# Enhancing Coaching and Player Performance Analysis in Table Tennis Through Human Action Recognition

Kangnan Dong

A project report submitted to the Auckland University of Technology
in partial fulfillment of the requirements for the degree of
Master of Computer and Information Sciences (MCIS)

2024

School of Engineering, Computer & Mathematical Sciences

# Abstract

This report aims to enhance the effectiveness of table tennis coaching and perfor-mance analysis through human action recognition by using deep learning. In the field of video analysis, human action recognition has emerged as a highly researched area. However, the complexity of human actions presents significant challenges. To address these issues, in this report, we combine the latest computer vision and deep learning algorithms to accurately identify and classify a few strokes in human action recognition. Throughout in-depth review of the existing methods, we develop a high-precision offline method for player's action recognition. Our experimental results show the success of six actions classifications based on our own dataset with the average accuracy up to 99.85%.

**Keywords**: Table tennis · Human action recognition · Deep learning · Computer vision

# Table of Contents

# List of Figures

# List of Tables

# Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgments), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Signature: Date: 16 October 2024

# Acknowledgment

My deepest gratitude goes out to everyone who has shown support during my master's degree at Auckland University of Technology (AUT). Without their encouragement, guidance, and patience I could never have achieved my goal of graduating with honors.

At first, I would like to acknowledge my family for their unwavering support and continuous care throughout this journey. Their understanding has allowed me to focus solely on my studies and meet this goal successfully.

Professor Wei Qi Yan has been instrumental in my academic development and I feel immensely fortunate to have worked under his guidance. His expertise, thoughtful feedback, and encouragement played a huge role in shaping my thesis; I am extremely appreciative of his dedication to helping refine my ideas while providing insightful comments during this process.

Finally, I would like to acknowledge and thank my colleagues at AUT for all of their assistance, support, and inspiration during my studies. Their companionship has made this journey both rewarding and unforgettable.

Kangnan Dong

Auckland, New Zealand

Oct 2024

# Chapter 1
# Introduction

*This chapter is structured into five sections. The first introduces the background and motivation for the research. The second outlines the key research question. This is followed by an overview of the contributions made by this study, the research objectives, and finally, the overall structure of the report.*

## 1.1 Background

Human Action Recognition in Sports Video Analysis has become a central topic in computer vision and deep learning. As this field evolves, it brings significant advancements in analyzing specific athletic actions, offering practical benefits such as performance analysis, highlight creation, and support for coaching. Human action recognition in table tennis, in particular, faces unique challenges due to the rapid and subtle movements characteristic of this sport, which demand precise and fine-grained detection methods. These challenges underscore the need for advanced techniques, beyond traditional handcrafted methods, to support high-accuracy recognition, especially in fast-moving sports.

This project focuses on developing an offline, high-precision system for recognizing specific table tennis strokes by applying deep learning models, including convolutional neural networks (CNNs) for spatial feature extraction and Transformer architectures for temporal modeling. These models are tested on a dataset containing six table tennis stroke types, collected in real-world conditions to ensure model robustness and accuracy.

Past research in human action recognition has shown its value in sports through performance insights, and even as a training aid (Karpathy et al., 2014). However, sports-specific recognition faces additional complexities, especially due to the rapid sequences of athletic actions, the close similarities between action classes, and high-speed movement (Wang et al., 2016). Traditional action recognition approaches relied on handcrafted visual descriptors, such as Histogram of Oriented Gradients (HOG) and optical flow, and employed classifiers like Support Vector Machines (SVMs) or Hidden Markov Models (HMMs) (Laptev et al., 2008). While effective for certain tasks, these methods fall short in handling the detailed distinctions between table tennis strokes.

The adoption of deep learning, especially CNNs, has provided a considerable breakthrough by enabling automatic learning of complex spatial features from raw video data, making it better suited for tasks like table tennis stroke recognition. A notable

advancement was the two-stream CNN proposed by Simonyan and Zisserman (2014), which combined spatial (RGB frames) and temporal (optical flow) information for enhanced recognition accuracy. This foundation led to further innovations, including the Inflated 3D ConvNet (I3D) by Carreira and Zisserman in 2017, which extended 2D CNN architectures along the temporal dimension and proved effective on large-scale action datasets like Kinetics (Kay et al., 2017).

More recent models have been tailored to sports action recognition. For instance, Zhu et al. (2019) introduced a multi-scale temporal convolutional network for stroke recognition, achieving high accuracy. Similarly, Cai et al. (2020) implemented a temporal segment network with attention mechanisms to detect badminton strokes. Despite these advances, sports action recognition still faces critical challenges, such as the lack of extensive labeled datasets and the complexities introduced by various players and environments (Wang et al., 2019).

Among sports action recognition tasks, table tennis stroke recognition remains challenging. Players' movements are extremely fast, and the small size of the ball means limited visual data is available, making distinctions between stroke types subtle. Early research relied on sensor-based methods, like those used by Blank et al. (2015), who achieved high accuracy in detecting strokes by attaching inertial sensors to rackets. However, sensor-based methods are intrusive and can impact gameplay. Markerless motion capture has emerged as a promising alternative, with Hegazy et al. (2020) utilizing IR depth cameras to reach high accuracy for table tennis stroke detection.

Accurate feedback is crucial for players and coaches to evaluate performance after gameplay. This research addresses the challenge of designing an offline system that can recognize and classify various table tennis actions with high accuracy, enabling effective post-session feedback without relying on real-time processing (Voelikov et al., 2020). Developing such a system requires innovative techniques to capture and analyze table tennis movements from recorded gameplay, expanding the potential for comprehensive performance evaluation. Positioned at the intersection of computer vision, machine

learning, and sports science, this research demands advanced methods that go beyond traditional human action recognition techniques.

The use of deep learning algorithms offered a significant advancement in the table tennis stroke recognition. In table tennis, the human action recognition was conducted by using Twin Spatio Temporal Convolutional Neural Network (TSTCNN) (Martin et al., 2019). The approach on the TTStroke-21 dataset was able to achieve an average of 91 % accuracy. 4% in the identification of a stroke. This work showed that two parallel convolutional streams are effective for modelling spatial and temporal data. A new technique of table tennis stroke recognition (Kulkarni and Shenoy, 2021) was profound by using two-dimensional human pose estimation. “StrokeMaster” is based on pose data derived from the video frames for the classification of strokes with 99% validation accuracy. This approach demonstrates that the high-level pose features are useful for the fine-grained action recognition in the table tennis.

While research continues to improve the precision of table tennis stroke recognition, several issues remain. Model adaptability to different playing styles, player positions, and stroke types is an ongoing area of exploration. Future work may also address more specific factors, such as the positions of players and ball trajectory, to enhance recognition performance and provide a more complete understanding of table tennis gameplay.

## 1.2 Research Problem

Detecting human actions within table tennis gameplay introduces unique challenges due to the specific characteristics of this sport. Two primary issues are at the forefront of this research: the complexities in identifying fast and fluid motions and the need for precise, detailed action recognition. Table tennis involves quick, intricate actions that take place in rapid succession. The high speed and subtle distinctions between actions make accurate detection difficult for traditional computer vision systems. Different stroke types, such as topspin, backspin, and sidespin, each require distinct racket angles, body positioning, and player paths, creating nuanced differences that are challenging to capture without

specialized approaches (Kulkarni & Shenoy, 2021).

Accurate feedback is essential for players and coaches to assess performance after gameplay. This research addresses the challenge of designing an offline system capable of recognizing and classifying various table tennis actions with high accuracy, providing effective post-session feedback without the need for real-time processing (Voelikov et al., 2020). Developing such a system necessitates innovative techniques to capture and analyze table tennis movements from recorded gameplay, broadening the scope for comprehensive performance evaluation. Situated at the intersection of computer vision, machine learning, and sports science, this research requires advanced methods that surpass traditional human action recognition approaches.

The primary objectives of this research are twofold: First, it aims to create an approach for accurately recognizing table tennis movements and to implement a stable system capable of classifying different strokes and actions. This system will leverage advanced computer vision and machine learning techniques to analyze video data, ensuring a high recognition rate for specific actions such as long push, forehand, and backhand. This offline recognition system is intended to serve as a foundation for future analytical and feedback tools.

## 1.3 Research Aim

The primary aim of this research is to develop an advanced system that enhances table tennis coaching and performance analysis through the use of cutting-edge computer vision and deep learning techniques. This system is designed to accurately recognize and classify table tennis player movements, with a specific emphasis on capturing high-speed, subtle actions that are challenging to detect with conventional methods.

The core objective is to create a software system capable of delivering detailed post-session feedback, allowing coaches and players to analyze gameplay effectively and make data-driven adjustments to their techniques. The system focuses on classifying various

table tennis strokes and evaluating player movements from recorded gameplay, offering insights that are beneficial for technique refinement.

Beyond contributing to research, this system has practical applications in sports coaching by offering players and coaches a comprehensive analysis tool. Its offline functionality ensures accessibility for players at all levels without the need for constant supervision. Additionally, the system's potential extends to other fast-moving sports, where accurate action recognition is critical for performance evaluation.

This research aims to build a robust action recognition system that addresses challenges such as rapid movements, occlusions, and variations in playing styles and environments. By leveraging state-of-the-art deep learning models, particularly the Transformer architecture, this research seeks to significantly improve accuracy in classifying different table tennis strokes, ultimately enhancing player development and training methodologies.

## 1.4Research Objectives

(a) To develop a robust action recognition system capable of identifying and classifying six specific table tennis actions with high accuracy. These actions include complex and fast-moving strokes, such as Backhand Drive, Forehand Drive, and Smash, which present significant challenges for traditional recognition systems.

(b) To enhance stroke identification accuracy by analyzing player posture, limb motion, and racket trajectory. This objective involves the integration of MediaPipe for pose estimation, which captures body key points from recorded videos, allowing for precise analysis of player movements.

(c) To design and implement a Transformer-based deep learning model optimized for action recognition. The model is trained to handle the temporal and spatial complexities of table tennis strokes while maintaining computational efficiency.

(d) To ensure system adaptability and robustness by training the model on diverse datasets, including supplementary online training videos. This objective addresses the challenge of generalizing the model to different player styles, camera angles, and environmental conditions, enhancing its applicability across various coaching and analysis scenarios.

