Thesis Proposal for MPhil Degree

Student Name Yijie XU
ID Number 50004147
Email yxu409@connect.hkust-gz.edu.cn

Group Project Research and Implementation of Multimodal Human-Computer Interaction Perception Technology
Project Mentor Dr. Li CHEN

Individual Project Signal Processing and Modeling of Multimodal Human-Computer Interaction Perception Technology
Prime Supervisor Prof. Hui XIONG
Hub/Thrust AI Thrust / Information Hub
Co-supervisor Prof. Xiaojuan MA
Hub/Thrust Department of CSE / School of Engineering

Division of RBM

Exploring Multi-modal Human Motion Perception for Human-Computer Interaction Systems

Abstract

Human body motion perception is a crucial research area with significant potential applications, including rehabilitation, human-computer interaction, and sports training. Traditional approaches relying on single modalities, such as video, accelerometer, and gyroscopes, have inherent limitations that can result in inaccuracies and inconsistencies in data collection. To overcome these challenges, we propose using multimodal-based models, specifically Wi-Fi, surface electromyography (sEMG), and video data, which provide complementary information on body posture and movement. Our study explores Wi-Fi signals for indoor positioning and larger-scale movement and sEMG sensors for precise muscle activity, including subtle movements, despite their lack of precise location data due to being directly attached to the skin. My research will focus on the latter two modalities, combined with data collection and model design, to produce outstanding results and contribute to the field.

Contents

1 Source of Research Topic, Research Purpose and Significance 4

1.1 The Source of the Topic . . . 4

1.2 Research Background and Significance . . . 5

1.3 Drawbacks . . . 7

2 Literature Review and Analysis 8

2.1 Wi-Fi Sensing for Motion Perceptions . . . 8

2.2 sEMG Signals for Motion Perception . . . 9

2.3 Vision Signals for Motion Perception . . . 11

2.4 Analysis of the Literature Review . . . 12

3 Main Content of Research 14

3.1 Research Avenues . . . 14

3.1.1 Wi-Fi-based Motion Perception . . . 14

3.1.2 sEMG-based Motion Perception . . . 14

3.1.3 Vision-based Motion Perception . . . 15

3.1.4 Fusion of Multi-modal Signals for Motion Perception . . . 15

3.1.5 Evaluation and Validation . . . 16

3.2 Implementation Strategies . . . 16

3.2.1 Real-Time Multi-Modal Data Collection and Analysis Platform . . . 17

3.2.2 Multi-Modal Data Collection and Categorization . . . 17

3.2.3 Multi-Modal Model Development for Motion Perception . . . 17

3.2.4 Future Directions and Considerations . . . 18

4 Accomplished Work 19

4.1 Wi-Fi Signals . . . 19

4.1.1 Hardware Side . . . 19

4.1.2 Software Side . . . 20

4.2 Vision Signals . . . 20

4.3 sEMG signals . . . 21

5 Research Plan, Expected Objectives, and Results 23

5.1 Research Plan 23

5.2 Expected Objectives and Research Achievements 23

5.3 Time Scheme 24

6 Required Conditions and Resources 25

7 Anticipated Problems and Solutions 26

1 Source of Research Topic, Research Purpose and Significance

1.1 The Source of the Topic

Motion perception is a rapidly growing field of research with numerous potential applications in areas such as rehabilitation, human-computer interaction, and sports training. The ability to accurately perceive and interpret human body posture and movement is essential for developing advanced technologies to improve our daily lives.

Traditional approaches to motion perception have relied on single modalities such as video, accelerometer, and gyroscopes. While these methods have proven effective in certain situations, they also have inherent limitations that can result in inaccuracies and inconsistencies in data collection. For example, video-based methods can be affected by lighting conditions and occlusions, while the accelerometer and gyroscope-based methods may not provide sufficient information about subtle movements.

To overcome these challenges, researchers have begun to explore the use of multimodal-based models that combine multiple sources of information to provide a more comprehensive understanding of human body posture and movement. One promising approach is using Wi-Fi signals for indoor positioning and larger-scale movement detection. Wi-Fi signals can penetrate walls and other obstacles, making them well-suited for indoor environments where traditional methods may struggle.

Another promising modality is surface electromyography (sEMG), which measures the electrical activity muscles produce during contraction. sEMG sensors can provide detailed information about muscle activity, including subtle movements that may be difficult to detect using other methods. However, sEMG sensors lack precise location data due to being directly attached to the skin.

This study proposes combining Wi-Fi, sEMG, and video data to create a multimodal-based model for human motion perception. Our approach leverages the strengths of each modality to provide complementary information about body posture and movement. We will focus on using sEMG and Wi-Fi signals, combined with advanced data collection and model design techniques, on producing outstanding results that contribute to the field.

1.2 Research Background and Significance

Human motions are important. It has played a critical role in human communication and social interactions throughout history. Before the invention of writing and language, motions were the primary means of conveying ideas, emotions, and intentions. Even today, motions continue to play an essential role in human communication, adding nuance and meaning to spoken language. Additionally, motion perception is crucial in non-verbal communication, such as interpreting body language and facial expressions. Moreover, the ability to recognize and interpret motions is essential in fields such as psychology, anthropology, and communication studies. Overall, human motion perception is vital to understanding and navigating human interactions and has been integral to human communication throughout history.

After introducing the research background and significance of human motions and motion perceptions, it is important to delve into the development path of motion recognition and the history of using different modalities of sensors and data in human motion perception.

In the domain of human motion recognition, the development path has experienced significant advancements over the years, spanning a range of techniques and methodologies. The early attempts at motion recognition can be traced back to the late 20th century, when researchers initially focused on the analysis of basic motions, such as walking and running, using optical **Mo**tion **Cap**ture systems (Mocap). These systems relied on the precise tracking of markers placed on the subject's body, enabling the reconstruction of human motion in three-dimensional space. Although effective, these systems were constrained by their high cost, intrusiveness, and limitations to controlled environments.

With the advent of computer vision and machine learning techniques in the early 21st century, researchers began exploring alternative approaches for motion recognition, which included the use of 2D videos and depth sensors. These methods facilitated the extraction of motion features from image sequences, allowing for the recognition of human motions in more natural and uncontrolled environments. The introduction of the Microsoft Kinect in 2010 further propelled the field, as it provided an affordable, non-intrusive, and easy-to-use depth sensor, which enabled the extraction of accurate skeleton information for human motion analysis.

In parallel with these developments, researchers also began investigating the potential of multimodal approaches, incorporating various sensor modalities and data types for enhanced motion recognition. The integration of **I**nertial **M**easurement **U**nits (IMUs), for instance, allowed for the capture of motion data with high temporal resolution, compensating for the shortcomings of vision-based methods in dynamic and occluded scenarios. Furthermore, the fusion of data from multiple sensors, such as RGB cameras, depth sensors, and IMUs, has demonstrated the potential to improve motion recognition performance, particularly in challenging environments and situations.

One of the most important modalities is inertial sensing, which involves using sensors such as accelerometers and gyroscopes to measure the motion of the body. Inertial sensors are particularly useful for tracking the motion of limbs and joints, and have been used in a variety of applications, from sports training to medical rehabilitation.

Another important modality is **e**lectro**m**yo**g**raphy (EMG), which involves measuring the electrical activity of muscles. EMG has been used to develop systems that can detect subtle muscle movements, such as those involved in sign language or facial expressions.

In recent years, there has also been a growing interest in using wearable devices such as smartwatches and fitness trackers to track human motion. These devices typically incorporate a variety of sensors, including accelerometers, gyroscopes, and heart rate monitors, and can provide a wealth of data about the user's activity levels, sleep patterns, and overall health.

The emergence of deep learning techniques, especially convolutional neural networks (CNNs) and recurrent neural networks (RNNs), has also significantly impacted the field of human motion recognition. These methods have facilitated the automatic learning of hierarchical features from raw sensor data, eliminating the need for manual feature engineering. Researchers have successfully applied these techniques to various modalities, including RGB videos, depth maps, IMU and EMG data, achieving state-of-the-art performance in multiple motion recognition tasks.

In summary, the development path of motion recognition has witnessed considerable advancements over the years, progressing from marker-based optical systems to sophisticated deep learning techniques. The incorporation of multimodal sensor data and fusion techniques has further enhanced the field, allowing for improved motion recognition performance in a wide range of scenarios. As researchers continue to explore new modalities, techniques, and applications, human motion recognition remains a vibrant and evolving area of study, with significant potential for future breakthroughs.

1.3 Drawbacks

However, the modalities above have inherent limitations that can result in inaccuracies and inconsistencies in data collection, leading to potential issues in motion recognition. This chapter will cover problems that can occur when using single-modality sensors:

1. **Limited sensing range**: Most single-modality sensors have limited sensing range, which means they can only capture a limited area of the body or environment. For example, a camera-based system can only capture visible motions, while an accelerometer-based system can only capture movements that involve acceleration.

2. **Limited accuracy**: Single-modality sensors can also have limited accuracy, especially when it comes to capturing small or subtle movements. For example, accelerometer-based systems may struggle to distinguish between slight variations in movement or sudden stops, leading to inaccuracies in the captured data.

3. **Environmental interference**: Single-modality sensors can also be affected by environmental interference, such as lighting conditions, noise, and other sources of interference. For example, a camera-based system may struggle to recognize motions in low-light conditions, while an accelerometer-based system may be affected by vibrations from external sources.

4. **Limited depth of analysis**: Single-modality sensors typically capture only one aspect of the motion, such as motion or muscle activity. As a result, the study may be limited regarding the depth and scope of the captured data.

To overcome these challenges, we proposed using multimodal-based models, which combine data from multiple sources to provide a more comprehensive understanding of Human motion recognition. Various sensors, such as Wi-Fi, surface electromyography (sEMG), and video data, can capture complementary information on body posture and movement, leading to more accurate and precise motion recognition results.

2 Literature Review and Analysis

To introduce our method of combining multi-modal data for motion perception, it is vital to have a research review of using tensors of single modality for motion perception. In this section, we will first cover Wi-Fi, sEMG and vision signals, then have a review of using multi-modal signals.

2.1 Wi-Fi Sensing for Motion Perceptions

Wi-Fi sensing for motion perception has gained significant attention in recent years due to the widespread deployment of Wi-Fi networks and the increasing demand for **H**uman-**C**omputer **I**nteraction (HCI) technologies. This section presents a comprehensive review of the state-of-the-art in Wi-Fi sensing for motion perception, including key techniques, methodologies, and applications.

Early works on Wi-Fi sensing for gesture recognition primarily focused on the analysis of **R**eceived **S**ignal **S**trength (RSS) variations caused by human movements. Adib et al. [1] proposed WiTrack, a through-wall 3D motion tracking system that leverages radio frequency signals to track the motion of a person. Similarly, another work of Adib et al. [2] utilized Wi-Fi signal reflections to accurately track the position of a person indoors. These systems demonstrated the potential of Wi-Fi signals for motion perception but were limited by the coarse spatial resolution of RSS.

