Action Recognition in Videos using Deep Learning approaches

MS (Research) Thesis

By

NEELESH GHANGHORIYA

2004101006

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

May 2024

Action Recognition in Videos using Deep Learning approaches

A THESIS

Submitted in partial fulfillment of the
requirements for the award of the degree
of
Master of Science (Research)

by

NEELESH GHANGHORIYA
2004101006

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE
May 2024

INDIAN INSTITUTE OF TECHNOLOGY INDORE

I hereby certify that the work which is being presented in the thesis entitled **Action Recognition in Videos using Deep Learning Approaches** in the partial fulfillment of the requirements for the award of the degree of **MASTER OF SCIENCE (RESEARCH)** and submitted in the **DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from July 2022 to May 2024 under the supervision of Dr. Aruna Tiwari, professor , Indian Institute of Technology Indore, India and Dr Sanjay Singh, Principal Scientist, CSIR – Central Electronics Engineering Research Institute (CSIR – CEERI), Pilani, India.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

neelesh
ghanghoriya

Signature of the student with date
NEELESH GHANGHORIYA

This is to certify that the above statement made by the candidate is correct to the best of my/our knowledge.

24.05.2024
Signature of Thesis Supervisor with date

Dr. Aruna Tiwari

संजय
Signature of Thesis Supervisor with date

Dr. Sanjay Singh

NEELESH GHANGHORIYA has successfully given his MS Research Oral Examination held on ________________________

Signature of Chairperson (OEB)
Date:

Signature of Thesis Supervisor
Date:

Signature of Convener DPGC
Date:

Signature of Head of Department
Date:

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my gratitude to people who, in one or the other way, contributed by making this time learnable, enjoyable, and bearable. At first, I would like to thank my supervisors **Dr. Aruna Tiwari** and **Dr. Sanjay Singh**, who was a constant source of inspiration during my work. With their constant guidance and research directions, this research work has been completed. Their continuous support and encouragement have motivated me to remain streamlined in my research work. I am also grateful to **Dr. Ranveer Singh**, HOD of Computer Science, for all his help and support.

I am thankful to **Dr. Ram Bilas Pachori** and **Dr. Ranveer Singh**, my research progress committee members, for taking out some valuable time to evaluate my progress all these years. Their valuable comments and suggestions helped me to improve my work at various stages.

My sincere acknowledgment and respect to **Prof. Suhas S. Joshi**, Director, Indian Institute of Technology Indore, for providing me the opportunity to explore my research capabilities at the Indian Institute of Technology Indore.

I would also like to thank **Mr. Sumeet Saurav, Mr. Rituraj**, and my batchmates for discussions and help during my thesis work.

I would like to express my heartfelt respect to my parents, my sisters, and my friends for the love, care, and support they have provided to me throughout my life.

Neelesh Ghanghoriya

To my family and friends

ABSTRACT

Given the inherent complexity of video data, action recognition in videos poses a formidable challenge in computer vision. The 3D space-time volume encompassing frame sequences contains substantial redundant information, diverting the model from acquiring a discriminative representation of the performed action class. Although 3D Convolutional Neural Networks (3D CNNs) exhibit exceptional spatio-temporal feature learning capabilities, leading to state-of-the-art action recognition performance on various large-scale benchmark video datasets, a naive 3D CNN architecture comes with drawbacks. Firstly, it demonstrates incompetence in modeling long-range dependencies due to the fixed and limited receptive field of the 3D convolutional kernel. Secondly, its demand for a substantial amount of data and extensive computational time during training arises from the number of parameters involved.

Recently, much research has focused on alleviating the limitation of 3D CNNs. Various techniques have tried to increase the 3D CNN model's depth by stacking multiple convolutional layers. Although expanding the depth has compensated for the 3D kernel's limited receptive field, it has exploded the model's parameter, making its need for training data and computation time consumption critical. In this thesis, we tackle various constraints of the 3D CNN architecture to perform action recognition in the limited availability of training data. We propose a Self-Attention Convolution Neural Network named – SAC3D, as it incorporates the 3D self-attention mechanism in the popular C3D model baseline. The 3D Self-attention mechanism guides the 3D convolutional layers of the model by providing information on the pairwise similarity of pixels that exists between them. The correlation strength of each pixel-to-pixel relationship in 3D space-time helps the model map the underlying action better and enhance its discriminative feature learning capability. To further improve the enhancement in the feature representation of the 3D CNN, instead of using an RGB representation of the video, we use a 3D Discrete Wavelet Transform (3D DWT) as a pre-processing step to obtain a motion-salient representation that localizes action in space and frequency. Thus, the model is presented with the localized information of action occurring in the

video and can filter out unnecessary information present in the video. We evaluate our model on the benchmark UCF11 and UCF Sports action datasets. We have employed widely recognized performance metrics, including classification accuracy, precision, recall, F1 score, and AUC score, to assess the effectiveness of the models proposed in our study. We have adopted the fine-tuning scheme, a transfer learning approach, to train our model effectively. The experimental results in sections 3.2 and 4.2 show the effectiveness of our proposed approaches.

List of Publications

Publications from M.S. Research Thesis Work:

Neelesh Ghanghoriya, Rituraj Singh, Sumeet Saurav, Aruna Tiwari and Sanjay Singh "3D Self Attention Convolutional Neural Network with 3D Discrete Wavelet Transform preprocessing for Action Recognition in videos." (Under preparation)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Video action recognition is one of the most crucial and challenging problems in the computer vision domain. Action recognition is fundamental for many real-world problems, such as video retrieval, intelligent surveillance, healthcare monitoring, behavior analysis, etc. Video is an ordered sequence of continuous images playing over time. In the computer vision area, an image is a 2D matrix of pixel values representing the scene's spatial information. A video is a 3D matrix of pixel values representing the information in the space-time volume compared to a still 2D image. A human action can be defined as a unique and ordered sequence resulting from the coordinated movements of various body parts unfolding over a specific period. Action recognition is a classification problem that requires identifying the video's action category based on the action performed. The action recognition task in videos is broadly divided into two fundamental steps. Firstly, defining a representation for the 3D space-time volume of video that fairly represents discrimination in features of different actions. Second, determining a model architecture that can learn these feature representations of actions effectively and efficiently. This thesis centers on the pursuit of an efficient representation of video and the development of a deep learning framework for action recognition in videos, particularly when confronted with limited training data availability. Our primary focus is on achieving effectiveness and efficiency in video representation and deep learning framework construction, addressing the challenges posed by limited training data.

The rest of this chapter is organized as follows. Section 1.1 provides background information on video action recognition. In addition, section 1.2 elaborates on the motivation for our work. The research objectives are formulated in Section 1.3. In Section 1.4, we present the summary of thesis contributions, and an outline of the rest of the thesis is described in Section 1.5.

1.1 Background

Video is a complex 3D volume of pixel information in space and time. The physical constraints while capturing a video, such as illumination changes, occlusion, background clutter, variation in viewpoint, camera motion, etc., make the video data representation complex. Action in the video is a complex ordered sequence of body movements that occur inconsistently over space and time. Furthermore, considerable variation and unpredictability exist in postures and the execution of a particular action among different individuals. The inherently complex nature of videos and actions makes action recognition tasks in videos challenging. As discussed, the general steps involved in solving the action recognition in videos are first to define a representation for the input video and second, classification of the video based on the defined representation. Most traditional approaches [34, 63] consist of the following three steps - feature extraction, feature descriptor computation, and feature classification. It first extracts handcrafted low-level features from videos that undergo preprocessing. Following the feature preprocessing step, the feature representation and feature encoding step form a video-level representation. Finally, classification is performed based on the formed video representation to determine the action category. The complete pipeline of traditional approaches is computationally extensive and complex and requires human intervention at each step. While traditional approaches exhibited commendable recognition performance, they encountered challenges when scaling up for large-scale datasets. On the other hand, deep learning approaches [15, 29, 31, 44, 68] are end-to-end trainable algorithms. These approaches automatically learn feature representation from raw RGB data and have a trainable classifier to perform action recognition. The advent of deep learning approaches began with the adoption of a convolutional neural network (CNN) for action recognition in videos[15, 29, 31, 44, 68]. Due to inductive bias properties such as local connectivity and translation equivalence, Convolutional Neural Networks (CNN) [35, 64, 66] achieved outstanding performance in various image domain tasks such as recognition, segmentation, detection, and captioning. Considering video as a sequence of 2D images over time and CNN's excellent visual feature learning capability in images, many researchers Donahue et al. [15], Ji et al. [29], Karpathy et al. [31], Liu et al. [44], Simonyan and Zisserman [68] have successfully adopted the CNN architectures for action recognition tasks in videos. The early trend began with using two-stream CNN architectures [68], which consisted of two independent 2D CNNs for spatial and temporal streams. The spatial stream network takes RGB frames as input and learns to represent the appearance information of frames. In contrast, the temporal stream network learns to represent the motion information from the optical flow representation of RGB frames. Though this approach achieved good action recognition accuracy, the expensive computation of the optical flow of the RGB frames and storing it beforehand were among the significant drawbacks. To overcome this drawback, researchers Donahue et al. [15], Liu et al. [44] proposed to use a hybrid framework of 2D CNN and sequential models such as Long Short-Term Memory (LSTM). This hybrid framework used LSTM networks on the top of a 2D CNN for temporal modeling of the extracted spatial CNN features. This approach only models the high-level convolutional features from the top layers, while the low-level features from the earlier layers of 2D CNN are not processed explicitly. With the advent of 3D CNN architecture, action recognition methods can directly process and extract the spatio-temporal features in the 3D volume of consecutive video frames. It can simultaneously enable both low-level and high-level spatial and temporal features.

1.2 Motivation

The 3D CNN has colossal parameters that require extensive training data. The introduction of various large-scale datasets such as Sports1M[31] and Kinetics[9] have successfully allowed 3D CNN architecture to showcase their powerful capabilities in learning spatio-temporal features in videos. In the past few years, Extensive research [9, 23, 24, 70, 71, 80, 81] has been done in identifying and resolving various issues of the 3D CNN architecture for its application to video action recognition. A prominent limitation of 3D CNN-based models lies in their extensive computational time and data requirements due to the substantial parameters inherent in the architecture. To mitigate this, researchers Carreira and Zisserman [9], Hara et al. [23, 24], Tran et al. [71], Wei et al. [81] have suggested factorizing the 3D kernel into a combination of 2D kernels in spatial dimensions and 1D kernels in temporal dimensions, reducing model parameters and complexity. However, this factorization may compromise the 3D CNN's ability to effectively capture spatio-temporal features simultaneously, especially in scenarios with limited data availability. Consequently, we explore an alternative approach to address the high computational training time requirement. Another inherent drawback of 3D CNN architecture is the susceptibility to overfitting, exacerbated by limited training data and high model complexity. Researchers Carreira and Zisserman [9], Hara et al. [24] have successfully mitigated this concern through transfer learning, specifically fine-tuning. They initialize their model's weights with pre-trained weights from another model on larger datasets, strategically leveraging pre-existing knowledge to enhance generalization capability. Despite these efforts, a universally defined methodology for applying transfer learning in deep learning models remains elusive. In our work, we adopt a fine-tuning approach to train our proposed 3D CNN model, aiming to improve generalization performance and address overfitting concerns associated with limited training data. Another significant and less explored limitation in 3D CNN architecture is a fixed and limited kernel's receptive field that only processes small local neighborhoods around each pixel in videos to build spatio-temporal feature dependencies. Due to the small and fixed kernel size usually, $3 \times 3 \times 3$ or $5 \times 5 \times 5$, 3D CNN can capture the local spatio-temporal features but fails to capture the global dependencies that exist between pixels situated at more considerable distances. Conventionally, the limited receptive field of the 3D kernel is extended by increasing the depth of the model via stacking multiple convolutional layers. However, this approach results in a massive increase in the parameters, increasing the complexity of the model and making the model more prone to overfitting problems. Another limitation of 3D CNN is that the feature representations derived from many of the 3D CNN-based models may not be sufficiently discriminative despite the impressive feature learning capability inherent in 3D CNNs. The representation learned by 3D CNNs is adversely affected by the inclusion of unnecessary and redundant information, leading to a deterioration in its overall discriminative feature quality.