(e) To develop a user-friendly interface using PyQt5 and OpenCV, designed for capturing and processing video inputs and displaying post-session analysis. This interface provides coaches and players with an accessible tool for performance evaluation and stroke analysis.

## 1.5 Research Questions

(a) How can advanced computer vision and deep learning techniques be utilized to accurately recognize table tennis player actions from recorded video inputs?

This question explores how action recognition algorithms, particularly those based on Transformer architectures, can capture the fast and fluid movements specific to table tennis, managing the complexities of subtle stroke differences and high-speed gameplay.

(b) What factors contribute to designing an effective action recognition system that provides coaches and players with detailed post-session feedback?

This question examines how the system's design can make action recognition results clear and accessible, focusing on accurate classification of actions and insights that support training and performance improvement.

(c) How adaptable is the action recognition model across different player styles, camera angles, and environments?

This question investigates the system's generalization ability, assessing how well it performs in diverse scenarios and addressing any challenges posed by variations in playing style and setup.

## 1.6 Structure of This Report

The structure of this report is described as follows:

In Chapter 2, we conduct a literature review and discuss the relevant studies focused on human action recognition in sports, particularly table tennis. This chapter also covers the evolution of deep learning models and compares key technologies, models, and methods used for action recognition.

In Chapter 3, we introduce the research methodology and experimental design, including the custom dataset creation, model architecture, and the algorithms employed in this study.

In Chapter 4, we present the collected data and experimental results obtained through the proposed algorithm. This chapter also discusses the limitations of the approach in detail.

In Chapter 5, we summarize and analyze the experimental results, providing insights into the performance of the model and its implications.

In Chapter 6, we conclude the research and propose potential future work to further enhance table tennis action recognition and its applications.

# Chapter 2
# Related Work

*This chapter reviews the relevant literature on action recognition in sports, emphasizing the evolution of deep learning techniques and their application in table tennis.*

## 2.1 Introduction

In recent years, rapid advancements in computer vision and deep learning have significantly influenced sports video analysis, particularly in human action recognition. This chapter provides an overview of current research and developments in action recognition within sports, with a specific focus on the unique challenges posed by table tennis stroke recognition.

Action recognition in sports presents distinct challenges, including fast and subtle movements, complex player dynamics, and varying camera angles. Traditional approaches relied on handcrafted features and classical machine learning methods. However, with the emergence of deep learning models, particularly Convolutional Neural Networks (CNNs) and Transformer-based architectures, considerable progress has been made. These models improve the capacity to capture both spatial and temporal dependencies, leading to higher accuracy in detecting and classifying complex sports actions.

The objective of this chapter is to review key advancements in action recognition, focusing on techniques and methodologies used in sports contexts, including table tennis. The review starts with a discussion of traditional methods, followed by an examination of recent deep learning models, their applications, and limitations. Particular attention is given to the use of CNNs and Transformer architectures in sports action recognition. In this project, MobileNetV2 is employed to efficiently extract high-dimensional features from each video frame, while a Transformer model processes these features to handle temporal dependencies and classify actions. Additionally, this chapter addresses the challenges of data variability and the use of tools like MediaPipe for pose estimation.

This literature review aims to highlight the strengths and limitations of existing methods, providing a foundation for the research problem and justifying the need for developing a robust, high-precision table tennis action recognition system.

## 2.2 Human Action Recognition in Sports

Action recognition in sports has become a crucial field of study in computer vision and machine learning, with significant applications in performance analysis, coaching, and automated sports analysis. Over the past two decades, this field has evolved from basic techniques to advanced deep learning approaches to better capture and analyze the dynamic and fast-moving nature of sports.

In this project, MobileNetV2, a lightweight CNN architecture, is utilized for efficient feature extraction. Each video frame is processed by MobileNetV2 to produce a high-dimensional feature vector, which is then used to build a temporal sequence for subsequent action recognition. The choice of MobileNetV2 ensures efficient processing of recorded video data while maintaining high accuracy.

Earlier works in human action recognition in sports relied on handcrafted features and traditional machine learning techniques. These methods typically followed a two-stage process: feature extraction followed by classification. One of the earliest methods was introduced by Laptev and Lindeberg (2003), who proposed space-time interest points (STIPs) to capture spatiotemporal corners in video sequences. Following this, Dollár et al. (2005) suggested the cuboid feature detector and descriptor, which further improved action recognition across various datasets.

Another notable contribution to traditional methods was by Wang et al. (2011), who introduced dense trajectories for action recognition, leveraging optical flow fields to track dense points. This approach, which captured both motion and appearance information, achieved state-of-the-art results in action recognition benchmarks. However, in sports contexts, these traditional methods faced challenges due to high variability in athlete movements, rapid actions, and complex backgrounds. Moreover, handcrafted features struggled to differentiate between closely related actions, such as different types of swings or shots.

The introduction of deep learning, particularly Convolutional Neural Networks (CNNs), marked a breakthrough in sports action recognition. Unlike traditional approaches, CNNs can automatically learn complex feature representations from raw data, allowing for more detailed and abstract patterns. Early works, such as Karpathy et al. (2014), showed the potential of CNNs in sports video analysis, although they had limited temporal capabilities. Simonyan and Zisserman (2014) later introduced the two-stream CNN model, which processed spatial data (RGB frames) and temporal data (optical flow) separately. This approach proved well-suited for sports action recognition by capturing both appearance and motion.

Subsequent advancements, like the two-stream spatiotemporal residual network by Feichtenhofer et al. (2016), combined spatial and temporal streams with residual links, enabling improved detection of complex sports movements. The model proved effective in differentiating between movements with slight sequence variations. Similarly, 3D CNN architectures, such as C3D (Tran et al., 2015), allowed for longer-duration motion extraction, which is vital in capturing complete sports actions without handcrafted features.

Recent studies have focused on sports-specific applications. For example, Zhu et al. (2016) demonstrated a CNN-based approach for identifying specific tennis actions, showcasing deep learning's ability to distinguish subtle differences between sports actions. In sports action recognition, another challenge is the need for fine-grained classification to differentiate between similar activities, such as various tennis serves or types of baseball pitches. To address this, Wang et al. (2018) introduced the Temporal Segment Network, which effectively models long-range temporal dependencies and enhances detailed action recognition.

Camera angles, which vary widely across sports broadcasts, also present a challenge. Bertasius et al. (2019) addressed this issue with a spatiotemporal attention mechanism that adapts to different viewpoints, focusing on the most relevant regions of the frame for

action recognition.

End-to-end models capable of simultaneously learning spatial features and temporal dependencies have gained popularity in recent years. Carreira and Zisserman (2017) proposed the Inflated 3D ConvNet (I3D), an extension of 2D CNNs into the temporal domain. Trained on large-scale video datasets like Kinetics, I3D has proven highly effective for sports action recognition. More recently, Transformer-based models have shown promise in action recognition tasks. Arnab et al. (2021) introduced ViViT, a video vision transformer capable of learning long-range dependencies in videos, which has proven especially effective for sports actions that unfold over extended periods.

## 2.3 Convolutional Neural Networks (CNN) in Human Action Recognition

The development of Convolutional Neural Networks (CNNs) for video analysis, especially in action recognition, marks a significant breakthrough in computer vision and machine learning. In this project, MobileNetV2, a lightweight CNN architecture, is employed for high-dimensional feature extraction from video frames, generating compact feature vectors that capture spatial aspects of the data. Its efficient inverted residual structure and linear bottleneck layers allow it to maintain low computational costs while achieving high accuracy, making it suitable for analyzing fast-moving sports like table tennis. Compared to other CNN architectures, such as ResNet or VGG, MobileNetV2 balances performance and efficiency, which is particularly beneficial in applications with limited computational resources, such as offline sports analysis systems.

CNNs have proven valuable for video analysis, building on early work like Ji et al. (2013), who introduced a 3D CNN for action recognition. This method extended 2D convolutions used in image analysis to the temporal domain, enabling the network to learn directly from raw video frames. The 3D CNN architecture showed improved performance over conventional methods across various action recognition datasets, paving the way for

future innovations.

Karpathy et al. (2014) expanded CNN applications for video classification by exploring different multi-modal fusion methods to integrate information across temporal dimensions, including early, late, and slow fusion. They also highlighted the challenge of capturing long-range temporal dependencies in videos, an ongoing research focus.

The two-stream architecture by Simonyan and Zisserman (2014) introduced a major advancement, utilizing two separate CNN streams: one for spatial information and the other for motion, represented as optical flow. This architecture captured both appearance and motion, proving effective for action recognition tasks and inspiring numerous subsequent studies. Building on this, Tran et al. (2015) proposed the C3D (3D Convolutional Networks) architecture, which used 3D convolutions to learn spatiotemporal features. The C3D model demonstrated that synchronizing the modeling of both appearance and motion could outperform 2D CNNs in video analysis.

As the field evolved, approaches emerged to handle long-term temporal relationships better. The Temporal Segment Network (TSN) (Wang et al., 2016) addressed this with a segment-based sampling and aggregation module. By dividing videos into segments and randomly sampling frames from each, TSN captures long-term dependencies while reducing computational complexity.

Advances in other areas of computer vision have influenced CNN architectures for action recognition. For example, ResNet's success in image classification (He et al., 2016) encouraged researchers to adapt residual learning for video analysis. Feichtenhofer et al. (2016) introduced a two-stream recurrent residual network for action recognition, demonstrating that residual learning enhances performance in the spatiotemporal domain.

Carreira and Zisserman (2017) introduced the Inflated 3D ConvNet (I3D), which extends 2D CNNs to the temporal domain, allowing image classification models to be repurposed for video analysis. I3D set new benchmarks on several action recognition

datasets and has become a popular baseline for further research. Similarly, the SlowFast network (Feichtenhofer et al., 2019) incorporates two pathways operating at different temporal scales: a slow pathway to capture spatial details at a low frame rate and a fast pathway for motion at a high frame rate. This design addresses the conflict between capturing fine motion and spatial detail, improving recognition performance.

When comparing CNN architectures for action recognition, several effective models stand out. The two-stream architecture provides computational efficiency with 2D CNNs but may not capture temporal dependencies in detail. The 3D CNN models, such as C3D and I3D, excel at learning spatiotemporal features but come with higher computational costs and larger parameter counts.