To overcome the limitations of RSS-based systems, researchers began exploring the use of **C**hannel **S**tate **I**nformation (CSI) as a more fine-grained indicator of human motion. Pu et al. [3] developed WiSee, a system that exploits CSI to recognize a set of predefined gestures. WiSee demonstrated the feasibility of using Wi-Fi signals for gesture recognition, and numerous researchers have since built upon this work. For example, Wei et al. [4] proposed WiGest, which leverages CSI phase information to detect the presence of a human and recognize basic gestures. Wu et al. [5] introduced PhaseBeat, a CSI-based system that detects micro-movements, such as heartbeat and respiration, through phase information analysis.

Deep learning techniques have also been employed to improve the performance of Wi-Fi sensing systems for gesture recognition. Wang et al. [6] proposed CiFi, which combines

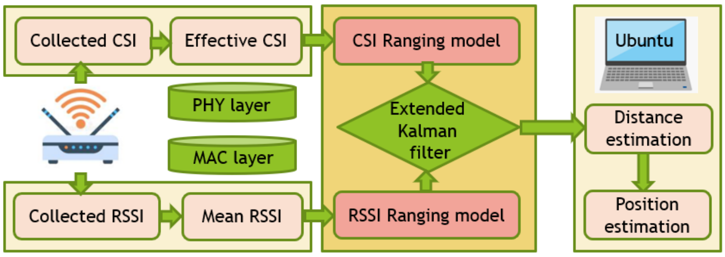


Figure 1: Wi-Fi sensing by CSI and RSI

Convolutional **N**eural **N**etworks (CNN) and CSI to achieve high recognition accuracy across multiple environments and devices. Li et al. [7] presented WiFinger, a system that leverages deep learning to recognize finger gestures based on CSI data, opening the door to a wide range of HCI applications.

In addition to gesture recognition, Wi-Fi sensing has been explored for various other motion perception applications, such as fall detection, human counting, and localization. Wang et al. [8] designed WiFall, a system that detects falls based on CSI amplitude variations caused by human motion. Wang et al. [9] developed WiHear, which enables Wi-Fi signals to hear our talks without deploying any devices.

In summary, Wi-Fi sensing for motion perception has evolved significantly over the past decade, with advancements in both hardware and software techniques. From early works using RSS to more recent studies employing CSI and deep learning, Wi-Fi sensing has shown great potential for a wide range of applications, including gesture recognition, fall detection, and localization. Future research should continue to focus on improving the accuracy, robustness, and generalizability of Wi-Fi sensing systems, as well as exploring novel applications for this promising technology.

2.2 sEMG Signals for Motion Perception

Over the past decade, the investigation of **s**urface **e**lectro**m**yo**g**raphy (sEMG) signals has played a pivotal role in the advancement of motion perception research [10]. Surface electromyography (sEMG) has been widely utilized for various applications, including rehabilitation engineering, human-machine interfaces (HMIs), and motion perception [11], [12]. The development of sEMG-based systems for gesture recognition and motion perception can be traced back to the early 1980s, when researchers began to investigate the potential of sEMG for controlling prosthetic devices [13]. As the field progressed, advanced signal processing techniques and machine learning algorithms were introduced to improve the performance of sEMG-based systems.

During the 1990s, researchers focused on improving the accuracy and robustness of sEMG-based gesture recognition. Time-domain features, such as mean absolute value, waveform length, and zero-crossing rate, were commonly used for extracting relevant information from sEMG signals [14]. The use of **A**rtificial **N**eural **N**etworks (ANNs) for the classification of sEMG signals started to gain popularity, leading to significant improvements in the recognition of hand gestures and movements [15].

The rise of support vector machines (SVM) in the 2000s further improved the performance of sEMG-based gesture recognition systems. Alongside the increasing computational power, the combination of SVM classifiers and various feature extraction methods, such as wavelet analysis, allowed for the identification of multiple degrees of freedom in hand and arm movements with high accuracy [16]. The introduction of wearable technology and miniaturization of sEMG sensors further facilitated the development of portable and user-friendly HMIs [17].

In the 2010s, deep learning techniques, particularly **C**onvolutional **N**eural **N**etworks (CNNs) and **R**ecurrent **N**eural **N**etworks (RNNs), were introduced to sEMG-based gesture recognition and motion perception research. CNNs demonstrated exceptional performance in the classification of sEMG signals, surpassing traditional feature extraction and classification methods [18]. RNNs, specifically **L**ong **S**hort-**T**erm **M**emory (LSTM) networks, showed potential for modeling the temporal dynamics of sEMG signals, leading to improved recognition of complex and time-varying gestures [19].

Throughout the years, researchers have also explored different approaches for improving the usability and adaptability of sEMG-based systems. Techniques such as transfer learning and domain adaptation have been proposed to address issues related to subject-specificity and variations in sensor placements [20]. Additionally, the integration of sEMG with other modalities, such as inertial measurement units (IMUs) and vision-based systems, has shown promising results in enhancing gesture recognition and motion perception capabilities [21].

In summary, the historical development of using sEMG signals for gesture recognition and motion perception has seen significant advancements over the years. From the early applications in prosthetic control to the recent developments in deep learning and multimodal systems, sEMG-based gesture recognition and motion perception research continue to evolve and expand, offering new opportunities for rehabilitation, human-machine interaction, and beyond.

2.3 Vision Signals for Motion Perception

In this section, we review the literature on the use of vision signals for motion perception. Over the years, computer vision techniques have evolved significantly in their ability to extract and process motion information from visual data, leading to improved performance in gesture recognition tasks.

One of the early developments in the field of computer vision was the introduction of optical flow algorithms for estimating motion in image sequences. These algorithms, such as the Lucas-Kanade method [22] and the Horn-Schunck method [23], enabled the extraction of motion information from RGB images.

The emergence of depth sensors, such as the Microsoft Kinect, provided a breakthrough in the field of gesture recognition and motion perception. The Kinect captured depth information in addition to RGB data, allowing researchers to develop more robust algorithms that relied on three-dimensional information. Shotton et al. [24] proposed a body part labeling algorithm using depth information, which laid the groundwork for subsequent research on gesture recognition.

Another significant development was the introduction of **C**onvolutional **N**eural **N**etworks (CNNs) for motion perception. The success of CNNs in image classification tasks [25] led researchers to investigate their application for gesture recognition. Ji et al. [26] introduced 3D CNNs for action recognition, utilizing both spatial and temporal information from video sequences. This approach was further improved by the use of two-stream CNNs, which separately processed RGB and optical flow data before fusing them for action recognition [27].

Depth-based approaches were also enhanced by the introduction of skeleton-based methods for gesture recognition. These methods utilized the 3D positions of human joints to recognize gestures, which provided a more compact and computationally efficient representation of motion information. Yang and Tian [28] proposed the EigenJoints method, while Wang et al. [29] introduced the Actionlet Ensemble Model, both of which achieved state-of-the-art performance in gesture recognition tasks.

With the advancements in deep learning, researchers began to explore the combination of

RGB and depth data for motion perception. One such approach is the fusion of multi-modal data at different levels of the network architecture. For example, Song et al. [30] presented an end-to-end trainable deep architecture that fused RGB and depth data in a hierarchical manner for gesture recognition.

More recently, researchers have begun to explore the use of **R**ecurrent **N**eural **N**etworks (RNNs) for motion perception. RNNs, such as **L**ong **S**hort-**T**erm **M**emory (LSTM) networks [31] and **G**ated **R**ecurrent **U**nits (GRUs) [32], are well-suited for modeling temporal sequences and have shown promising results in gesture recognition tasks [33].

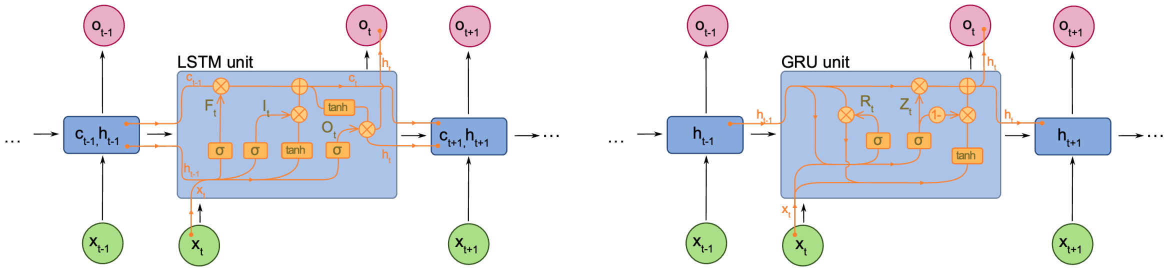


Figure 2: Long Short-Term Memory networks and Gated Recurrent Units

In conclusion, the history of computer vision techniques for motion perception has seen significant advancements, from early optical flow algorithms to the use of deep learning methods that combine RGB and depth information. These developments have led to increasingly accurate and robust algorithms for motion perception, paving the way for further research in this domain.

2.4 Analysis of the Literature Review

The literature review above provides valuable insights into the research advancements and limitations of using Wi-Fi, sEMG, and vision signals for motion perception. In recent years, researchers have extensively investigated the use of Wi-Fi signals for motion perception, providing non-intrusive and ubiquitous sensing technology for indoor monitoring and localization applications. While Wi-Fi signals present certain limitations, including poor environmental adaptability and low resolution, they remain a promising instrument for researchers and developers.

Surface **e**lectro**m**yo**g**raphy (sEMG) signals have also gained significant attention for motion perception research, owing to their advantages of electromechanical delay and muscle activation information. Various studies have utilized sEMG signals to recognize discrete motion intentions, hand gestures, and motion recognition. Deep learning models have been employed to augment the precision of motion recognition, and these advancements hold significant potential for improving prosthesis control, rehabilitation, and human-computer interaction in individuals with disabilities, such as upper-limb amputees.

Computer vision technology has facilitated the recognition of hand gestures and body movements, with techniques from image processing, machine learning, and deep learning employed to recognize and classify gestures. The development of computer vision technology for motion perception holds the potential to revolutionize human interactions with machines and virtual environments. However, researchers face challenges such as variability in hand and body appearance, motion, size, and shape.

3 Main Content of Research

Our principal endeavor is to **pioneer** the investigation and implementation of multi-modal human-computer interaction perception technology, with a focus on its application in diverse scenarios, such as human motion perception. This objective encompasses two primary components: the exploration of viable research avenues and the determination of suitable implementation strategies. The initial subsection will address the core aspects of employing Wi-Fi, sEMG and vision signals for motion perception, while the subsequent subsection will delineate the advancements in implementation.

3.1 Research Avenues

In this subsection, we will explore the main research directions for multi-modal human-computer interaction perception technology, focusing on the integration of Wi-Fi, sEMG, and vision signals for motion perception. The primary goal of this research is to develop a comprehensive framework that combines the strengths of each modality to enhance the overall performance of the system in terms of accuracy, reliability, and efficiency.

