanisms for processing natural signals, deep learning has gained much attention in the domain of Artificial Intelligence(AI) due to its state-of-the-art performance in areas such as speech recognition [7], natural language processing [8], and computer vision [9]. Figure 1.1 shows that with the emergence of technologies such as cloud computing, big data storage, and high-speed networks, the convenience of data transactions has enabled deep learning models to use more diverse and comprehensive data sets. These advances lead to more effective and efficient data processing and deep learning models, enabling new and innovative applications in various fields.

The use of visual information is deeply ingrained in the human problem-solving process. Many real-world problems require visual data to be effectively addressed, and information processed by humans is also visual. Over the decades, computer vision has been applied to various scenarios, including medical imaging, surveillance, and self-driving vehicles [10]. As a typical application of deep learning in

human-computer interaction(HCI), HAR is a data-driven task that requires a large amount of data for training and evaluation. Hence, the abundant data gathered from sensors or cameras streamlines the automatic learning and feature extraction processes in deep learning models. We will leverage deep learning technology to introduce a HAR model integrated with IoT infrastructure.

1.1.1 Unlocking the Benefits of HAR

The advancements in deep learning have significantly improved the accuracy and efficiency of HAR systems [11], leading to new opportunities for their applications. In video surveillance, the application of HAR enhances the performance of surveillance systems by automating the detection and analysis of human motion. The technology enables the identification and classification of various activities, detecting anomalies or significant events to improve the overall efficiency and accuracy of the system [12]. Furthermore, by combining human activity recognition and face recognition technology, the real-time tracking and analysis of human activities in public spaces can better understand human behavior and improve security measures accordingly.

In addition, studying HAR can deepen our understanding of human behavior and movement, leading to new insights and knowledge in computer vision. By analyzing complex and dynamic human behavior, researchers can gain new insights into various aspects of human behavior, such as the relationship between body posture and movement, and the interaction of individuals with their surroundings. Therefore, exploring and advancing this field can lead to significant advances in our understanding of human behavior and enhance the intelligence of machines by providing them with a comprehensive understanding of human behavior and movements.

By leveraging the power of AI, machines can learn to interpret and understand problems visually, leading to improved decision-making and problem-solving capabilities. For instance, in healthcare settings, researchers can monitor patients' movements and detect anomalies, providing timely alerts to medical professionals. Furthermore, it can also contribute to smart cities, where urban infrastructure and services are optimized through advanced technologies [4].

Overall, HAR has the potential to drive the world towards increased intelligence by providing machines with a more comprehensive understanding of human behavior and movements. Although it may deliver more concerns about privacy leakage, this still can lead to the development of more positive and effective solutions across a wide range of industries and applications. Additionally, it has the potential to enhance the intelligence of machines by providing them with a more comprehensive understanding of human behavior and movements. The utilization of HAR has the potential to propel the world towards intelligent automation and facilitate the development of more efficient and effective solutions across a diverse

range of industries and applications. This serves as the ultimate aim of the present study.

1.2 Main approaches to HAR and Ways of Data Collection

HAR is a challenging research area due to several factors, including the need for accurate and efficient recognition algorithms, the robustness required to handle changes in lighting, perspective, and clothing, and the need to have large annotated datasets. To overcome these, researchers have investigated two main approaches: vision-based and sensor-based HAR [13].

In order to classify the various activities performed by humans, they are usually divided into several groups. The first of these groups consists of gestures, which are simple movements of the hands or other parts of the body used to convey a specific thought or meaning. The second category is actions, which are simple activities, often involving multiple gestures. Interactions represent another activity category, characterized by the participation of two agents, one of which is always a person. The other agent can be the object or another person, resulting in human-object or human-human interaction.

Finally, group activities are the most complex, requiring more than two people and often involving one or more objects. As such, research related to HAR is inherently intertwined with different methodologies, scientific statistical analyses, and meticulous documentation of data generated by human activities that can illuminate the distinct features of human culture [14].

Despite the advancements, the methodologies employed in human activity recognition remain limited.

1. Wearable sensors

   The advent of the IoT and mobile computing in recent years has fostered an environment for wearable sensors. These sensors have emerged as a preeminent form of HAR that can be integrated into portable and wearable devices. Several standard wearable sensors are facilitating human activity detection with the ability to measure signal differences before and after human activities, like magnetometers and gyroscopes.

2. Environmental sensors

   In contrast to wearable sensors, environmental sensors are typically deployed in the surrounding environment or affixed to particular objects to monitor alterations in environmental parameters during physical activities, to capture human activities. Experimental results have indicated that sensors generate highly detailed data, resulting in rising accuracy in activity classification. However, environmental sensors are less prevalent than wearable sensors due to their complex setup requirements.

3. Kinds of videos

   Vision-based HAR has garnered considerable attention due to its widespread usage in real-world applications. Examples of such applications include closed-circuit television systems deployed in public spaces and various online video sites. These facilities, either physical or virtual, with video data as the primary medium, offer a rich source of HAR data.

Data collection plays a pivotal role in HAR systems, as the quality of the input data directly affects the subsequent analytical steps. Over time, plenty of datasets have been employed to validate the efficacy of HAR models. To capture the uniform patterns, video recording devices, and diverse physical measurement sensors have been utilized for an extended period. Wearable sensors have garnered significant popularity due to their affordability, portability, and compatibility with various devices. However, hybrid sensors that consist of both wearable and environmental sensors, are also prevalent. Notably, hybrid sensors have gained increased prominence in intricate activity recognition applications owing to the ability to improve model robustness and performance through multiple sensors [15]. Furthermore, researchers often generate their datasets by collecting data in various environments and activity types in alignment with their research objectives. These datasets are frequently made accessible to the public to encourage further exploration in this domain.

1.2.1 Challenges in Conventional Human Activity Detection

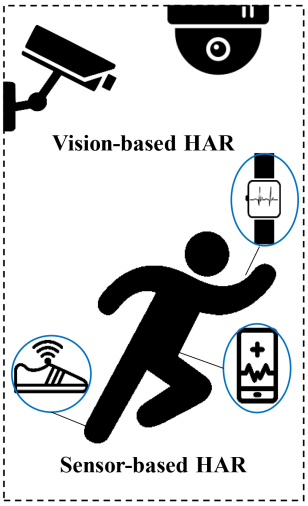


Figure 1.2: Distinguished by different methods of data collection [13].

As shown in Figure1.2, there are diverse ways of collecting human activity data. Vision-based HAR involves analyzing video streams to recognize human activities. Relevant features such as motion, shape, and color are extracted from the video stream using computer vision techniques. These features are then used to classify the activity, such as walking, running, or sitting. Sensor-based HAR, on the other hand, relies on data from wearable or environmental sensors such as accelerometers, gyroscopes, and magnetic sensors. The data is processed using signal processing techniques and machine learning algorithms to identify and classify activity. Sensor-based approaches are less affected by environmental factors such as lighting and viewpoint and can provide more detailed information about body movement.

Vision-based and sensor-based HAR have their strengths and weaknesses, and the choice of approach depends on the specific application requirements. Vision-based systems are useful in applications where visual information is important, such as sports analysis or security monitoring. Sensor-based HAR is well-suited for applications where privacy is a concern, such as in healthcare or home monitoring, or where wearable devices are more appropriate, such as sports training or rehabilitation. Video-based human activity recognition methods invariably introduce irrelevant information that is difficult to eliminate, such as intricate backgrounds or non-target objects.

Recently, HAR has gained prominence due to the proliferation of wearable devices and wireless network technologies. HAR has proven its value by enhancing various application areas, particularly improving the quality of life for the elderly and disabled. While numerous machine learning methods have been explored for HAR, the ongoing challenge is the most effective approach that balances high accuracy and interpretability with scalability and efficiency. Equally important is the need to examine the advantages and drawbacks and to propose a framework that efficiently utilizes these sensors for precise human activity recognition.

Several recent studies have contributed to HAR research using various sensor-based methods. In 2016, researchers created a dataset to investigate the impact of wearable device positioning on activity recognition, utilizing acceleration data from smartphones and smartwatches [16]. However, a drawback lies in the potential discomfort and inconvenience of wearing multiple devices simultaneously. In 2018, another study focused on measuring human biological signals and identifying activities by attaching wearable sensors to subjects' lower extremities, joints, and waist [17]. While providing detailed physiological data, this approach necessitates participants to wear multiple sensors, which may not be practical for extended use. In a 2021 study, the Biosignalsplux Researcher Kit was employed, capturing biosignals from channels to enhance HAR [18]. However, this method may involve the burden of carrying additional equipment, and if a smartphone serves as the data source, it can be constrained by battery limitations. Thus, while sensor-based approaches offer valuable insights into HAR, they raise concerns about user comfort

and device practicality.

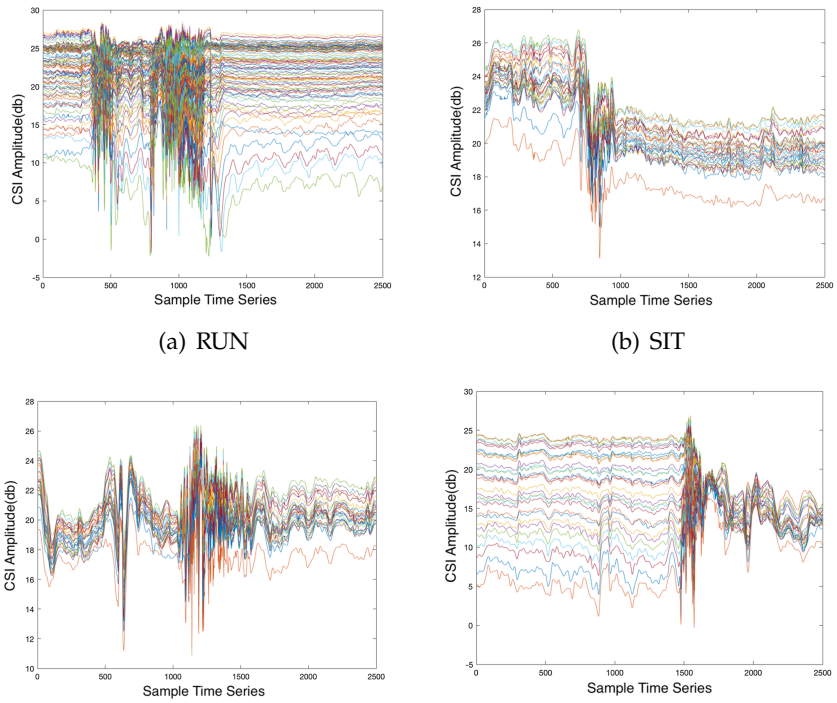
Wearable sensors have gained popularity for their high recognition accuracy. However, such systems require users to wear and carry extra devices, which can be inconvenient and cumbersome. Another popular option is smartphones, equipped with numerous embedded sensors. Nevertheless, activity recognition may cease when users forget their smartphones, and sensor usage can deplete the battery. Consequently, there is a growing interest in identifying novel data sources for delineating human activities, injecting fresh vitality into this evolving field.

1.3 An Alternative: Wi-Fi CSI

HAR is increasingly important in healthcare for elderly and impaired people, smart homes, and IoT-based solutions. Wearable and visual-based solutions for HAR can be limited in residential environments. Thus, device-free sensing technologies have been investigated, such as WiFi-based approaches. WiFi-based solutions [19] which take advantage of the fact that human actions between WiFi transmitters and receivers will influence WiFi signal characteristics. WiFi signals can be generated without additional cost, and passive activity recognition systems based on WiFi do not require wearable devices.

Informative characteristics of WiFi have been widely accepted due to the abundant and stable information [20]. CSI provides fine-grained physical layer information such as amplitude/phase information for each sub-carrier, making it a possible candidate for sensor information input for HAR. CSI is measured from radio links per orthogonal frequency division multiplexing (OFDM) sub-carriers for each received packet [21]. However, the high noise ratio of raw measurements makes them less representative of human activities.

The raw CSI data for activity recognition is collected using transmitters and receivers. For each packet reception, the data values are extracted into an $NT \cdot NR \cdot NS$ dimensional matrix, where NT and NR represent the number of transmitters and receivers, respectively, and NS denotes the number of sub-carrier groups. The CSI matrix is then flattened into a column vector, with each column representing the time series of values for each sub-carrier group. If only the magnitude is considered for activity recognition, ignoring the phase, the example in Figure 1.3 illustrates the raw data for running, sitting, walking, and standing activities. The raw data can train machine learning models based on the magnitude. We postulate that investigating the method of utilizing CSI to capture human activities is a highly worthwhile pursuit. Firstly, the underlying hardware is exceedingly easy to set up, potentially even ubiquitous within households. Secondly, CSI offers a unique capability to mitigate static object interference enabling robust tracking.