This work effectively addresses the aforementioned challenges by proposing a comprehensive solution for action recognition tasks, demonstrated on small-scale benchmarks such as UCF11[45] and UCF Sports Action[60] datasets. Firstly, we introduce a 3D Self-Attention Convolutional Neural Network named SAC3D. Our proposed SAC3D integrates a 3D self-attention module into the baseline architecture of the C3D model [70]. The embedded 3D self-attention module computes attention maps based on weighted correlations among features in hidden maps produced by 3D Convolutional layers. These attention maps guide SAC3D to focus on capturing the strong local and global feature dependencies, irrespective of their spatial distance in the feature map. Leveraging the 3D self-attention mechanism, SAC3D captures pairwise pixel similarity, addressing limitations imposed by fixed and limited kernel sizes. The module empowers SAC3D to distinguish foreground actors from the background by strategically assigning more weight to pairwise relationships among foreground pixels, reducing irrelevant information and enhancing discernment capability. To further enhance foreground-background separation, we introduce a 3D discrete wavelet transform (3D DWT)-based preprocessing step on RGB frame sequences. This step generates a salient representation, localizing actions in spatial and frequency domains. Training SAC3D on this representation enables the model to learn better discriminative features in the RGB frame sequence, improving accuracy by focusing attention on relevant regions while suppressing background noise. Another advantage of employing this salient representation is its contribution to mitigating the high computation time associated with 3D CNNs. While its use does not directly reduce the model's parameters and complexity, the sparse nature of the model's input data leads to decreased computation time requirements. This valuable optimization enhances the efficiency of our proposed 3D CNN architecture without compromising its overall structure and intricacy. Our proposed SAC3D model is tailored for efficient training on smaller datasets, addressing substantial data and computational time demands. Adopting a transfer learning approach, we fine-tune SAC3D using pretrained weights from the C3D model trained on the Sports1M dataset. Experimental evaluations on the benchmark UCF11 [45] and UCF-Sports action [60] datasets highlight SAC3D's superior performance over state-of-the-art approaches, affirming its efficiency in addressing challenges associated with action recognition in videos.

1.3 Research Objectives

This thesis addresses the challenges in action recognition within the constraints of limited training data and the drawbacks of traditional 3D CNN architectures. Introducing a novel approach named SAC3D, a Self-Attention Convolution Neural Network, the study integrates a 3D self-attention mechanism into the C3D model baseline. This mechanism enhances the discriminative feature learning capability by guiding the 3D convolutional layers based on pairwise pixel similarity. Additionally, the research incorporates a 3D Discrete Wavelet Transform (3D DWT) as a preprocessing step, using motion-salient representations to localize action in space and frequency. The proposed model is evaluated on UCF11 and UCF Sports action datasets, demonstrating effectiveness through widely recognized performance metrics and employing a fine-tuning scheme for efficient training. We have divided the thesis objectives into three key research tasks, outlined below:

1. We propose a novel framework for action recognition in videos based on 3D CNN for small-scale datasets.

2. We propose to use a 3D self-attention to mitigate the various issues in 3D CNN architecture (SAC3D) effectively.

3. We further enhance the feature representation of the 3D CNN model by using

a 3D DWT as a preprocessing step. The detection and representation of salient spatiotemporal regions of the RGB frame sequences improve the accuracy of our proposed framework.

1.4 Thesis Contribution

The summary of our research contributions is presented below, with more discussion available in the subsequent chapters.

Contribution I: 3D Self Attention Convolutional Neural Network for action recognition in videos In videos, action happens in a small portion of the frame and across fewer frames. Therefore, in identifying the action category in the video, it is crucial to focus more on a few specific parts of the video rather than the entire video. In this thesis, we focused on improving the feature learning capabilities of 3D CNN architecture. We developed a novel action recognition framework, 3D Self-Attention Convolutional Neural Network, named SAC3D. We tackled the incompetence of the C3D baseline to model long-range dependencies by incorporating a 3D self-attention mechanism. 3D self-attention mechanism allows the model to focus on important information and ignore redundant irrelevant information. We successfully trained our SAC3D model effectively and efficiently, conducting action recognition on the UCF11 and UCF Sports action datasets. The evaluation utilized well-established performance metrics such as classification accuracy, precision, recall, F1 score, and AUC score to assess the effectiveness of the proposed models in our study.

Contribution II: 3D Self Attention Convolutional Neural Network with 3D Discrete Wavelet Transform preprocessing for Action Recognition in videos In this contribution, we introduced an additional preprocessing step utilizing 3D Discrete Wavelet Transform (3D DWT) on the input RGB frame sequences. The application of 3D DWT generates various wavelet subbands, offering localized information in both frequency and space domains. Leveraging these selected subbands, we constructed a novel representation that effectively separates the foreground human action from background scenes. This refined representation enhances feature learning, contributing to the improved performance of our proposed SAC3D model. In summary, this work advances the quality of feature representation in SAC3D by training the model on a salient motion representation instead of the RGB frame sequence from UCF11 and UCF sports action video clips.

1.5 Organization of Thesis

We have organized this thesis into five chapters. A summary of each chapter is provided below:

Chapter 1 **(Introduction)**

The present chapter provides an overview of the background knowledge about video action recognition tasks, elucidates the motivation behind our work, and outlines the contributions made by this thesis.

Chapter 2 **(Literature Survey)**

This chapter provides a detailed literature survey on various deep learning approaches for solving video action recognition tasks and a survey on the self-attention mechanism and its application to action recognition in the video. It also surveys various approaches that utilize 3D Discrete wavelet transform for the action recognition task. Finally, it presents a detailed account of the various metrics employed for evaluating the performance of the proposed methods.

Chapter 3 **(3D Self Attention CNN (SAC3D) for action recognition in videos)**

In this chapter, we proposed an action recognition framework - a 3D self-attention convolutional neural network (SAC3D)- for video action recognition. We proposed incorporating a 3D self-attention mechanism in Convolutional 3D (C3D) architecture. We evaluate the performance of the proposed approach on the benchmark UCF11 and UCF sports action datasets and compare our approach with state-of-the-art methods.

Chapter 4 **(3D self-attention CNN (SAC3D) with 3D DWT preprocessing for action recognition in videos)**

In this chapter, we proposed an action recognition framework - a 3D self-attention convolutional neural network (SAC3D) with a 3D Discrete wavelet transform (3D DWT) for video action recognition. We used an additional preprocessing step based on 3D DWT. We incorporated a 3D self-attention mechanism in Convolutional 3D (C3D) architecture. We evaluated the performance of the proposed approach on the benchmark UCF11 and UCF sports action datasets and compared our approach with state-of-the-art methods.

Chapter 5 **(Conclusion and Future Work)**

This chapter provides a conclusion with the contribution of this thesis and the potential future directions of our work.

Chapter 2

Literature Survey

This chapter provides a detailed literature survey in four sections. Section 2.1 discusses the deep learning approaches that successfully solve video action recognition problems. Section 2.2 provides the survey of the self-attention mechanism and its application to action recognition in the video. Section 2.3 surveys various approaches that successfully applied discrete wavelet transform (DWT) for video action recognition. Finally, Section 2.4 provides a study on the different performance metrics that researchers have used to analyze the existing action recognition approaches and discusses the metrics used for performance evaluation of the proposed methods.

2.1 Convolutional Neural Network (CNN)

Action recognition in videos has been an extensively researched subject in the computer vision community for decades. The video's action recognition task typically involves two primary steps [4, 94]. Firstly, defining a representation for the 3D space-time volume of video that represents discrimination in features of different actions well. Second, determining a model architecture that can learn these feature representations of actions effectively and efficiently. From the early traditional approaches to modern deep learning approaches, all have thrived to improve and efficiently implement the above two steps. Traditional approaches mainly focused on obtaining handcrafted video features by applying various image processing techniques, such as

HOG, HOG3D[34], SIFT, SIFT3D[63], etc., to form a fixed-level representation of the video. Subsequently, utilizing the generated video representation to discern the action within the video, it was categorized into a specific action class. The complete pipeline of traditional approaches is computationally extensive and complex and requires human intervention at each step. While traditional approaches demonstrated commendable recognition performance, their scalability to large-scale datasets posed significant challenges. Inspired by the success of the Convolutional Neural Network (CNN) in the image domain[35, 64, 66], early deep learning approaches efforts concentrated on adopting 2D CNN architecture in the video domain[15, 31, 44, 68]. Karpathy et al. [31] designed single stream 2D CNN architecture and explored multiple fusion strategies to fuse temporal information from consecutive frames. Their attempt to capture local spatio-temporal features could not effectively model the temporal information across the frames in the video. Another solution to modeling the temporal information that became a trend in action recognition architecture was two-stream network architecture [68], which consisted of two separate independent 2D CNN architectures for spatial stream network and temporal stream network. The spatial stream network takes RGB frames as input and learns to represent the appearance features of the frames. At the same time, the temporal stream network learns to represent the motion information from the optical flow representation of the stacked RGB frames. Though these approaches achieved good action recognition accuracy, the expensive computation of optical flow representation from the RGB frames and storing it beforehand were among the significant drawbacks. In attempts to replace the optical flow for motion representation, researchers proposed using a hybrid framework of 2D CNN and sequential models such as Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM). Donahue et al. [15], Liu et al. [44] used LSTM networks on the top of the 2D CNN for temporal modeling from the extracted spatial features. Their approaches only modeled the high-level convolutional features from the top layers of the 2D CNN, While the low-level features from the earlier layers of the 2D CNN are not processed explicitly. Video is a 3D volume of 2D frame sequences across time. The natural extension of the CNN model in the video domain should have a 3D kernel that can simultaneously process spatial and temporal information. Thus, the CNN model that consists of the 3D kernel can exclude modeling temporal information from spatial features as required previously with 2D CNN architectures. Ji et al. [29], inspired by this ideology, attempted to develop 3D CNN architecture as a more suitable extension of CNN to the video domain. Despite correct reasoning, the model performance was unsatisfactory due to the high complexity of the developed 3D CNN network and the limited availability of training data. Tran et al. [70] designed a modular 3D CNN architecture with a $3 \times 3 \times 3$ kernel: C3D, which performed at par with existing state-of-the-art methods. Despite the increased boost in the action recognition performance with C3D[70], researchers continued exploring two-stream networks for their model architecture due to the parameter complexity of C3D[70] model. Feichtenhofer et al. [19] presented various fusion strategies for two-stream CNN networks to fuse spatial and temporal streams, and Wang et al. [79] suggested various good practices to train two-stream CNN networks. Owing to them, the performance of two-stream CNN architecture achieved new state-of-the-art results. Carreira and Zisserman [9] in 2017, developed a new 3D CNN architecture, Inflated 3D (I3D), in which 3D kernels were made by stacking 2D kernels. It enabled 3D CNN architecture to leverage the learning of pretrained 2D CNN on large-scale ImageNet dataset [14]. Due to the outstanding performance of I3D and the new publication of a large-scale video dataset Kinetics-400[9], researchers started focusing on adopting 3D CNN in their model architecture. Research in improving the 3D CNN architecture has advanced quickly in the past few years. Hara et al. [23] developed ResNet3D model for video domain as an extension of ResNet[26] in image domain. It attempted to tackle the overfitting issue of 3D CNN architecture by implementing the residual connection. Hara et al. [24] validated the available dataset for training ResNet3D [23] without overfitting and Hara et al. [25] suggested good practices to train the ResNet3D [23] architecture. P3D [81], and R2+1D [71] explored the idea of factorizing the 3D kernel into a 2D spatial kernel and 1D temporal kernel to reduce the complexity of the 3D CNN architecture. A less explored drawback of the 3D CNN architecture, as indicated in the literature, is the fixed and limited receptive field of the 3D kernel. This limitation constrains its ability to effectively capture long-range dependencies inherent in the spatio-temporal features of video clips. Addressing this concern, Wang et al. [80] introduced a non-local operation-based building block as a convolutional architecture extension, aiming to capture long-range dependencies. Their approach involves calculating the weighted average of all positions as the response at each position. This limitation in 3D CNN's receptive field and the analogous work by Wang et al. [80] form crucial considerations in developing our proposed 3D self-attention convolutional neural network.