3.1.1 Wi-Fi-based Motion Perception

Wi-Fi signals have shown great potential in human motion perception due to their ability to penetrate walls and non-metallic objects, thus enabling **N**on-**L**ine-**O**f-**S**ight (NLOS) monitoring. Research in this area will focus on developing novel algorithms and techniques to accurately estimate human body movements based on Wi-Fi signal variations. This includes investigating machine learning and deep learning approaches for modeling and interpreting Wi-Fi signal characteristics, as well as exploring methods for mitigating the impact of environmental factors on Wi-Fi-based motion perception.

3.1.2 sEMG-based Motion Perception

Surface electromyography (sEMG) offers a direct and precise means of capturing muscle activities associated with human movements. Research in this area will concentrate on enhancing the quality of sEMG signals through advanced signal processing techniques, such as filtering, feature extraction, and noise reduction. Additionally, we will investigate machine learning algorithms, including supervised and unsupervised learning, to classify and interpret sEMG signals for various motion perception tasks. This research will also explore the potential of incorporating wearable devices and sensors to improve the overall performance of sEMG-based motion perception systems.

3.1.3 Vision-based Motion Perception

Vision-based motion perception techniques rely on visual information from cameras or other optical devices to estimate human movements. Research in this domain will focus on improving the robustness and accuracy of vision-based methods by employing advanced computer vision techniques, such as optical flow, feature matching, and deep learning-based approaches like convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Furthermore, we will explore methods for addressing challenges associated with occlusions, varying lighting conditions, and real-time processing requirements.

3.1.4 Fusion of Multi-modal Signals for Motion Perception

The integration of Wi-Fi, sEMG, and vision signals is crucial for developing a comprehensive multi-modal human-computer interaction perception system. Research in this area will focus on designing efficient fusion algorithms that can effectively combine information from different modalities to enhance the overall performance of the system. This includes investigating various fusion strategies, such as data-level, feature-level, and decision-level fusion, as well as exploring the use of machine learning techniques, such as ensemble learning and multi-task learning, to optimize the fusion process. We will try to use the State-of-the-Art transformer architecture in this problem. One existing example is **UniT**[34], a unified Transformer model to simultaneously learn the most prominent tasks across different domains, ranging from object detection to natural language understanding and multimodal reasoning. Additionally, we will examine methods for selecting and weighting the contributions of each modality based on their reliability and relevance to specific motion perception tasks.

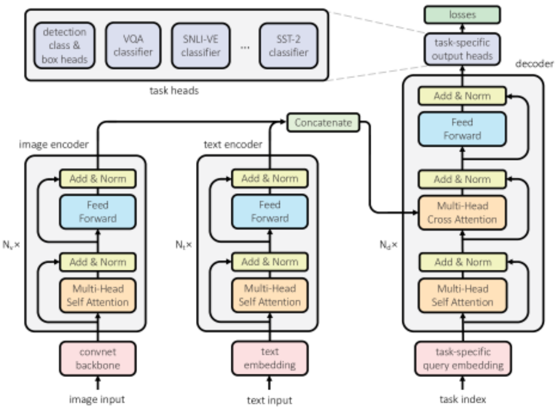

detection class & box heads
VQA classifier
SNLI-VE classifier
...
SST-2 classifier
task heads
losses
task-specific output heads
decoder
Concatenate
image encoder
text encoder
Add & Norm
Feed Forward
Add & Norm
Multi-Head Self Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Self Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Cross Attention
Add & Norm
Multi-Head Self Attention
N_v×
N_t×
N_d×
convnet backbone
text embedding
task-specific query embedding
image input
text input
task index

Figure 3: An overview of our UniT model, which jointly handles a wide range of tasks in different domains with a unified transformer encoder-decoder architecture.

3.1.5 Evaluation and Validation

A thorough evaluation of the proposed multi-modal human-computer interaction perception system is essential for ensuring its effectiveness and reliability in real-world scenarios. Research in this area will include the development of suitable performance metrics and benchmark datasets for assessing the accuracy, robustness, and efficiency of the integrated system. Moreover, we will conduct rigorous validation studies, including comparative analyses with existing single-modality and multi-modal systems, as well as user studies to evaluate the usability and user satisfaction of the proposed framework.

In summary, this research aims to advance the state-of-the-art in multi-modal human-computer interaction perception technology by focusing on the integration of Wi-Fi, sEMG, and vision signals for motion perception. The proposed research directions will collectively contribute to the development of a comprehensive and reliable framework that can effectively facilitate various human motion perception tasks and applications.

3.2 Implementation Strategies

In this section, some scheduled implementions will be introduced. We will also discuss some practical scenarios and further working directions.

3.2.1 Real-Time Multi-Modal Data Collection and Analysis Platform

In this proposal, we aim to design and implement a real-time multi-modal data collection and analysis platform, which will integrate three distinct modalities: surface electromyography (sEMG), Wi-Fi-based positioning systems, and computer vision. The purpose of this platform is to facilitate the seamless collection, processing, and analysis of multi-modal data for motion perception and classification tasks in various applications, such as rehabilitation, sports performance analysis, and human-computer interaction.

3.2.2 Multi-Modal Data Collection and Categorization

To ensure the effectiveness and robustness of our proposed multi-modal model, it is crucial to collect adequate data representing different actions and modalities. The data collection process will involve recruiting participants from diverse backgrounds, performing a wide range of actions, and capturing the corresponding sEMG, Wi-Fi, and vision signals. These data samples will then be stored into different categories, which will be defined based on the type of action, the context of the action, and the modality being analyzed. This systematic categorization will enable the development of more accurate and efficient motion perception and classification algorithms, as well as facilitate the evaluation of the model's performance across different modalities and action types. An example of such dataset [35] could be studied and such publication is expected.

3.2.3 Multi-Modal Model Development for Motion Perception

Upon the completion of the data collection and categorization phase, we will proceed to develop a multi-modal model for motion perception and classification. The proposed model will leverage state-of-the-art transformer-based architectures, which have demonstrated remarkable success in various machine learning tasks, including natural language processing, computer vision, and time-series analysis. These transformer-based models are particularly well-suited for our application, as they can effectively capture the complex relationships and dependencies among the different modalities.

In order to develop a comprehensive and robust multi-modal model, we will also investigate existing works in the field. Some notable studies to consider include those that focus on fusing sEMG and computer vision signals for gesture recognition [35] (which is currently my focus), as well as those that explore Wi-Fi-based human activity recognition [36]. By building upon these existing works and incorporating transformer-based architectures, our proposed model aims to achieve high performance in motion perception and classification across all three modalities.

3.2.4 Future Directions and Considerations

As we progress in our research, we will continuously evaluate our multi-modal model's performance and discuss potential improvements and extensions. One possible direction for future work is to explore the integration of additional modalities, such as **I**nertial **M**easurement **U**nits (IMUs) or depth sensors, which is a popular direction to form RGB-D signals, to further enhance the model's accuracy and robustness. Another avenue to consider is the development of adaptive algorithms that can dynamically adjust the model's parameters based on the specific context or application, ensuring optimal performance across various settings.

Moreover, we will investigate the potential of our multi-modal model for real-world applications, such as rehabilitation programs, sports training, and virtual or augmented reality systems. By closely collaborating with domain experts and end-users, we aim to ensure that our research contributes to the development of practical, effective, and user-friendly solutions that address pressing challenges in motion perception and classification.

4 Accomplished Work

This section outlines the work completed since the initiation of this research project in February 2023. The progress thus far includes three distinct chapters, each focusing on one of the three modalities employed in our multi-modal human-computer interaction perception technology.

4.1 Wi-Fi Signals

I have successfully developed a Wi-Fi signal capturing platform [fig:4], comprising both hardware and software components. This platform enables the efficient collection and processing of Wi-Fi signals, facilitating the investigation of their potential applications in human motion perception tasks.

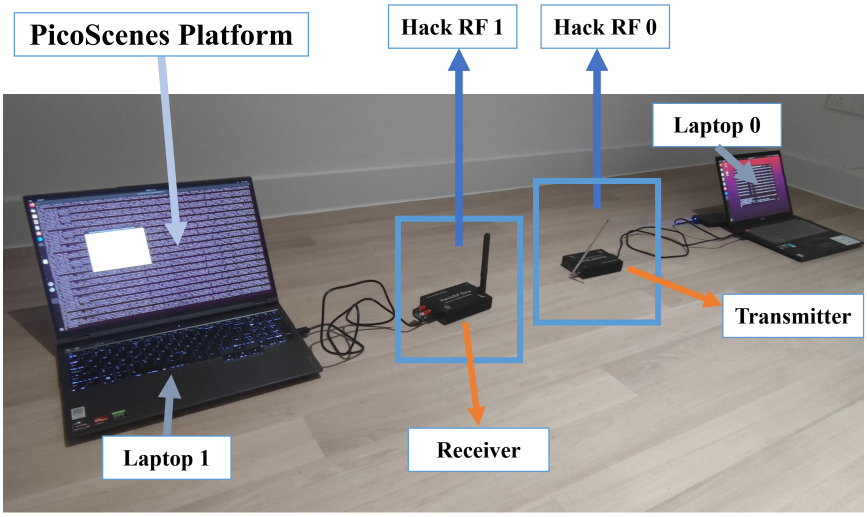


Figure 4: Wi-Fi signals capturing system

4.1.1 Hardware Side

For the hardware component, we have selected the HackRF One hardware, a sophisticated software-defined radio (SDR) peripheral that facilitates the transmission and reception of radio signals within the frequency range of 1 MHz to 6 GHz [37]. The HackRF One boasts numerous advantages, such as an extensive frequency range, high resolution, and compatibility with a diverse array of open-source software. Furthermore, it features an open-source hardware design, which affords the opportunity for potential customization and enhancement in future research endeavors. We have procured two HackRF One units, designating one for transmission and the other for reception. Despite encountering challenges in optimizing the hardware settings for specific frequency bands and mitigating interference, we have accomplished the foundational work necessary for environmental information collection.

In parallel, we have conducted an extensive literature review on Wi-Fi-based human activity recognition techniques [36], delving into methodologies for augmenting signal quality, such as adaptive filtering and noise reduction algorithms [38]. The insights gleaned from this review will prove invaluable for refining our Wi-Fi signal collection platform and improving the quality of the data obtained.