(a) RUN (b) SIT

(c) WALK (d) STAND

Figure 1.3: CSI Phase of Different Activities [22]

1.3.1 Widar System Overview

With the advancement of signal processing technology, Wi-Fi devices have demonstrated commendable performance within specific data domains. Researchers introduced Widar 3.0 in 2021 as a Wi-Fi-based human activity recognition system to facilitate cross-domain recognition. In contrast to their prior works, namely Widar and Widar 2.0, which primarily tracked general human motion states, Widar 3.0 aspires to achieve a more holistic recognition of overall activities while discerning movement trends specific to different body parts. The core innovation of this system centers on domain-independent features related to human posture at a lower-level signal, capturing the unique dynamics of gestures that remain consistent across various data domains. As a result, the model necessitates only a single training iteration but can be flexibly adapted to diverse data domains. Experimentation across multiple domain factors, including environment, location, and human orientation, yielded an impressive recognition accuracy.

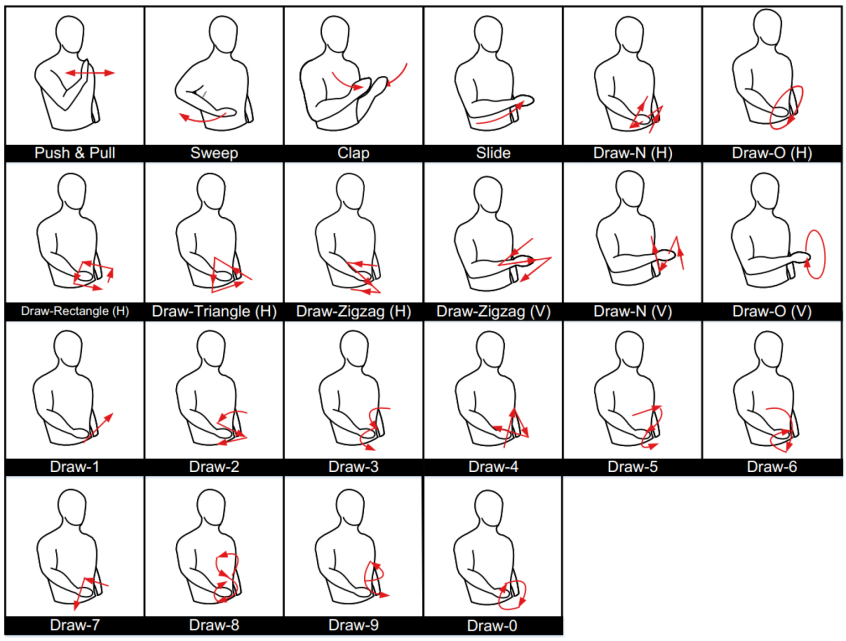
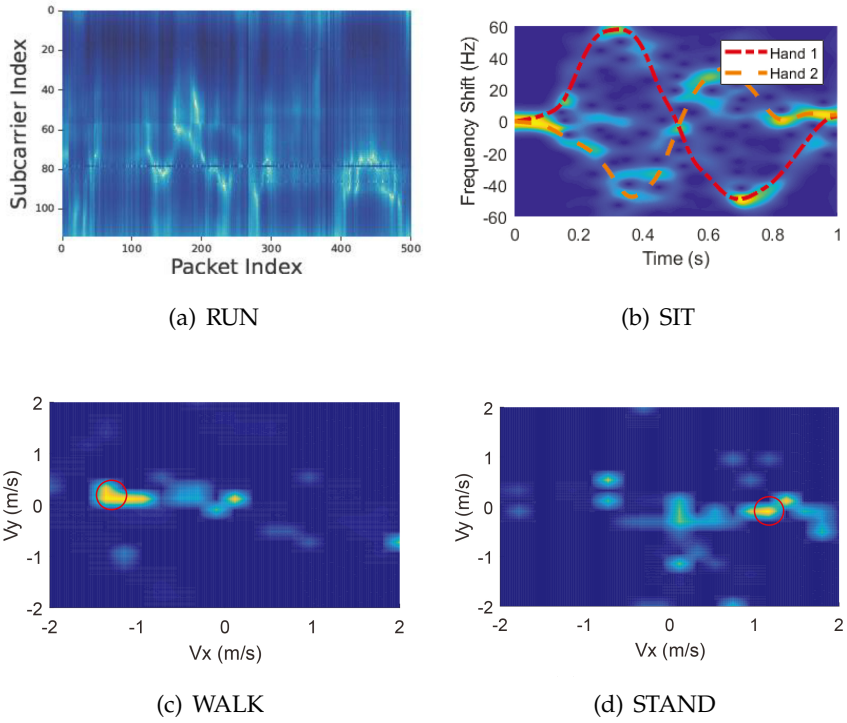


Figure 1.4: Common Gestures Utilized in Human-Computer Interaction [23].

In these experiments, gesture data was gathered from five distinct locations and orientations within each sensing area. The dataset comprises common gestures in human-computer interaction, such as push, pull, swipe, tap, circle, and zigzag. Figure 1.4 illustrates sketches of all the gestures in Widar 3.0. The dataset encompasses



(a) RUN (b) SIT

(c) WALK (d) STAND

Figure 1.5: CSI Phase of Different Activities [22]

a total of 12,000 gesture samples, incorporating contributions from 16 users, spanning five locations, encompassing five orientations, entailing six distinct gestures, and involving five instances for each combination.

The analysis delves into the distribution of velocity components within the human body coordinate system by examining CSI obtained from Wi-Fi Network Interface Cards (NICs). The CSI data encapsulates critical frequency and amplitude information relating to electromagnetic waves, as depicted in Fig. 1.5(a). Each body part exhibits a unique velocity distribution, serving as a distinctive indicator of human activity. This underscores the significance of human body reflection signals, such as Doppler frequency shift (DFS), in conveying essential information about the dynamic characteristics of human activities, as evidenced in Fig. 1.5(b). However, DFS is closely tied to an individual's position and direction, rendering it challenging to identify specific activities solely through DFS profiles. To tackle this, the authors developed a technique to derive a body velocity profile (BVP) from a DFS profile.

Fig. 1.5(c) and Fig. 1.5(d) showcase the BVP as a matrix that quantifies the power distribution of physical velocity within the body coordinate system. Signal power contributed by any velocity component in the human frame is mapped to a specific frequency component in the DFS profile of the link connecting the human to the

Wi-Fi device. This mapping relies on coefficients associated with transmitter and receiver positions. Deriving the BVP from the DFS profile involves extracting the principal components of the CSI stream using a PCA-based algorithm, followed by a short-term Fourier transform to generate the power distribution in the time and Doppler frequency domains, ultimately yielding the BVP within the DFS profile.

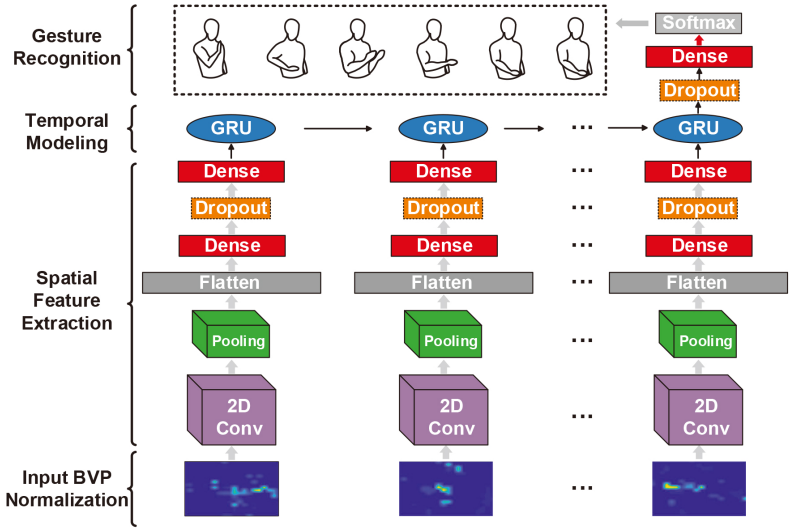


Figure 1.6: Structure of deep learning model for activity recognition [13].

Widar 3.0 incorporates its recognition mechanism, depicted in Figure 1.6, encompassing BVP normalization, spatial feature extraction, temporal modeling, and outlier detection. The normalization phase is pivotal for eliminating extraneous factors that could disrupt the stability of gesture indicators. Spatial features are extracted systematically through a Convolutional Neural Network (CNN) applied to each snapshot of the BVP data. Subsequently, temporal modeling is accomplished through a Recurrent Neural Network (RNN), with the employment of Gated Recurrent Units (GRUs) to capture temporal dynamics. Lastly, outlier detection is employed to identify and discard any new activity classes that do not align with the predefined set. However, it tends to exhibit lower accuracy for activities such as "push and pull" and "circle." This discrepancy may be attributed to factors like hand occlusion from particular angles or information loss during the extraction of BVP data for body parts characterized by significant vertical variations.

1.4 Exploration of HAR Through Wi-Fi CSI

In the experiment utilizing a WiFi-based system, the WiFi environment's transmitters and receivers are always equipped with network interface cards (NICs) [19], which enable capturing the CSI values within the recorded transmissions. To capture the effect of moving or stationary human bodies on wireless signal propagation in the channel, NIC tools capture CSI data packets exchanged between transmitters and receivers. CSI data for human sensing has three dimensions: the number of antennas, subcarrier data packets per antenna, and timestamp numbers.

The values of each subcarrier represent how the signal propagates through diffraction, reflection, and scattering, which provides information about the spatial environment. The amount of sub-carriers is determined by the NIC tool's bandwidth, and more subcarriers naturally result in higher-resolution data. Additionally, for each subcarrier, its temporal dynamics indicate environmental changes.

In wireless communication systems, noise can corrupt CSI data, reducing its accuracy and usefulness. To address this issue, it is crucial to filter the noise and extract relevant features from CSI data. Multiple studies have identified that Principal Component Analysis (PCA) denoising [24] is a common method used for this purpose. PCA is a statistical technique that enables dimensionality reduction and feature extraction. It works by identifying the linear combination of raw features that captures the most significant variation in the data and removes noise from data by discarding the principal components corresponding to the noise. The remaining principal components can then be used for feature extraction and classification.

In 2017, a CSI-based system was proposed for Activity Recognition and Monitoring (CARM) [19]. The authors addressed the limitations of existing systems that use machine learning to discover statistical patterns by building a velocity and an activity model. The former describes the relationship between the frequency of CSI power changes and the speed of human motion, and the latter describes the relationship between the speed of motion of different body parts and specific human activities. The experimental results proved that CARM provides an accurate WiFi-based approach to identifying human activities.

The authors also address technical challenges, such as noise and environmental changes. PCA is the main denoising method used in this research. In the context of CSI streams, PCA can be used to extract the underlying signal corresponding to the motion of a person or object. According to the analysis of the researchers, among the multiple principal components obtained, the first principal component contains most of the noise, and due to the orthogonality of the phase, removing the first component will not cause a significant loss of human motion information.

During feature extraction, CARM extracts activity features from the frequency components of different activities at different time scales. These features capture the duration and frequency of an activity, where duration represents the time spent performing the activity, and frequency represents the multipath velocity due to

body motion during the activity. Additionally, CARM also estimates the trunk and leg velocities using the percentile method introduced in Doppler radar. The experiment achieved an average accuracy rate of 96%. Commercialization attempts were also made, and the overall success indicates the direction for future research.

Another research proposed constructing a public WiFi-based activity dataset named WiAR [25] to reduce time and labor costs, share large amounts of activity data, and promote the development of wireless sensing in practical applications. The study considers four factors that influence the dataset: indoor environment, activity type, activity diversity, and the relative position between transmitters and receivers. The data is collected from three indoor environments (empty rooms, conference rooms, and offices). Sixteen actions are included, which are categorized into upper, lower, and overall actions based on the position of key joints. The diversity of human activities is analyzed in terms of differences between the same activities performed by different volunteers.