2.2 Self Attention Mechanism

Videos contain irrelevant, redundant information that is not useful for deciding the underlying action performed in videos. For the action recognition task, background information present in multiple frames of the video is irrelevant and redundant, whereas the foreground information of the video is mostly the conclusive information needed to categorize action in videos. Therefore, it becomes essential to have an attention mechanism that can enhance and highlight the salient regions of the videos for the model to focus and reduce the influence of irrelevant and redundant information on the model's feature learning ability. Vaswani et al. [75] introduced a self-attention operation-based Transformer architecture that achieved state-of-the-art results in the machine translation tasks. The extraordinary performance of transformer architecture in many natural language processing tasks is due to the effectiveness of modeling long-range dependencies by the rooted self-attention operation. Inspired by the success of the transformer architecture, researchers Arnab et al. [5], Bertasius et al. [6], Carion et al. [8], Chen et al. [11], Fan et al. [18], Kim et al. [33], Liu et al. [46, 47], Patrick et al. [56], Sun et al. [69], Yan et al. [83], Yang et al. [85], Zhang et al. [91, 92], Zhu et al. [93] have successfully adopted the transformer architecture for many computer vision tasks and proved the effectiveness of the self-attention mechanism in modeling long-range visual dependencies as well. Dosovitskiy et al. [16] introduced vision transformer (Vit), a pure transformer architecture design, to achieve state-of-the-art results in image classification. Arnab et al. [5], Bertasius et al. [6] adopted the extension of the Vit [16] for video action classification task and achieved state-of-the-art performance on many benchmark datasets. Many research works Arnab et al. [5], Bertasius et al. [6], Chen et al. [11], Fan et al. [18], Kim et al. [33], Liu et al. [46, 47], Mazzia et al. [48], Patrick et al. [56], Sun et al. [69], Yan et al. [83], Zhang et al. [91, 92] have focused on refining the transformer architectural design for performing better on video action recognition. Although the transformer outperformed the preferred architecture, CNN, in many visual tasks, many works Cordonnier et al. [12], Kim et al. [33], Ramachandran et al. [57], Tuli et al. [73] have tried comparing the suitability of Transformer with CNN architecture for visual tasks. Various researchers have argued in favor of transformer architecture that the power of self-attention operation and multi-headed attention mechanism given sufficient training data have supremacy over the convolutional architecture in having better learning capability in visual tasks. Transformer architectures have colossal data requirements and heavy computational training time compared to CNN architecture. The transformer architecture also has the disadvantage of not explicitly using the local connectivity properties of visual data to model Spatio-temporal feature dependencies, which is an essential aspect of learning visual tasks. Due to these drawbacks, Researchers in the computer vision community continue to opt for CNN for their architectures. Nevertheless, the power of self-attention networks in modeling long-ranging visual dependencies, which is essential for any vision task, cannot be overlooked. Many recent works in natural language processing share the idea of using self-attention in cooperation with convolution operation [84]. Researchers Essa and Abdelmaksoud [17], Guo et al. [22], Jiang et al. [30], Li et al. [38, 39, 40, 41], Yang et al. [84], Zeng and Li [88], Zhang et al. [90] in computer vision have successfully leveraged the power of the self-attention mechanism with convolutional neural network architecture design to obtain a better representation of the visual data and achieve better performance in many visual tasks.

As mentioned earlier, the video action recognition task comprises two essential steps. Up to this point, our survey has focused on approaches dedicated to constructing effective and efficient model architectures. Another critical facet involves devising a representation for the 3D space-time volume of a video that captures the discriminative features distinguishing various actions. In addressing this, we propose the utilization of Discrete Wavelet Transform (DWT) as a key component of our representation. DWT extracts motion saliency, offering a nuanced depiction of actions in both spatial and frequency domains, thereby enhancing the discriminative quality of our model's feature representation.

2.3 Discrete Wavelet Transform

Researchers in the field of action recognition in videos have taken great advantage of the Discrete Wavelet Transform (DWT) in improving action recognition performance. DWT localizes changes in video pixels over space and frequency to enhance the spatio-temporal feature extraction and representation. Researchers have mainly used 2D DWT or 3D DWT to gather information about the changing features from individual 2D frames or the 3D volume of frame sequences. Shao and Gao [65] explored the wavelet transform-based feature descriptor. Imtiaz et al. [27] presented a multi-resolution feature extraction algorithm using multilevel 2D DWT. It operates within the frames of video sequences to extract features. Khare and Jeon [32] designed a multi-resolution approach to represent human objects using 2D DWT coefficients. Li and Liu [42] developed a novel scene analysis algorithm that categorizes scene changes into short, motion, gradual, and static scenes based on 3D DWT. Their investigation of the spatial and temporal distributions suggested that statistical features of 3D DWT coefficients can describe the correlation among adjacent frames. Rapantzikos et al. [58] showed the potential of 3D DWT in the detection and representation of spatio-temporal saliency of the frame sequence. The framework measured the saliency based on the orientation-selective bands of the 3D DWT and represented different events using simple features of salient regions. Al-Berry et al. [3] built a multiscale representation of human actions using a 3D multiscale stationary wavelet analysis technique. The boost in recognition performance was due to fusing spatio-temporal information that got highlighted at different scales and orientations. Mohammadi et al. [50] developed a hybrid classifier that compressed the features and classified them using SVM with polynomial or sigmoid kernels. It evaluated the effect of 3D DWT as the pre-processing step in detecting saliency that strengthens motion feature extraction in the bag of visual words framework. Chang et al. [10] utilized DWT in the Wavelet-Attention Decoupling (WAD) module to effectively disentangle salient and subtle motion features in the time-frequency domain, significantly enhancing the recognition of fine-grained actions in skeleton-based action recognition. Bhuiyan et al. [7] integrated DWT into human activity recognition on smartphones, facilitating robust feature extraction by capturing local features from raw activity signals, enhancing the recognition process's accuracy and computational efficiency. Zhang et al. [89] incorporated DWT into dense trajectory models to enhance the recognition of human actions in video by providing more descriptive features through the separation of dominant and secondary motions across different frequency bands and directions, leading to improved performance in complex spatial-temporal domains.

Inspired by the notable advancements in action recognition achieved through the Discrete Wavelet Transform (DWT), particularly in localizing spatio-temporal features, we integrate the powerful tool of 3D DWT as a preprocessing step for RGB frame sequences. This enhances the model's learning capability by providing foreground-background separation and complements our proposed 3D self-attention convolutional neural network, leading to improved recognition accuracy.

2.4 Performance Metrics

As already discussed, Video action recognition is a classification problem. Most researchers evaluate the performance of their approaches based on classification accuracy. We have also used classification accuracy as the performance measure to assess and compare our approaches with various state-of-the-art methods. We have also tested our approaches on other commonly used performance metrics such as precision, recall, f_1 score, and AUC score. The standard formulations for these are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FPFN} \tag{2.1}$$

$$Precision(P) = \frac{TP}{TP + FP} \tag{2.2}$$

$$Recall(R) = \frac{TP}{TP + FN} \tag{2.3}$$

$$F_1 - score = \frac{2 \cdot P \cdot R}{P + R} \tag{2.4}$$

$$AUC - score = \frac{2P \cdot R}{P + R} \tag{2.5}$$

In the above equations,

- True Positive(TP) signifies how many positive class samples the model predicted correctly.

- True Negative(TN) signifies how many negative class samples the model predicted correctly.

- False Positive(FP) signifies how many negative class samples the model predicted incorrectly.

- False Negative(FN) signifies how many positive class samples the model predicted incorrectly.

Accuracy measures how often the classifier correctly predicts. We can define accuracy as the ratio of correct predictions and the total number of predictions. Precision is the ratio of true positives and total positives predicted. A precision score towards one will signify that the model does not miss any true positives and can classify well between correct and incorrect labels. A recall is essentially the ratio of true positives to all the positives in the ground truth. Recall measures the proportion of actual positives that the model correctly identified. The F1 score is the harmonic mean of precision and recall. The highest possible value of an F1 score is 1.0, indicating perfect precision and recall, and the lowest possible value is 0 if either the precision or the recall is zero. In conclusion, the performance evaluation of our approaches in video action recognition encompasses a comprehensive set of metrics, including classification accuracy, precision, recall, F1 score, and AUC score. These metrics provide a robust assessment of our models, ensuring a thorough understanding of their effectiveness across various aspects of classification performance.

Chapter 3

3D Self Attention Convolutional Neural Network for action recognition in videos

In computer vision, a video is a 3D signal that evolves over space-time, capturing changes in appearance information over time. Human action is a complex, ordered sequence of body movements recorded across multiple video frames. Within videos, actions unfold in small portions of the frame and span fewer frames, implying that human action in a video represents the part of the 3D signal that undergoes rapid changes in time and is localized in space. Hence, it is crucial to prioritize this specific segment rather than the entire footage to identify the action category in a video accurately. Consequently, an attention mechanism becomes essential to assign relative importance to different parts of the video.

3.1 Proposed Method

We propose a 3D Self Attention Convolutional Neural Network named SAC3D with an RGB video clip as input. Figure 3.2 shows the block diagram of our proposed model architecture. We have incorporated a 3D self-attention module [41] in Convolutional 3D (C3D) [70] model's baseline architecture. The embedded 3D self-attention module computes attention maps based on weighted correlations among features in hidden maps produced by 3D Convolutional layers. The attention maps guide SAC3D to capture robust feature dependencies while disregarding weaker local dependencies. This addresses the limitations associated with the fixed and limited kernel size of the 3D convolutional layer. The 3D self-attention module enables our SAC3D to pay weighted attention to different parts of the video, enhancing the model's ability to focus on significant features. We propose to purposely train our SAC3D model on small-scale video datasets, addressing the overfitting issue of 3D CNN architectures on limited data availability. We have used a transfer learning - fine-tuning scheme for training our SAC3D model to avoid overfitting problems. We have adopted the weights of the pretrained C3D model on the Sport1M[31] dataset for instantiating baseline convolutional layers and fully connected layers of our SAC3D model. We train and validate our model on the benchmark UCF11 [45] and UCF sports action [60] datasets. Following this, we delve into the specifics of the components comprising our proposed action recognition framework. Subsequently, we elaborate on the details of the 3D self-attention module and provide the configuration details of the proposed SAC3D model.

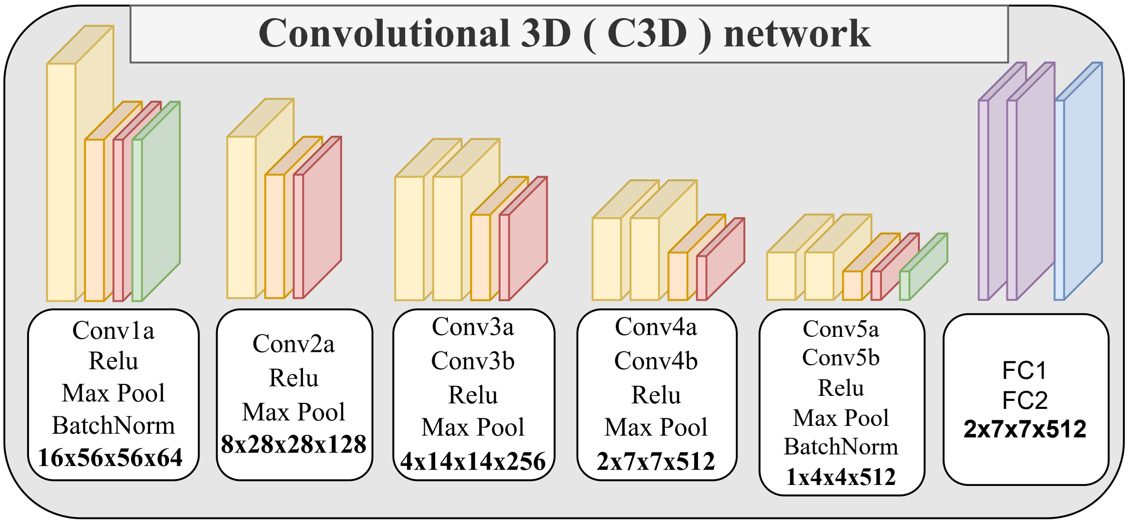


Figure 3.1: Block diagram of Convolutional 3D (C3D) neural network proposed by Tran et al. [70]