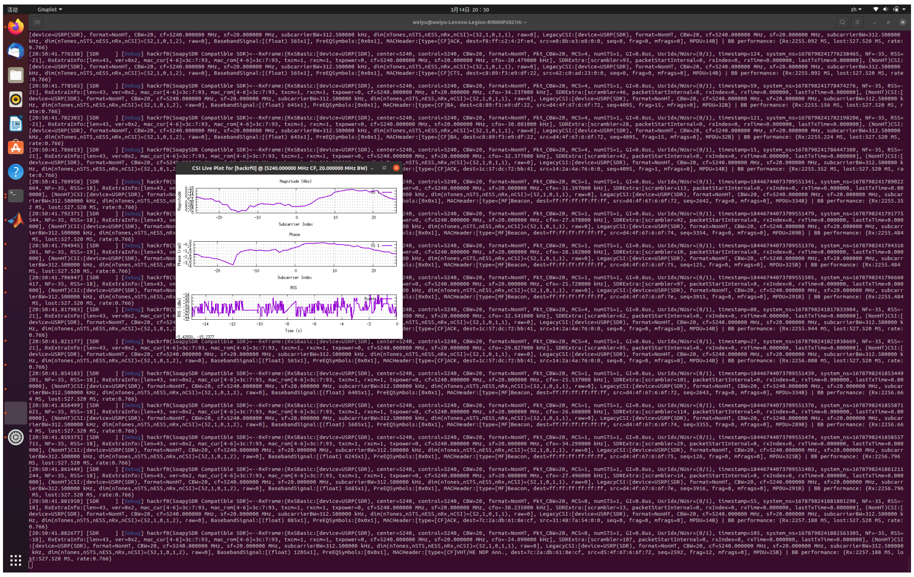
4.1.2 Software Side

For the software side, we selected the PicoScenes software [fig:5], a versatile and high-performance Wi-Fi sensing platform designed for capturing **C**hannel **S**tate **I**nformation (CSI) [39]. We have familiarized ourselves with various APIs and learned how to utilize MATLAB and Python toolboxes for parsing the CSI files captured. Some challenges of using this software include limited documentation, closed-source nature, and compatibility issues. Nevertheless, we will continue exploring more APIs to customize the software for additional use cases.

To further improve the software side, we have investigated various techniques for Wi-Fi signal processing and feature extraction, such as **P**rincipal **C**omponent **A**nalysis (PCA) and **D**iscrete **W**avelet **T**ransform (DWT) [40]. This will enable us to extract meaningful information from the collected Wi-Fi signals and facilitate their integration with other modalities in the multi-modal model.