Similar studies have gradually emerged, signaling a growing recognition of WiFi-based HAR. These investigations have converged on a shared view regarding the immense potential of CSI data in advancing HAR. WiFi, which involves the analysis of CSI values during wireless signal transmission and reception, has risen as a prominent research focal point within the HAR domain. While researchers have made substantial strides in this arena, current models grapple with certain limitations. They face formidable challenges when tackling the complexities of deep learning and HAR tasks, as these demand the processing of intricate multidimensional data, encompassing both temporal and spatial dimensions, to accurately discern human activities. Furthermore, the omnipresent specters of noise and interference can erode data accuracy, necessitating the development of more robust processing methodologies to surmount these challenges.

In response to these formidable challenges, this paper introduces the Spatial-Temporal Attention-Enhanced Network(ST-ANet), meticulously designed to harness WiFi CSI data for human activity recognition. The ST-ANet bolsters HAR performance by adeptly capturing spatiotemporal multipath information, thus surmounting the formidable obstacles posed by data complexity and noise interference. Specifically, it elevates accuracy by finely delineating dynamic shifts and signal propagation routes within human activities through spatiotemporal data modeling. While existing research has made commendable progress, challenges remain on the horizon. The ST-ANet, which we delve into in the next section, promises to open new vistas in deep learning and HAR. In Chapter 2, we will explore Deep Learning Recognition within WiFi-based HAR. Our journey will encompass the innovative "SlowFast" Two-pathway Action Recognition Model with its related studies, and the attention mechanisms enhancing the learning of temporal features.

Chapter 2

Related Works

In this chapter, we look deep into the realm of Deep Learning Recognition in the context of WiFi-based HAR. We commence by exploring the intricate landscape of deep learning methodologies applied to WiFi-based HAR, shedding light on the transformative potential of these techniques. Within this context, we introduce the Two-pathway Action Recognition Model known as "SlowFast," which stands as an exemplar of deep learning innovation in HAR. Moreover, we turn our focus to the critical facet of enhancing the learning of temporal features, a pivotal element in the domain of HAR. This enhancement is achieved through attention mechanisms, which play a role in refining the model's ability to discern temporal nuances within activity recognition. Within this sphere, we also study the intricacies of self-attention modules, unveiling their profound impact on temporal feature learning. Following this, we put our attention to related studies that have harnessed the power of SlowFast for HAR applications, elucidating the diverse array of insights and advancements that have emerged from this research avenue.

2.1 Deep Learning Recognition on WiFi-based HAR

Numerous types of deep learning models have been applied to the HAR task. Currently, two primary paradigms dominate video-based action recognition algorithms: CNN-based and RNN-based methods [11]. CNN-based algorithms are renowned for their utilization of spatiotemporal information for encoding. This process entails the direct application of convolution operations to extract temporal information, resulting in 3D convolutions that encapsulate both the 2D spatial and temporal features. Notably, they excel in implementing multi-stream network designs that segregate temporal and spatial information extraction [9]. On the other hand, RNN-based models encompass a diverse range of architectures tailored for processing sequential data, like time series or text. They incorporate loops within their structure to facilitate the transfer of information from one sequence step to the next. In addition to CNN and RNN approaches the Transformer architecture has

emerged as a noteworthy contender in the HAR field [26]. Transformers leverage attention mechanisms and excel at capturing spatial and temporal relationships in data.

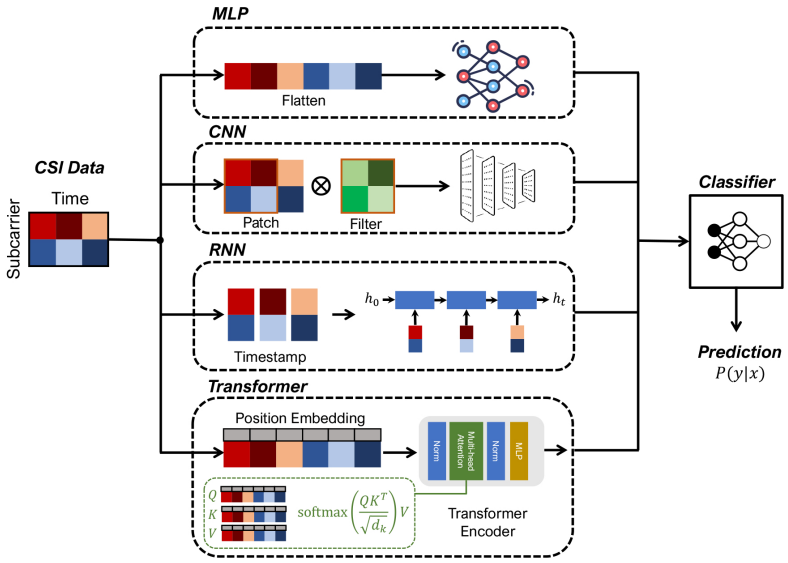


Figure 2.1: Structure of deep learning model for activity recognition [27].

Figure 2.1 illustrates the differences between the three methods and the fundamental concepts underlying the Transformer architecture. Deep learning models are composed of multiple processing layers that can automatically learn high-level representations without requiring heavy feature engineering or domain knowledge. This is especially beneficial for HAR tasks, where features of interest can be complex and challenging to describe. Deep learning models can learn to extract these features from raw data, such as sensor readings or video frames [27]. For instance, CNNs are particularly suitable for HAR tasks because they can capture local and scale-invariant features from time series data.

By using convolutional layers, the network can learn to extract features from different time steps simultaneously, thus capturing temporal dependencies and patterns. As shown in Figure 2.1, CNNs have multiple layers of neurons that can learn to detect both simple and complex features in input data, similar to the specialized cells in the brain's visual cortex that detect edges, lines, and other visual features. The four key ideas of CNN are local connection, parameter sharing, pooling, and multi-layer [28]. Local connectivity means that each neuron is only connected to a small part of the input, allowing the network to learn spatially

local features. Parameter sharing means that all neurons in a given layer use the same weights, reducing the number of parameters and improving generalization. Pooling is a down-sampling operation that reduces the size of feature maps and helps make the network invariant to small spatial translations. Multi-layer refers to multiple convolutional and fully connected layers to enable the network to learn more and more abstract features.

In addition, the success of RNNs in NLP has drawn researchers' attention to their potential in HAR. For instance, long short-term memory (LSTM) networks [29], are used to explore temporal relationships in data, capturing how activity evolves. RNNs are great for modeling time series because they can extract both temporal and semantic information. They can remember previous information and use it to influence the output of subsequent nodes in the short term. However, to capture long-term dependencies, the LSTM structure includes three special gates: the input, the output, and the forget gate.

RNNs and their variants can improve prediction accuracy with more data [30], which can only handle data of a predefined size. Predictions vary over time, making RNNs more sensitive to changes in input data. In HAR, RNNs and their variants are good at exploiting temporal correlations in human activities [31], which is crucial for recognizing them. Researchers have exploited LSTMs to learn complex dependencies across time in features extracted by deep learning models from range-Doppler maps [32]. Through the cooperation of CNN and LSTM, the task of simultaneously exploring spatial and temporal information is completed.

Furthermore, the recent Transformer architecture is based on an attention mechanism that enables the model to selectively focus on different parts of the input sequence during computation [33]. This architecture can be utilized for feature extraction from images by dividing the image into smaller blocks, which are concatenated and enriched with positional embeddings to determine their spatial location. The attention mechanism in the Transformer architecture uses a dot product to calculate the attention between any two patches, which measures the cosine similarity between them. This allows the Transformer to capture spatial-temporal features. With sufficient training data, the Transformer can interconnect with every patch of the image, making it a powerful tool for computer vision applications. However, the Transformer architecture has several parameters, making the training process computationally expensive.

2.1.1 Two-pathway Action Recognition Model: SlowFast

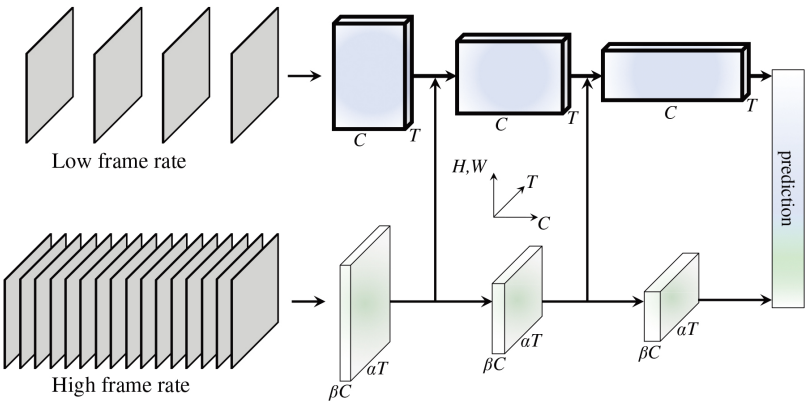


Figure 2.2: Dual-path networks: SlowFast [34].

The statistics of natural spatiotemporal signals suggest that we should not treat space and time symmetrically in video recognition tasks. Slow motions are more common than fast motions, and our perceptual system is biased towards slow movements [35]. Therefore, it may be beneficial to factor the architecture to treat spatial structures and temporal events separately. Recognizing categorical spatial semantics of visual content, such as object identities, can evolve slowly and be refreshed relatively slowly. In contrast, the motion being performed can evolve much faster than their subject identities, such as clapping, waving, shaking, walking, or jumping. This can be achieved by using fast refreshing frames (high temporal resolution) to effectively model potentially fast-changing motion while recognizing the categorical semantics at a slower pace.

The authors of a 2019 research [34] propose a new architecture for video recognition called SlowFast networks. The SlowFast network architecture offers a promising approach to capturing spatial and temporal information, resulting in improved accuracy. It consists of two pathways: a Slow pathway that operates at a low frame rate to capture spatial semantics and a Fast pathway that operates at a high frame rate to capture motion at fine temporal resolution.

The Slow pathway is a convolutional model in the SlowFast network that operates on a video clip as a spatiotemporal volume. It processes only one key frame out of several frames due to a large temporal stride on input frames. The Fast pathway, on the other hand, is designed to capture motion at a fine temporal resolution and has a higher input resolution than the Slow pathway [34]. It also maintains high-resolution features throughout the network hierarchy and does not use temporal downsampling layers. However, its lower channel capacity makes it more

lightweight and computationally efficient, indicating a tradeoff between temporal and spatial modeling abilities.

In essence, the Fast pathway works in parallel with the Slow ones and aims to achieve a fine representation along the temporal dimension, by using a smaller temporal stride and a higher frame rate than the Slow pathway, enabling it to sample frames at a higher density.

Lateral connections play a vital role in fusing information from the two pathways in the SlowFast network to ensure that each path is aware of the representation learned by the other [36]. This technique has been extensively used in optical flow-based two-stream networks and image object detection to incorporate different levels of spatial resolution and semantics. Figure 2.2 shows a lateral connection between the two paths in each stage. These connections allow the features of the Fast pathway to be fused into the Slow pathway [37]. However, since the two pathways have different time dimensions, lateral connections need to perform transformations to align them. Finally, global average pooling is performed on the output of each pathway, and the resulting pooled feature vectors are concatenated to form the input to a fully connected classifier layer.

The SlowFast method is gradually becoming the mainstream approach to video-based action recognition, owing to its ability to separate temporal and spatial information and enhance recognition accuracy. It has gained increasing popularity in recent years and is now considered one of the leading methods in this field [38]. Its success has inspired further research into developing more sophisticated algorithms capable of extracting and processing video data more effectively and efficiently. An interesting idea to explore further is the tradeoff between temporal and spatial modeling abilities and their impact on the performance of the SlowFast network. Consequently, this paper aims to investigate novel and innovative approaches for leveraging its functionality and enhancing the accuracy of video-based action recognition algorithms. It is anticipated that the impact of the SlowFast methodology will continue to expand in the future.