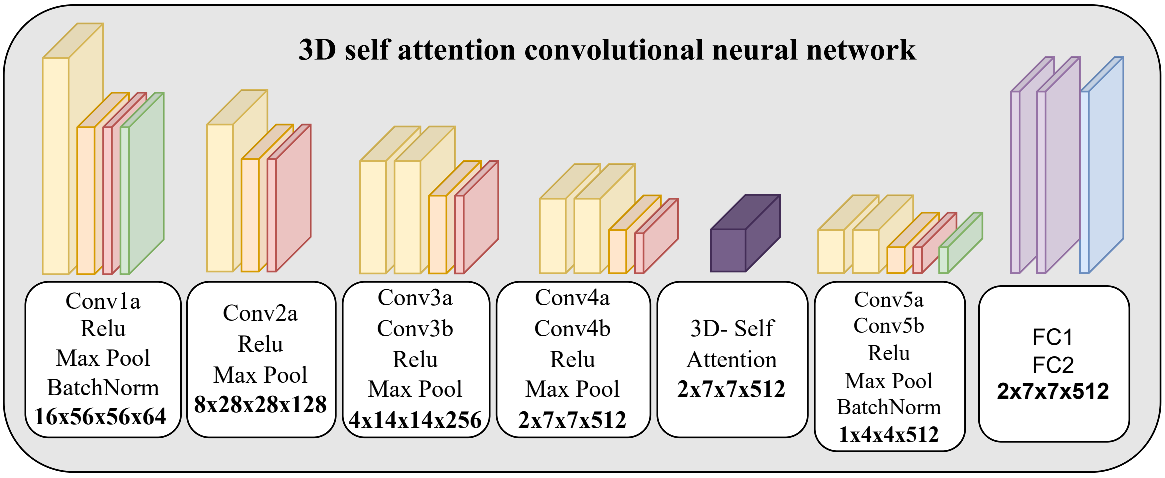


Figure 3.2: Block Diagram of proposed 3D Self Attention Convolutional Neural Network (SAC3D)

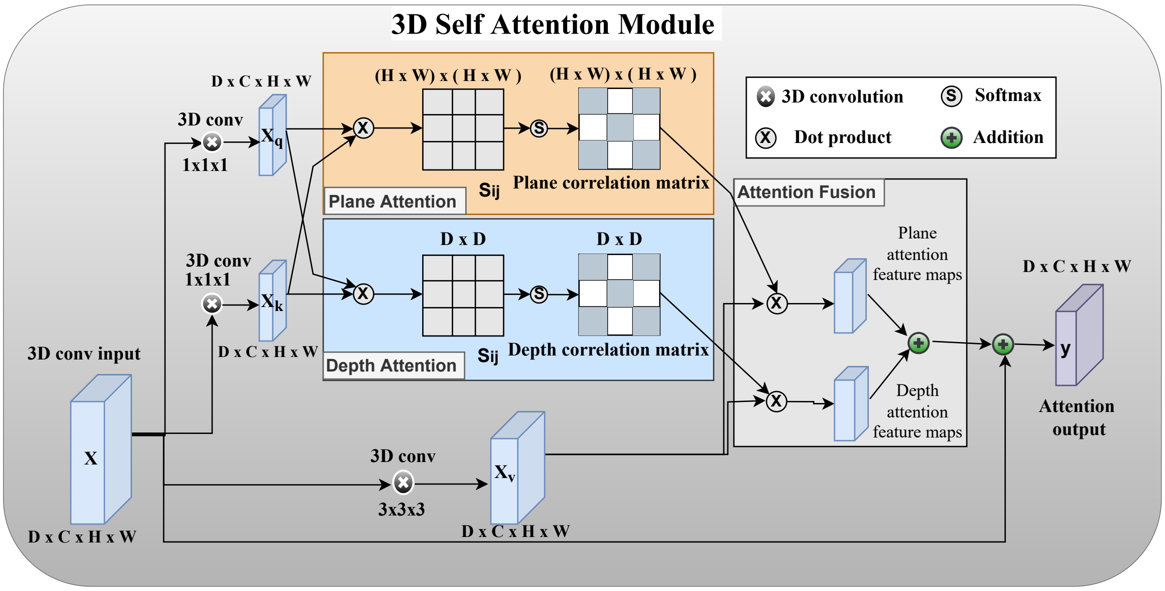


Figure 3.3: Proposed 3D Self Attention module

3.1.1 3D Self Attention module

Self-Attention is a powerful attention mechanism that effectively captures dependencies between two input positions. It computes a correlation matrix along different dimensions of the input sequence to capture the pairwise relationship between feature positions. The correlation value at each position is calculated through the weighted sum of all other positions. Considering all the positions while computing the similarity between pairwise positions of the input, the self-attention mechanism can identify and represent the strong and weak relationships in the input feature map. Based on the strength of the similarity relationships among pixels, it can infer the strong global dependencies and weak local dependencies in the input sequence. Therefore, a self-attention mechanism aids in assigning the necessary relative importance to different parts of the video. Mathematically, a self-attention mechanism involves mapping a query, a key, and a value to the input feature map, where the query, key, and value are all derived from the input feature map. We categorize dependencies present in the spatiotemporal volume of the hidden feature map into intra-frame and inter-frame dependencies. Intra-frame dependencies are relationships in the pixels in spatial dimensions within each frame. In contrast, inter-frame dependencies are the relationships in the features across different frames along the temporal dimension. In this work, we capture the long-range spatiotemporal feature dependencies using a 3D Self-Attention mechanism. The proposed 3D self-attention mechanism gives specific attention to the intra-frame dependencies of the feature maps by the plane attention module; likewise, it provides specific attention to inter-frame dependencies by the temporal attention module. Let $X \in R^{D \times H \times W \times C}$ denote the hidden convolutional feature map that is input to the 3D self-attention module, where D is frame depth, H and W are the height, and width and C corresponds to the number of channels in the input feature map. We first apply $1 \times 1 \times 1$ convolution to the input feature map to form embedding vectors query $X_q \in R^{C \times D \times H \times W}$, *key* $X_k \in R^{C \times D \times H \times W}$ and *value* $X_v \in R^{C \times D \times H \times W}$. These embedded vectors further served as input to the Plane attention module and Temporal attention module, which are defined below.

3.1.1.1 Plane attention module

To capture intra-frame dependencies within the feature map, we initially flatten the 2D spatial dimensions of the query X_q, key X_k, and value X_v vectors into 1D vectors. We transform each query X_q, key X_k, and value X_v vectors from $(C \times D \times H \times W)$ to flatten 1D vectors $\in (C \times D \times N)$ feature space, where N is the total number of pixels. We next perform the inner dot product of the transposed query vector ${X_q}^T(i)$ and key vector $X_k(j)$ to compute the correlation matrix s_{ij}.

$$s_{ij} = X_q(i)^T \times X_k(j) \tag{3.1}$$

The correlation matrix describes the similarity-relationship between the features within each frame of the input feature map. The softmax function is employed to derive the self-attention map plane, denoted as a_{ij}, by normalizing the correlation matrix s_{ij}.

$$a_{ij} = SOFTMAX(s_{ij}) = \frac{\exp^{s_{ij}}}{\sum_{i=1}^{k} \exp^{s_{ij}}} \tag{3.2}$$

The final output of the plane attention module is calculated by the dot product of the value vector X_v to the weight matrix of self-attention map a_{ij}.

$$P_{att} = \sum_{i=1}^{N} X_v(i) \times a_{ij} \tag{3.3}$$

The output of the plane attention module P_{att} signifies the weighted strength of the pairwise feature dependencies within each frame of the input feature map.

3.1.1.2 Temporal attention module

It captures the intra-frame dependencies present in the features across the different frames of the input feature map over the temporal dimension. Similar to the plane attention module, we begin by transforming each of the queries X_q, key X_k, and value X_v vectors from $(C \times D \times H \times W)$ to $((H \times W) \times D \times C)$. We next perform the inner dot product of the transposed query vector ${X_q}^T(i)$ and key vector $X_k(j)$ to compute the correlation matrix s_{ij}.

$$s_{ij} = X_q(i)^T \times X_k(j) \tag{3.4}$$

The correlation matrix s_{ij} describes the similarity-relationship between the features that exist across the different frames of the input feature map. The softmax function obtains the temporal self-attention map $a_(ij)$ upon normalizing the correlation matrix s_{ij}.

$$a_{ij} = SOFTMAX(s_{ij}) = \frac{\exp^{s_{ij}}}{\sum_{i=1}^{k} \exp^{s_{ij}}} \tag{3.5}$$

The final output of the temporal attention module is calculated by the dot product of the value vector X_v to the weight matrix of self-attention map a_{ij}.

$$D_{att} = \sum_{i=1}^{N} X_v(i) \times a_{ij} \tag{3.6}$$

The output of the temporal attention module signifies the strength of the pairwise feature dependencies that exist across different frames of the input feature map.

3.1.1.3 Attention fusion module

The output of the plane attention module and temporal attention module presents the strength of the feature dependencies within each frame and across the frame sequence of the input feature map. We perform a weighted fusion of the outputs of both the plane and temporal attention modules to obtain the overall strength of the spatio-temporal feature dependencies of the input feature map. We use a trainable weight parameter γ to provide the best fusion of the attention to input feature map based on the performance. We add the results of the plane and depth attention module output to the input feature map through a residual connection to obtain the final output Y of the 3D self-attention module.

$$Y = \left[P_{att} + D_{att} \right] * \gamma + X \tag{3.7}$$

3.1.2 3D self-attention convolutional neural network

An overview of our proposed 3D self-attention convolutional neural network model (SAC3D) is shown in figure 3.2. We have extended the popular standard 3D CNN architecture - C3D[70] by incorporating a 3D self-attention mechanism. An overview of C3D architecture[70] is shown in figure 3.1. We followed the modular design of the C3D architecture. Our SAC3D model comprises five convolutional blocks with a total of eight 3D convolutional layers, featuring 64, 128, 256, 256, 512, 512, 512, and 512 channels, respectively. Each convolutional block includes a 3D max-pooling layer followed by ReLU activation. In addition to the convolutional blocks, our proposed model integrates two batch normalization layers, two fully connected (FC) layers, a 3D self-attention layer, and a dense softmax output layer. We have enhanced the C3D baseline by inserting two batch normalization layers after the first convolutional block and another before the initiation of the FC layers. Following experimental analysis, we introduced a 3D self-attention layer between the fourth and fifth convolutional blocks. The weights of the new layers, including the 3D self-attention layer, dense output layer, and batch normalization layers, were initialized randomly. We have adopted the weights of the 3D convolutional layers and FC layers from pretrained C3D on the sport1M [31] dataset.

3.2 Experiments

In this section, we have evaluated our proposed action recognition framework – a 3D self-attention convolutional neural network (SAC3D) with RGB frame sequences as input. We have conducted extensive experiments to determine the effectiveness of the proposed augmentation to C3D architecture upon encapsulating the proposed 3D self-attention module to improve recognition accuracy as the performance measure. We showcase the results of our SAC3D model for action recognition on the UCF11[45] and UCF sports action[60] datasets. We further provide our SAC3D model's performance in evaluation metrics such as confusion matrix, precision, recall, f1-score, and AUC score. The subsequent part of this section provides information about the