4.2 Vision Signals

We have successfully implemented the FFmpeg API to capture videos with different parameters, such as video codec, width, height, resolution, and bitrate. The advantages of FFmpeg include its open-source nature, cross-platform compatibility, and extensive functionality for processing



```
[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,1), raw=0], LegacyCSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000
kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,2), raw=0], BasebandSignal:[(float) 565x1], PreEQSymbols:[0x0x1], MACHeader:[type=[CF]ACK, dest=f8:ff:c2:4:2f:e4, src=e0:8b:e3:e8:0:0, seq=0, frag=0, mfrags=0], MPDU=14B} | BB performance: {Rx:2255.092 MS, lost:527.520 MS, rate:
0.766}
[20:50:41.776338] [SDR     ] [Debug] hackrf0(SoapySDR Compatible SDR)<--RxFrame:{RxSBasic:[device=USRP(SDR), center=5240, control=5240, CBW=20, format=NonHT, Pkt_CBW=20, MCS=4, numSTS=1, GI=0.8us, UsrIdx/NUsr=(0/1), timestamp=124, system_ns=1678798241776238465, NF=-35, RSS=
-21], RxExtraInfo:[len=43, ver=0x2, mac_cur[4-6]=3c:7:93, mac_rom[4-6]=3c:7:93, txcm=1, rxcm=1, txpower=0, cf=5240.000000 MHz, sf=20.000000 MHz, cfo=-20.479000 kHz], SDRExtra:[scrambler=93, packetStartInternal=0, rxIndex=0, rxTime=0.000000, lastTxTime=0.000000], (NonHT)CSI:
[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,1), raw=0], LegacyCSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000
kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,2), raw=0], BasebandSignal:[(float) 565x1], PreEQSymbols:[0x0x1], MACHeader:[type=[CF]CTS, dest=c8:89:f3:e9:df:22, src=62:c0:ad:23:0:0, seq=0, frag=0, mfrags=0], MPDU=14B} | BB performance: {Rx:2255.092 MS, lost:527.520 MS, rate
:0.766}
[20:50:41.778565] [SDR     ] [Debug] hackrf0(SoapySDR Compatible SDR)<--RxFrame:{RxSBasic:[device=USRP(SDR), center=5240, control=5240, CBW=20, format=NonHT, Pkt_CBW=20, MCS=4, numSTS=1, GI=0.8us, UsrIdx/NUsr=(0/1), timestamp=59, system_ns=1678798241778474276, NF=-35, RSS=-
21], RxExtraInfo:[len=43, ver=0x2, mac_cur[4-6]=3c:7:93, mac_rom[4-6]=3c:7:93, txcm=1, rxcm=1, txpower=0, cf=5240.000000 MHz, sf=20.000000 MHz, cfo=-34.237000 kHz], SDRExtra:[scrambler=46, packetStartInternal=0, rxIndex=0, rxTime=0.000000, lastTxTime=0.000000], (NonHT)CSI:[
device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,1), raw=0], LegacyCSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000 k
Hz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,2), raw=0], BasebandSignal:[(float) 645x1], PreEQSymbols:[0x0x1], MACHeader:[type=[CF]BA, dest=c8:89:f3:e9:df:22, src=d4:4f:67:6:6f:72, seq=4095, frag=15, mfrags=0], MPDU=32B} | BB performance: {Rx:2255.156 MS, lost:527.520 MS, r
ate:0.766}
[20:50:41.782303] [SDR     ] [Debug] hackrf0(SoapySDR Compatible SDR)<--RxFrame:{RxSBasic:[device=USRP(SDR), center=5240, control=5240, CBW=20, format=NonHT, Pkt_CBW=20, MCS=4, numSTS=1, GI=0.8us, UsrIdx/NUsr=(0/1), timestamp=121, system_ns=1678798241782198206, NF=-35, RSS=
-21], RxExtraInfo:[len=43, ver=0x2, mac_cur[4-6]=3c:7:93, mac_rom[4-6]=3c:7:93, txcm=1, rxcm=1, txpower=0, cf=5240.000000 MHz, sf=20.000000 MHz, cfo=-30.081000 kHz], SDRExtra:[scrambler=28, packetStartInternal=0, rxIndex=0, rxTime=0.000000, lastTxTime=0.000000], (NonHT)CSI:
[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,1), raw=0], LegacyCSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000
kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,2), raw=0], BasebandSignal:[(float) 645x1], PreEQSymbols:[0x0x1], MACHeader:[type=[CF]BA, dest=c8:89:f3:e9:df:22, src=d4:4f:67:6:6f:72, seq=4095, frag=15, mfrags=0], MPDU=32B} | BB performance: {Rx:2255.224 MS, lost:527.520 MS,
rate:0.766}
[20:50:41.786613] [SDR     ] [Debug] hackrf0(SoapySDR Compatible SDR)<--RxFrame:{RxSBasic:[device=USRP(SDR), center=5240, control=5240, CBW=20, format=NonHT, Pkt_CBW=20, MCS=4, numSTS=1, GI=0.8us, UsrIdx/NUsr=(0/1), timestamp=15, system_ns=1678798241786447380, NF=-35, RSS=-
21], RxExtraInfo:[len=43, ver=0x2, mac_cur[4-6]=3c:7:93, mac_rom[4-6]=3c:7:93, txcm=1, rxcm=1, txpower=0, cf=5240.000000 MHz, sf=20.000000 MHz, cfo=-32.377000 kHz], SDRExtra:[scrambler=42, packetStartInternal=0, rxIndex=0, rxTime=0.000000, lastTxTime=0.000000], (NonHT)CSI:[
device=USRP(SDR), format=NonHT, CBW=20, cf=524                                                    ,nSTS,nESS,nRx,nCSI)=(52,1,0,1,1), raw=0], LegacyCSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000 k
Hz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,2                                                    der:[type=[CF]ACK, dest=1c:57:dc:72:bb:41, src=14:2a:4a:76:0:0, seq=4095, frag=1, mfrags=0], MPDU=14B} | BB performance: {Rx:2255.352 MS, lost:527.520 MS, ra
te:0.766}
[20:50:41.789958] [SDR     ] [Debug] hackrf0(S                                                    240, control=5240, CBW=20, format=NonHT, Pkt_CBW=20, MCS=1, numSTS=1, GI=0.8us, UsrIdx/NUsr=(0/1), timestamp=18446744073709551341, system_ns=1678798241789022
664, NF=-35, RSS=-18], RxExtraInfo:[len=43, ve                                                    wer=0, cf=5240.000000 MHz, sf=20.000000 MHz, cfo=-31.397000 kHz], SDRExtra:[scrambler=113, packetStartInternal=0, rxIndex=0, rxTime=0.000000, lastTxTime=0.00
0000], (NonHT)CSI:[device=USRP(SDR), format=No                                                    000 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,1), raw=0], LegacyCSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcar
rierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nR                                                    bols:[0x0x1], MACHeader:[type=[MF]Beacon, dest=ff:ff:ff:ff:ff:ff, src=d4:4f:67:6:6f:72, seq=2642, frag=0, mfrags=0], MPDU=334B} | BB performance: {Rx:2255.35
2 MS, lost:527.520 MS, rate:0.766}
[20:50:41.792371] [SDR     ] [Debug] hackrf0(S                                                    240, control=5240, CBW=20, format=NonHT, Pkt_CBW=20, MCS=1, numSTS=1, GI=0.8us, UsrIdx/NUsr=(0/1), timestamp=18446744073709551479, system_ns=1678798241791775
544, NF=-35, RSS=-18], RxExtraInfo:[len=43, ve                                                    wer=0, cf=5240.000000 MHz, sf=20.000000 MHz, cfo=-27.678000 kHz], SDRExtra:[scrambler=92, packetStartInternal=0, rxIndex=0, rxTime=0.000000, lastTxTime=0.000
000], (NonHT)CSI:[device=USRP(SDR), format=Non                                                    00 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,1), raw=0], LegacyCSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarr
ierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx                                                    ols:[0x0x1], MACHeader:[type=[MF]Beacon, dest=ff:ff:ff:ff:ff:ff, src=d4:4f:67:6:6f:74, seq=3354, frag=0, mfrags=0], MPDU=289B} | BB performance: {Rx:2255.484
 MS, lost:527.520 MS, rate:0.766}
[20:50:41.794945] [SDR     ] [Debug] hackrf0(S                                                    240, control=5240, CBW=20, format=NonHT, Pkt_CBW=20, MCS=1, numSTS=1, GI=0.8us, UsrIdx/NUsr=(0/1), timestamp=18446744073709551376, system_ns=1678798241794318
201, NF=-35, RSS=-19], RxExtraInfo:[len=43, ve                                                    wer=0, cf=5240.000000 MHz, sf=20.000000 MHz, cfo=-28.182000 kHz], SDRExtra:[scrambler=10, packetStartInternal=0, rxIndex=0, rxTime=0.000000, lastTxTime=0.000
000], (NonHT)CSI:[device=USRP(SDR), format=Non                                                    00 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,1), raw=0], LegacyCSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarr
ierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx                                                    ols:[0x0x1], MACHeader:[type=[MF]Beacon, dest=ff:ff:ff:ff:ff:ff, src=d4:4f:67:6:6f:75, seq=125, frag=0, mfrags=0], MPDU=325B} | BB performance: {Rx:2255.484
MS, lost:527.520 MS, rate:0.766}
[20:50:41.796947] [SDR     ] [Debug] hackrf0(S                                                    240, control=5240, CBW=20, format=NonHT, Pkt_CBW=20, MCS=1, numSTS=1, GI=0.8us, UsrIdx/NUsr=(0/1), timestamp=18446744073709551505, system_ns=1678798241796660
417, NF=-35, RSS=-18], RxExtraInfo:[len=43, ve                                                    wer=0, cf=5240.000000 MHz, sf=20.000000 MHz, cfo=-25.720000 kHz], SDRExtra:[scrambler=97, packetStartInternal=0, rxIndex=0, rxTime=0.000000, lastTxTime=0.000
000], (NonHT)CSI:[device=USRP(SDR), format=Non                                                    00 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,1), raw=0], LegacyCSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarr
ierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx                                                    ols:[0x0x1], MACHeader:[type=[MF]Beacon, dest=ff:ff:ff:ff:ff:ff, src=d4:4f:67:6:6f:7e, seq=3915, frag=0, mfrags=0], MPDU=291B} | BB performance: {Rx:2255.484
 MS, lost:527.520 MS, rate:0.766}
[20:50:41.817983] [SDR     ] [Debug] hackrf0(S                                                    240, control=5240, CBW=20, format=NonHT, Pkt_CBW=20, MCS=4, numSTS=1, GI=0.8us, UsrIdx/NUsr=(0/1), timestamp=80, system_ns=1678798241817833904, NF=-35, RSS=-
21], RxExtraInfo:[len=43, ver=0x2, mac_cur[4-6                                                    000 MHz, sf=20.000000 MHz, cfo=-32.541000 kHz], SDRExtra:[scrambler=62, packetStartInternal=0, rxIndex=0, rxTime=0.000000, lastTxTime=0.000000], (NonHT)CSI:[
device=USRP(SDR), format=NonHT, CBW=20, cf=524                                                    ,nSTS,nESS,nRx,nCSI)=(52,1,0,1,1), raw=0], LegacyCSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000 k
Hz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,2                                                    der:[type=[CF]ACK, dest=1c:57:dc:72:bb:41, src=14:2a:4a:76:0:0, seq=0, frag=0, mfrags=0], MPDU=14B} | BB performance: {Rx:2255.944 MS, lost:527.520 MS, rate:
0.766}
[20:50:41.821177] [SDR     ] [Debug] hackrf0(SoapySDR Compatible SDR)<--RxFrame:{RxSBasic:[device=USRP(SDR), center=5240, control=5240, CBW=20, format=NonHT, Pkt_CBW=20, MCS=4, numSTS=1, GI=0.8us, UsrIdx/NUsr=(0/1), timestamp=27, system_ns=1678798241821038669, NF=-35, RSS=-
21], RxExtraInfo:[len=43, ver=0x2, mac_cur[4-6]=3c:7:93, mac_rom[4-6]=3c:7:93, txcm=1, rxcm=1, txpower=0, cf=5240.000000 MHz, sf=20.000000 MHz, cfo=-29.827000 kHz], SDRExtra:[scrambler=95, packetStartInternal=0, rxIndex=0, rxTime=0.000000, lastTxTime=0.000000], (NonHT)CSI:[
device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,1), raw=0], LegacyCSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000 k
Hz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,2), raw=0], BasebandSignal:[(float) 565x1], PreEQSymbols:[0x0x1], MACHeader:[type=[CF]ACK, dest=1c:57:dc:72:bb:41, src=14:2a:4a:76:0:0, seq=0, frag=0, mfrags=0], MPDU=14B} | BB performance: {Rx:2256.008 MS, lost:527.520 MS, rate:
0.766}
[20:50:41.854183] [SDR     ] [Debug] hackrf0(SoapySDR Compatible SDR)<--RxFrame:{RxSBasic:[device=USRP(SDR), center=5240, control=5240, CBW=20, format=NonHT, Pkt_CBW=20, MCS=1, numSTS=1, GI=0.8us, UsrIdx/NUsr=(0/1), timestamp=18446744073709551439, system_ns=1678798241853449
201, NF=-35, RSS=-18], RxExtraInfo:[len=43, ver=0x2, mac_cur[4-6]=3c:7:93, mac_rom[4-6]=3c:7:93, txcm=1, rxcm=1, txpower=0, cf=5240.000000 MHz, sf=20.000000 MHz, cfo=-25.537000 kHz], SDRExtra:[scrambler=101, packetStartInternal=0, rxIndex=0, rxTime=0.000000, lastTxTime=0.00
0000], (NonHT)CSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,1), raw=0], LegacyCSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcar
rierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,2), raw=0], BasebandSignal:[(float) 6405x1], PreEQSymbols:[0x0x1], MACHeader:[type=[MF]Beacon, dest=ff:ff:ff:ff:ff:ff, src=d4:4f:67:6:6f:72, seq=2643, frag=0, mfrags=0], MPDU=334B} | BB performance: {Rx:2256.66
4 MS, lost:527.520 MS, rate:0.766}
[20:50:41.856499] [SDR     ] [Debug] hackrf0(SoapySDR Compatible SDR)<--RxFrame:{RxSBasic:[device=USRP(SDR), center=5240, control=5240, CBW=20, format=NonHT, Pkt_CBW=20, MCS=1, numSTS=1, GI=0.8us, UsrIdx/NUsr=(0/1), timestamp=18446744073709551376, system_ns=1678798241855871
621, NF=-35, RSS=-18], RxExtraInfo:[len=43, ver=0x2, mac_cur[4-6]=3c:7:93, mac_rom[4-6]=3c:7:93, txcm=1, rxcm=1, txpower=0, cf=5240.000000 MHz, sf=20.000000 MHz, cfo=-26.608000 kHz], SDRExtra:[scrambler=114, packetStartInternal=0, rxIndex=0, rxTime=0.000000, lastTxTime=0.00
0000], (NonHT)CSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,1), raw=0], LegacyCSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcar
rierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,2), raw=0], BasebandSignal:[(float) 5605x1], PreEQSymbols:[0x0x1], MACHeader:[type=[MF]Beacon, dest=ff:ff:ff:ff:ff:ff, src=d4:4f:67:6:6f:74, seq=3355, frag=0, mfrags=0], MPDU=289B} | BB performance: {Rx:2256.66
4 MS, lost:527.520 MS, rate:0.766}
[20:50:41.859375] [SDR     ] [Debug] hackrf0(SoapySDR Compatible SDR)<--RxFrame:{RxSBasic:[device=USRP(SDR), center=5240, control=5240, CBW=20, format=NonHT, Pkt_CBW=20, MCS=1, numSTS=1, GI=0.8us, UsrIdx/NUsr=(0/1), timestamp=18446744073709551474, system_ns=1678798241858557
711, NF=-35, RSS=-18], RxExtraInfo:[len=43, ver=0x2, mac_cur[4-6]=3c:7:93, mac_rom[4-6]=3c:7:93, txcm=1, rxcm=1, txpower=0, cf=5240.000000 MHz, sf=20.000000 MHz, cfo=-34.299000 kHz], SDRExtra:[scrambler=29, packetStartInternal=0, rxIndex=0, rxTime=0.000000, lastTxTime=0.000
000], (NonHT)CSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,1), raw=0], LegacyCSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarr
ierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,2), raw=0], BasebandSignal:[(float) 6245x1], PreEQSymbols:[0x0x1], MACHeader:[type=[MF]Beacon, dest=ff:ff:ff:ff:ff:ff, src=d4:4f:67:6:6f:75, seq=126, frag=0, mfrags=0], MPDU=325B} | BB performance: {Rx:2256.796
MS, lost:527.520 MS, rate:0.766}
[20:50:41.861449] [SDR     ] [Debug] hackrf0(SoapySDR Compatible SDR)<--RxFrame:{RxSBasic:[device=USRP(SDR), center=5240, control=5240, CBW=20, format=NonHT, Pkt_CBW=20, MCS=1, numSTS=1, GI=0.8us, UsrIdx/NUsr=(0/1), timestamp=18446744073709551403, system_ns=1678798241861211
244, NF=-35, RSS=-18], RxExtraInfo:[len=43, ver=0x2, mac_cur[4-6]=3c:7:93, mac_rom[4-6]=3c:7:93, txcm=1, rxcm=1, txpower=0, cf=5240.000000 MHz, sf=20.000000 MHz, cfo=-27.496000 kHz], SDRExtra:[scrambler=14, packetStartInternal=0, rxIndex=0, rxTime=0.000000, lastTxTime=0.000
000], (NonHT)CSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,1), raw=0], LegacyCSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarr
ierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,2), raw=0], BasebandSignal:[(float) 5685x1], PreEQSymbols:[0x0x1], MACHeader:[type=[MF]Beacon, dest=ff:ff:ff:ff:ff:ff, src=d4:4f:67:6:6f:7e, seq=3916, frag=0, mfrags=0], MPDU=291B} | BB performance: {Rx:2256.796
 MS, lost:527.520 MS, rate:0.766}
[20:50:41.881950] [SDR     ] [Debug] hackrf0(SoapySDR Compatible SDR)<--RxFrame:{RxSBasic:[device=USRP(SDR), center=5240, control=5240, CBW=20, format=NonHT, Pkt_CBW=20, MCS=0, numSTS=1, GI=0.8us, UsrIdx/NUsr=(0/1), timestamp=55, system_ns=1678798241881801290, NF=-35, RSS=-
18], RxExtraInfo:[len=43, ver=0x2, mac_cur[4-6]=3c:7:93, mac_rom[4-6]=3c:7:93, txcm=1, rxcm=1, txpower=0, cf=5240.000000 MHz, sf=20.000000 MHz, cfo=-38.235000 kHz], SDRExtra:[scrambler=13, packetStartInternal=0, rxIndex=0, rxTime=0.000000, lastTxTime=0.000000], (NonHT)CSI:[
device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,1), raw=0], LegacyCSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000 k
Hz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,2), raw=0], BasebandSignal:[(float) 885x1], PreEQSymbols:[0x0x1], MACHeader:[type=[CF]ACK, dest=7c:2a:db:61:8e:cf, src=31:48:7a:54:0:0, seq=0, frag=0, mfrags=0], MPDU=14B} | BB performance: {Rx:2257.188 MS, lost:527.520 MS, rate:
0.766}
[20:50:41.882677] [SDR     ] [Debug] hackrf0(SoapySDR Compatible SDR)<--RxFrame:{RxSBasic:[device=USRP(SDR), center=5240, control=5240, CBW=20, format=NonHT, Pkt_CBW=20, MCS=0, numSTS=1, GI=0.8us, UsrIdx/NUsr=(0/1), timestamp=105, system_ns=1678798241882563305, NF=-35, RSS=
-18], RxExtraInfo:[len=43, ver=0x2, mac_cur[4-6]=3c:7:93, mac_rom[4-6]=3c:7:93, txcm=1, rxcm=1, txpower=0, cf=5240.000000 MHz, sf=20.000000 MHz, cfo=-24.090000 kHz], SDRExtra:[scrambler=107, packetStartInternal=0, rxIndex=0, rxTime=0.000000, lastTxTime=0.000000], (NonHT)CSI
:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,1), raw=0], LegacyCSI:[device=USRP(SDR), format=NonHT, CBW=20, cf=5240.000000 MHz, sf=20.000000 MHz, subcarrierBW=312.500000
 kHz, dim(nTones,nSTS,nESS,nRx,nCSI)=(52,1,0,1,2), raw=0], BasebandSignal:[(float) 1205x1], PreEQSymbols:[0x0x1], MACHeader:[type=[CF]VHT/HE NDP Ann., dest=7c:2a:db:61:8e:cf, src=d5:4f:67:6:6f:72, seq=2592, frag=12, mfrags=0], MPDU=25B} | BB performance: {Rx:2257.188 MS, lo
st:527.520 MS, rate:0.766}
```

Figure 5: Real-time CPI capturing by PicoScenes

video and audio files [41]. The flexibility and efficiency of FFmpeg make it an ideal choice for capturing vision signals in various settings, as it allows for real-time video processing and easy integration with other tools and libraries.