2.2 Related Studies Using SlowFast

After Facebook's AI team proposed the SlowFast network structure, they went on to propose the Audiovisual SlowFast network [36], which aims to integrate audio and visual perception by developing a unified representation of sound and vision. This approach is based on the idea of having two visual paths - a slow and a fast path - which are deeply integrated with a faster audio path, as shown in Figure 2.3. The AVSlowFast network fuses audio and visual features at multiple levels, allowing audio to contribute to a hierarchical audio-visual concept. To enhance the synchronization between audio and visual modalities, hierarchical audiovisual synchronization is performed, inspired by previous neuroscience research. This technique enables the network to learn joint audio-visual features [39]. The

AVSlowFast network is evaluated on six video action classification and detection datasets, achieving state-of-the-art results. One challenge in training AVSlowFast networks is the different learning dynamics between audio and visual modalities. To address this challenge, the authors introduce DropPathway, a regularization technique that randomly drops audio paths during training. The authors also conduct detailed ablation studies to understand the impact of different components on performance. Overall, the AVSlowFast architecture represents an important step towards integrating audio and visual perception into a unified representation. It is also pointed out that the SlowFast network can flexibly increase the path to increase the overall number of feature extractions, and the added feature channels can further strengthen the learning effect.

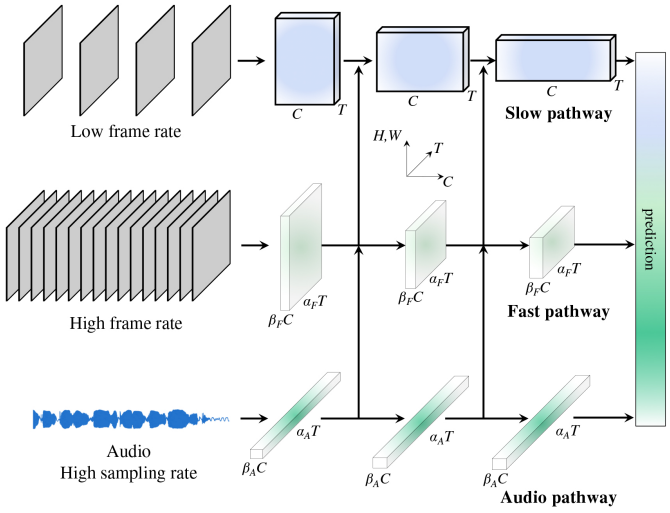


Figure 2.3: Add a audio path to SlowFast [36].

SlowFast provides a lightweight and effective spatiotemporal feature learning network for video-based action recognition research. In a recent study, a new architecture called spatiotemporal ResNets is introduced for human action recognition in videos, which combines the advantages of SlowFast two-stream convolutional networks and residual networks [40]. The model is initialized with pre-trained ResNets for image classification, and the convolutional dimension map filters are converted to temporal filters, allowing the network to operate over a large temporal range of inputs. The entire model is trained end-to-end for hierarchical learning of complex spatiotemporal features and is evaluated on two standard action recognition benchmarks, where it outperforms previous state-of-the-art results.

In another study, the importance of object detection algorithms in vision tasks

was highlighted, and the limitations of SlowFast's detection algorithm were pointed out in terms of both detection accuracy and speed [41]. To address this issue, the paper proposes utilizing YOLOv3, YOLOX, and CascadeRCNN to improve detection accuracy and speed. The author points out that SlowFast's detection algorithm FasterRCNN is not perfect. As a two-stage object detection algorithm, FasterRCNN generates region proposals through a Region Proposal Network (RPN). The RPN structure generates region proposals, maps proposals to feature maps, and then obtains the corresponding feature matrix. The generated feature map is then classified, and outputted by the fully connected layer to obtain the prediction. In contrast, YOLOv3 is a one-stage object detection algorithm that takes the entire image as input and directly predicts the location and class of the bounding box using a neural network. CascadeRCNN, an extension of FasterRCNN, uses multiple stages of classification and regression to improve accuracy. The paper emphasizes the importance of enhancing SlowFast's detection algorithm.

In a 2020 study, the authors propose a joint utilization of 3D convolution and post-temporal modeling for action recognition in videos [42]. They emphasize the significance of spatial and temporal information in action recognition, where spatial information denotes static information in a scene and temporal information captures the dynamic nature of actions. The authors critique Temporal Global Average Pooling(TGAP), which fails to make full use of temporal information and ignores the ordering of temporal features. They suggest using an attention mechanism to determine which temporal features are more important. To better use temporal information, the authors replace the traditional temporal global average pooling layer with a Bidirectional Encoder Representation from Transformer (BERT) layer [43], which utilizes BERT's attention mechanism. Their experiments show that BERT's attention mechanism outperforms the traditional temporal global average pooling layer used in 3D CNN architectures such as ResNeXt, I3D, and SlowFast.

To implement BERT on the SlowFast architecture, the authors propose two alternative solutions: early-fusion BERT and late-fusion BERT. In the early-fusion BERT, temporal features are concatenated before BERT layers, and a single BERT module is used. In late-fusion BERT, two different BERT modules are used, one for each stream, and the outputs of the two BERT modules from both streams are concatenated. Both BERT solutions outperform the standard SlowFast architecture, but the improvement of early-fusion methods is limited due to the destruction of the temporal richness of fast streams [43]. Moreover, the higher temporal resolution in the SlowFast architecture and the implementation of the two-way structure requires consideration of the increase in model parameters.

The SlowFast architecture offers numerous advantages, particularly in accelerating video recognition research. In a 2022 study, researchers used a dual-stream model represented by SlowFast and proposed a Temporal Correlation Module (TCM) for action recognition in videos [44]. The TCM extracts action visual tempo, which characterizes action dynamics and temporal scale. It comprises two main

components: a Multi-scale Temporal Dynamics Module (MTDM) and a Temporal Attention Module (TAM).

The MTDM is a computer vision component that extracts efficient temporal dynamic features for fast- and slow-paced motion. It involves three main steps: feature source utilization, visual similarity computation, and motion estimation. In the feature source utilization step, the MTDM extracts slow-paced and fast-paced motion visual rhythm features by sampling deep features at different rates. This can efficiently utilize single-layer deep features and is not constrained by the backbone network. On the other hand, the TAM aims to capture local cross-temporal interactions between adjacent frames to enhance temporal information while reducing the influence of non-important features during training [45]. This module learns temporal attention using a band matrix, where the weights of temporally aggregated features are computed by considering only the temporal interactions with their k neighbors. The scope of temporal interactions is determined by a function of the characteristic temporal dimension, which expands the temporal receptive field and captures both fast-paced and slow-paced information. TAM can better enhance useful slow and fast visual beat information and suppress unwanted information, making it more effective in video recognition.

In summary, TCM combines SlowFast and attention mechanisms to increase the model's sensitivity to temporal features. The researchers also verified the method on multiple datasets, and the results demonstrated that it can accurately identify the temporal characteristics of human action data [11]. This provides a valuable reference for future research.

2.3 Enhance Learning of Temporal Features Using Attention Mechanism

In recent years, there has been an increasing number of studies that enhance the ability of spatial-temporal SlowFast networks to capture both the local and overall context of data for a better understanding of human actions. To this end, a proposed model has integrated the self-attention mechanism to extract four important features in video information: spatial information, temporal information, slow motion information, and fast motion information. By utilizing the self-attention mechanism, the network can extract global semantic context, which can greatly improve the accuracy of action recognition [46]. Overall, incorporating the self-attention mechanism into the SlowFast two-way network has shown promising results and can be applied to various applications.

2.3.1 Self-attention Modules

Self-attention, also known as intra-attention, is an attention mechanism that associates different positions of a single sequence to compute a representation of

the sequence. This attention mechanism has been successfully used in various natural language processing tasks, such as reading comprehension, abstract summarization, textual entailment, and learning task-independent sentence representations [46].

Vaswani and his team proposed the Transformer [47], a transduction model that relies entirely on self-attention to compute representations of its inputs and outputs, without using sequence-aligned RNNs or convolutions. The Transformer is a significant advancement in exploiting attention mechanisms because it efficiently preserves long-range dependencies between distant locations. We aim to apply self-attention to video feature extraction and study how to extract critical information from continuous action frames containing temporal semantic relationships. This includes considering the spatial structure of the human body and continuous action changes.

HAR is of great importance in computer vision and pattern recognition, which involves detecting and classifying human actions. A new short-term HAR model has been proposed [33], aimed at classifying actions over short time steps in the past, which is crucial for real-time applications such as robotics.

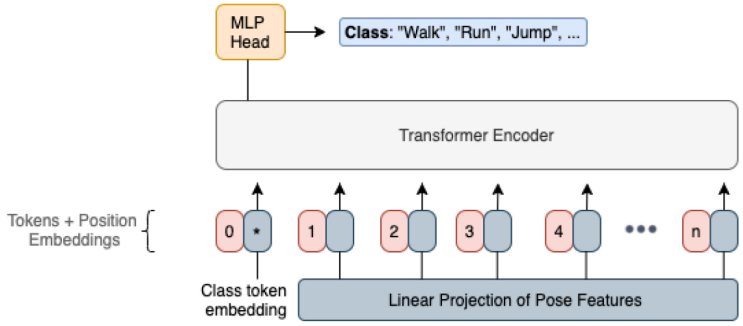


Figure 2.4: Learnable positional embedding in Transformer architecture [33].

This study is inspired by the Transformer architecture and proposes a unique approach to solving the problem. The Transformer-encoder takes 2D human pose estimation, linearly projects it to the dimensions of the model, and adds a class label before passing it through the encoder, as shown in Figure 2.4. The ViT model's learnable positional embedding is also included with each input token. The encoder is built based on a multi-layer multi-head self-attention and feed-forward network, and the output class tokens are then passed through a multi-layer perception head to obtain the final class prediction. The architecture is designed for short-duration human action recognition, with a focus on real-time applications.

This new approach to HAR has several benefits, including improved accuracy, robustness, and faster performance. The proposed model can be implemented in

real-time applications such as robotics, where quick response times are crucial [48]. Overall, this study provides a promising solution to the problem of HAR and demonstrates the potential of the Transformer architecture in the field of computer vision and pattern recognition.

2.3.2 SlowFast HAR Models Utilizing Attention Mechanism

In 2020, a study proposed a Spatial-Temporal SlowFast Self-Attention Network for action recognition [26], which combined global context and long-term dependencies using a self-attention mechanism. The study highlights that human behavior can be divided into human movement and human action, which require considering both the movement and the movement of limb parts to solve the action recognition problem. The proposed network integrates spatial and temporal attention mechanisms to focus on relevant regions and action times. On the spatial scale, human limbs are regarded as spatial features, while on the temporal scale, the duration of each action and the resulting spatial coherence are considered as temporal features for action recognition [49]. The main contribution of this method is the combination of the SlowFast network and the self-attention mechanism, which addresses the limitations of local feature-based methods of convolutional neural networks. Furthermore, this approach combines global context and long-term dependence to improve accuracy.

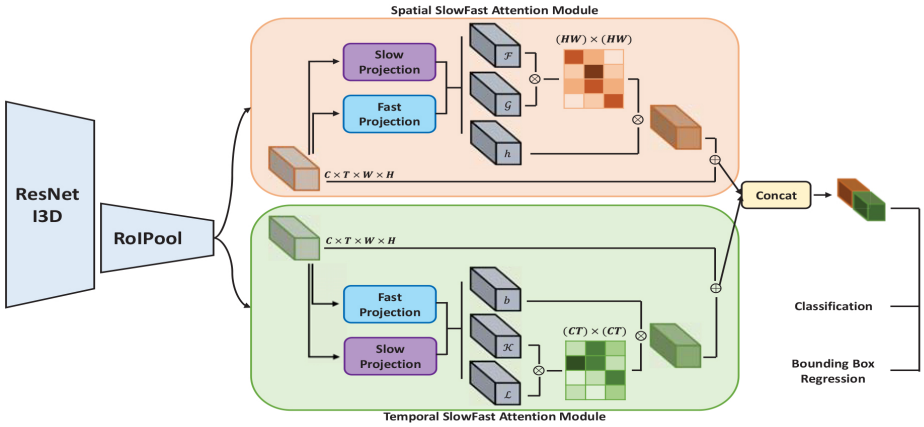


Figure 2.5: Overall architecture of The Spatio-Temporal SlowFast Self-Attention Network [26].

This paper focuses on two main topics: action recognition and self-attention. Regarding action recognition, the paper highlights the importance of understanding the relationships between people and objects in video data. However, capturing

long-term dependencies using CNNs remains a challenge. To address this, the paper proposes using self-attention to focus on meaningful regions that are crucial to the target and can learn long-term dependencies while focusing on important features. Self-attention has been successfully used in image generation tasks, such as the Self-Attention GAN (SAGAN). The paper applies this module to video understanding tasks to enhance the model's ability to detect long-range interactions and important contexts, as shown in Figure 2.5. By doing so, the proposed Spatio-Temporal SlowFast Self-Attention Network can extract four types of features, including spatial information, temporal information, slow action information, and fast action information, to enable better action recognition performance.