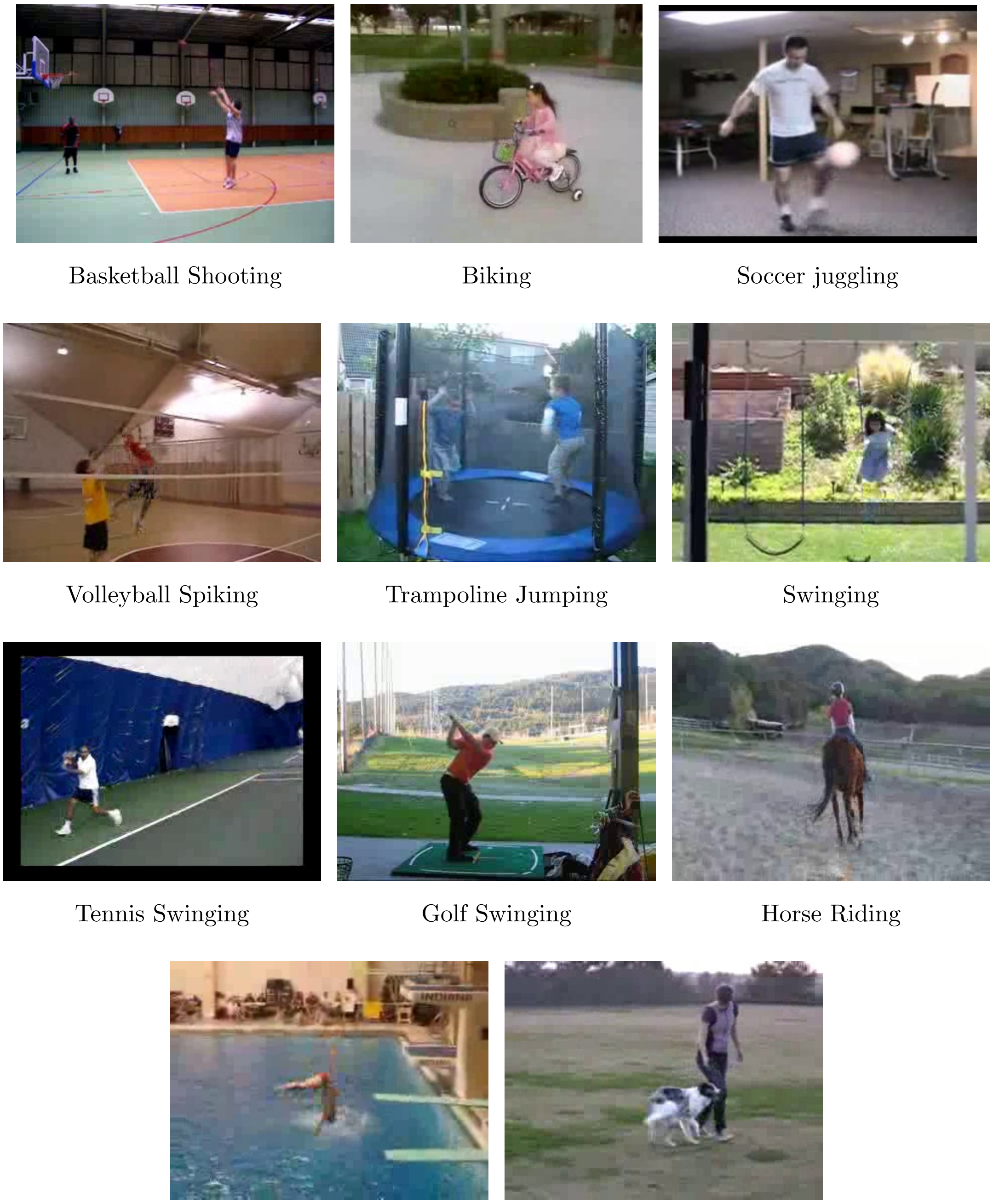
UCF11 dataset, including experimental details and results, along with a comparative analysis of our SAC3D model against state-of-the-art methods on the UCF11 dataset. Additionally, the section encompasses details related to the UCF Sports Action dataset, including experimental particulars, results, and a comparative evaluation of our SAC3D model with state-of-the-art methods on the UCF Sports Action dataset.

3.2.1 UCF11 Dataset

UCF11 (YouTube actions) [45] is a small benchmark video dataset for action recognition. The task of performing action recognition on the UCF11 dataset becomes challenging due to significant variations in camera motion, viewpoint, illumination conditions, the inconsistent appearance of variant scale objects, cluttered backgrounds, etc. The dataset has 11 sports action categories, including basketball shooting, soccer juggling, volleyball spiking, trampoline jumping, swinging, tennis swinging, golf swinging, biking/cycling, horseback riding, trampoline jumping, and walking with a dog. The dataset contains 25 groups with more than four video clips for each group in each category. The video clips within the same group share similar characteristics, such as having the same actor, similar backgrounds, similar viewpoints, etc.

3.2.1.1 Training and Testing details

In this work, we have used the Leave-One-Group-Out cross-validation scheme as in [45] as a train and test dataset split for the experiments. The proposed methodologies begin by extracting all the frames from each video. The next step is the formulation of continuous frame sequences from the extracted frames. We have used 16 frames as depth for each frame sequence of a video; thus, we divide the video into continuous segments with 16 frames. For each video segment, We first resize each frame to 128×171 spatial dimension, and then a center cropping is performed to limit spatial dimension 112×112. All the pixels in the cropped frames are scaled to limit the values $0-1$. We stack all preprocessed 16 consecutive frames to form an RGB frame sequence with dimension $16 \times 112 \times 112 \times 3$. The representation of each input video segment is a



Basketball Shooting

Biking

Soccer juggling

Volleyball Spiking

Trampoline Jumping

Swinging

Tennis Swinging

Golf Swinging

Horse Riding

Diving

Walking a dog

Figure 3.4: UCF11 Dataset[45]

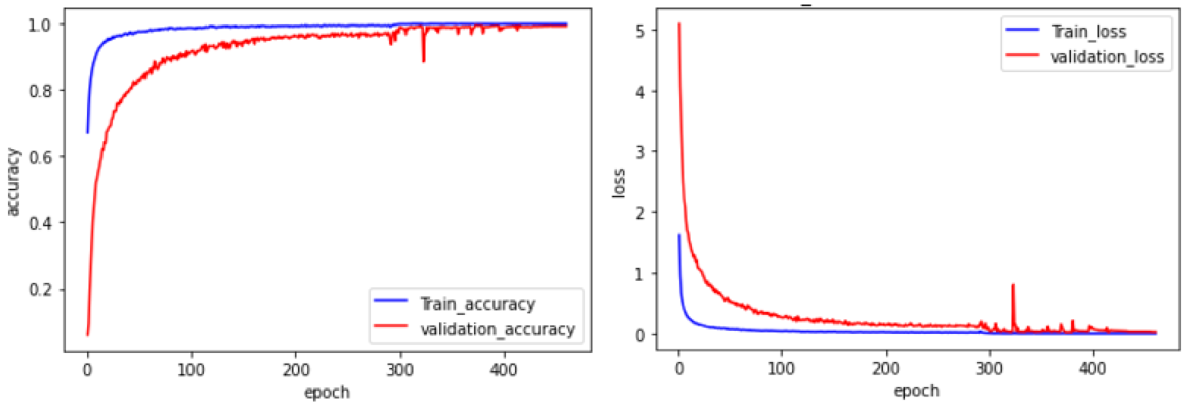
fixed size 4D tensor of dimension $16 \times 112 \times 112 \times 3$. The batch size of 8 frame sequences has been used for the experiment. Finally, 8 of the frame sequence representations are batched together to obtain the 5D tensor of dimension $8 \times 16 \times 112 \times 112 \times 3$ as the final input to the model.

The next step in the experiment is to input the final representation of the video segment to the proposed model for feature learning. We have optimized our SAC3D model with an ADAM optimizer with an initial learning rate of 10^{-4} and using Sparse Categorical Cross Entropy as the loss function. We have adopted a transfer learning-a fine-tuning scheme by adopting the weights of pretrained C3D on the Sports1M dataset [31] for initializing the weight of 3D convolutional layers and fully connected layers. First, we freeze all model layers except the last dense output layer and train the model for 20 epochs. We unfreeze the FC layers of the model and train the model for another 30 epochs. We finally make all layers trainable and train the model for 370 epochs. We have experimented with other strategies of freezing and unfreezing layers with different epochs but have experimentally found this combination strategy to result in optimal performance accuracy for the model. Figure 3.6 presents the model's training accuracy and loss plots and validation accuracy and loss plots. We have also trained model baseline architecture, C3D, with the same training scheme mentioned above. Figure 3.5 presents the C3D model's training accuracy and loss plots and validation accuracy and loss plots.

We adopt the same steps for the input test video as the training video. We first segment the input test video into continuous video segments. We next obtain the final input representation for each test video segment using the same preprocessing steps as during training. Finally, we predict action categories for each test video segment and evaluate our model performance based on the results obtained by comparing the predictions with the ground truth label of each video segment.

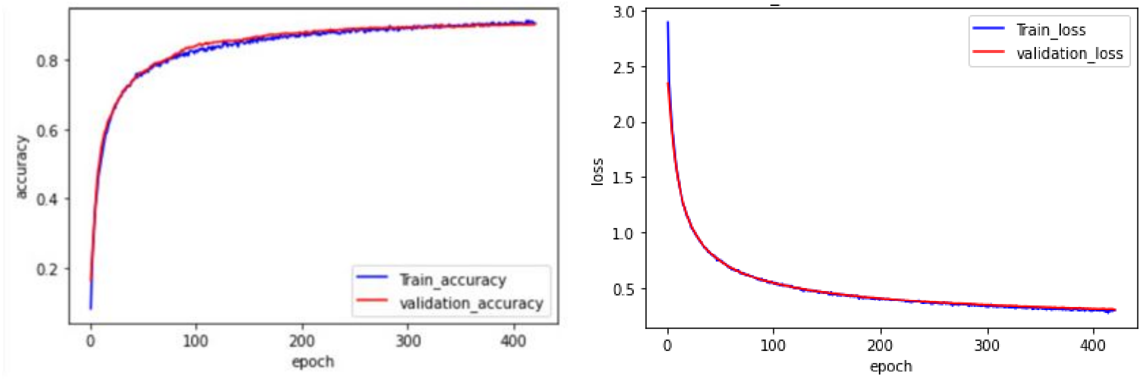
3.2.1.2 Results

Our proposed model achieves an accuracy of 93.20% during testing. Our model excels in the performance with the baseline C3D model's accuracy of 89.07% by roughly



a Training and Validation accuracy **b** Training and Validation Loss

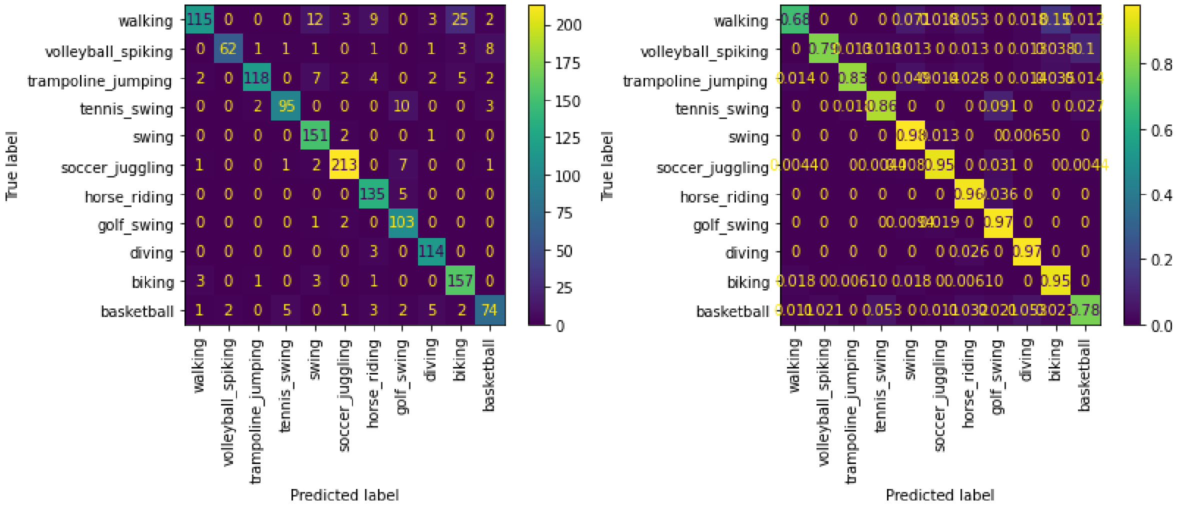
Figure 3.5: (a) Training accuracy (b) Training loss for C3D model on UCF11 dataset with RGB representation.



a Training and Validation accuracy **b** Training and Validation loss

Figure 3.6: (a) Training accuracy (b) Training loss for SAC3D model on UCF11 dataset with RGB representation.