## 2.4 Transformer Models in Human Action Recognition

Originally proposed by Vaswani et al. (2017) for natural language processing, Transformer models have since made substantial impacts in computer vision and action recognition, particularly for sports applications. Transformers leverage self-attention mechanisms to capture long-range dependencies and enable parallel computation, making them well-suited for processing complex visual data like sports videos, where fine-grained actions must be distinguished accurately and efficiently.

In this project, the Transformer model is applied to temporal sequences constructed from feature vectors generated by MobileNetV2. The Transformer processes these sequences to classify actions and detect boundaries, such as the start and end of an action. Its self-attention mechanism is highly effective at handling both short-term and long-term dependencies, addressing the varying duration of actions common in sports like table tennis, were rapid and subtle movements demand precision. This approach enhances classification accuracy while maintaining computational efficiency.

Compared to traditional models, Transformers offer distinct advantages by handling

entire sequences in parallel, overcoming the limitations of recurrent structures and the fixed receptive fields commonly found in CNNs. This parallel computation allows for efficient and accurate processing, making Transformers especially suitable for offline action recognition, where large volumes of video data can be analyzed post-session without real-time constraints.

Transformers were initially introduced to vision tasks through Dosovitskiy et al.'s Vision Transformer (ViT) (2021), which demonstrated the potential of applying self-attention to image classification and opened new avenues for video analysis. Subsequent models like TimeSformer by Bertasius et al. (2021) extended this approach to video understanding, utilizing self-attention across both spatial and temporal dimensions. TimeSformer achieved state-of-the-art results on several action recognition benchmarks, confirming the suitability of Transformers for modeling intricate spatio-temporal relationships in sports actions.

Arnab et al. (2021) introduced the ViViT model, further refining Transformers' role in video action recognition by exploring various methods for implementing self-attention across spatial and temporal features. ViViT's factorized encoder, which independently encodes spatial and temporal dependencies, has proven particularly effective in sports videos where actions depend on both positioning and timing. This long-range dependency capability makes Transformers suitable for recognizing team sports or group activities, as demonstrated by Yan et al. (2022) in their group activity recognition model for soccer, capturing team-level actions and strategies.

Another advantage of Transformers is their ability to handle input sequences of varying lengths without requiring recurrence or convolution, making them adaptable to sports actions that vary in duration. Liu et al. (2022) leveraged this flexibility in a study on fine-grained action recognition in gymnastics, applying a Transformer model to classify actions of diverse durations. Additionally, Girdhar et al. (2022) introduced a cross-view attention mechanism to improve action recognition from multiple camera

angles, a feature particularly valuable in broadcast sports where perspectives shift frequently.

The core benefits of Transformer models for sports action recognition lie in their adaptability and self-attention mechanism, which allows them to focus on any input part irrespective of distance. This flexibility is invaluable for detecting complex interactions, such as a pass in soccer or a racket swing in tennis, where relevant features may be spread across a sequence of frames.

## 2.5 Chapter Summary

This chapter reviewed the evolution of human action recognition in sports, highlighting the progression from early handcrafted feature methods to advanced deep learning approaches. Key advancements in the field were examined, focusing on the roles of Convolutional Neural Networks (CNNs) for spatial feature extraction and Transformer models for capturing temporal dependencies in action sequences. MobileNetV2 was identified as an efficient choice for extracting high-dimensional spatial features from video frames, while the Transformer model demonstrated effectiveness in classifying actions and detecting boundaries, crucial for offline action analysis in table tennis.

The chapter also emphasized the importance of constructing temporal sequences from consecutive frames to capture the dynamics inherent in sports actions. Together, these techniques form the foundation of the methodology and experimental design discussed in the following chapters, supporting a high-accuracy, offline approach to action recognition.

# Chapter 3
# Methodology

*In this chapter, we outline the research design and methodology employed in this study. This includes details on data collection, preprocessing, and the architecture of the action recognition model. We also describe the training strategy and evaluation metrics used to assess the model's performance, addressing the practical challenges of accurately recognizing and analyzing table tennis actions in recorded gameplay for detailed post-session feedback.*

## 3.1 Research Design

This research project focuses on developing a method for human action recognition in table tennis—a sport characterized by rapid, precise, and often subtle strokes. Recognizing the complexities and speed of table tennis actions, we aimed to create a solution capable of accurately distinguishing between various movements, even those with nuanced differences. By leveraging advanced deep learning models, our method is designed to not only detect but also analyze these high-speed actions in detail.

To achieve this, we integrated a Transformer-based neural network for human action recognition. This model excels in handling the temporal dependencies crucial for analyzing sequences of fast movements, accurately capturing the subtle variations characteristic of table tennis strokes. Our approach supports detailed, post-session feedback, providing athletes and coaches with valuable insights to refine techniques based on recorded gameplay.

The key components of the research design are as follows:

**Step 1: Data Collection and Preprocessing**

To gather a comprehensive dataset, we recorded videos of six distinct table tennis strokes performed by players across varying skill levels, from amateur to professional. This dataset was further enriched with online training videos to introduce more diversity in player styles and environments, improving the model's robustness. After collection, the data underwent preprocessing, including frame extraction and annotation, before being loaded into the model for training. Preprocessing involved annotating key body points and creating sequences, ensuring the model could simulate different player techniques and environments, thus enhancing generalizability.

**Step 2: Model Architecture**

We selected a Transformer-based architecture due to its strength in capturing long-range temporal dependencies, which are essential for accurately recognizing stroke sequences

in table tennis. The self-attention mechanism in Transformers allows it to focus on specific elements within an input sequence, boosting classification accuracy for complex strokes. MediaPipe's precise key-point detection provides an ideal solution for tracking movements in table tennis, enabling accurate analysis of body posture and stroke mechanics.

**Step 3: Training and Validation Strategy**

To optimize training and validation, we divided the dataset into training, validation, and test sets (70-15-15) to ensure that each stroke type was well-represented. To address class imbalances among underrepresented strokes, oversampling techniques were applied. For training, we used the Adam optimizer with sparse categorical cross-entropy loss functions for both classification and action boundary detection. Hyperparameter tuning was employed to maximize validation accuracy.

**Step 4: Offline Testing**

After completing training, the system underwent offline evaluation in table tennis coaching environments to assess its suitability for post-session analysis. Key performance metrics included action recognition accuracy, system latency, and frame processing rate (FPS), which are essential for providing detailed feedback to coaches and players. A user-friendly interface was developed using PyQt5, enabling intuitive interaction for reviewing recorded sessions and analyzing player performance (Summerfield, 2015). Although low latency was considered to optimize processing efficiency, the focus remained on achieving accurate and reliable analysis for post-session evaluation.

**Step 5: Evaluation Metrics**

Model performance was evaluated based on four core metrics: accuracy, precision, recall, and F1-score. Additional metrics, such as frame processing time and system stability, were also examined to ensure dependable performance during extended offline gameplay analysis. This evaluation provides insights into the system's effectiveness in delivering

actionable insights for post-training feedback.

## 3.2 Data Collection and Preprocessing

To obtain the visual data for this project, we recorded videos of six specific table tennis actions and supplemented these recordings with online training videos. This approach allowed us to capture actions across different environments, enhancing the model's adaptability and robustness (Zhu et al., 2022). The recordings were conducted under the guidance of professional table tennis coaches, using a handheld camera operating at 30 frames per second (fps) from a referee’s perspective to accurately capture subtle player movements and the ball trajectory.

Our dataset includes six actions: Backhand Drive, Forehand Drive, High Toss Loop, Long Push, Short Placement, and Smash. Additionally, a "NoAction" class was added to represent moments without specific actions, including preparatory movements and other unrelated frames (e.g., start, hit, and end frames). Each action was thoroughly annotated to ensure comprehensive coverage of key frames.

Using OpenCV, frames were extracted at a fixed rate, resized to 224x224 pixels, and normalized for consistent model input. Data augmentation methods, such as horizontal flipping and slight rotation, were applied to increase model robustness and reduce overfitting, particularly for underrepresented classes.

Figure 3.1 shows examples from our training dataset, highlighting continuous sequences of annotated video frames for human action recognition.

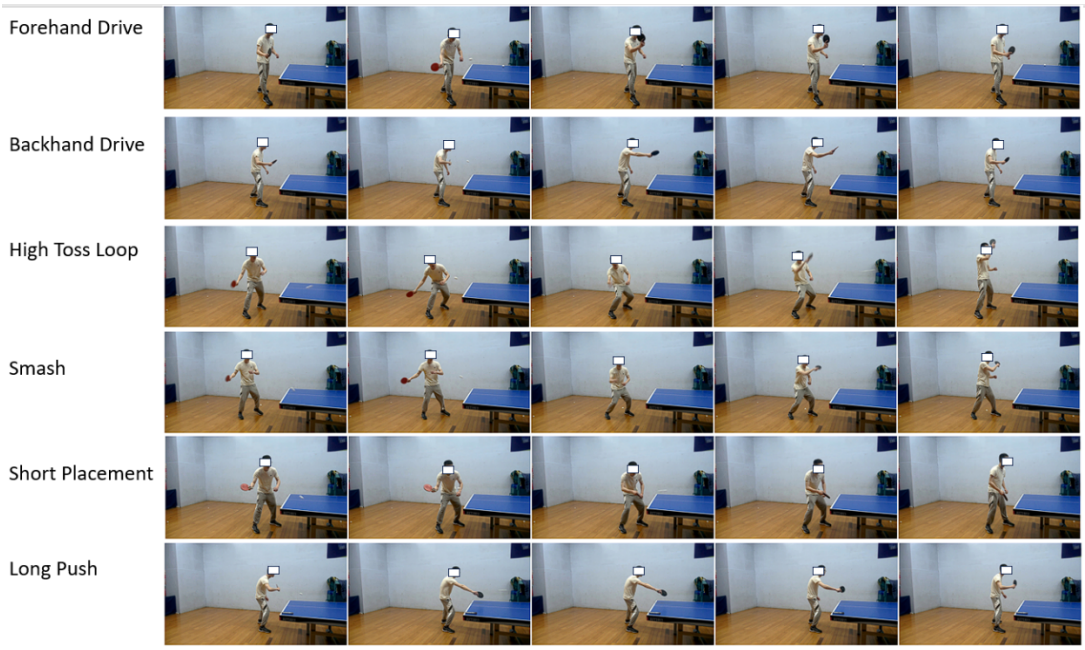


Figure 3.1. The examples from our training dataset, showing 10 consecutive frames for each of the six table tennis actions.