The authors have introduced a spatial attention module to capture contextual information such as hands and faces, which is crucial for determining human behavior. This module is designed to focus on spatial features as well as other contextual information, such as hands and faces, and is based on the self-attention module for image understanding. However, it has been modified to find spatially significant parts of the entire video features. Additionally, the temporal attention module has been introduced to focus on important regions of the temporal axis. These two modules are capable of separating slow and fast path features, as the amount of feature information differs between them. To compute attention maps, video features are first projected into two new feature spaces in the attention module. The output of the attention layer is then multiplied by a scale parameter and added to the initial input feature map.

The self-attention mechanism employs spatial encoding to enhance the deep learning model's comprehension of spatial features [33]. However, attention is computationally expensive because of the intricate input sequence length. Current techniques for diminishing the number of tokens produced during spatial encoding have limitations, including structured spatial compression and unstructured token pruning. Methods aimed at reducing the computational burden of models linked with Vision Transformers (ViT) have garnered significant research interest in recent years.

In a 2022 study, the author proposes Evo-ViT [50], a self-motivated slow-fast token evolution approach for dynamic Vision Transformers (ViT) that addresses the inefficiency problem of modeling long-range dependencies among tokens in ViTs from the beginning of the training process while suitable for structured compression methods, as shown in Figure 2.6. The proposed approach distinguishes informative tokens from placeholder tokens for each instance in an unstructured and dynamic way and updates the two types of tokens with different computation paths.

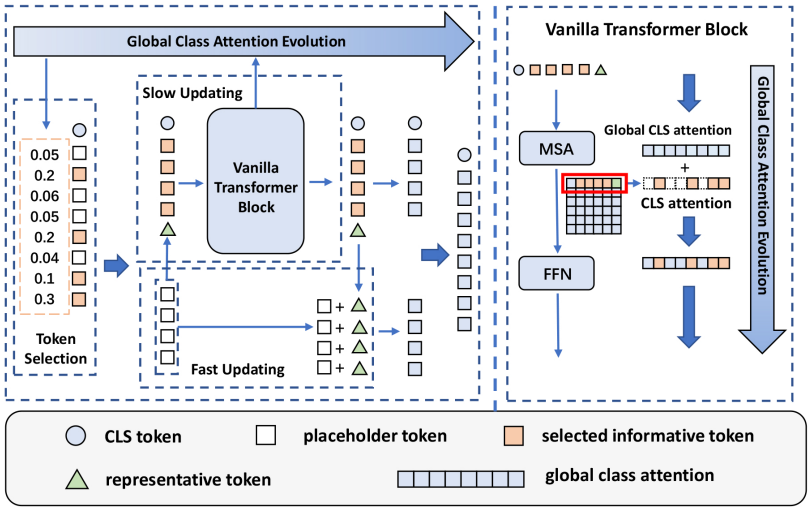


Figure 2.6: The pipeline of Evo-ViT: token selection and SlowFast token updating [50].

ViT is an image classification model that utilizes the Transformer architecture [51]. It divides images into blocks and linearly projects them into positional embeddings, which are combined with extra class tokens (CLS) to create a global image representation. These tokens, including the CLS token, are then fed through stacked transformer encoders for final classification. Each encoder contains a multi-head self-attention (MSA) module and a feed-forward network (FFN) module. The MSA module is an extension of the self-attention module, where queries, keys, and values are split and computed in parallel alongside the image block operation.

Compared with ViT, Evo-ViT solves the problem of slow training speed by dynamically distinguishing informative tokens from placeholder tokens [52]. This is achieved by processing each input instance from the very beginning of the training process. The architecture of Evo-ViT consists of two main modules - a structure-preserving token selection module and a slow token update module. The structure-preserving token selection module determines informative tokens and placeholder tokens by evolving global class attention and leveraging residual connections to regularize the attentional information flow. In the subsequent slow-fast token update module, the informative token is carefully evolved through the MSA and FFN modules. In contrast, the placeholder token is roughly summarized and updated through the representative tokens.

In conclusion, the proposed method has a unique approach to preserving all tokens, which ensures complete information flow throughout the training process. This enables the model to update tokens at both slow and fast rates, rather

than simply discarding placeholder tokens. This research provides a novel way of synthesizing temporal and spatial features, and effectively integrates self-attention and position encoding, leading to improved model performance in dealing with complex objects [53]. As a result, this method can enhance the model's ability to provide more complete and accurate functions. At the same time, several studies have examined the effectiveness of the SlowFast model in extracting complex features from video data. These studies have shown that the SlowFast network can not only effectively integrate spatio-temporal features, but also demonstrates good flexibility, making it a promising option for various applications. Building on this phenomenon, this paper aims to propose an attention mechanism that can accurately identify spatial features and address the statistical noise generated by the SlowFast model's two-way architecture. The proposed mechanism is expected to improve the ability of feature extraction and enhance its overall performance.

Chapter 3

Methodology and Evaluation

3.1 Proposed ST-ANet: Architecture and Multipath Components

Our particular dataset presented challenges because it contains 3D data, featuring consecutive frames of BVP data, with dimensions $H \times W$ and T frame rates. Therefore, the input to our network is represented as $X_{in} \in \mathbb{R}^{C \times T \times H \times W}$, laying the foundation for the introduction of ST-ANet, an architectural innovation designed to meet these demands. ST-ANet is in its ability to expertly convert continuous BVP frame data into video dimensions for processing. Additionally, it utilizes slow frames, residual blocks, and self-attention encoders. The network is not only customized for the complexity of the Widar dataset but also engineered to efficiently extract and process critical information from the wireless signal data. Furthermore, due to the generality of the self-attention mechanism, it possesses the versatility to process video data.

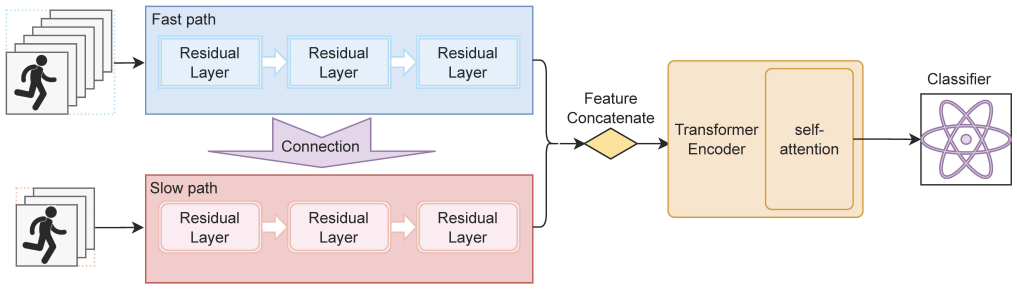


Figure 3.1: Overall architecture of the ST-ANet.

At the heart of ST-ANet is the Slow-Fast framework, as shown in Figure 3.1, the framework operates at two different frame rates, capturing complex temporal nuances and providing high temporal resolution. It consists of two basic paths: the slow path and the fast path.

The slow path is specialized for processing wireless signal data into a spatiotemporal representation. By taking more time steps (τ) over the input frames, it carefully analyzes each τ frame, allowing it to capture long-term temporal dependencies. In this case, $T = 11$ frames represent the period of slow-path processing. In contrast, the fast path specializes in preserving fine temporal details by exploiting smaller time steps (τ/α), where $\alpha > 1$ represents the frame rate ratio between the fast and slow paths. This approach produces densely sampled αT frames, ensuring excellent temporal fidelity. In our experiments, the typical value of α is set to 8. However, considering the limited period in the dataset and the short-lived nature of the action duration, α is set to 2, and the fast path processes $\alpha T = 22$ frames.

Both pathways benefit from the introduction of "horizontal connections", implemented to facilitate the exchange of information between them. These connections enable the slow path to leverage insights gained by the fast path, promoting a complete understanding of the data. By fusing temporal features extracted by the fast path, these connections prevent over-fitting while preserving the overall nature of the data.

In the complex framework of ST-ANet, "residual blocks" play a crucial role in enhancing feature extraction. These modules draw inspiration from the success of Residual Networks and are positioned in the architecture. What sets them apart is their adaptation to the slow paths. It is worth noting that in the slow path, these residual blocks are carefully integrated, but with the introduction of temporal convolutions in the lateral connections, as opposed to relying solely on 2D convolution kernels. This design choice meets the need for a larger spatial receptive field, especially when analyzing fast-moving objects. At the same time, it ensures that the features to be fused can also be converted into the same dimension size.

To further enhance the processing capabilities of cascaded features derived from dual-path CNN cascades, ST-ANet integrates a "self-attention encoder". This addition endows the network with the ability to focus on relevant temporal and spatial contexts in the data. Thanks to the self-attention mechanism, self-attention encoders excel at capturing complex data relationships. By applying self-attention to concatenated features, the network acquires the ability to discern and emphasize relevant temporal and spatial cues. This enhancement enables the network to excel in complex tasks, including but not limited to video understanding, precise object recognition, and accurate action recognition. Therefore, the self-attention encoder positions ST-ANet as a powerful tool in the field of deep learning, especially in the fields of video analysis and data understanding.

3.1.1 Fast Pathway: Capturing Temporal Information

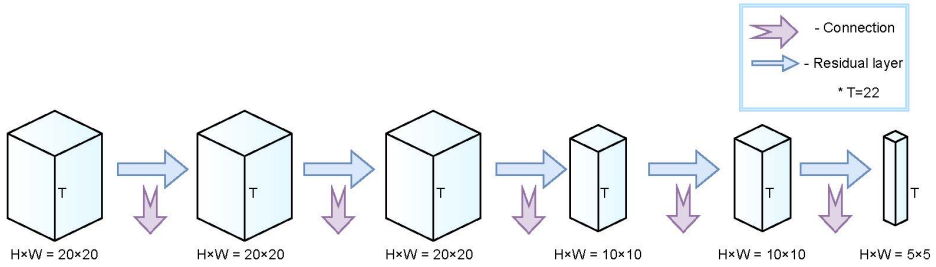


Figure 3.2: The fast pathway operates at higher frame rates.

The fast path emerges as a dynamic counterpart to its slower sibling, the slow path. Unlike the slow path, the fast path operates at a higher frame rate, enabling it to process video frames at an accelerated pace. This acceleration is a key advantage as it allows the fast path to capture temporal information.

Within this accelerated journey, the fast path encounters a sequence of video frames, denoted as $x_f = X_{in} \in \mathbb{R}^{C \times T \times H \times W}$. It's worth highlighting that this path excels in handling consecutive frames, which is crucial when operating at a higher frame rate, typically at T = 22 frames per second.

The distinguishing feature of the fast path lies in its convolutional layers, as shown in Figure 3.2. Much like its slower counterpart, these layers play a fundamental role in extracting features from individual frames. However, their primary focus within the fast path is the detection of motion and dynamic patterns within the video data. To enhance ST-ANet's capacity for robust performance and ensure stable training, we introduce the concept of a residual network, often referred to as ResNet. This concept has significantly impacted various domains, including image classification, object detection, and speech recognition. The core innovation of a residual network addresses the challenges of vanishing and exploding gradients encountered during the training of deep neural networks. It accomplishes this feat by incorporating residual blocks and skip connections, leading to more effective training and improved model performance.

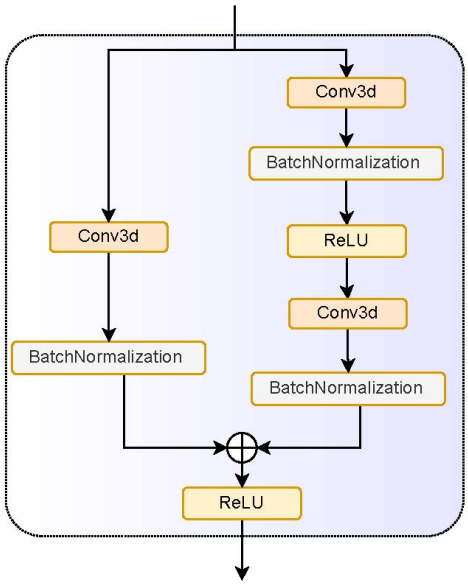


Figure 3.3: The content of a residual block.