Additionally, we have utilized the Google MediaPipe library to recognize human hands. MediaPipe offers several advantages compared to other projects, such as open-source access, high performance, and a modular framework for building multimodal machine learning applications [42]. By leveraging the capabilities of MediaPipe, we can efficiently detect and track hand gestures and poses, which will be crucial for our multi-modal model's motion perception and classification tasks.

Furthermore, we have reviewed recent advancements in computer vision techniques for human activity recognition, such as **C**onvolutional **N**eural **N**etworks (CNNs) and **R**ecurrent **N**eural **N**etworks (RNNs) [43]. These insights will help inform the development of our vision-based motion perception and classification algorithms and ensure that our multi-modal model leverages state-of-the-art techniques.

4.3 sEMG signals

We have rewritten the API provided by Biometrics Ltd, thereby enhancing performance and enabling access to more information from the underlying hardware. Moreover, we are currently reviewing research articles on using sEMG signals for gesture recognition and motion perception [44]. This will inform our proposed models, leading to higher performance and accuracy in these tasks.

To ensure the quality of the collected sEMG data, we have investigated various techniques for signal preprocessing, such as filtering, normalization, and segmentation [45]. Implementing these techniques will help minimize noise and artifacts in the sEMG signals and facilitate extracting meaningful features for motion perception and classification.

In addition, we have explored feature extraction methods specifically tailored for sEMG signals, such as time-domain, frequency-domain, and time-frequency domain features [46]. Understanding the advantages and drawbacks of these methods will enable us to select the most appropriate features for our multi-modal model and ensure its effectiveness in capturing the nuances of muscle activation patterns.

Lastly, we have examined machine learning algorithms and deep learning architectures that have been successfully applied to sEMG-based gesture recognition and motion perception tasks, such as **S**upport **V**ector **M**achines (SVMs), **A**rtificial **N**eural **N**etworks (ANNs), and **C**onvolutional **N**eural **N**etworks (CNNs) [47]. This knowledge will inform the development of our sEMG-based motion perception and classification algorithms and contribute to the overall performance of the multi-modal model.

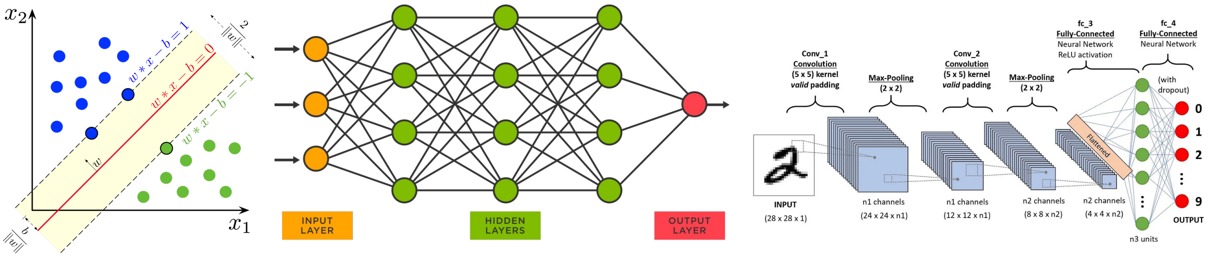

x_2
x_1
$w * x - b = 1$
$w * x - b = 0$
$w * x - b = -1$
$\frac{2}{\|w\|}$
$\frac{b}{\|w\|}$
w
INPUT LAYER
HIDDEN LAYERS
OUTPUT LAYER
Conv_1 Convolution (5 x 5) kernel valid padding
Max-Pooling (2 x 2)
Conv_2 Convolution (5 x 5) kernel valid padding
Max-Pooling (2 x 2)
fc_3 Fully-Connected Neural Network ReLU activation
fc_4 Fully-Connected Neural Network
(with dropout)
Flattened
INPUT (28 x 28 x 1)
n1 channels (24 x 24 x n1)
n1 channels (12 x 12 x n1)
n2 channels (8 x 8 x n2)
n2 channels (4 x 4 x n2)
n3 units
0
1
2
9
OUTPUT

Figure 6: Structure of SVM, ANN and CNN

We have also completed the first-stage establishment of the online data collection and analysis platform. It currently captures real-time sEMG and vision signals through web interfaces. Integration of Wi-Fi signals is still in progress. As we continue to refine and enhance each platform component, we expect to create a robust and efficient system capable of facilitating high-quality multi-modal data collection and analysis for a wide range of applications.

5 Research Plan, Expected Objectives, and Results

5.1 Research Plan

The research plan for the multi-modal human-computer interaction perception technology project will be carried out in a systematic manner. The plan includes the following steps:

1. Literature review and state-of-the-art analysis: Thoroughly review existing literature and research on Wi-Fi, sEMG, and vision-based motion perception, as well as multi-modal fusion techniques.

2. Development of methodologies and algorithms: Design and implement novel algorithms and techniques for each modality and the fusion process, leveraging the insights gained from the literature review.

3. Evaluation and validation: Develop performance metrics, benchmark datasets, and conduct comparative analyses with existing methods to assess the effectiveness and reliability of the proposed framework.

4. User studies and applications: Evaluate the usability and user satisfaction of the integrated system through user studies and explore potential applications in various scenarios.

5. Dissemination of results: Publish research findings in peer-reviewed journals and conferences, and contribute to open-source software projects related to multi-modal human-computer interaction perception technology.

5.2 Expected Objectives and Research Achievements

Upon the completion of this research project, we expect to achieve the following objectives and research accomplishments:

- A comprehensive understanding of the state-of-the-art in Wi-Fi, sEMG, and vision-based motion perception techniques, as well as multi-modal fusion strategies.

- Development of novel algorithms and techniques for accurate and efficient motion perception using Wi-Fi, sEMG, and vision signals, individually and in combination.

- A robust and reliable multi-modal human-computer interaction perception framework that can effectively integrate Wi-Fi, sEMG, and vision signals for various motion perception tasks.
- Rigorous validation of the proposed framework, demonstrating its superior performance compared to existing single-modality and multi-modal systems.
- Identification of potential applications of the integrated system in various scenarios, such as rehabilitation, sports, and entertainment.
- Publication of research findings in high-impact journals and conferences, and contributions to the broader research community through open-source projects.

5.3 Time Scheme

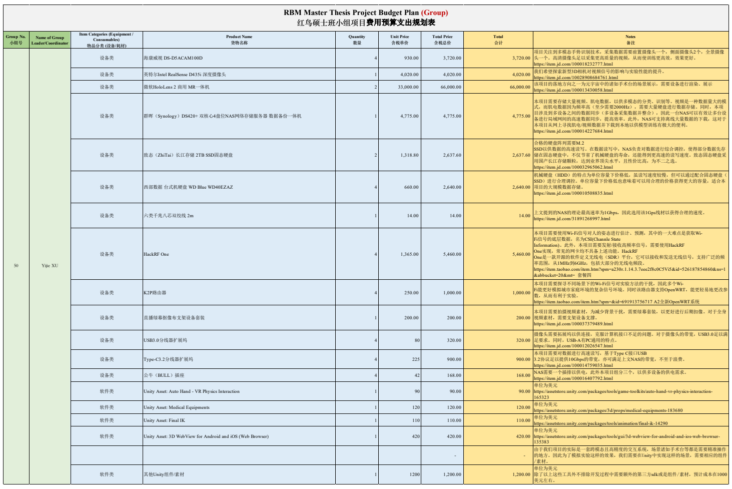
The following time scheme outlines the timeline for the research project, starting on March 19th, 2023, and concluding by August 2024.

- **March 2023 - June 2023:** Conduct literature review and state-of-the-art analysis.
- **July 2023 - December 2023:** Develop methodologies and algorithms for each modality and the fusion process.
- **January 2024 - April 2024:** Evaluate and validate the proposed framework, including comparative analyses and user studies.
- **May 2024 - June 2024:** Explore potential applications and finalize research findings.
- **July 2024 - August 2024:** Prepare and submit research publications, and contribute to open-source projects.

By following this structured research plan and time scheme, we aim to complete the proposed research project within the designated timeframe, leading to significant advancements in multi-modal human-computer interaction perception technology and contributing to the body of knowledge in this domain.

6 Required Conditions and Resources

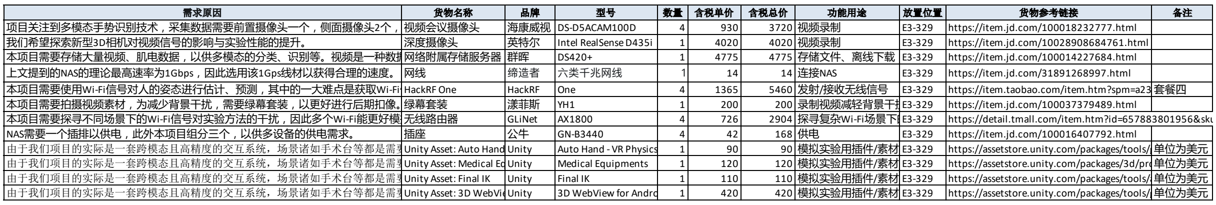
Having our budget plan [fig:7] approved by RBM, we have commenced the procurement process [fig:8] and have submitted a comprehensive list of first-stage products and resources required for the successful execution of our project. Prior to submission, these resources underwent thorough scrutiny and evaluation, based on their quality and competitive pricing. Currently, we are awaiting the Division of RBM's approval to proceed to the next phase of procurement.