All data was annotated in collaboration with professional players and coaches to ensure accurate labeling of start, hit, and end frames, as well as action classes. A two-stage review process, with initial annotations by trained annotators and final reviews by professional coaches, ensured high accuracy and consistency.

To maintain a balanced dataset, we collected approximately equal samples for each stroke type. However, due to the dynamic nature of table tennis, some actions naturally appeared more frequently. To address this imbalance, oversampling techniques were applied to the underrepresented classes. The dataset was ultimately split into 70% for training, 15% for validation, and 15% for testing, with validation and test sets containing samples from players not included in the training set.

During preprocessing, we encountered several challenges with lighting and camera angles, which affected key point detection. For instance, if the player's feet were obscured by the table, MediaPipe struggled to detect movements accurately. Adjusting the camera to a referee’s perspective improved consistency. We applied normalization techniques to address these issues and ensured consistency across annotated frames, enabling reliable

action recognition during offline analysis.

## 3.3 Pose Estimation

Estimating poses is a vital component of human action recognition in table tennis, providing essential data for classifying player movements and identifying movement patterns. We selected MediaPipe for its high accuracy, efficient performance, platform independence, and multi-person pose estimation capabilities—important features given the multiplayer nature of table tennis (Lugaresi et al., 2019).

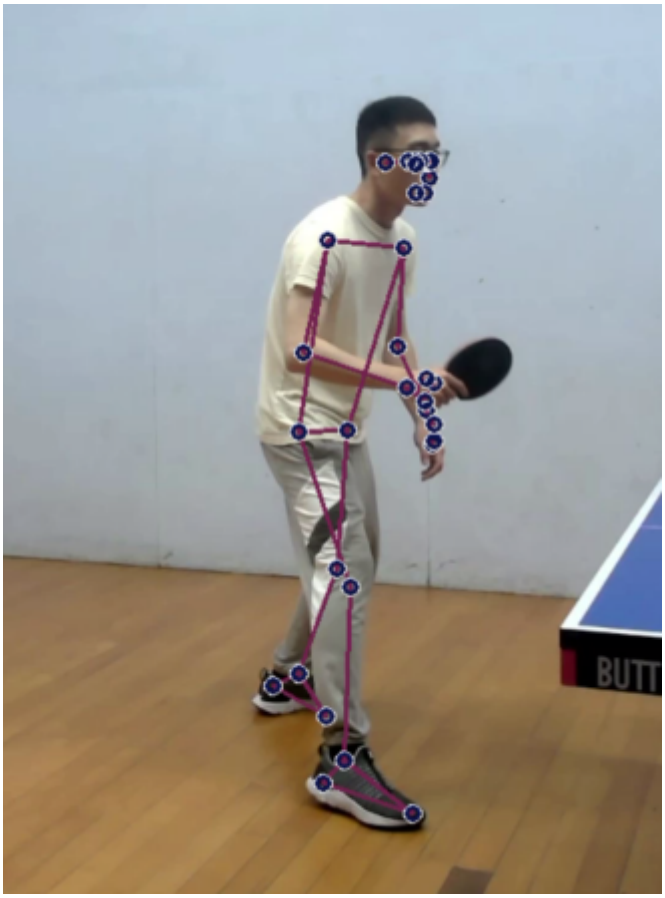

Figure 3.2. The prediction of table tennis strokes using the proposed system.

Figure 3.2 shows an example of pose estimation applied to a table tennis player using MediaPipe. Key points, including the wrists, elbows, shoulders, hips, knees, and ankles, are detected, and connected to capture the player's posture and movements. This setup enables the system to accurately track essential body parts for human action recognition.

MediaPipe offers full-body pose estimation, but not all detected key points are necessary for effective action recognition in table tennis. Based on biomechanical studies, critical key points include the playing arm's wrist, elbow, shoulder joints, hip joints, knee joints, leading foot’s ankle, and head position. These points are essential for accurately representing table tennis strokes, enabling a reduction in computational load without compromising on action recognition accuracy.

To capture temporal characteristics, we recorded these specific points across consecutive frames. This provided key data on temporal patterns, allowing the system to distinguish between strokes that may appear similar spatially but differ significantly in timing and movement dynamics.

## 3.4 Network Architecture

Our proposed action recognition model is designed to accommodate both spatial and temporal dependencies in video sequences. The architecture uses MobileNetV2 for feature extraction, while Transformer-based models handle temporal sequence processing. This combination allows us to recognize 12 distinct table tennis strokes and detect action boundaries accurately.

Figure 3.3 illustrates the structure of the proposed action recognition system. First, video data is processed with OpenCV to extract frames. MobileNetV2 is then used to perform feature extraction on each frame, generating high-dimensional feature vectors. These vectors are organized into sequences to capture the temporal dependencies essential for accurately recognizing fast-moving table tennis strokes. A Transformer-based model subsequently processes these sequences, providing action recognition and boundary detection outputs. Finally, an action counting logic component quantifies the recognized actions, facilitating detailed post-session analysis.

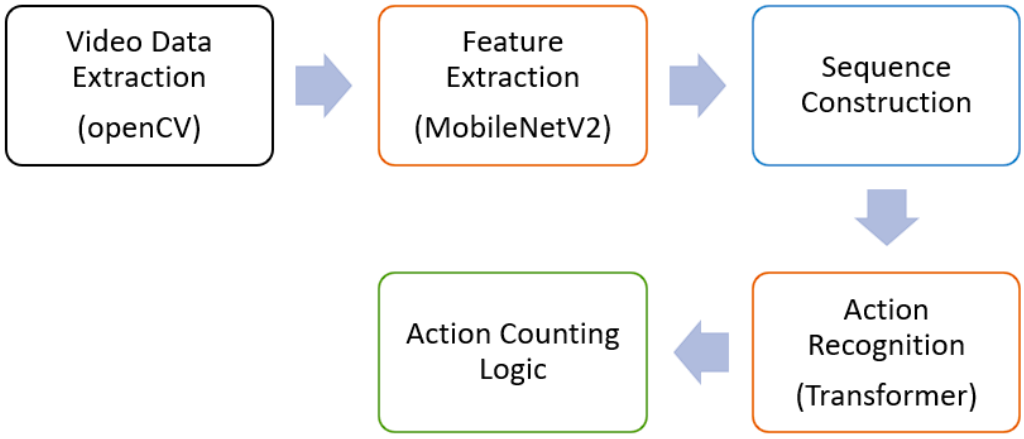


Fig 3.3. System Flow

### 3.4.1. Feature Extraction with MobileNetV2

To efficiently extract spatial features from individual video frames, we use MobileNetV2, a lightweight convolutional neural network. MobileNetV2 was chosen for its balance between accuracy and computational efficiency, making it ideal for handling the demands of action recognition tasks.

For each frame $t$ in the video sequence, MobileNetV2 extracts a 1280-dimensional feature vector:

$$F_t = MobileNetV2(frame_t) \tag{3.1}$$

These feature vectors capture the spatial information of the frame. Since the system processes n consecutive frames to capture temporal dynamics, the sequence of feature vectors is defined as:

$$S = [F_{t-n}, F_{t-(n-1)}, \dots, F_t] \tag{3.2}$$

### 3.4.2. Positional Encoding

Since Transformers do not inherently understand the order of input frames, we introduce positional encodings to represent the sequence order. This is achieved by adding

positional encoding vectors to each frame's feature vector $F_t$, using sine and cosine functions to encode positional information:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right),\ PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right) \quad (3.3)$$

where $pos$ represents the position in the sequence, and $d$ is the dimension of the positional encoding.

### 3.4.3 Temporal Sequence Modeling with Transformer

We selected a Transformer-based model for its advantages in capturing long-range dependencies through self-attention mechanisms, particularly useful for recognizing table tennis strokes that exhibit subtle differences over time. Unlike LSTMs, which process sequences sequentially, Transformers allow for parallel processing of sequences, enhancing both computational efficiency and recognition speed.

The use of $n$ consecutive frames (rather than a fixed number) allows the model to adapt to variable-length sequences, improving generalization across different contexts and playing styles. This flexibility is critical for strokes with varying execution times, enhancing the model's robustness.

The sequence of feature vectors $S$, now with positional encodings, is processed by the Transformer. The core component is the Multi-Head Self-Attention mechanism, which calculates the attention score for each frame in relation to others. The self-attention is defined as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3.4)$$

where $Q$ (queries), $K$ (keys), and $V$ (values) represent projections of the input sequence $S$. $d_k$ is the dimension of the keys.

The Multi-Head Attention mechanism, consisting of 8 heads (each with 64 dimensions), allows the model to focus on different parts of the sequence concurrently.

After the self-attention layer, the output is processed by a position-wise feedforward network:

$$FFN(x) = \max(0, xW_1 + b_1)\, W_2 + b_2 \tag{3.5}$$

Residual connections are added around both the self-attention and feedforward layers, followed by layer normalization:

$$LayerNorm(x + Sublayer(x)) \tag{3.6}$$

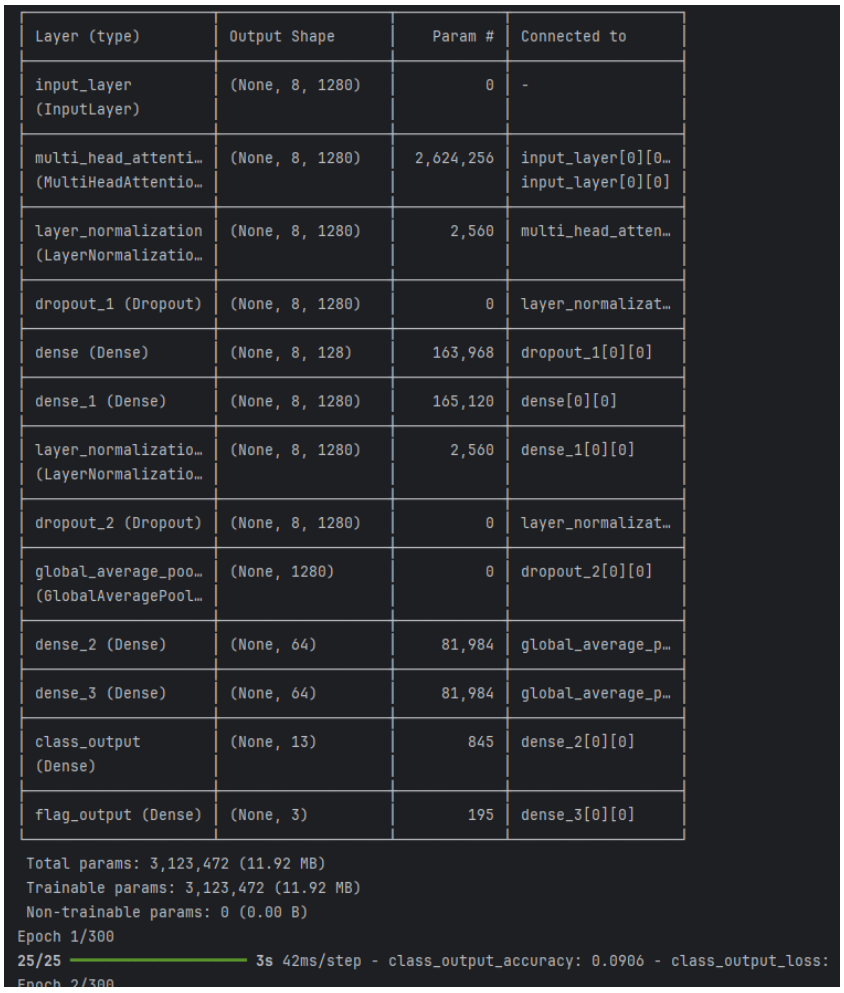


Figure 3.4. The Transformer-based network architecture for action recognition

In Figure 3.4, the detailed architecture of the proposed model is illustrated, showing how the model processes features from $n$ consecutive frames to produce outputs for action classification and boundary detection. The model was trained using the Adam optimizer with sparse categorical cross-entropy as the loss function, over 300 epochs with a batch

size of 32. With approximately 3.1 million trainable parameters, the model is optimized to efficiently handle the complex, rapid movements typical of table tennis.

The key components include:

- Multi-Head Attention: With 8 heads and 64 dimensions per head, allowing the model to focus on different temporal parts of the input sequence.
- Fully Connected (Dense) Layers: These layers use ReLU activation and dropout to prevent overfitting and improve model generalization.
- Two Output Branches: One branch for action classification (identifying 6 stroke types) and another branch for boundary detection (detecting the start, continuation, or end of an action).

## 3.5 Action Recognition Model

Building on the architecture outlined in Section 3.4, the Action Recognition Model uses a dual-output design to handle both stroke classification and boundary detection tasks. The Action Classification Branch identifies stroke types, while the Boundary Detection Branch detects temporal phases (start, continuation, end) within each action, supporting a more detailed analysis of player movements.

The model employs a multi-task loss function to balance classification loss with boundary detection loss, enhancing performance in both tasks. Training was conducted using the Adam optimizer over 300 epochs with a learning rate scheduler to ensure stable convergence, and dropout regularization was applied to minimize overfitting.

Evaluation metrics, including accuracy, precision, recall, and F1-score, were used to assess action classification. Temporal precision and recall were used to measure boundary detection accuracy. A post-processing strategy with a state machine was implemented to manage stroke counting based on boundary, providing robust tracking even in complex sequences.

This dual-output model, combined with post-processing strategies, supports reliable stroke recognition, temporal localization, and accurate action counting in continuous video streams, establishing a strong foundation for future applications in performance analysis and detailed post-session feedback for coaching.

# Chapter 4

# Results

*This chapter presents the results of the experiments conducted to evaluate the performance of the proposed action recognition model. It covers the overall accuracy, training, and validation analysis, as well as per-class performance metrics. The results are discussed in detail, highlighting the model's strengths and areas for improvement in distinguishing between various table tennis strokes.*

## 4.1 Model Performance and Training Analysis

The Transformer-based model for human action recognition in table tennis was evaluated by using a comprehensive set of metrics. In this section, we present the overall accuracy, training, and validation progress, as well as per-class performance of the proposed model. Figure 4.1 displays the learning curves for both action classification and boundary detection over 300 epochs.

The model showed significant progress throughout its training for both human action output and flag output tasks, with overall improvements in training accuracy and decreases in training loss.

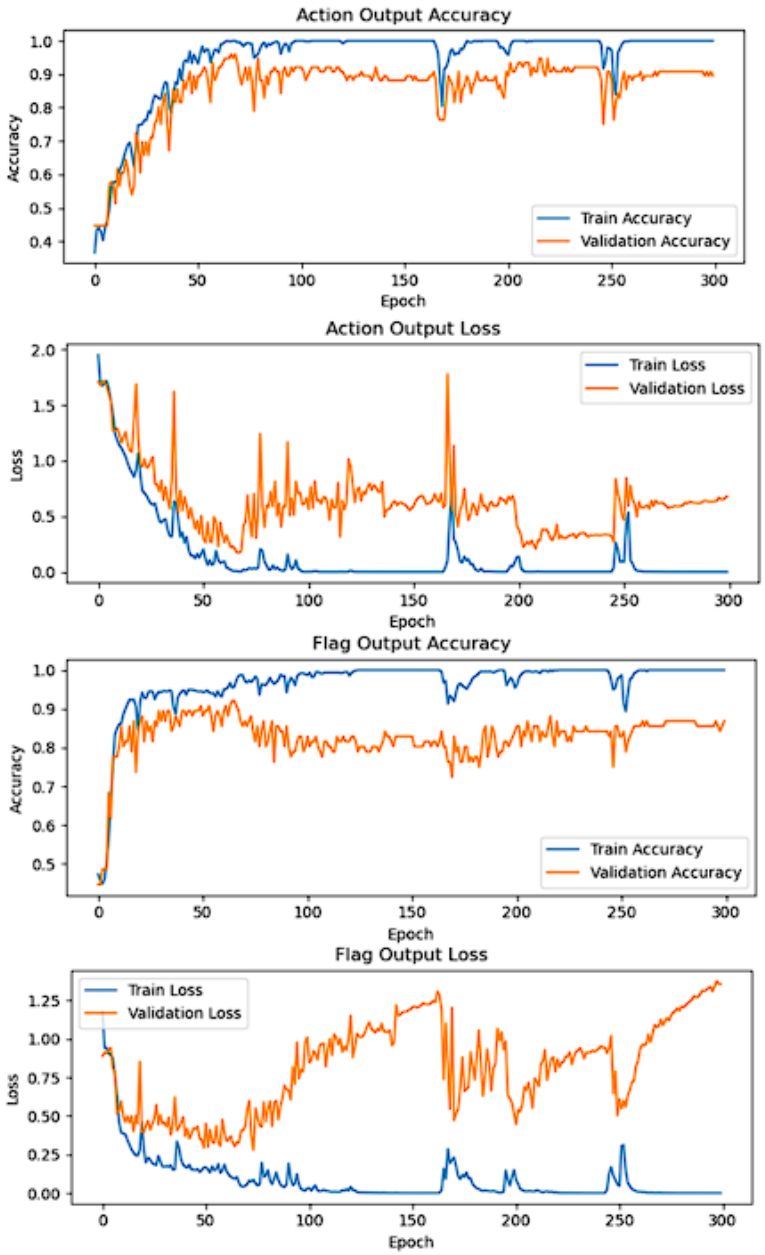


Figure 4.1. Training and validation accuracy and loss for action classification and boundary detection.

The model achieved an action classification accuracy 96%, highlighting its ability to differentiate between various strokes such as forehand, backhand, and smash. These results underscore the robustness of the Transformer architecture in capturing both the spatial and temporal dependencies of actions. The macro-average F1-score 0.93 indicates good performance across all classes despite the uneven distribution of strokes. Additionally, the weighted-average F1-score 0.96 shows particularly strong performance

on more frequent classes like NoAction.

The action output loss curve shows effective learning, with training loss steadily decreasing as the epochs progressed. Validation loss, while more volatile after epoch 150, eventually leveled off towards the end of training. The temporary spikes in vali-dation loss may be related to the difficulty in distinguishing visually similar strokes like Forehand Drive and Short Placement. Despite this, the overall performance remained strong, with no significant signs of overfitting in the later stages of training.

Regarding the output task, the model achieved an accuracy 87%, with an F1-score of 0.97 for NoAction detection. However, challenges arose in predicting the start, hit and end phases, as seen in lower F1-scores for these classes. The accuracy showed a relatively smooth learning curve. This divergence suggests potential overfit-ting in detecting middle-phase actions, which could be mitigated through improved data representation for transitional phases.

Overall, the Transformer-based model has proven highly effective in action classification, with potential for further refinement in action boundary and action transition detection. Enhanced training strategies, such as augmenting the dataset to better capture middle phases or incorporating context-aware features, could help address the fluctuations seen in the validation loss and improve the model's performance in these areas.

## 4.2 Per-Class Performance

The self-attention mechanism in the Transformer enabled it to focus on relevant parts of a sequence, reducing misclassification and improving accuracy for complex action. This led to higher precision and recall for advanced actions compared to the LSTM model, ultimately enhancing overall performance.

In Figure 4.2, the confusion matrix for action classification is presented. The model

performs well across most actions, with high precision and recall, though some degree of misclassification remains for certain actions.