As shown in the figure 3.3. The process begins with an input feature map, represented as $x_n = H_{n-1}(x_{n-1})$. This represents the feature map at a particular layer, denoted by n, which depends on the previous layer $n-1$. A series of operations, including convolution, batch normalization, and activation functions, are applied to the input feature map. These operations result in the learning of a residual map, denoted as $F_n(x_n)$, capturing essential features related to motion and dynamics. Finally, the learned residual map $F_n(x_n)$ is added back to the original input x_n, which can be mathematically expressed as $H_n(x_n) = F_n(x_n) + x_n$. The result of this addition forms the output of the residual block.

A critical choice arises between traditional temporal pooling layers and a novel approach involving convolutional kernels with a stride of two. This shift in strategy revolutionizes the game when preserving essential temporal information. Temporal pooling layers aim at reducing the temporal dimension of time series data, curbing computational demands, and paring down model parameters. However, temporal pooling selectively cherry-picks representative samples within predefined time windows.

As the convolution kernel traverses the data, it skips over two time steps at each slide, efficiently achieving down-sampling within the temporal dimension. This approach enhances the network's grasp of temporal patterns and dynamics. It operates efficiently with fewer parameters, allowing for an increase in network depth without an undue computational burden. The adaptive nature of stride-two

convolution empowers the network to autonomously unearth time-related features, eliminating the need for manual intervention in specifying time pool window sizes.

In summary, replacing temporal pooling with stride-two convolutions enhances its capacity to represent intricate patterns while mitigating information loss. With this shift in technique, the Fast Path sets its sights on the potent realm of 3D convolutions, honing in on temporal features intricately interwoven with motion and dynamics within videos. These temporal nuances extend their reach to the domain of action recognition, where the Fast Path excels at profiling actions by scrutinizing alterations in object position and shape over time. It also shows the ability to decipher intricate movement patterns like gestures, a critical skill for recognizing complex actions such as sign language. The fast path in this article comprises a total of five residual layers, culminating in the final output expressed as $F_{\text{Fast}} = H(x_f)$.

3.1.2 Slow Pathway: Capturing Spatial Information

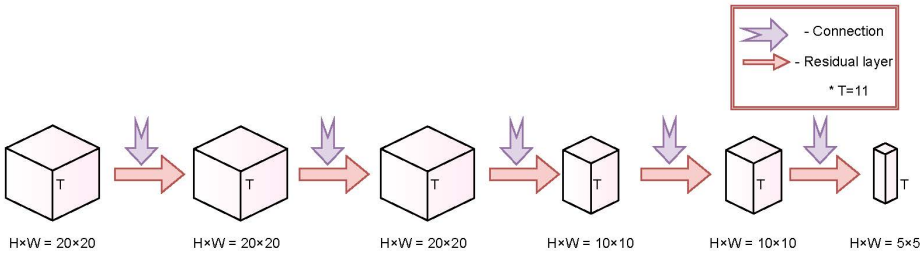


Figure 3.4: The slow pathway to capture static spatial information.

In the context of the slow-fast architecture, the slow path assumes a role, primarily geared toward capturing prolonged temporal dependencies within video data. The slow path's operational strategy involves the utilization of a larger time step parameter denoted as (τ). This distinctive approach facilitates a meticulous examination of each time step, effectively ensnaring extensive temporal dependencies present in the video sequence. In this study, we have opted for a τ value of 2, signifying the processing of $x_s = X_{in} \in \mathbb{R}^{C \times T \times H \times W}$ with $T = 11$ frames.

The slow path, as shown in Figure 3.4 adopts a convolutional model for treating video clips as spatiotemporal volumes, a configuration well-suited for spatial feature extraction. This augmentation significantly enhances our grasp of the spatial intricacies intrinsic to video content.

A salient characteristic of the slow path lies in the integration of lateral connections, a feature for seamless information exchange between the slow and fast

paths. These lateral connections are meticulously tailored to align with the temporal dimension, facilitating the amalgamation of fast path features into the slow path. This amalgamation empowers us with a comprehensive understanding of the video dataset.

The slow path exhibits prowess in efficiently capturing protracted temporal dependencies within videos, a capability of paramount importance for tasks demanding consideration of object transformations over extended periods. By deploying convolutional models on video clips, the slow path furnishes us with a deeper comprehension of the spatial and temporal nuances inherent in videos. This enriched understanding proves invaluable for analyzing intricate video data, such as object motion and action recognition. The inclusion of lateral connections further elevates the model's overall performance, countering the risk of over-fitting, preserving data fidelity, and reinforcing model resilience.

Slow paths hold a role within slow-fast architectures, characterized by their ability to encapsulate prolonged temporal dependencies and bestow profound insights into video content. This attribute proves indispensable when dealing with video analysis tasks. The incorporation of lateral connections stands as a testament to its capabilities, effectively addressing the multifaceted challenges posed by the analysis of complex video data. Owing to its slower frame processing speed, the slow path excels in learning spatial features. This proficiency encompasses object identification and detection within each frame, along with the discernment of object relationships. For instance, it adeptly identifies players, balls, and courts in basketball game videos, providing invaluable static spatial context for subsequent tasks such as action recognition. Consequently, the output of the Slow Path is succinctly expressed as $F_{\text{slow}} = H(x_s)$. Through the slow path, the model furnishes a profound analysis and comprehension of the static spatial information inherent in the video dataset.

3.1.3 Feature Fusion Using Attention Mechanisms

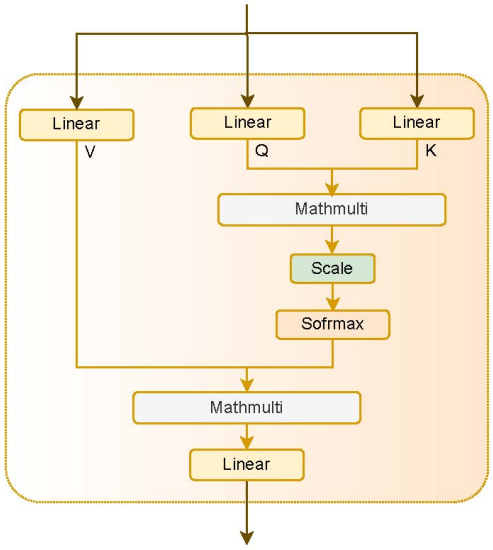


Figure 3.5: Overall architecture of the ST-ANet.

In the field of computer vision, the SlowFast network architecture has become a strong contender, known for its ability to effectively combine spatial and temporal features. This fusion capability is at the core of its huge success in tasks such as action recognition. In the context of SlowFast networks, the incorporation of attention mechanisms constitutes a key innovation, ushering in a new era of feature fusion and improved prediction accuracy. Its main function is to enable neural networks to selectively focus on specific areas of input data while reducing the relevance of less relevant information. This selective attention mechanism, when integrated into the SlowFast network, will be key to seamlessly aligning and merging spatial and temporal features.

The basic principles underpinning the operation of the attention mechanism can be articulated through mathematical formalism. Let us represent the input feature tensor as

$$X_a = x_f \bigoplus x_s \in \mathbb{R}^{C \times T \times H \times W}$$

, where T represents the time dimension, C represents the number of channels, H and W correspond to the height and width of the spatial dimension, respectively. In the context of SlowFast networks, spatial and temporal features are represented in this tensor.

The attention mechanism, as shown in Figure 3.5, introduces the transformation of the input tensor X to calculate the query (Q), key (K), and value (V) tensors, which are the essential components of the attention mechanism:

$$Q = X_a W_Q, \quad K = X_a W_K, \quad V = X_a W_V \tag{3.1}$$

Here, W_Q, W_K, and W_V represent learnable weight matrices that project the input tensor into query, key, and value spaces. The dimensions of these weight matrices are often adjusted to fit the specific design of the network.

The attention weight (A) is calculated as a function of the query (Q) and key (K) tensors:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{3.2}$$

In this equation, softmax represents the softmax activation function, and d_k represents the dimension of the key (K) vector, which determines the scale of the attention score.

Subsequently, the weighted sum of the value (V) tensors is calculated using the attention weights (A):

$$\text{Attention}(X_a) = AV \tag{3.3}$$

This process is performed in all spatial and temporal dimensions, facilitating the alignment and fusion of features within the SlowFast network.

The deployment of attention mechanisms in the SlowFast architecture enables precise alignment of spatial and temporal features, allowing the model to identify key patterns and relationships between these dimensions. This judicious feature fusion ensures that spatial and temporal information contributes harmoniously to the final prediction output, thus improving the efficiency of the network in the field of action recognition.

3.2 Evaluation Metrics and Experimental Setup

Cross-entropy loss is the main tool for evaluating the model's competence. In the context of classification, the mathematical formulation of the cross-entropy loss is as follows:

$$H(y, p) = -\left[y \cdot \log(p) + (1 - y) \cdot \log(1 - p)\right] \tag{3.4}$$

Here, $H(y, p)$ represents the cross-entropy loss, y embodies the actual class labels, which can take values of either 0 or 1, and $p = Linear(Attention(X_a))$ symbolizes the model's predicted probability that the input belongs to class 1, which is typically the positive class of interest.

One of the attributes of cross-entropy loss lies in its seamless integration with gradient-based optimization algorithms, including the venerable stochastic gradient descent (SGD) and its numerous variations. This compatibility streamlines the computation of gradients and facilitates the adjustment of model parameters during the training process. Consequently, our training regimen attains computational efficiency, leading to accelerated model convergence and more effective learning.

The issue of class imbalance, a common challenge in many real-world datasets, particularly in human activity recognition tasks, looms large. Fortunately, cross-entropy loss offers an elegant solution by imposing more significant penalties on misclassified instances from the minority class, thus addressing the imbalance issue with finesse. This property encourages our model to assign greater importance to the accurate identification of rare and critical activities within the dataset. In essence, cross-entropy loss fosters the generation of well-calibrated probability estimates, a critical facet with far-reaching implications, including improved uncertainty estimation and more robust risk assessment.

In summation, the selection of cross-entropy loss as our primary loss function stems from a myriad of advantages: its innate interpretability, harmonious synergy with optimization algorithms, prowess in addressing class imbalance, support for probabilistic outputs, and a well-documented history of success across various classification domains. Empirical experiments provide resounding evidence, underscoring the astute decision behind this choice. It significantly elevates the overall efficacy of our model in navigating the intricacies of human activity recognition grounded in the rich tapestry of Wi-Fi CSI data.

3.3 Widar Dataset

The Widar dataset is a crucial resource within the field of wireless signal analysis, with a particular emphasis on human gesture recognition, indoor positioning, and wireless sensing. Researchers and developers frequently turn to this dataset for its comprehensive insights. At its core, the Widar dataset relies on Wi-Fi CSI measurements, denoted as $\hat{H}$, meticulously captured by readily available Wi-Fi devices. These measurements delve into the intricate multi-path effects of wireless signals within indoor environments, revealing crucial information about their behavior at specific time instances (t) and frequencies (f). The equation representing CSI measurements is as follows [54]:

$$\hat{H}(f,t) = \sum_{l=1}^{L} \alpha_l(f,t) e^{-j2\pi f \tau_l(f,t)} e^{j\epsilon(f,t)} \tag{3.5}$$

Here, L signifies the number of propagation paths, while α_l and τ_l represent the complex attenuation and propagation delay of the l-th path, respectively. Furthermore, $\epsilon(f,t)$ accounts for the phase error introduced by factors such as timing

alignment offset, sampling frequency offset, and carrier frequency offset.

To comprehensively understand these CSI measurements in the context of human activities, DFS (Doppler Frequency Shift) profiles are meticulously calculated. These profiles provide a perspective by representing the distribution of signal power over Doppler frequencies. The equation for DFS representation is given as [54]:

$$\hat{H}(f,t) = Hs(f) + \sum_{l \in P_d} \alpha_l(t) e^{j2\pi \int_{-\infty}^{t} D_l(u)du} \tag{3.6}$$

In this equation, Hs encapsulates the cumulative effect of static signals with zero DFS (e.g., line-of-sight signals), while P_d characterizes dynamic signals with non-zero DFS, such as signals reflected by a person. The term D_l encapsulates the Doppler frequency shift associated with the l-th path.