4%. Tables 3.1 and 3.2 present the performance of the C3D and our proposed SAC3D model. Our proposed model, SAC3D, beats the baseline C3D model on all of the performance measures of accuracy, precision, recall, f1-score, and AUC score. We further present in the figures 3.7 and 3.8 the confusion matrix and normalized confusion matrix the class-wise performance of the C3D and our SAC3D model, respectively. We present the comparison of our SAC3D model with other state-of-the-art methods on the UCF11 dataset in table3.3.



a Confusion Matrix

b Normalized Confusion Matrix

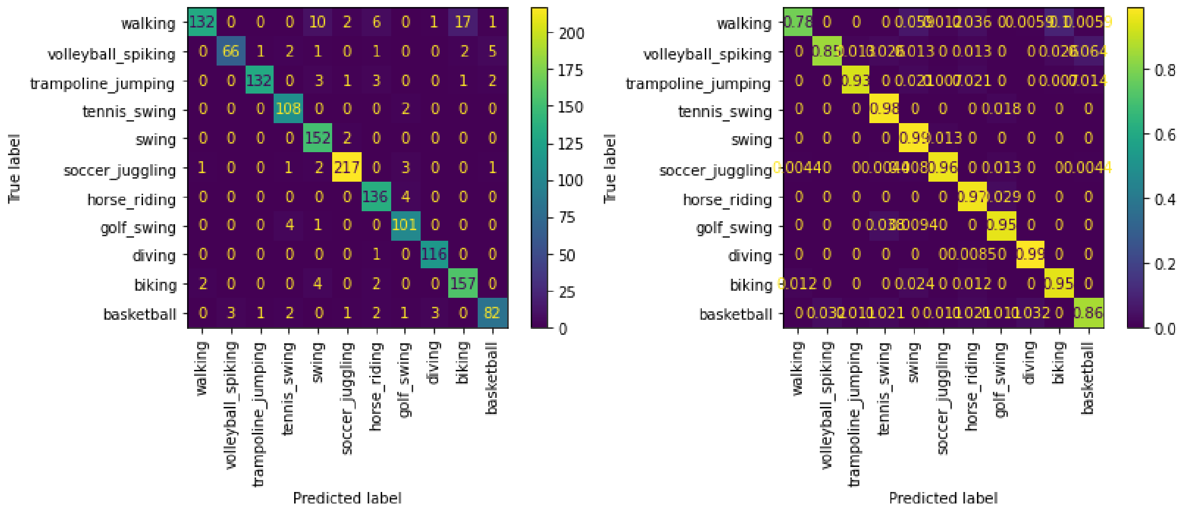
Figure 3.7: (a) Confusion Matrix (b) Normalized Confusion Matrix of C3D model on test set of UCF11[45] dataset model with RGB representation.

Performance Measure	**Score**
Accuracy	0.8907
Precision	0.8974
Recall	0.8907
F1 - score	0.8887
AUC - score(one vs one)	0.9947
AUC - score(one vs rest)	0.9947

Table 3.1: Performance of C3D model UCF11[45]

3.2.2 UCF Sports Action Dataset

UCF Sports action dataset[60] consists of 150 sequences with a resolution of 720 × 480 collected from broadcast television channels such as BBC and ESPN. The dataset includes 10 actions like diving, golf swing, kicking, lifting, riding-horse, running, skateboarding, swing-bench, swing-side, and walking. Video frames from the UCF Sports



a Confusion Matrix

b Normalized Confusion Matrix

Figure 3.8: (a) Confusion Matrix (b) Normalized Confusion Matrix of our SAC3D model on test set of UCF11[45] dataset model with RGB representation.

Performance Measure	Score
Accuracy	0.9320
Precision	0.9348
Recall	0.9312
F1 - score	0.9978
AUC - score(one vs one)	0.9976
AUC - score(one vs rest)	0.9996

Table 3.2: Performance of SAC3D on UCF11[45]

action dataset represent natural and genuine actions from different perspectives in a wide variety of scenes.

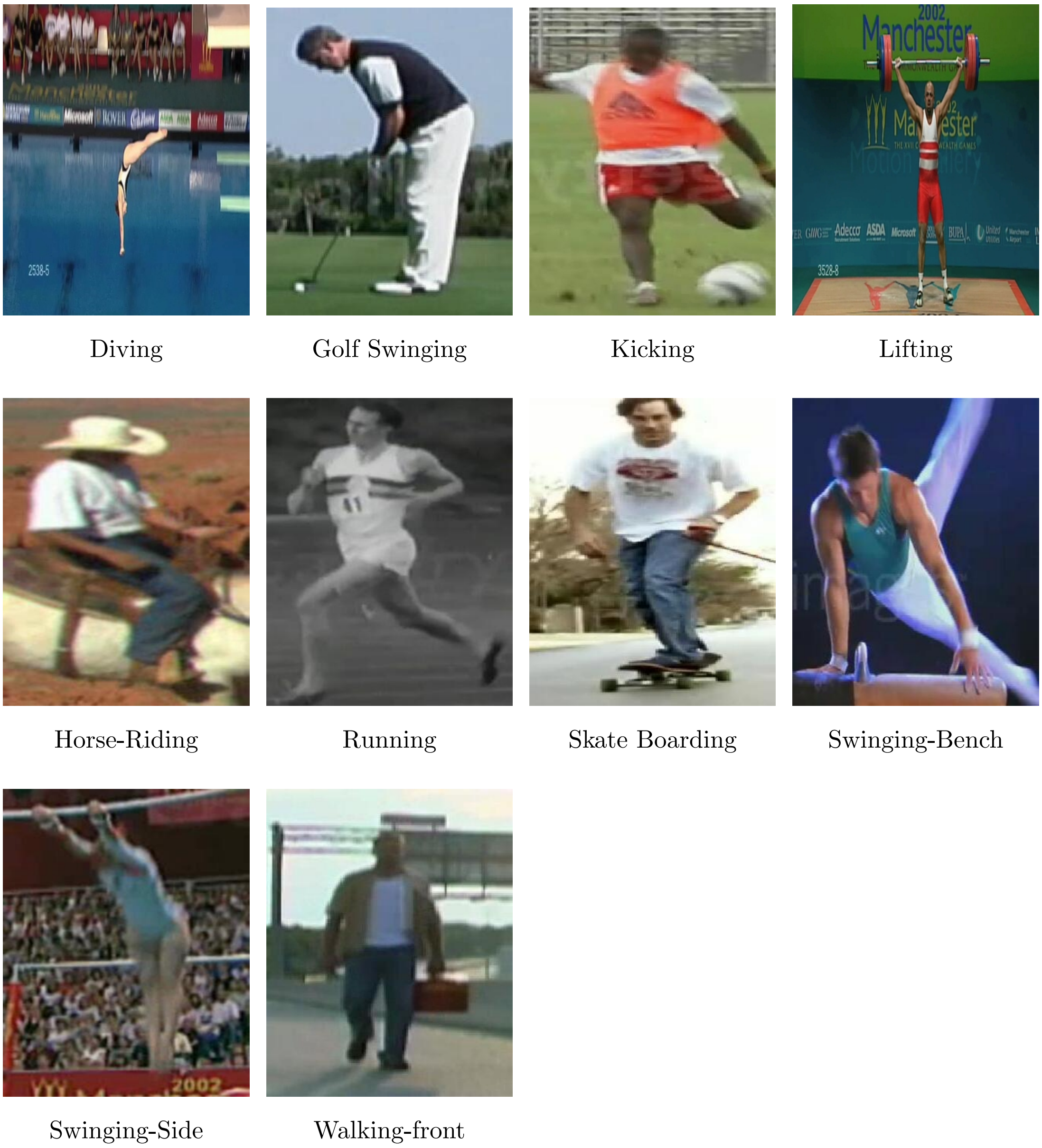
3.2.2.1 Training and Testing details

In this work, to decrease background correlation between the training and test sets, we split the dataset into 107 training samples and 43 test samples, as suggested by

Method	*Accuracy*
Liu et al. [45]	71.20
Ravanbakhsh et al. [59]	77.10
Sharma et al. [67]	84.96
Liu et al. [43]	89.70
Pan et al. [53]	86.90
Patel et al. [55]	89.43
Meng et al. [49]	89.70
Gharaee et al. [21]	89.50
Gammulle et al. [20]	89.20
Pan and Li [54]	89.24
Ullah et al. [74]	92.84
Wang et al. [76]	93.7
Javidani and Mahmoudi-Aznaveh [28]	95.73
Dai et al. [13]	96.90
Zebhi et al. [86]	93.4
Abdelbaky and Aly [1]	81.4
Muhammad et al. [51]	96.60
Akbar et al. [2]	100
Zebhi et al. [87]	97.13
Xiao et al. [82]	98.91
Saif et al. [62]	95.44
C3D*	89.07
SAC3D*	93.20

Table 3.3: Comparison with the state of the art methods on UCF11[45]

[37], ensuring a 70:30 balance for each action class. Additionally, clips in the training and testing sets could not come from the same video file. These sets were built to



Diving

Golf Swinging

Kicking

Lifting

Horse-Riding

Running

Skate Boarding

Swinging-Bench

Swinging-Side

Walking-front

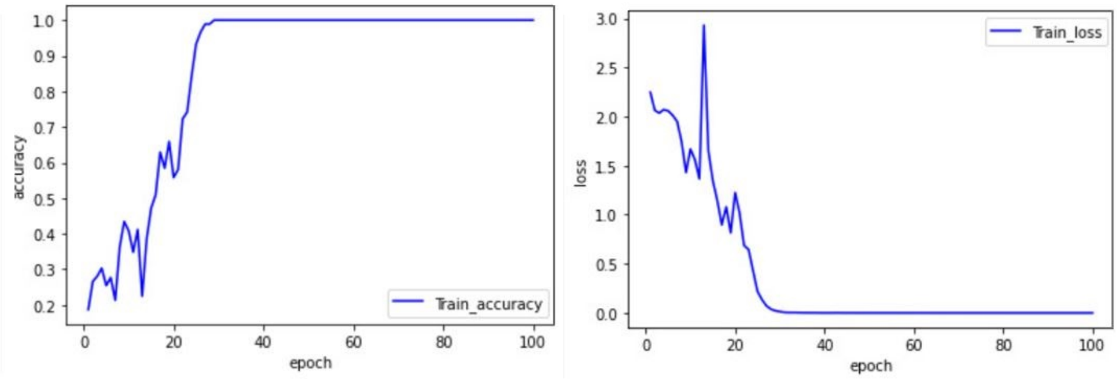
Figure 3.9: UCF Sports Action Dataset[60]

ensure that clips from the same video were not used for both training and testing. The proposed methodologies begin by extracting all the frames from each video. The next step is the formulation of continuous frame sequences from the extracted frames. We have used 16 frames as depth for each frame sequence of a video; thus, we divide the video into continuous segments with 16 frames. For each video segment, We first resize each frame to 128 × 171 spatial dimension, and then a center cropping is performed to limit spatial dimension 112 × 112. All the pixels in the cropped frames are scaled to limit the values $0 - 1$. We stack all preprocessed 16 consecutive frames to form an RGB frame sequence with dimension $16 \times 112 \times 112 \times 3$. The representation of each input video segment is a fixed size 4D tensor of dimension $16 \times 112 \times 112 \times 3$. The experiment was conducted with a batch size of 8 frame sequences. Finally, 8 of the frame sequence representations are batched together to obtain the 5D tensor of dimension $8 \times 16 \times 112 \times 112 \times 3$ as the final input to the model.

The next step in the experiment is to input the final representation of the video segment to the proposed model for feature learning. We have optimized our proposed model with an ADAM optimizer with an initial learning rate of 10^{-4} and using Sparse Categorical Cross Entropy as the loss function. We have proposed a transfer learning-a fine-tuning scheme by adopting the weights of pretrained C3D on the Sports1M dataset [31] for initializing the weight of 3D convolutional layers and fully connected layers. First, we freeze all model layers except the last dense output layer and the FC layers of the model and train the model for 30 epochs. We make all layers trainable and train the model for 70 epochs. We have experimented with other strategies of freezing and unfreezing layers with different epochs but have experimentally found this combination strategy to result in optimal performance accuracy for the model. Figure 3.11 presents the model's training accuracy and loss plots. We also train our model baseline architecture, C3D, with the same training scheme mentioned above. Figure 3.10 presents the C3D model's training accuracy and loss plots and validation accuracy and loss plots.

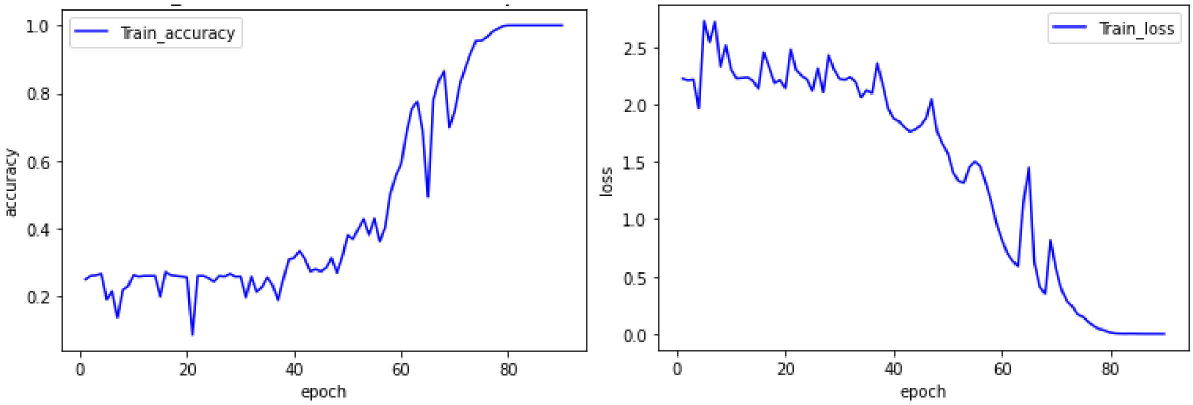
We adopt the same steps for the input test video as the training video. We first segment the input test video into continuous video segments. We next obtain the final input representation for each test video segment using the same preprocessing steps as during training. Finally, we predict action categories for each test video segment and evaluate our SAC3D model performance based on the results obtained by comparing the predictions with the ground truth label of each video segment.