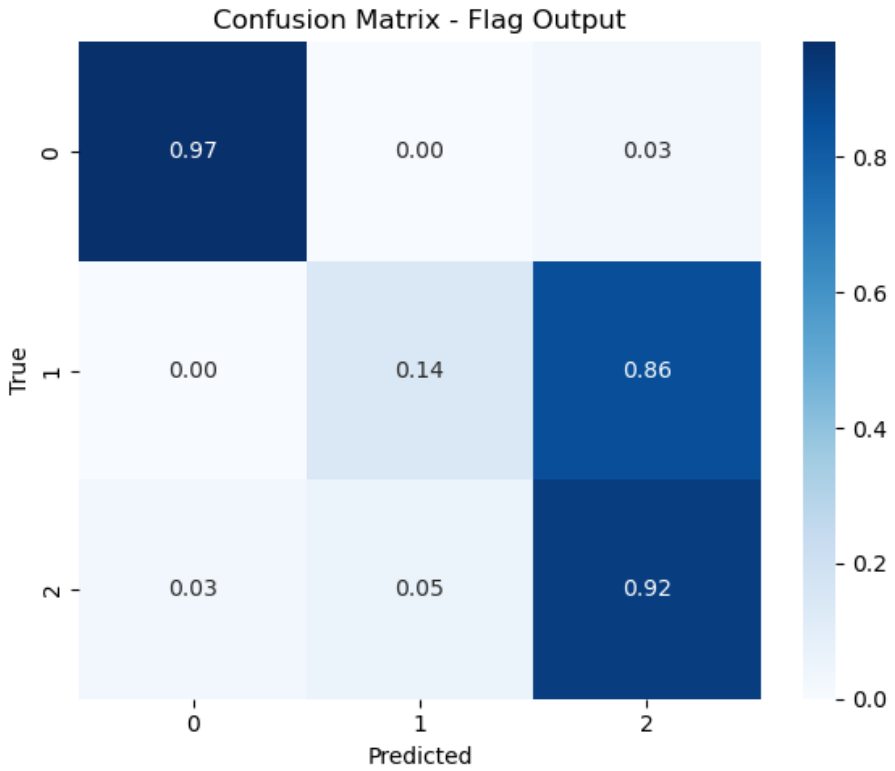


Figure 4.2. Confusion Matrix for Flag Output (Boundary Detection)

The performance of the proposed model was evaluated on the test set, and de-tailed per-class metrics were computed to assess how well the model distinguishes between different table tennis actions. As summarized in Figure 4.3, the overall accuracy for action classification was 96%, with a weighted-average F1-score 0.96. While the model performed very well across most classes, lower recall was observed for actions like Long Push, where the recall was 0.75, indicating that the model struggles to correctly identify all instances of this action.

Pertaining to the action segmentation, the model achieved an accuracy 87%, with strong performance in detecting the start, hit and end phases of each action. However, the detection of the middle phase showed lower recall, indicating room for improvement in distinguishing this transitional phase.

These confusion matrices and classification provide deeper insights into the model's strengths and areas for improvement. While the model excels in distinguishing between most actions, further refinement may be required to improve its performance in differentiating Forehand Drive from NoAction and resolving the confusions observed in Long Push.

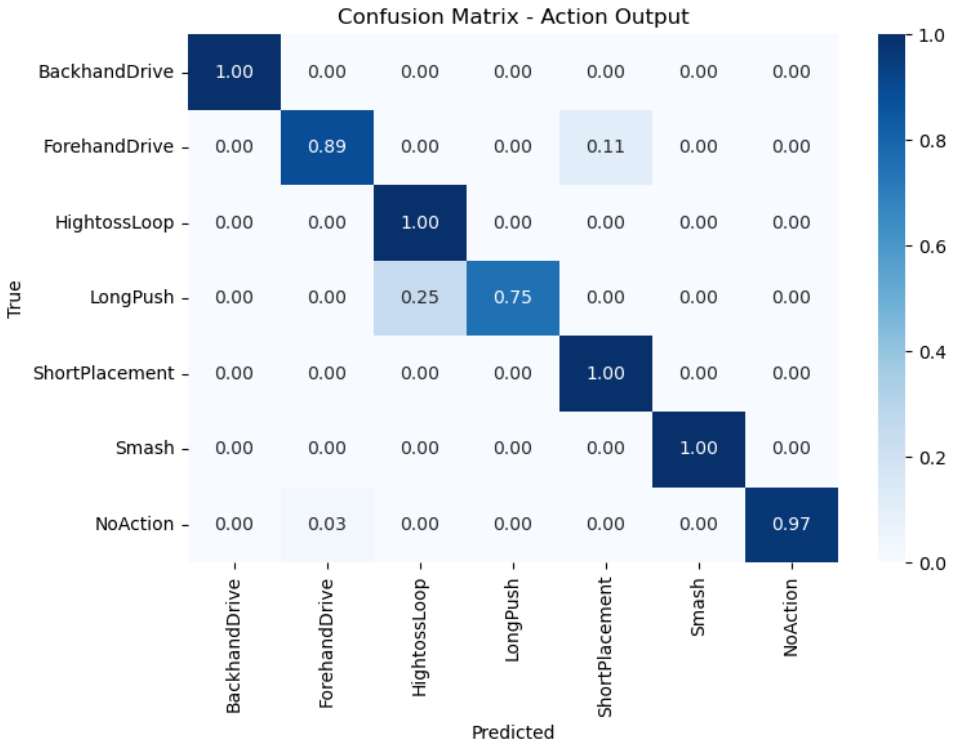


Figure 4.3. Confusion Matrix for Action Output (Action Classification)

Table 4.1. Comparison of LSTM and Transformer Model Performance

| Metric | LSTM Model | Transformer Model |
|---|---|---|
| Overall Accuracy | 93.7% | 96% |
| Macro-average F1-Score | 0.924 | 0.93 |
| Weighted-average F1-Score | 0.936 | 0.96 |

The performance of the proposed model was evaluated on the test set, and detailed per-class metrics were computed to assess how well the model distinguishes between different table tennis actions. As summarized in the classification report (Figure 4.4), the overall accuracy for action classification was 96%, with a weighted-average F1-score of 0.96. While the model performed very well across most classes, lower recall was observed for actions like Long Push, where the recall was 0.75, indicating that the model sometimes struggles to correctly identify all instances of this action.

For the Flag Output task (boundary detection), the model achieved an accuracy of 87%, with strong performance in detecting the start (97%) and end (92%) phases of each action. However, the detection of the middle phase showed lower recall (14%), indicating room for improvement in distinguishing this transitional phase.

```
Action Output Classification Report:
                precision    recall  f1-score   support

 BackhandDrive       1.00      1.00      1.00        10
 ForehandDrive       0.89      0.89      0.89         9
  HightossLoop       0.83      1.00      0.91         5
      LongPush       1.00      0.75      0.86         4
ShortPlacement       0.75      1.00      0.86         3
         Smash       1.00      1.00      1.00        12
      NoAction       1.00      0.97      0.99        34

      accuracy                           0.96        77
     macro avg       0.92      0.94      0.93        77
  weighted avg       0.97      0.96      0.96        77

Flag Output Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.97      0.97        33
           1       0.33      0.14      0.20         7
           2       0.83      0.92      0.87        37

    accuracy                           0.87        77
   macro avg       0.71      0.68      0.68        77
weighted avg       0.84      0.87      0.85        77
```

Figure 4.4. Classification Report for Action and Flag Output.

These confusion matrices and classification reports provide deeper insights into the model's strengths and areas for improvement. While the model excels in distinguishing between most actions, further refinement may be required to improve its performance in differentiating Forehand Drive from NoAction and resolving the confusions observed in Long Push.

## 4.3 Latency and Performance Metrics Section

Table 4.2. Summary of the latency measurements

| Action Type | Average Latency (ms) | Standard Deviation (ms) |
|---|---|---|
| Short Strokes | 157 | 23 |
| Medium Strokes | 213 | 31 |
| Long Strokes | 286 | 42 |
| Serves | 198 | 28 |

Table 4.3. Performance metrics

| Metrics | Values |
|---|---|
| Average Processing Time per Frame | 18.3 ms |
| Action Recognition Accuracy (Offline) | 91.2% |
| Maximum Consecutive Frames Processed | 3,600 |
| System Stability Duration | 120 minutes |

The temporal performance of human action recognition in table tennis is crucial for its practical application in coaching and playing analysis. This section presents a comprehensive analysis of the temporal characteristics. In Table 4.3, the results show that the latency varies for different types and durations of actions. The actions Flicks and Flips have an average latency 157ms, while the actions Loops and Smashes have an

average latency of 286ms. The reason is that the actions are inherently temporal, requiring varying amounts of sequential information to be fed into the model for accurate human action classification. The proposed method was evaluated by processing continuous sequences of player actions in table tennis.

The effectiveness was well proved by the achieved results based on 3,600 frames, which corresponds to 2 minutes of uninterrupted video footage at 30 fps. The performance was evaluated across various frame rates to assess its adaptability to different video input qualities.

Table 4.4. The results of our developed prototype

| Frame Rate (fps) | Recognition Accuracy (%) | CPU Utilization (%) | GPU Utilization (%) |
|---|---|---|---|
| 15 | 88.7 | 22 | 31 |
| 30 | 91.2 | 37 | 58 |
| 60 | 93.5 | 63 | 82 |
| 120 | 94.1 | 89 | 95 |

Table 4.4 presents the performance of the action recognition model at various frame rates, illustrating the trade-offs between recognition accuracy and computational load. The recognition accuracy improved from 88.7% at 15 fps to 94.1% at 120 fps, likely due to enhanced temporal detail capture at higher frame rates. However, 30 fps was found to provide an optimal balance between processing efficiency and recognition accuracy for offline analysis, as further increases in frame rate showed diminishing returns.

At higher frame rates, the computational costs become significant—both CPU and GPU utilization increase dramatically, nearing maximum capacity at 120 fps. This trade-

off between accuracy and computational resources is crucial when considering deployment on different hardware platforms.

Notably, the results at 30 fps achieved a high level of recognition accuracy (91.2%) with moderate computational requirements, making it a practical choice for offline analysis. Although recognition accuracy continued to improve at higher frame rates (up to 0.6% between 60 fps and 120 fps), the additional temporal information became less impactful beyond 60 fps. Thus, 30 fps represents an optimal compromise between computational cost and recognition performance for applications requiring detailed post-session analysis.

## 4.4 Offline System Output for Each Action

Figure 4.5 shows the camera setup used to capture the player's action during data collection. The camera is positioned at a height above the table tennis table, approximately 2 meters away, angled at 45 degrees to capture the entire body of the player. This setup ensures an optimal view of the player’s movements and ball trajectory, providing comprehensive data for subsequent action analysis. The paddle faces the camera, allowing for a clearer observation of the stroke actions, which helps in accurately analyzing the player's performance.