RBM Master Thesis Project Budget Plan (Group)
红鸟硕士班小组项目费用预算支出规划表

Group No. 小组号	Name of Group Leader/Coordinator	Item Categories (Equipment / Consumables) 物品分类 (设备/耗材)	Product Name 货物名称	Quantity 数量	Unit Price 含税单价	Total Price 含税总价	Total 合计	Notes 备注
50	Yijie XU	设备类	海康威视 DS-D5ACAM100D	4	930.00	3,720.00	3,720.00	项目关注到多模态手势识别技术，采集数据需要前置摄像头一个，侧面摄像头2个，全景摄像头一个。高清摄像头足以采集更高质量的视频，从而使训练更高效，效果更好。 https://item.jd.com/100018232777.html
		设备类	英特尔Intel RealSense D435i 深度摄像头	1	4,020.00	4,020.00	4,020.00	我们希望探索新型3D相机对视频信号的影响与实验性能的提升。 https://item.jd.com/10028908684761.html
		设备类	微软HoloLens 2 商用 MR一体机	2	33,000.00	66,000.00	66,000.00	该项目的落地方向之一为元宇宙中的诸如手术台的场景展示，需要设备进行渲染、展示 https://item.jd.com/100013430058.html
		设备类	群晖（Synology）DS420+ 双核心4盘位NAS网络存储服务器 数据备份一体机	1	4,775.00	4,775.00	4,775.00	本项目需要存储大量视频、肌电数据，以供多模态的分类、识别等。视频是一种数据量大的模式，而肌电数据因为频率高（至少需要2000Hz），需要大量硬盘进行数据存储。同时，本项目涉及到多设备之间的数据同步（多设备采集数据并整合），因此一台NAS可以有效让多台设备进行局域网间的高速数据同步，提高效率。此外，NAS可支持离线大量数据的下载，这对于本项目从网上寻找肌电/视频数据并下载到本地以供模型训练有极大的便利。 https://item.jd.com/100014227684.html
		设备类	致态（ZhiTai）长江存储 2TB SSD固态硬盘	2	1,318.80	2,637.60	2,637.60	合格的硬盘阵列需要M.2 SSD以供数据的高速读写。在数据读写中，NAS负责对数据进行综合调控，使得部分数据先存储在固态硬盘中，不仅节省了机械硬盘的寿命，还能得到更高速的读写速度。致态固态硬盘采用国产长江存储颗粒，达到业界顶尖水平，且性价比高，为不二之选。 https://item.jd.com/100032965062.html
		设备类	西部数据 台式机硬盘 WD Blue WD40EZAZ	4	660.00	2,640.00	2,640.00	机械硬盘（HDD）的特点为单位容量下价格低，虽读写速度较慢，但可以通过配合固态硬盘（SSD）进行合理调控。单位容量下价格低也意味着可以用合理的价格获得更大的容量，适合本项目的大规模数据存储。 https://item.jd.com/100010508835.html
		设备类	六类千兆八芯双绞线 2m	1	14.00	14.00	14.00	上文提到的NAS的理论最高速率为1Gbps，因此选用该1Gps线材以获得合理的速度。 https://item.jd.com/31891268997.html
		设备类	HackRF One	4	1,365.00	5,460.00	5,460.00	本项目需要使用Wi-Fi信号对人的姿态进行估计、预测，其中的一大难点是获取Wi-Fi信号的底层数据，名为CSI(Channle State Information)。此外，本项目需要发射/接收高频率信号，需要使用HackRF One实现，常见的网卡均不具备上述功能。HackRF One是一款开源的软件定义无线电（SDR）平台，它可以接收和发送无线信号，支持广泛的频率范围，从1MHz到6GHz，包括大部分的无线电频段。 https://item.taobao.com/item.htm?spm=a230r.1.14.3.7eee2f8c0C5Vi5&id=526187854860&ns=1&abbucket=20&mt= 套餐四
		设备类	K2P路由器	4	250.00	1,000.00	1,000.00	本项目需要探寻不同场景下的Wi-Fi信号对实验方法的干扰，因此多个Wi-Fi能更好模拟城市家庭环境的复杂信号环境。同时该路由器支持OpenWRT，能更轻易地更改参数，从而有利于实验。 https://item.taobao.com/item.htm?spm=&id=691913756717 A2全新OpenWRT系统
		设备类	直播绿幕抠像布支架设备套装	1	200.00	200.00	200.00	本项目需要拍摄视频素材，为减少背景干扰，需要绿幕套装，以更好进行后期扣像。对于全身视频素材，需要支架设备支撑。 https://item.jd.com/100037379489.html
		设备类	USB3.0分线器扩展坞	4	80	320.00	320.00	摄像头需要拓展坞以供连接，克服计算机接口不足的问题。对于摄像头的带宽，USB3.0足以满足要求。同时，USB-A有PC通用的特点。 https://item.jd.com/100012026547.html
		设备类	Type-C3.2分线器扩展坞	4	225	900.00	900.00	本项目需要对数据进行高速读写，基于Type C接口USB 3.2协议足以提供10Gbps的带宽，亦可满足上文NAS的带宽，不至于浪费。 https://item.jd.com/100014759035.html
		设备类	公牛（BULL）插座	4	42	168.00	168.00	NAS需要一个插排以供电，此外本项目组分三个，以供多设备的供电需求。 https://item.jd.com/100016407792.html
		软件类	Unity Asset: Auto Hand - VR Physics Interaction	1	90	90.00	90.00	单位为美元 https://assetstore.unity.com/packages/tools/game-toolkits/auto-hand-vr-physics-interaction-165323
		软件类	Unity Asset: Medical Equipments	1	120	120.00	120.00	单位为美元 https://assetstore.unity.com/packages/3d/props/medical-equipments-183680
		软件类	Unity Asset: Final IK	1	110	110.00	110.00	单位为美元 https://assetstore.unity.com/packages/tools/animation/final-ik-14290
		软件类	Unity Asset: 3D WebView for Android and iOS (Web Browser)	1	420	420.00	420.00	单位为美元 https://assetstore.unity.com/packages/tools/gui/3d-webview-for-android-and-ios-web-browser-135383
						-	-	由于我们项目的实际是一套跨模态且高精度的交互系统，场景诸如手术台等都是需要精准操作的地方。因此为了模拟实验这样的效果，我们需要在Unity中实现这样的场景，需要相应的组件/素材。
		软件类	其他Unity组件/素材	1	1200	1,200.00	1,200.00	单位为美元 除了以上这些工具外不排除开发过程中需要额外的第三方sdk或是组件/素材，预计成本在1000美元左右。

Figure 7: Part of the budget plan to RBM



需求原因	货物名称	品牌	型号	数量	含税单价	含税总价	功能用途	放置位置	货物参考链接	备注
项目关注到多模态手势识别技术，采集数据需要前置摄像头一个，侧面摄像头2个，	视频会议摄像头	海康威视	DS-D5ACAM100D	4	930	3720	视频录制	E3-329	https://item.jd.com/100018232777.html	
我们希望探索新型3D相机对视频信号的影响与实验性能的提升。	深度摄像头	英特尔	Intel RealSense D435i	1	4020	4020	视频录制	E3-329	https://item.jd.com/10028908684761.html	
本项目需要存储大量视频、肌电数据，以供多模态的分类、识别等。视频是一种数	网络附属存储服务器	群晖	DS420+	1	4775	4775	存储文件、离线下载	E3-329	https://item.jd.com/100014227684.html	
上文提到的NAS的理论最高速率为1Gbps，因此选用该1Gps线材以获得合理的速度。	网线	缔造者	六类千兆网线	1	14	14	连接NAS	E3-329	https://item.jd.com/31891268997.html	
本项目需要使用Wi-Fi信号对人的姿态进行估计、预测，其中的一大难点是获取Wi-Fi	HackRF One	HackRF	One	4	1365	5460	发射/接收无线信号	E3-329	https://item.taobao.com/item.htm?spm=a23	套餐四
本项目需要拍摄视频素材，为减少背景干扰，需要绿幕套装，以更好进行后期扣像。	绿幕套装	漾菲斯	YH1	1	200	200	录制视频减轻背景干	E3-329	https://item.jd.com/100037379489.html	
本项目需要探寻不同场景下的Wi-Fi信号对实验方法的干扰，因此多个Wi-Fi能更好模	无线路由器	GLiNet	AX1800	4	726	2904	探寻复杂Wi-Fi场景下	E3-329	https://detail.tmall.com/item.htm?id=657883801956&sku	
NAS需要一个插排以供电，此外本项目组分三个，以供多设备的供电需求。	插座	公牛	GN-B3440	4	42	168	供电	E3-329	https://item.jd.com/100016407792.html	
由于我们项目的实际是一套跨模态且高精度的交互系统，场景诸如手术台等都是需	Unity Asset: Auto Hand	Unity	Auto Hand - VR Physics	1	90	90	模拟实验用插件/素材	E3-329	https://assetstore.unity.com/packages/tools/	单位为美元
由于我们项目的实际是一套跨模态且高精度的交互系统，场景诸如手术台等都是需	Unity Asset: Medical Eq	Unity	Medical Equipments	1	120	120	模拟实验用插件/素材	E3-329	https://assetstore.unity.com/packages/3d/pr	单位为美元
由于我们项目的实际是一套跨模态且高精度的交互系统，场景诸如手术台等都是需	Unity Asset: Final IK	Unity	Final IK	1	110	110	模拟实验用插件/素材	E3-329	https://assetstore.unity.com/packages/tools/	单位为美元
由于我们项目的实际是一套跨模态且高精度的交互系统，场景诸如手术台等都是需	Unity Asset: 3D WebVi	Unity	3D WebView for Andr	1	420	420	模拟实验用插件/素材	E3-329	https://assetstore.unity.com/packages/tools/	单位为美元

Figure 8: Part of the procurement application to RBM

Moreover, the effective implementation of our experimental design necessitates the availability of spacious environments. Specifically, each of the three modalities of sensors we plan to utilize requires adequate space to operate optimally. For instance, the transmission of Wi-Fi signals requires a designated room. Similarly, the operation of individuals with surface electromyography (sEMG) sensors entails ample room to perform different actions. Additionally, the procurement of a green screen necessitates a sizable area for its placement and the installation of accompanying cameras.

7 Anticipated Problems and Solutions

In the course of undertaking this research project, we anticipate encountering several challenges that may impede our progress. However, we have proactively devised strategies to mitigate these issues and facilitate a smooth research process.

1. Hardware-Software Compatibility Issues with Wi-Fi Signals: We anticipate potential compatibility issues between the hardware (HackRF One[37]) and software (PicoScenes[39]) components of our Wi-Fi signal processing system. These challenges may arise due to limited documentation and tutorials available for integrating these components. To address this, we plan to engage the software author on GitLab by submitting detailed issues outlining our concerns. Moreover, we will explore relevant literature, technical forums, and online communities to gather insights and potential solutions from experts in the field.

2. Uneven Task Distribution within the Research Group: A heterogeneous distribution of tasks among the group members may lead to inconsistencies in the completion status of individual assignments. To mitigate this, we will establish a structured communication framework that includes regular group meetings, progress updates, and collaborative task management. This will enable us to synchronize our efforts, maintain transparency, and ensure that all group members are actively contributing to the project's objectives.