One of the components of the Widar dataset is the BVP, which serves as a distinctive indicator of human activities. BVP is derived from DFS profiles and represents the distribution of signal power over velocity components within the body coordinate system. This relationship between DFS profiles and BVP is encapsulated in the equation [54]:

$$D(i) = c(i)A(i)V \tag{3.7}$$

Here, $D(i)$ signifies the DFS profile emanating from the i-th link, $c(i)$ accounts for the scaling factor related to propagation loss, $A(i)$ denotes the assignment matrix facilitating the connection between DFS profiles and BVP, and V represents the BVP matrix embodying velocity components.

BVP estimation from DFS profiles is a critical step in the dataset's creation and analysis. This process involves an optimization approach aimed at minimizing the Earth Mover's Distance (EMD) between the estimated BVP and the observed DFS profiles. The optimization problem can be framed as follows [54]:

$$\min_{V} \left(\sum_{i=1}^{M} |\text{EMD}(A(i)V, D_i)| + \eta \|V\|_0 \right) \tag{3.8}$$

In this formulation, M corresponds to the number of Wi-Fi links, D_i represents the observed DFS profile from the i-th link, η serves as a sparsity coefficient and $\|V\|_0$ quantifies the number of non-zero velocity components within V.

Utilizing BVP files, as opposed to raw CSI, offers several distinct advantages in the context of human activity recognition. BVP data encapsulates physiological information related to blood volume changes, making it more informative for discerning human activities. This physiological data provides a richer source of information, enabling a deeper understanding of activities such as walking, running, or even subtle gestures. Moreover, BVP matrices are less susceptible to interference and noise, which can often obscure the underlying patterns in raw CSI data. By leveraging BVP, we can enhance the accuracy and reliability of our human activity

recognition models, ultimately leading to more robust and effective applications in various fields, from healthcare to smart home automation.

The input data utilized in this article consists of a $22 \times 20 \times 20$ matrix derived from the BVP data, as shown in Figure 1.5(c) and Figure 1.5(d). This matrix contains vital information essential for subsequent analysis and serves as a foundational component for researchers delving into the intricacies of gesture recognition, indoor tracking, and wireless sensing. It's important to note that, in the experiments conducted for this study, we do not take into account the noise and imperfections that may arise during the collection and computation of BVP data by the Widar System. Consequently, our approach focuses solely on normalizing the raw data, and subsequent experiments are carried out on this pre-processed dataset.

Chapter 4

Experiment and Results

4.1 Analysis and Ablation Study

Our research evaluated the ST-ANet model in Human Activity Recognition (HAR). We optimized hyperparameters and refined the training regimen. We implemented learning rate scheduling for further refinement and eventually set the learning rate to 0.01 to balance speed and stability. Using a batch size of 64, we minimized overfitting and utilized GPU resources efficiently. We trained the model for 20 iterations for convergence. We used the SGD algorithm to minimize the loss function. To assess model performance, we relied on the accuracy metric. This metric provided a reliable measure of classification ability. We conducted experiments on the WIDAR dataset, comprising 22 actions with 8000 samples each. We divided the dataset into 80% for training and 20% for validation. Our experiments aimed to compare different models on the WIDAR dataset and highlight ST-ANet's performance in HAR. This approach will provide valuable insights into the competitive landscape of HAR models.

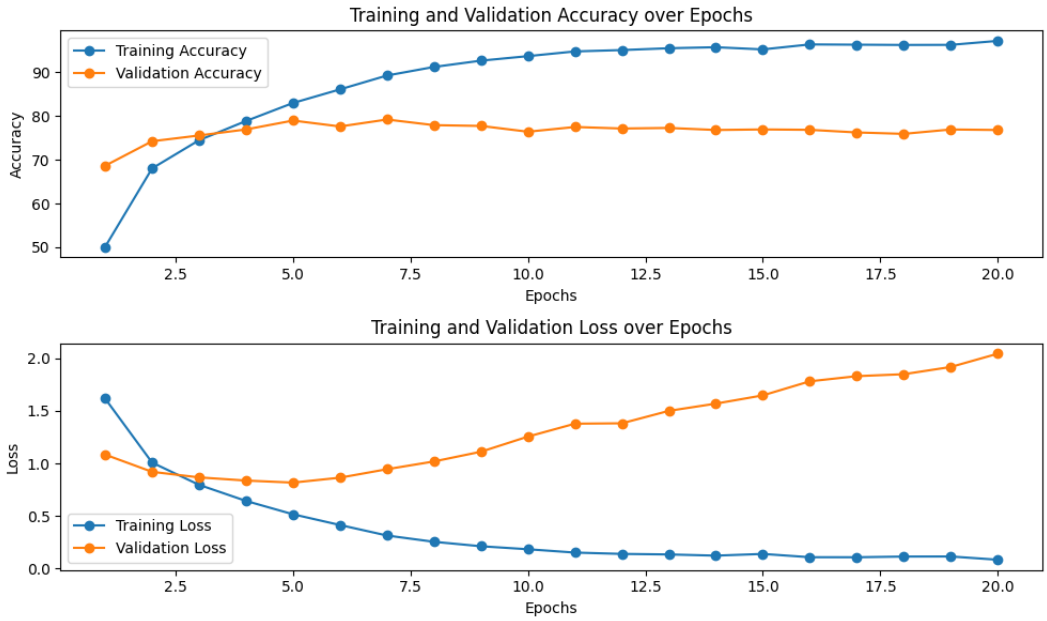


Figure 4.1: Training Performance of ST-ANet.

As shown in Figure 4.1, the model was trained over 20 epochs, utilizing a split dataset into training and validation subsets to facilitate a detailed evaluation of its learning progress and generalization capability. Key metrics such as training and validation accuracy and loss were recorded meticulously after each epoch to monitor performance and identify potential issues like overfitting.

Training involved iterative adjustments of model parameters, optimized through the backpropagation of errors derived from a loss function. The model began with a training accuracy of 50.1% and a loss of 1.6179, both of which improved substantially throughout the training process; by the 20th epoch, the training accuracy had risen to 97.21%, while the loss had decreased to 0.0823. This demonstrates the model's ability to effectively learn and adapt based on the feedback from the training data.

However, the performance on the validation set told a slightly different story. The initial validation accuracy was 68.66%, peaking at 79.36% during the 7th epoch before displaying fluctuations and generally stabilizing in the mid-70s range towards the later epochs. This highest point of validation accuracy did not align with the peak training accuracy, suggesting the model might be overfitting as training progressed. Similarly, validation loss decreased initially to 0.81706 by the 5th epoch but then began to increase, suggesting diminishing returns in the model's ability to generalize to new data as training continued.

This discrepancy between training and validation performance suggests that

while the model has strong learning capabilities, it struggles with generalization, potentially due to overfitting. The peak in validation accuracy around the 7th epoch indicates that implementing early stopping could help mitigate this issue. Future improvements might include incorporating techniques such as dropout or regularization, modifying the model's architecture, adjusting the learning rate, or exploring different optimization algorithms to enhance both performance and generalization capabilities.

In conclusion, the model demonstrated significant learning potential; however, its ability to generalize effectively remains a challenge. Addressing these issues is crucial to ensure that the high performance observed during training translates more effectively to practical applications, thus enhancing the model's utility in real-world settings.

4.1.1 Performance Comparison with Existing Models

We provide a comparative analysis of the ST-ANet's performance against several established models using the Widar dataset. This comparison is vital for understanding how ST-ANet stands relative to traditional and contemporary approaches in the field of human activity recognition.

The comparisons presented in the experiment were drawn from a variety of sources to ensure accuracy and reliability. For the Multilayer Perceptron (MLP) model, foundational insights were sourced from Gardner and Dorling's work [55]. The Convolutional Neural Network with 5 layers (CNN-5) draws inspiration from LeCun et al.'s seminal paper [56]. In the case of Recurrent Neural Networks (RNN), Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM) models, the structural frameworks were primarily adapted from Hochreiter and Schmidhuber's pioneering research [57]. Additionally, insights into the GRU model's implementation were gleaned from Dua et al.'s recent study on multi-task learning [58]. Finally, the Vision Transformer (ViT) architecture was referenced from Dosovitskiy et al.[51]. While the Multilayer Perceptron and CNN-5 models were implemented independently, adjustments were made to the RNN, GRU, and LSTM architectures to align with the dimensions of the dataset under consideration.

Furthermore, it's important to note that all models were executed independently, and the results presented are derived from individual runs. This approach was necessitated by the BVP processing of the WiDAR dataset, which precluded a direct comparison with traditional image or video datasets. As a result, each model underwent separate training and evaluation processes tailored to the dataset's characteristics, ensuring that the comparative analysis accurately reflects the performance of each architecture within the context of WiDAR data. This methodology underscores the reliability and validity of the findings presented in the table. The table below presents a detailed comparison of model performance, quantifying each model's

accuracy, computational complexity (measured in millions of floating-point operations, or Flops), and model size (measured in millions of parameters, or Params).

Table 4.1: Widar Dataset - Model Performance

Method	Accuracy (%)	Flops (M)	Params (M)
MLP [55]	67.24	9.15	9.150
CNN-5 [56]	70.19	3.38	0.299
RNN [57]	46.77	0.66	0.031
GRU [58]	62.50	1.98	0.091
LSTM [57]	63.35	2.64	0.121
ViT [51]	64.85	9.28	0.106
ST-ANet	79.36	5.93	0.411

As illustrated, ST-ANet significantly outperforms all other tested models in accuracy, achieving a 79.36% accuracy rate. Notably, it maintains a moderate level of computational demand, with 5.93 million Flops, and a relatively low parameter count at 0.411 million, showcasing an efficient balance between performance and computational efficiency.

The traditional MLP, while having a high parameter and Flop count, lags in accuracy at 67.24%. Similarly, CNN-5, despite being more efficient in terms of parameters, achieves only 70.19% accuracy. Recurrent models like the RNN, GRU, and LSTM display lower accuracies and variable computational costs, reflecting the challenges of using such architectures for complex spatial-temporal feature extraction in activity recognition tasks. The Vision Transformer (ViT), a newer model type, shows moderate performance in both computational cost and accuracy.

This comparison highlights ST-ANet's superior capability in accurately recognizing human activities with a more optimized balance of computational overhead and model complexity, suggesting its suitability for deployment in real-world applications where both accuracy and efficiency are critical.

4.1.2 Ablation Study

In our investigation of the proposed ST-ANet, we systematically evaluate the impact of three critical network components: the fast-path, slow-path, and self-attention mechanism. Our analysis, as presented in Table 4.2, sheds light on the effectiveness of these components in the context of human activity recognition on the Widar dataset.

The fast-path focuses on swift data processing, emphasizing speed in feature extraction. Its operation alone achieved a baseline accuracy of 73.19%, demonstrating its effectiveness in handling simpler, high-frequency features without extensive temporal or contextual analysis. This establishes the importance of the fast-path in

Table 4.2: Comparison of Individual Components' Performance

Fast-path	Slow-path	Self-Attention	Accuracy (%)
✓	×	×	73.19
✓	✓	×	76.61
✓	×	✓	75.43
✓	✓	✓	79.36

scenarios where rapid response is crucial. Adding the slow-path to the fast-path improved accuracy to 76.61%. The slow-path processes data more thoroughly, allowing for the extraction of detailed, lower-frequency features over longer timescales. This increase in accuracy highlights the slow-path's role in capturing more complex aspects of human activities that unfold gradually.

The integration of the self-attention mechanism with the fast-path led to an accuracy of 75.43%. Although this was a smaller increase compared to the dual-path configuration, it underscores the self-attention's ability to dynamically refine the network's focus on relevant features. This mechanism adjusts feature weighting based on the contextual importance of the data, enhancing the model's precision. The combination of all three components resulted in the highest accuracy of 79.36%, demonstrating their synergistic effect. This configuration leverages the rapid initial processing of the fast path, the comprehensive temporal analysis of the slow path, and the contextual prioritization of the self-attention mechanism. Such interplay allows the ST-ANet to optimize feature extraction and selection, adapting to the varying complexities found within activity patterns.

This component-wise analysis not only confirms the effectiveness of each component individually but also highlights the improved performance achievable through their integration. It provides crucial insights for the optimization of deep learning architectures, particularly for tasks requiring a nuanced understanding and classification of human activities. Future research could expand this model's application across diverse datasets and real-world scenarios, adjusting component configurations to address more complex challenges. Exploring alternative architectures or newer forms of attention mechanisms may also yield enhancements in performance and efficiency. This systematic evaluation forms a strong foundation for further advancements in activity recognition systems, pointing towards a robust approach to designing more adaptable and efficient networks.