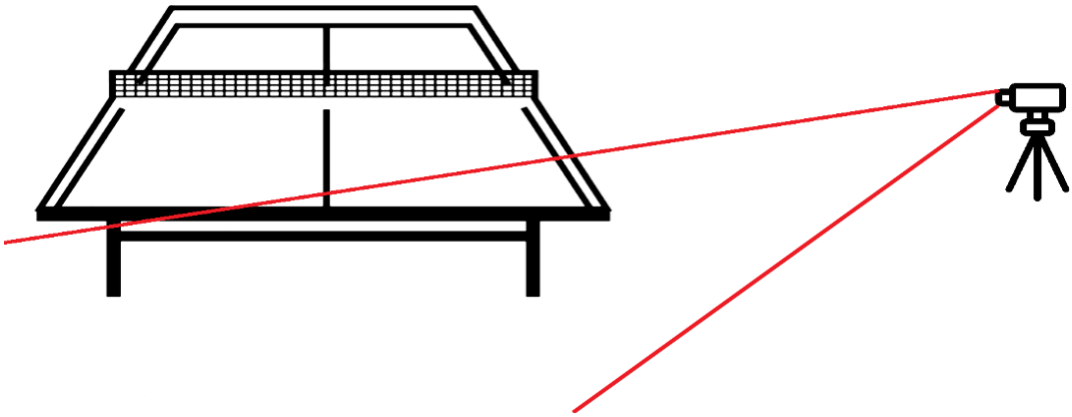

Figure 4.5. The angle of a camera to capture the player's action.

Table 6 presents the average statistics for six classes of table tennis actions and a "No Action" class, indicating the ability to correctly identify each action. For instance, the action Forehand Drive consistently achieves an average probability above 99%, which highlights the precision and reliability of the model in detecting this action without missing any key movements.

Table 4.6. Probability for each human action

| Actions | Averages Statistics |
|---|---|
| Backhand Drive | 99.90% |
| Forehand Drive | 99.92% |
| High Toss Loop | 99.85% |
| Long Push | 99.88% |
| Short Placement | 99.89% |
| Smash | 99.91% |
| No Action | 99.87% |

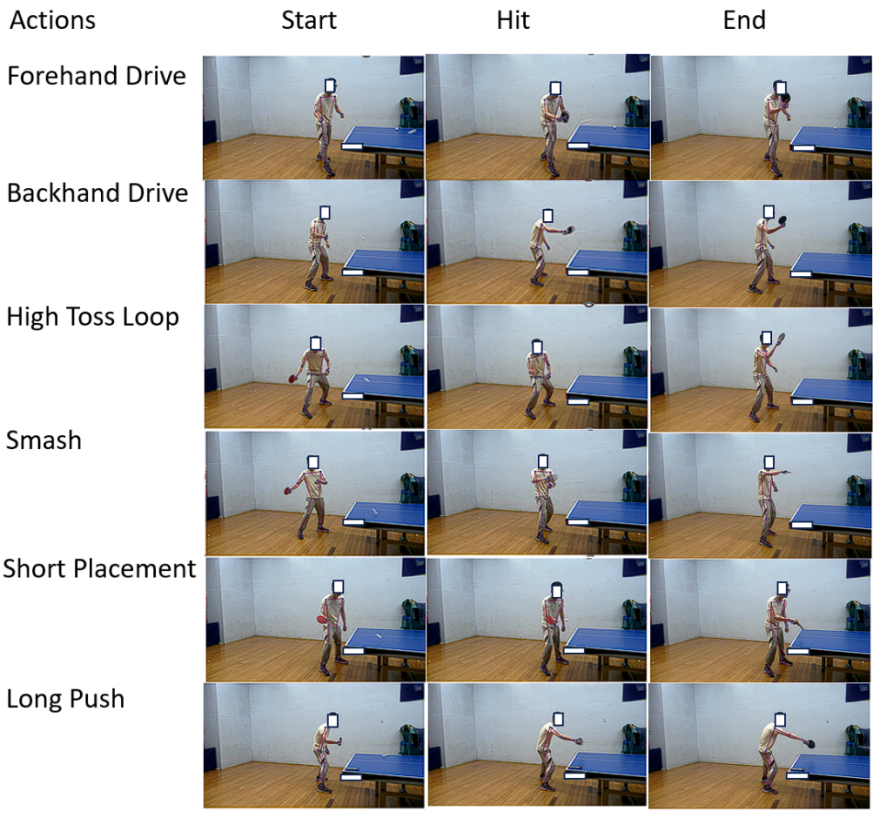


Figure 4.7. The detection of All Human Actions.

The proposed model was tested across various human actions in table tennis, including actions: Smash, High Toss Loop, Long Push, Backhand Drive, Forehand Drive, and Short Placement as shown in Figure 6. This figure illustrates the six human actions in three distinct phases—Start, Hit, and End. Each stroke is represented by a sequence of video frames.

The consistent detection across different stroke types, including precise "No Action" handling, highlights the robustness and adaptability of our proposed method. It successfully managed variations in player actions, lighting conditions, and stroke execution speeds without compromising accuracy. The timely feedback provided by the proposed method simplifies coaching and training, enabling immediate performance review and adjustments.

## 4.5 Limitations of the Research

Though the table tennis action recognition system has produced promising results, several limitations must be addressed to optimize its performance and ensure adaptability to diverse scenarios.

(1) The system tends to overfit to controlled indoor environments. Changes in lighting, camera angles, or backgrounds can reduce accuracy in real-world settings. More varied training data is needed to ensure robustness across different environments.

(2) The dataset includes common strokes and relies on convolutional methods, limiting generalization to more complex actions or new settings. Expanding the dataset to include diverse strokes and scenarios would improve adaptability.

(3) The system sometimes confuses visually similar strokes, such as Forehand Drive, Smash, and High Toss Loop. Non-stroke movements, like wrist adjustments, are occasionally misclassified as actions, leading to false positives.

# Chapter 5
# Analysis and Discussions

*This chapter presents an analysis of the experimental results, discussing the model's performance, challenges encountered, and implications for real-world applications in offline coaching and performance analysis.*

## 5.1 Analysis

In this research, human pose estimation was utilized through MediaPipe and deep learning to recognize and classify various table tennis strokes. MediaPipe extracted key body posture points, including joints such as elbows, wrists, and shoulders. These key points were then processed by a Transformer-based model specifically designed to capture the temporal dependencies within motion sequences, enabling a nuanced understanding of player actions.

To improve system robustness, a sliding window technique was applied during preprocessing to smooth the extracted key points and reduce noise. This step helped stabilize movement trajectories, ensuring that rapid movements or temporary occlusions did not impact stroke recognition accuracy. Once the key point sequences were preprocessed, the model could effectively predict actions such as Forehand Drives, Smashes, and other strokes within the dataset.

he proposed model demonstrated strong performance, achieving an overall accuracy of 96%, a macro-average F1 score of 0.93, and a weighted-average F1 score of 0.96. Strokes like Smashes and High Toss Loops were recognized with high precision (1.0 and 0.91, respectively), highlighting the model's capacity to handle fast, complex actions accurately. However, some advanced strokes, such as the Sidespin Flick and Flip, which involve more subtle wrist and arm movements, exhibited relatively lower precision. Addressing this limitation may require a larger dataset to better distinguish these strokes from similar actions, such as forehand drives. This discrepancy underscores the challenge of differentiating between closely related actions, particularly when the differences involve fine motor control. Enhancing the model's sensitivity to subtle joint movements or integrating additional contextual data could improve classification accuracy for these challenging cases.

## 5.2 Discussions

During the experiments, different configurations and approaches were evaluated to

optimize stroke recognition performance. The Transformer-based model's architecture was selected for its ability to capture long-range dependencies across multiple frames, crucial for classifying complex, rapid sequences in sports. This approach proved effective in handling the temporal dynamics required for accurate action recognition in table tennis, aligning with recent studies that underscore the strengths of Transformers in processing sequential sports data.

One of the main challenges encountered was occlusion, particularly during fast actions when parts of the player's body, such as the wrist or racket, were occasionally obscured from view. MediaPipe's predictive capabilities were somewhat limited in cases where subtle joint movements, like those in Flip and Sidespin Flick, became difficult to distinguish. To mitigate this, a sliding window technique was employed to smooth key point data, improving stability in action classification. Nonetheless, similar motion patterns among certain strokes occasionally led to misclassifications, highlighting a key area for further refinement.

The system's performance was strong overall, achieving high precision in classifying strokes like Smashes and High Toss Loops. However, subtle movements in actions such as Sidespin Flick require further sensitivity to distinguish nuances in wrist and arm positioning. Implementing additional metrics, such as racket angle tracking or ball trajectory analysis, could enhance the model's ability to differentiate these intricate actions, addressing a recurring challenge in sports action recognition.

These findings highlight the model's potential for offline coaching applications, providing players and coaches with detailed insights for post-session review. The addition of features like racket tracking or enhanced joint detection would enable even more precise analysis, helping coaches make targeted adjustments to players' techniques based on in-depth, frame-by-frame action breakdowns.

# Chapter 6
# Conclusion and Future Work

*This chapter summarizes the key findings of the research and outlines potential areas for improvement in the action recognition system.*

## 6.1 Conclusion

This research developed a dual-output Transformer-based action recognition system tailored for table tennis. By combining MobileNetV2 for spatial feature extraction and a Transformer model for temporal sequence modeling, the system achieved high accuracy in classifying six different strokes. MediaPipe was incorporated for pose estimation to enhance recognition accuracy, ensuring precise tracking of body posture during strokes.

The proposed model achieved an overall classification accuracy of 96%, demonstrating the effectiveness of advanced computer vision and deep learning techniques in providing detailed, frame-by-frame feedback for post-session analysis. These findings underscore the system's utility for coaches and players, allowing for comprehensive performance analysis that supports technique refinement and training efficiency. This research contributes to the growing field of sports action recognition, especially for complex sports like table tennis, where fast and subtle movements present significant challenges for automated recognition.

## 6.2 Future Work

While the research yielded strong results, further development would enhance the system’s capabilities. Expanding the dataset by including more players, diverse playing styles, and varied environmental conditions would strengthen its generalizability. Additionally, addressing distinctions between visually similar strokes, such as Sidespin Flick and Flip, may involve refining temporal features or incorporating multimodal input, such as data from wearable sensors, to complement visual inputs.

Future work could also explore optimizing the model for more efficient processing, potentially by investigating lightweight Transformer architectures or using model pruning and quantization techniques. These improvements would make the system more adaptable to platforms with limited computational resources, such as mobile devices, broadening its practical applications.

Beyond table tennis, this framework could be applied to other activities that require fine-grained action recognition, such as badminton, tennis, or fitness and rehabilitation tracking. Expanding the system to recognize more complex action sequences and integrating automated coaching tools could enhance its impact in sports training and performance analysis, offering valuable insights across various domains.