3. Procurement Process Delays: The procurement of essential equipment and components may be hindered by lengthy institutional procedures and administrative protocols, potentially delaying the establishment of our experimental platform. To circumvent this issue, we will engage in proactive communication with the relevant administrative personnel to expedite approval processes. Additionally, we will explore alternative sourcing strategies, such as reaching out to industry partners or leveraging existing resources within the institution, to ensure timely acquisition of the required items.

We possess a firm conviction that the obstacles we currently face can be surmounted. Our research project is aimed at achieving a dual purpose of academic excellence and practical engineering applications. We are confident that through our diligent efforts and systematic approach, we can attain this goal.

We recognize the challenges that may arise during the course of our research project, but we are committed to overcoming them with the utmost determination and perseverance. Our project embodies a rigorous and comprehensive investigation of the subject matter, which will be meticulously documented and evaluated to ensure its scholarly value.

Moreover, we are driven to make a significant impact beyond the academic realm by contributing to the engineering community with practical and applicable solutions. We are confident that our findings will be highly regarded and embraced by professionals in the field.

In summary, we are dedicated to executing our research project with excellence, rigor, and innovation. We firmly believe that our work will bridge the gap between academic excellence and practical engineering applications and will be highly regarded by both the academic and engineering communities.

References

[1] F. Adib, Z. Kabelac, D. Katabi, and R. C. Miller, “3D tracking via body radio reflections,” in *11th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 14)*, 2014, pp. 317–329.

[2] F. Adib and D. Katabi, “See through walls with WiFi!” In *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*, 2013, pp. 75–86.

[3] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, “Whole-home gesture recognition using wireless signals,” in *Proceedings of the 19th annual international conference on Mobile computing & networking*, 2013, pp. 27–38.

[4] H. Abdelnasser, M. Youssef, and K. A. Harras, “Wigest: A ubiquitous WiFi-based gesture recognition system,” in *2015 IEEE Conference on Computer Communications (INFOCOM)*, 2015, pp. 1472–1480. DOI: `10.1109/INFOCOM.2015.7218525`.

[5] X. Wang, C. Yang, and S. Mao, “PhaseBeat: Exploiting csi phase data for vital sign monitoring with commodity WiFi devices,” in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, IEEE, 2017, pp. 1230–1239.

[6] X. Wang, X. Wang, and S. Mao, “CiFi: Deep convolutional neural networks for indoor localization with 5 ghz Wi-Fi,” in *2017 IEEE International Conference on Communications (ICC)*, IEEE, 2017, pp. 1–6.

[7] H. Li, W. Yang, J. Wang, Y. Xu, and L. Huang, “Wifinger: Talk to your smart devices with finger-grained gesture,” in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 250–261.

[8] Y. Wang, K. Wu, and L. M. Ni, “Wifall: Device-free fall detection by wireless networks,” *IEEE Transactions on Mobile Computing*, vol. 16, no. 2, pp. 581–594, 2016.

[9] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, “We can hear you with Wi-Fi!” In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom ’14, Maui, Hawaii, USA: Association for Computing Machinery, 2014, pp. 593–604, ISBN: 9781450327831. DOI: `10.1145/2639108.2639112`. [Online]. Available: `https://doi.org/10.1145/2639108.2639112`.

[10] S. Kyeong, J. Feng, J. K. Ryu, J. J. Park, K. H. Lee, and J. Kim, “Surface electromyography characteristics for motion intention recognition and implementation issues in lower-limb exoskeletons,” *International Journal of Control, Automation and Systems*, vol. 20, no. 3, pp. 1018–1028, 2022, ISSN: 2005-4092. DOI: `10.1007/s12555-020-0934-3`. [Online]. Available: `https://doi.org/10.1007/s12555-020-0934-3`.

[11] A. Phinyomark, P. Phukpattaranont, and C. Limsakul, “Feature reduction and selection for emg signal classification,” *Expert systems with applications*, vol. 39, no. 8, pp. 7420–7431, 2012.

[12] M. Ortiz-Catalan, B. Håkansson, and R. Brånemark, “An osseointegrated human-machine gateway for long-term sensory feedback and motor control of artificial limbs,” *Science translational medicine*, vol. 6, no. 257, 257re6–257re6, 2014.

[13] P. Parker, K. Englehart, and B. Hudgins, “Myoelectric signal processing for control of powered limb prostheses,” *Journal of electromyography and kinesiology*, vol. 16, no. 6, pp. 541–548, 2006.

[14] M. Zecca, S. Micera, M. C. Carrozza, and P. Dario, “Control of multifunctional prosthetic hands by processing the electromyographic signal,” *Critical Reviews in Biomedical Engineering*, vol. 30, no. 4-6, 2002.

[15] L. Hargrove, K. Englehart, and B. Hudgins, “A training strategy to reduce classification degradation due to electrode displacements in pattern recognition based myoelectric control,” *Biomedical signal processing and control*, vol. 3, no. 2, pp. 175–180, 2008.

[16] K. Englehart, B. Hudgins, P. A. Parker, and M. Stevenson, “Classification of the myoelectric signal using time-frequency based representations,” *Medical engineering & physics*, vol. 21, no. 6-7, pp. 431–438, 1999.

[17] C. Castellini and P. Van Der Smagt, “Surface emg in advanced hand prosthetics,” *Biological cybernetics*, vol. 100, pp. 35–47, 2009.

[18] L. Diener, M. Janke, and T. Schultz, “Direct conversion from facial myoelectric signals to speech using deep neural networks,” in *2015 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2015, pp. 1–7.

[19] W. Geng, Y. Du, W. Jin, W. Wei, Y. Hu, and J. Li, "Gesture recognition by instantaneous surface emg images," *Scientific reports*, vol. 6, no. 1, p. 36 571, 2016.

[20] Y. Du, W. Jin, W. Wei, Y. Hu, and W. Geng, "Surface emg-based inter-session gesture recognition enhanced by deep domain adaptation," *Sensors*, vol. 17, no. 3, p. 458, 2017.

[21] A. Radmand, E. Scheme, and K. Englehart, "High-density force myography: A possible alternative for upper-limb prosthetic control.," *Journal of Rehabilitation Research & Development*, vol. 53, no. 4, 2016.

[22] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI'81: 7th international joint conference on Artificial intelligence*, vol. 2, 1981, pp. 674–679.

[23] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[24] J. Shotton, A. Fitzgibbon, M. Cook, *et al.*, "Real-time human pose recognition in parts from single depth images," in *CVPR 2011*, IEEE, 2011, pp. 1297–1304.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[26] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[27] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[28] X. Yang and Y. L. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, IEEE, 2012, pp. 14–19.

[29] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 2012, pp. 1290–1297.

[30] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, “An end-to-end spatio-temporal attention model for human action recognition from skeleton data,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[31] A. Graves and A. Graves, “Long short-term memory,” *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[32] K. Cho, B. Van Merriënboer, C. Gulcehre, *et al.*, “Learning phrase representations using rnn encoder-decoder for statistical machine translation,” *arXiv preprint arXiv:1406.1078*, 2014.

[33] J. Donahue, L. Anne Hendricks, S. Guadarrama, *et al.*, “Long-term recurrent convolutional networks for visual recognition and description,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[34] R. Hu and A. Singh, “Transformer is all you need: Multimodal multitask learning with a unified transformer,” *CoRR*, vol. abs/2102.10772, 2021. arXiv: `2102.10772`. [Online]. Available: `https://arxiv.org/abs/2102.10772`.

[35] M. Atzori, A. Gijsberts, C. Castellini, *et al.*, “Electromyography data for non-invasive naturally-controlled robotic hand prostheses,” *Scientific Data*, vol. 1, no. 1, p. 140 053, 2014, ISSN: 2052-4463. DOI: `10.1038/sdata.2014.53`. [Online]. Available: `https://doi.org/10.1038/sdata.2014.53`.

[36] W. Wang, A. X. Liu, and M. Shahzad, “Gait recognition using WiFi signals,” in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp ’16, Heidelberg, Germany: Association for Computing Machinery, 2016, pp. 363–373, ISBN: 9781450344616. DOI: `10.1145/2971648.2971670`. [Online]. Available: `https://doi.org/10.1145/2971648.2971670`.

[37] M. Ossmann, *HackRF One*, https://greatscottgadgets.com/hackrf/, 2018.

[38] J. Karedal, S. Wyne, P. Almers, F. Tufvesson, and A. F. Molisch, “A measurement-based statistical model for industrial ultra-wideband channels,” *IEEE Transactions on Wireless Communications*, vol. 6, no. 8, pp. 3028–3037, 2007. DOI: `10.1109/TWC.2007.051050`.

[39] Z. Jiang, T. H. Luan, X. Ren, *et al.*, "Eliminating the barriers: Demystifying Wi-Fi baseband design and introducing the picoscenes Wi-Fi sensing platform," in *IEEE Internet of Things Journal*, vol. 9, 2022, pp. 4476–4496. DOI: `10.1109/JIOT.2021.3104666`.

[40] S. Palipana, D. Rojas, P. Agrawal, and D. Pesch, "FallDeFi: Ubiquitous fall detection using commodity Wi-Fi devices," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 4, Jan. 2018. DOI: `10.1145/3161183`. [Online]. Available: `https://doi.org/10.1145/3161183`.

[41] S. Tomar, "Converting video formats with ffmpeg," *Linux Journal*, vol. 2006, no. 146, p. 10, 2006.

[42] C. Lugaresi, J. Tang, H. Nash, *et al.*, "Mediapipe: A framework for building perception pipelines," *CoRR*, vol. abs/1906.08172, 2019. arXiv: `1906.08172`. [Online]. Available: `http://arxiv.org/abs/1906.08172`.

[43] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Los Alamitos, CA, USA: IEEE Computer Society, Dec. 2015, pp. 4489–4497. DOI: `10.1109/ICCV.2015.510`. [Online]. Available: `https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.510`.

[44] A. Phinyomark, S. Thongpanja, H. Hu, P. Phukpattaranont, and C. Limsakul, "The usefulness of mean and median frequencies in electromyography analysis," in *Computational Intelligence in Electromyography Analysis*, G. R. Naik, Ed., Rijeka: IntechOpen, 2012, ch. 8. DOI: `10.5772/50639`. [Online]. Available: `https://doi.org/10.5772/50639`.

[45] A. H. Al-Timemy, G. Bugmann, J. Escudero, and N. Outram, "Classification of finger movements for the dexterous hand prosthesis control with surface electromyography," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 3, pp. 608–618, 2016.

[46] T. Tuncer, S. Dogan, and A. Subasi, "Surface emg signal classification using ternary pattern and discrete wavelet transform based feature extraction for hand movement recognition," *Biomedical Signal Processing and Control*, vol. 58, p. 101 872, 2020, ISSN: 1746-8094. DOI: `https://doi.org/10.1016/j.bspc.2020.101872`. [Online].

Available: https://www.sciencedirect.com/science/Article/pii/S1746809420300288.

[47] X. Zhai, B. Jelfs, R. H. Chan, and C. Tin, "Self-recalibrating surface emg pattern recognition for neuroprosthesis control based on convolutional neural network," *Frontiers in neuroscience*, vol. 11, p. 379, 2017.