Chapter 5

Conclusion

The deployment of Wi-Fi devices presents a compelling combination of cost controllability and privacy protection, rendering Wi-Fi-based datasets an invaluable resource for advancing the field of HAR. Within this context, the neural network model crafted within this signaling system emerges as a powerful and versatile tool, boasting commendable attributes including reliability, portability, and cross-platform compatibility. Our empirical results, derived from the training of neural networks on this dataset, underscore the profound impact of attention mechanisms on enhancing the accuracy and efficacy of HAR systems.

The integration of attention mechanisms, as observed in our research, has played a pivotal role in shaping the landscape of HAR, influencing multiple critical aspects:

Firstly, the attention mechanism adeptly prioritizes and selects pivotal features from the input data through the utilization of learned attention weights. This dynamic feature selection process ensures that the model concentrates its computational resources on the most informative aspects of the sensor data.

Moreover, the inclusion of the attention mechanism empowers our model with an innate understanding of temporal dynamics inherent within the data. This enhanced temporal modeling prowess significantly augments the model's ability to discern intricate patterns, even when they manifest over varying time scales.

Furthermore, the attention mechanism exerts a profound influence on the generalization capabilities of the model. It equips the network with the ability to adapt to real-world fluctuations, accommodating diverse environmental conditions, and gracefully handling variations among different actors. This inherent robustness ensures that the model remains a stalwart and dependable asset when deployed in real-world scenarios.

Expanding upon these technical innovations, the incorporation of the slow dual-channel architecture within our ST-ANet has emerged as a transformative advancement. This architectural innovation has significantly bolstered the model's accuracy in capturing the nuanced spatial and temporal characteristics present within time series datasets. It is a testament to the fusion of cutting-edge technology

with the nuances of human activity recognition.

Looking ahead to the future, the landscape of HAR using the Widar dataset promises further advancements, with an unwavering focus on the technical frontiers. Efficiency Enhancement remains paramount, with a steadfast commitment to optimizing the computational efficiency of our model. The ultimate goal is to enable seamless real-time deployment on resource-constrained devices. Techniques such as model quantization, model compression, and hardware acceleration will be rigorously explored to realize this ambitious objective.

Simultaneously, the exploration of Transfer Learning techniques will continue to be a cornerstone of our research. Transfer learning, characterized by pre-training models on one environment or participant and fine-tuning them for diverse contexts, holds the potential to significantly elevate the model's generalization prowess. This strategy also mitigates the data-intensive nature of deep learning.

To fortify our model's robustness in the face of signal noise and interference, the development of Generative Adversarial Network (GAN) technology will be a primary focus. GANs, serving as a powerful denoising tool, will play a pivotal role in purifying raw sensor data. This enhancement will allow our model to focus exclusively on the most salient and meaningful signal components. In the context of real-world deployment, the evolution of deployment strategies will persist as a focal point. Our commitment to practical implementation remains unwavering, encompassing adaptability to diverse scenarios, scalability considerations, and an unyielding commitment to data privacy and user-friendliness.

In summation, our ST-ANet research signifies a profound milestone in the domain of HAR utilizing Wi-Fi CSI data. The integration of attention mechanisms and the innovative dual-channel architecture catapults our model to the forefront of HAR technology. The outlined trajectory for future work not only extends the horizons of our research but also lays the foundational framework for the pragmatic implementation of HAR systems across a diverse spectrum of real-world environments and applications.

Bibliography

[1] Sandro Nižetić, Petar Šolić, Diego López-de-Ipiña González-De, Luigi Patrono, et al. Internet of things (iot): Opportunities, issues and challenges towards a smart and sustainable future. *Journal of Cleaner Production*, 274:122877, 2020.

[2] Manuel Eugenio Morocho-Cayamcela, Haeyoung Lee, and Wansu Lim. Machine learning for 5g/b5g mobile and wireless communications: Potential, limitations, and future directions. *IEEE access*, 7:137184–137206, 2019.

[3] Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, 6(1):1–25, 2019.

[4] Zaheer Allam and Zaynah A Dhunny. On big data, artificial intelligence and smart cities. *Cities*, 89:80–91, 2019.

[5] Guy Lansley and Paul Longley. Deriving age and gender from forenames for consumer analytics. *Journal of Retailing and Consumer Services*, 30:271–278, 2016.

[6] Mehdi Mohammadi, Ala Al-Fuqaha, Sameh Sorour, and Mohsen Guizani. Deep learning for iot big data and streaming analytics: A survey. *IEEE Communications Surveys & Tutorials*, 20(4):2923–2960, 2018.

[7] Li Deng, Geoffrey Hinton, and Brian Kingsbury. New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8599–8603. IEEE, 2013.

[8] Daniel W Otter, Julian R Medina, and Jugal K Kalita. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624, 2020.

[9] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, et al. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

[10] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022.

[11] Djamila Romaissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. Vision-based human activity recognition: a survey. *Multimedia Tools and Applications*, 79:30509–30555, 2020.

[12] Sarvesh Vishwakarma and Anupam Agrawal. A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, 29:983–1009, 2013.

[13] L Minh Dang, Kyungbok Min, Hanxiang Wang, Md Jalil Piran, Cheol Hee Lee, and Hyeonjoon Moon. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition*, 108:107561, 2020.

[14] Anna Ferrari, Daniela Micucci, Marco Mobilio, and Paolo Napoletano. Trends in human activity recognition using smartphones. *Journal of Reliable Intelligent Environments*, 7(3):189–213, 2021.

[15] Yan Wang, Shuang Cang, and Hongnian Yu. A survey on wearable sensor modality centred human activity recognition in health care. *Expert Systems with Applications*, 137:167–190, 2019.

[16] Timo Sztyler and Heiner Stuckenschmidt. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–9. IEEE, 2016.

[17] Blair Hu, Elliott Rouse, and Levi Hargrove. Benchmark datasets for bilateral lower-limb neuromechanical signals from wearable sensors during unassisted locomotion in able-bodied individuals. *Frontiers in Robotics and AI*, 5:14, 2018.

[18] Hui Liu, Yale Hartmann, and Tanja Schultz. Csl-share: A multimodal wearable sensor-based human activity dataset, 2021.

[19] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. Device-free human activity recognition using commercial wifi devices. *IEEE Journal on Selected Areas in Communications*, 35(5):1118–1131, 2017.

[20] Fuqiang Gu, Mu-Huan Chung, Mark Chignell, Shahrokh Valaee, Baoding Zhou, and Xue Liu. A survey on deep learning for human activity recognition. *ACM Computing Surveys (CSUR)*, 54(8):1–34, 2021.

[21] Mohammed Abdulaziz Aide Al-Qaness, Fangmin Li, Xiaolin Ma, Yong Zhang, and Guo Liu. Device-free indoor activity recognition system. *Applied Sciences*, 6(11):329, 2016.

[22] Neena Damodaran, Elis Haruni, Muyassar Kokhkharova, and Jörg Schäfer. Device free human activity and fall recognition using wifi channel state information (csi). *CCF Transactions on Pervasive Computing and Interaction*, 2:1–17, 2020.

[23] ZhengYang, YiZhang, GuidongZhang, and YueZheng. Widar-3.0: A wifi-based activity recognition dataset. *IEEE DataPort*, 2021.

[24] Kishor H Walse, Rajiv V Dharaskar, and Vilas M Thakare. Pca based optimal ann classifiers for human activity recognition using mobile sensors data. In *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 1*, pages 429–436. Springer, 2016.

[25] Linlin Guo, Lei Wang, Chuang Lin, Jialin Liu, Bingxian Lu, Jian Fang, Zhonghao Liu, Zeyang Shan, Jingwen Yang, and Silu Guo. Wiar: A public dataset for wifi-based activity recognition. *IEEE Access*, 7:154935–154945, 2019.

[26] Myeongjun Kim, Taehun Kim, and Daijin Kim. Spatio-temporal slowfast self-attention network for action recognition. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2206–2210. IEEE, 2020.

[27] Jianfei Yang, Xinyan Chen, Dazhuo Wang, Han Zou, Chris Xiaoxuan Lu, Sumei Sun, and Lihua Xie. Deep learning and its applications to wifi human sensing: A benchmark and a tutorial. *arXiv preprint arXiv:2207.07859*, 2022.

[28] Parisa Fard Moshiri, Mohammad Nabati, Reza Shahbazian, and Seyed Ali Ghorashi. Csi-based human activity recognition using convolutional neural networks. In *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*, pages 7–12. IEEE, 2021.

[29] Jin Zhang, Fuxiang Wu, Bo Wei, Qieshi Zhang, Hui Huang, Syed W Shah, and Jun Cheng. Data augmentation and dense-lstm for human activity recognition using wifi signal. *IEEE Internet of Things Journal*, 8(6):4628–4641, 2020.

[30] Deepika Singh, Erinc Merdivan, Ismini Psychoula, Johannes Kropf, Sten Hanke, Matthieu Geist, and Andreas Holzinger. Human activity recognition using recurrent neural networks. In *Machine Learning and Knowledge Extraction: First IFIP TC 5, WG 8.4, 8.9, 12.9 International Cross-Domain Conference, CD-MAKE 2017, Reggio, Italy, August 29–September 1, 2017, Proceedings 1*, pages 267–274. Springer, 2017.

[31] SU Park, JH Park, Mohammed A Al-Masni, Mugahed A Al-Antari, Md Z Uddin, and T-S Kim. A depth camera-based human activity recognition via deep learning recurrent neural network for health and social care services. *Procedia Computer Science*, 100:78–84, 2016.

[32] Saiwen Wang, Jie Song, Jaime Lien, Ivan Poupyrev, and Otmar Hilliges. Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 851–860, 2016.

[33] Vittorio Mazzia, Simone Angarano, Francesco Salvetti, Federico Angelini, and Marcello Chiaberge. Action transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition*, 124:108487, 2022.

[34] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[35] Yoël Forterre. Slow, fast and furious: understanding the physics of plant movements. *Journal of experimental botany*, 64(15):4745–4760, 2013.

[36] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

[37] Tariq Ahmad, Jinsong Wu, Imran Khan, Asif Rahim, and Amjad Khan. Human action recognition in video sequence using logistic regression by features fusion approach based on cnn features. *International Journal of Advanced Computer Science and Applications*, (11), 2021.

[38] Chun-Fu Richard Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, and Quanfu Fan. Deep analysis of cnn-based spatio-temporal representations for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6165–6175, 2021.

[39] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 105–121, 2018.

[40] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4768–4777, 2017.

[41] Wei Zeng, Junjian Huang, Wei Zhang, Hai Nan, and Zhenjiang Fu. Slowfast action recognition algorithm based on faster and more accurate detectors. *Electronics*, 11(22):3770, 2022.

[42] M Esat Kalfaoglu, Sinan Kalkan, and A Aydin Alatan. Late temporal modeling in 3d cnn architectures with bert for action recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 731–747. Springer, 2020.

[43] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[44] Yuanzhong Liu, Junsong Yuan, and Zhigang Tu. Motion-driven visual tempo learning for video-based action recognition. *IEEE Transactions on Image Processing*, 31:4104–4116, 2022.

[45] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. Temporal coherence or temporal motion: Which is more critical for video-based person reidentification? In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 660–676. Springer, 2020.

[46] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational visual media*, 8(3):331–368, 2022.

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[48] Woohyeon Moon, Taeyoung Kim, Bumgeun Park, and Dongsoo Har. Enhanced transformer architecture for natural language processing. *arXiv preprint arXiv:2310.10930*, 2023.

[49] Hongsong Wang and Liang Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 499–508, 2017.

[50] Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2964–2972, 2022.

[51] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[52] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021.

[53] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[54] Yi Zhang, Yue Zheng, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. Widar3. 0: Zero-effort cross-domain gesture recognition with wi-fi. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8671–8688, 2021.

[55] Matt W Gardner and SR Dorling. Artificial neural networks (the multilayer perceptron): A review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636, 1998.

[56] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[57] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[58] Nidhi Dua, Shiva Nand Singh, and Vijay Bhaskar Semwal. Multi-input cnn-gru based human activity recognition using wearable sensors. *Computing*, 103(7):1461–1478, 2021.