.



a Training accuracy **b** Training loss

Figure 3.10: (a) Training accuracy (b) Training loss for C3D model on UCF sports action[60] dataset with RGB representation.

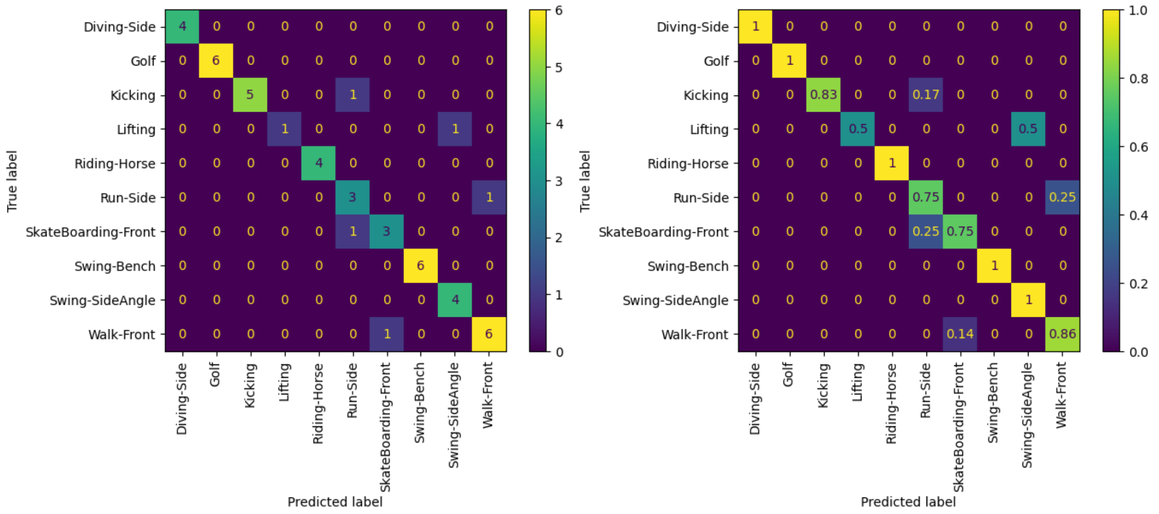


a Training accuracy **b** Training loss

Figure 3.11: (a) Training accuracy (b) Training loss for SAC3D model on UCF sports action[60] dataset with RGB representation.

3.2.2.2 Results

Our proposed model achieves an accuracy of 93.62% during testing. Our model excels in the performance with the baseline C3D model's accuracy of 89.37% by roughly 4%. Tables 3.4 and 3.5 present the performance of the C3D and our proposed SAC3D model. Our proposed model, SAC3D, beats the baseline C3D model on all of the performance measures of accuracy, precision, recall, f1-score, and AUC score. We further present in the figures 3.12 and 3.13 the confusion matrix and normalized confusion matrix the class-wise performance of the C3D and our SAC3D model, respectively. We compare our SAC3D model with other state-of-the-art methods on the UCF Sports action dataset in table 3.6.



True label	Diving-Side	Golf	Kicking	Lifting	Riding-Horse	Run-Side	SkateBoarding-Front	Swing-Bench	Swing-SideAngle	Walk-Front
Diving-Side	4	0	0	0	0	0	0	0	0	0
Golf	0	6	0	0	0	0	0	0	0	0
Kicking	0	0	5	0	0	1	0	0	0	0
Lifting	0	0	0	1	0	0	0	0	1	0
Riding-Horse	0	0	0	0	4	0	0	0	0	0
Run-Side	0	0	0	0	0	3	0	0	0	1
SkateBoarding-Front	0	0	0	0	0	1	3	0	0	0
Swing-Bench	0	0	0	0	0	0	0	6	0	0
Swing-SideAngle	0	0	0	0	0	0	0	0	4	0
Walk-Front	0	0	0	0	0	0	1	0	0	6

a Confusion Matrix

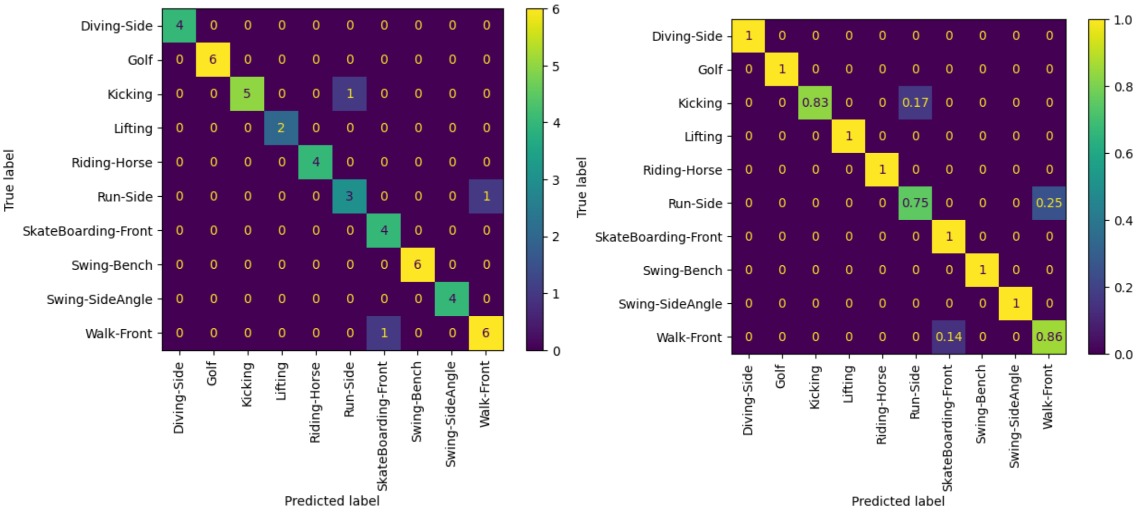
True label	Diving-Side	Golf	Kicking	Lifting	Riding-Horse	Run-Side	SkateBoarding-Front	Swing-Bench	Swing-SideAngle	Walk-Front
Diving-Side	1	0	0	0	0	0	0	0	0	0
Golf	0	1	0	0	0	0	0	0	0	0
Kicking	0	0	0.83	0	0	0.17	0	0	0	0
Lifting	0	0	0	0.5	0	0	0	0	0.5	0
Riding-Horse	0	0	0	0	1	0	0	0	0	0
Run-Side	0	0	0	0	0	0.75	0	0	0	0.25
SkateBoarding-Front	0	0	0	0	0	0.25	0.75	0	0	0
Swing-Bench	0	0	0	0	0	0	0	1	0	0
Swing-SideAngle	0	0	0	0	0	0	0	0	1	0
Walk-Front	0	0	0	0	0	0	0.14	0	0	0.86

b Normalized Confusion Matrix

Figure 3.12: (a) Confusion Matrix (b) Normalized Confusion Matrix of C3D model on test UCF sports action[60] dataset with RGB representation.

Performance Measure	**Score**
Accuracy	0.8936
Precision	0.9063
Recall	0.8936
F1 - score	0.8965
AUC- score(one vs one)	0.8929
AUC - score(one vs rest)	0.8956

Table 3.4: Performance of C3D model on UCF sports action[60]



a Confusion Matrix

b Normalized Confusion Matrix

Figure 3.13: (a) Confusion Matrix (b) Normalized Confusion Matrix of proposed SAC3D on test UCF sports action [60] dataset with RGB representation.

Performance Measure	**Score**
Accuracy	0.9362
Precision	0.9404
Recall	0.9361
F1 - score	0.9364
AUC - score(one vs one)	0.9363
AUC - score(one vs rest)	0.9367

Table 3.5: Performance of SAC3D on UCF sports action [60]

3.3 Summary

In this chapter, we have proposed and implemented an action recognition framework, a 3D Self-Attention Convolutional Neural Network named SAC3D, on the RGB video clips of the UCF11 and UCF sports action dataset. In this work, we chose a standard 3D CNN architecture, C3D, as our baseline model. We attempted to allevi-

Method	*Accuracy*
Rodriguez et al. [60]	69.20
Lan et al. [37]	83.70
Wang and Schmid [77]	88.20
Ravanbakhsh et al. [59]	88.11
Wang et al. [78]	91.89
Meng et al. [49]	93.20
Gharaee et al. [21]	97.80
Gammulle et al. [20]	92.20
Liu et al. [43]	95.00
Tu et al. [72]	97.50
Nazir et al. [52]	97.30
Dai et al. [13]	98.60
Zebhi et al. [86]	92.6
Abdelbaky and Aly [1]	92.67
Kumar et al. [36]	96.8
Muhammad et al. [51]	99.10
Russel and Selvaraj [61]	99.26
Akbar et al. [2]	99.8
Xiao et al. [82]	97.84
Saif et al. [62]	95.74
C3D*	89.36
SAC3D*	93.62

Table 3.6: Comparison with the state of the art methods on UCF Sports action[60]

ate various limitations of 3D CNN architectures. We tackle the incompetence of C3D architecture to model long-range dependencies by incorporating a 3D Self-attention mechanism in the C3D model baseline. Encapsulating the 3D self-attention mechanism in the C3D model also helps the model improve its feature learning by allocating weighted attention to the overall information content. 3D self-attention mechanism allows our SAC3D model to focus on important global information and ignore redundant, irrelevant information. We tackle the overfitting issues caused by large training data requirements by adopting a transfer learning approach - a fine-tuning scheme. Overall, in this work, we effectively and efficiently performed action recognition on the UCF11 and UCF sports action video datasets and achieved an accuracy of 93.20% and 93.62% comparable to the available state-of-the-art results.

Chapter 4

3D Self Attention Convolutional Neural Network with 3D Discrete Wavelet Transform preprocessing for Action Recognition in videos

As discussed in Chapter 3, Human actions are localized in space and time and occur within a small portion of the frame across fewer frames. Therefore, emphasis on these specific, fast-changing regions is crucial for accurate action categorization. We proposed an additional 3D Discrete Wavelet Transform (DWT) preprocessing step on the input RGB frame sequences to enhance our proposed model performance. The 3D DWT, adept at analyzing low and high-frequency information in 3D signals, generates different wavelet subbands providing localized information in both frequency and space domains. We recombine selected wavelet subbands to form a representation that focuses on fast-changing information, improving the salient representation of foreground human actions against background scenes. The subsequent section discusses the effectiveness of this 3D DWT preprocessing in achieving better separation between foreground and background in the salient representation. Notably, Figures 4.1 and 4.3 illustrate the original RGB frame sequences, while Figures 4.2 and 4.4

depict the corresponding preprocessed DWT frame sequences, showcasing enhanced foreground action separation. We are building upon our previous work in Chapter 3, where we introduced a 3D Self-Attention Convolutional Neural Network named SAC3D by incorporating a 3D self-attention module (SAC3D) in the convolutional 3D model.

4.1 Proposed Method

In this chapter, we introduce a novel action recognition framework with our proposed 3D self-attention convolutional neural network, SAC3D, with 3D Discrete Wavelet Transform preprocessing on RGB video clips, further advancing feature learning capabilities of our proposed SAC3D model. Figure 4.5 provides the block diagram of the proposed model architecture. Our proposed approach involved training SAC3D on the motion salient representation of RGB frame sequences obtained through 3D DWT preprocessing. We address the overfitting limitation on small-scale datasets by employing a transfer learning-fine-tuning scheme. We adopted pretrained C3D model weights from the Sport1M[31] dataset for baseline convolutional and fully connected layers. The model is then trained and validated on benchmark UCF11[45] and UCF Sports Action[60] datasets. Following this introduction, we explained the components within our proposed action recognition framework, starting with the 3D Discrete Wavelet Transform-based preprocessing step, followed by specifics on the 3D self-attention module and configuration details for the SAC3D model.