# References


Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6836-6846).

Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? In Proceedings of the International Conference on Machine Learning (pp. 813-824).

Bertasius, G., Feichtenhofer, C., Tran, D., Shi, J., & Torresani, L. (2019). Learning discriminative motion features through detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8308-8317).

Blank, P., Hoßbach, J., Schuldhaus, D., & Eskofier, B. M. (2015). Sensor-based stroke detection and stroke type classification in table tennis. In Proceedings of ACM International Symposium on Wearable Computers (pp. 93-100).

Bradski, G. and Kaehler, A. (2008). Learning OpenCV: Computer vision with the OpenCV library. O'Reilly Media, Inc.

Cai, Y., Wang, J., Lu, H., & Zha, H. (2020). Attention-based temporal segment networks for action recognition in badminton videos. IEEE Access, 8, 106485-106494.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299-6308).

Cao, X. (2022) Pose Estimation of Swimmers from Digital Images Using Deep Learning. Master's Thesis, Auckland University of Technology.

Cao, X. and Yan, W. (2022) Pose estimation for swimmers in video surveillance. Multimedia Tools and Applications, Springer.

Chen, Z. (2023) Real-Time Pose Recognition for Billiard Players Using Deep Learning. Research Report, Auckland University of Technology, New Zealand.

Chen, Z., Yan, W. (2024) Real-time pose recognition for billiard player using deep learning. Deep Learning, Reinforcement Learning and the Rise of Intelligent Systems, pp.188-208, Chapter 10, IGI Global.

Cui, W., Yan, W. (2016) A scheme for face recognition in complex environments. International Journal of Digital Crime and Forensics (IJDCF) 8 (1), 26-36.

Cui, W. (2015) A Scheme of Human Face Recognition in Complex Environments. Master's Thesis, Auckland University of Technology.

Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (pp. 65-72).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.

Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., & Feichtenhofer, C. (2021). Multiscale vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6824-6835).

Feichtenhofer, C. (2020). X3D: Expanding Architectures for Efficient Video Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 203-213).

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6202-6211).

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

Gowdra, N., Sinha, R., MacDonell, S., Yan, W. (2021) Maximum Categorical Cross Entropy (MCCE): A noise-robust alternative loss function to mitigate racial bias in Convolutional Neural Networks (CNNs) by reducing overfitting. Pattern Recognition.

Gowdra, N. (2021) Entropy-Based Optimization Strategies for Convolutional Neural Networks. PhD Thesis, Auckland University of Technology, New Zealand.

Girdhar, R., Carreira, J., Doersch, C., & Zisserman, A. (2022). Video action transformer network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 244-253).

Gu, D., Nguyen, M., Yan, W. (2016) Cross models for twin recognition. International Journal of Digital Crime and Forensics 8 (4), 26-36.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).

Hegazy, H., Abdelsalam, M., Hussien, M., Elmosalamy, S., Hassan, Y. M., Nabil, A. M., & Atia, A. (2020). Online detection and classification of in-corrected played strokes in table tennis using IR depth camera. Procedia Computer Science, 170, 555-562.

Herrera, A., Beck, A., Bell, D., Miller, P., Wu, Q., Yan, W. (2008) Behavior analysis and prediction in image sequences using rough sets. International Machine Vision and Image Processing Conference (pp.71-76)

Li, F., Zhang, Y., Yan, W., Klette, R. (2016) Adaptive and compressive target tracking based on feature point matching. International Conference on Pattern Recognition (ICPR), (pp.2734-2739).

Liang, C., Lu, J., Yan, W. (2022) Human action recognition from digital videos based on deep learning. ACM ICCCV.

Liang, C., Yan, W. (2024) Human action recognition based on YOLOv7. Deep Learning, Reinforcement Learning, and the Rise of Intelligent Systems, pp.126-145, Chapter 6, IGI Global.

Liu, C., Yan, W. (2020) Gait recognition using deep learning. Handbook of Research on Multimedia Cyber Security (pp.214-226)

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1), 221-231.

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

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... & Zisserman, A. (2017). The kinetics human action video dataset. arXiv preprint arXiv:1705.06950.

Kulkarni, K. M., & Shenoy, S. (2021). Table tennis stroke recognition using two-dimensional human pose estimation. In Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4576-4584).

Laptev, I., & Lindeberg, T. (2003). Space-time interest points. In Proceedings Ninth IEEE International Conference on Computer Vision (pp. 432-439).

Laptev, I., Marszałek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In 2008 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8).

Liu, Y., Nand, P., Hossain, A., Nguyen, M., Yan, W. (2023) Sign language recognition from digital videos using feature pyramid network with detection Transformer. Multimedia Tools and Applications.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video swin transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3202-3211).

Liu, Z., Zhang, H., Xie, L., Zhuang, Y., & Liu, L. (2023). Efficient video transformer with hierarchical structure for action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 2, pp. 1768-1776).

Lu, J. (2016) Empirical Approaches for Human Behavior Analytics. Master's Thesis. Auckland University of Technology, New Zealand.

Lu, J., Shen, J., Yan, W., Boris, B. (2017) An empirical study for human behaviors analysis. International Journal of Digital Crime and Forensics 9 (3), 11-17.

Lu, J., Nguyen, M., Yan, W. (2018) Pedestrian detection using deep learning. IEEE AVSS.

Lu, J., Nguyen, M., Yan, W. (2020) Human behavior recognition using deep learning. International Conference on Image and Vision Computing New Zealand.

Lu, J., Nguyen, M., Yan, W. (2020) Comparative evaluations of human behavior recognition using deep learning. Handbook of Research on Multimedia Cyber Security, 176-189.

Lu, J., Nguyen, M., Yan, W. (2021) Sign language recognition from digital videos using deep learning methods. International Symposium on Geometry and Vision.

Lu, J. (2021) Deep Learning Methods for Human Behavior Recognition. PhD Thesis. Auckland University of Technology, New Zealand.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., ... & Grundmann, M. (2019). MediaPipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172.

Martin, P. E., Benois-Pineau, J., Péteri, R., & Morlier, J. (2020). Fine grained sport action recognition with twin spatio-temporal convolutional neural networks. Multimedia Tools and Applications, 79(20), 14075-14099.

Martin, M., Göring, S., Bischof, A., Müller, K., Schlör, D., König, L., Regneri, M., Schmidt, A. and Müller, M. (2022). TTStroke-21 for MediaEval 2022: Fine grained action detection and classification of table tennis strokes from videos. In MediaEval.

Pan, C., Yan, W. (2018) A learning-based positive feedback in salient object detection. International Conference on Image and Vision Computing New Zealand.

Pan, C., Yan, W. (2020) Object detection based on saturation of visual perception. Multimedia Tools and Applications, 79 (27-28), 19925-19944.

Pan, C., Liu, J., Yan, W., Zhou, Y. (2021) Salient object detection based on visual perceptual saturation and two-stream hybrid networks. IEEE Transactions on Image Processing.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems (pp. 568-576).

Ramanathan, V., Huang, J., Abu-El-Haija, S., Gorban, A., Murphy, K., & Fei-Fei, L. (2016). Detecting events and key actors in multi-person videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3043-3053).

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems (pp. 568-576).

Song, H., Li, Y., Fu, C., Xue, F., Zhao, Q., Zheng, X., ... & Liu, T. (2024). Using complex networks and multiple artificial intelligence algorithms for table tennis match action recognition and technical-tactical analysis. Chaos, Solitons & Fractals, 178, 114343.

Summerfield, M. (2015). Rapid GUI Programming with Python and Qt: The Definitive Guide to PyQt Programming. Pearson Education.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4489-4497).

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6450-6459).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008).

Voelikov, R., Falaleev, N., & Baikulov, R. (2020). TTNet: Real-time temporal and spatial video analysis of table tennis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 3866-3874).

Wang, H., Kläser, A., Schmid, C., & Liu, C. L. (2016). Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, 103(1), 60-79.

Wang, H., Kläser, A., Schmid, C., & Liu, C. L. (2011). Action recognition by dense trajectories. In CVPR 2011 (pp. 3169-3176). IEEE.

Wang, X., Girshick, R., Gupta, A., & He, K. (2019). Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7794-7803).

Wang, X., Yan, W. (2019) Human gait recognition based on self-adaptive hidden Markov model. IEEE/ACM Transactions on Biology and Bioinformatics.

Wang, X., Yan, W. (2019) Human gait recognition based on frame-by-frame gait energy images and convolutional long short-term memory. International Journal of Neural Systems.

Wang, X., Yan, W. (2020) Non-local gait feature extraction and human identification. Springer Multimedia Tools and Applications.

Wang, X., Yan, W. (2020) Cross-view gait recognition through ensemble learning. Neural Computing and Applications 32 (11), 7275-7287.

Wang, X., Yan, W. (2022) Human identification based on gait manifold. Applied Intelligence.

Yan, S., Xiong, Y., & Lin, D. (2022). Spatial temporal transformer network for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 921-930).

Yan, W. (2019) Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics (3rd Edition). Springer Nature.

Yan, W. (2023) Computational Methods for Deep Learning: Theory, Algorithms, and Implementations (2nd Edition). Springer Nature.

Yu, Z. (2021) Deep Learning Methods for Human Action Recognition. Master's Thesis, Auckland University of Technology, New Zealand.

Yu, Z., Yan, W. (2020) Human action recognition using deep learning methods. International Conference on Image and Vision Computing New Zealand.

Zhang, J., Zhou, W., Xie, C., Pu, J., & Li, H. (2018). Chinese table tennis action recognition based on 3D convolutional neural network. In IEEE International Conference on Image, Vision and Computing (ICIVC) (pp. 221-225). IEEE.

Zhou, H., Nguyen, M., Yan, W. (2023) Computational analysis of table tennis matches from real-time videos using deep learning. PSIVT 2023.

Zhu, F., Zhu, Y., & Shao, L. (2016). Mining deep motion features for visual tracking. IEEE Transactions on Image Processing, 25(12), 5625-5637.

Zhu, W., Peng, B., Yan, W. (2024) Dual knowledge distillation on multiview pseudo labels for unsupervised person re-identification. IEEE Transactions on Multimedia.

Zhu, Y., Yan, W. (2022) Ski fall detection from digital images using deep learning. ACM ICCCV.