4.1.1 3D Discrete Wavelet Transform (3D DWT) based Preprocessing

The 3D DWT is a computationally efficient and effective tool for analyzing low and high-frequency information in the 3D signal. When subjected to any discrete signal, the discrete wavelet transform (DWT) decomposes into sets of wavelets that can provide localized information in the frequency and space domain. The 3D DWT

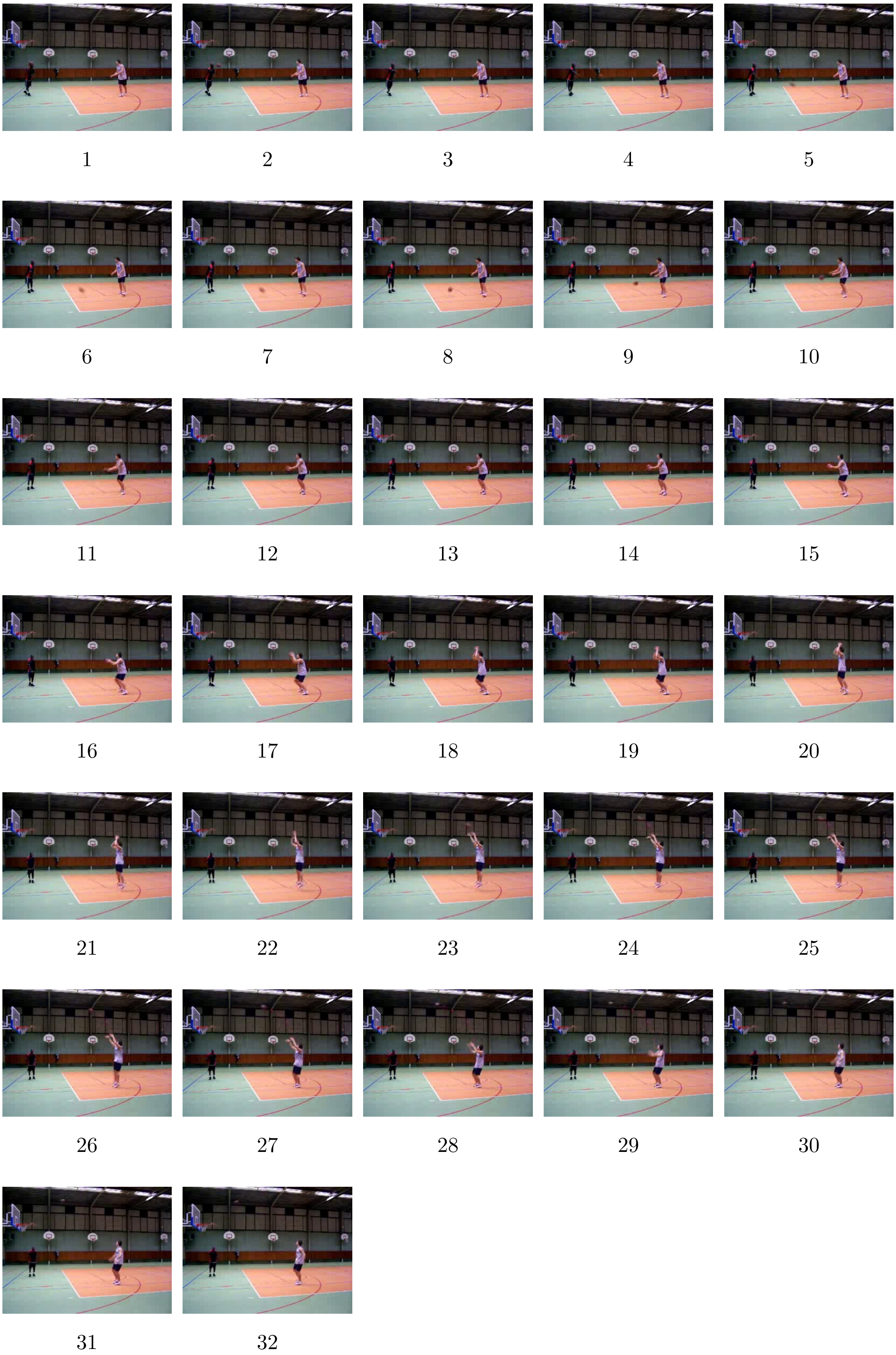


Figure 4.1: A sub-sampled 32-frame cube for a basketball video clip from UCF11[45]

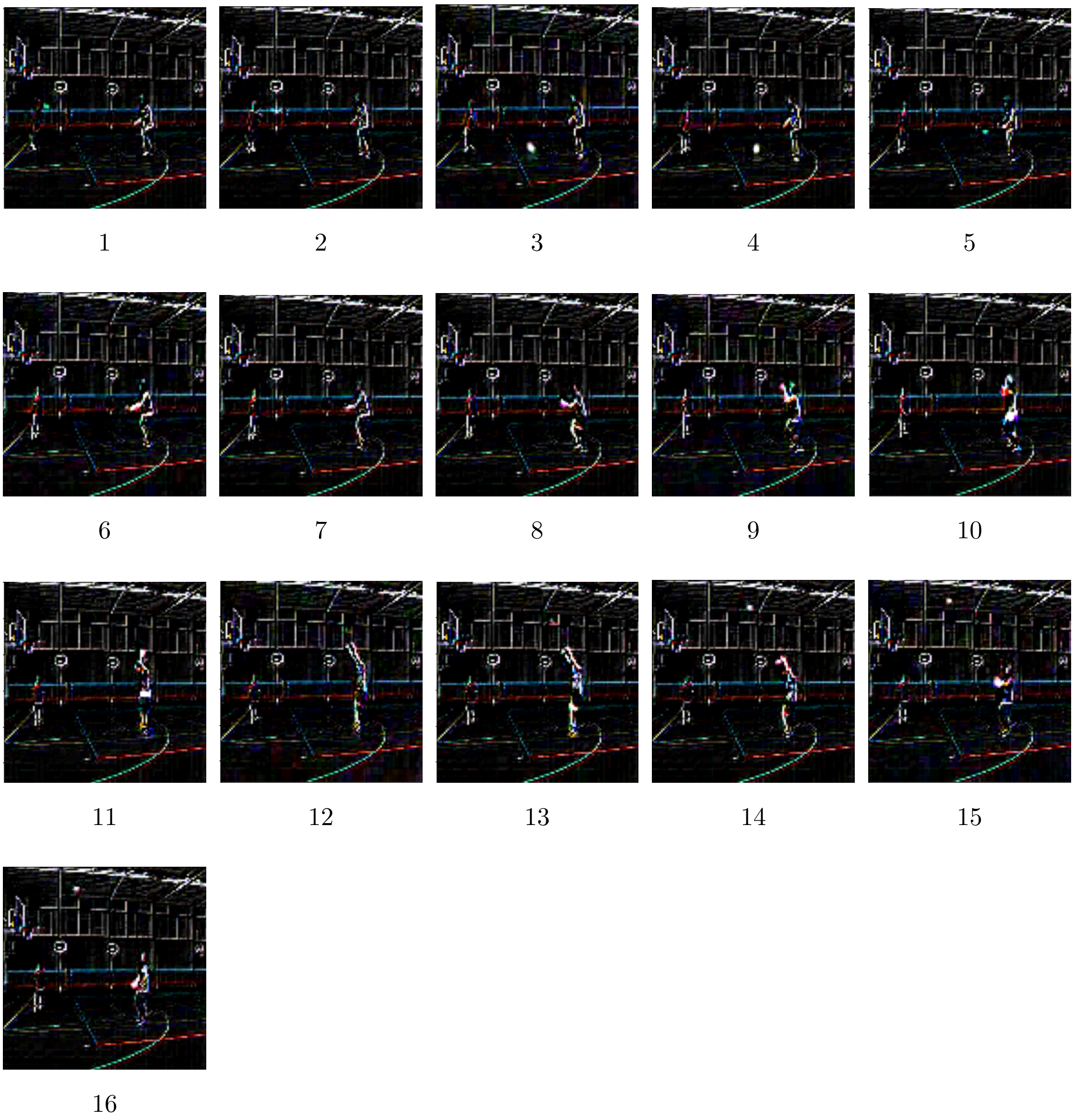


Figure 4.2: A 3D-DWT representation generated corresponding to the basketball video clip of UCF11[45] in figure 4.1

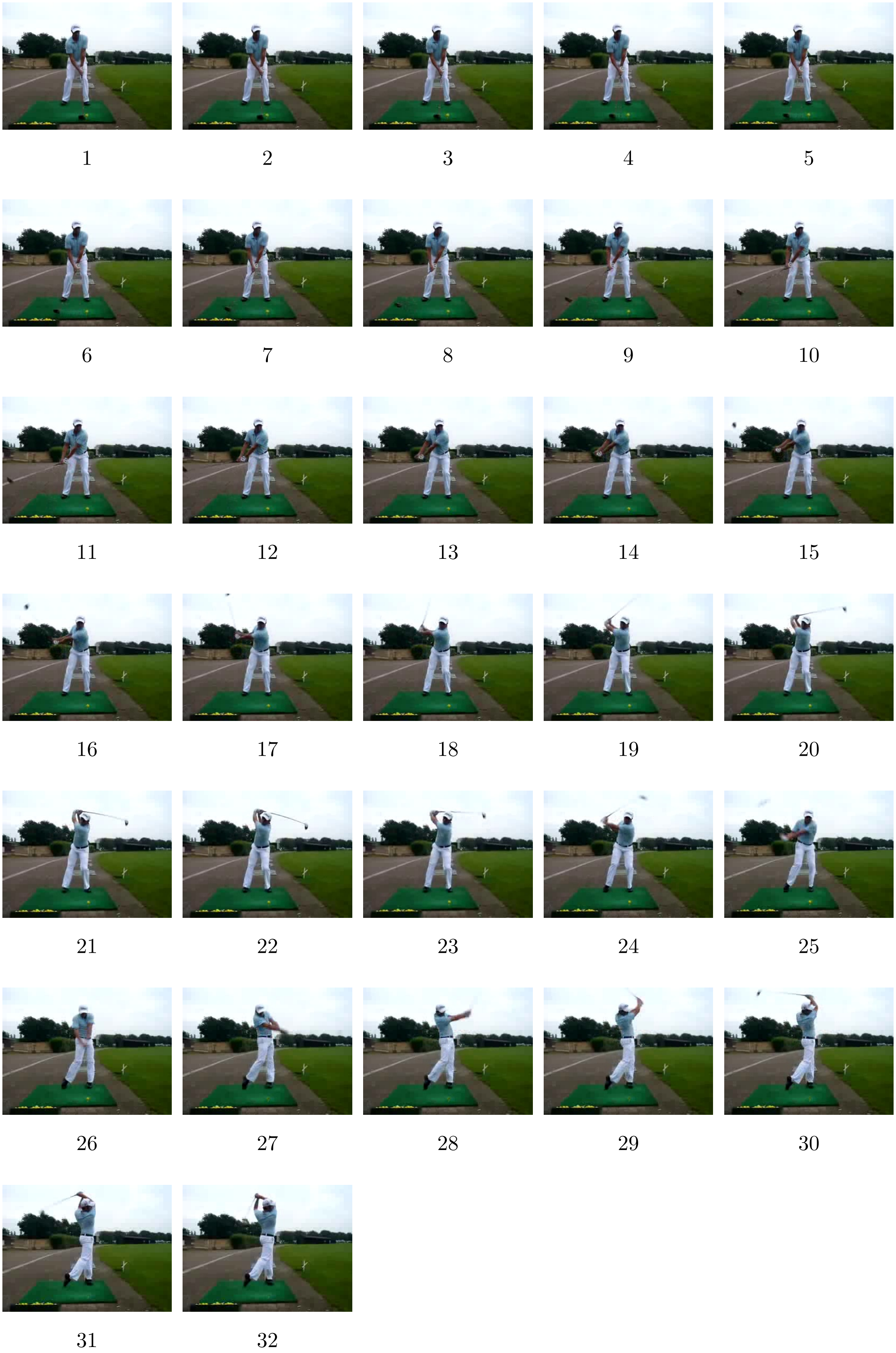
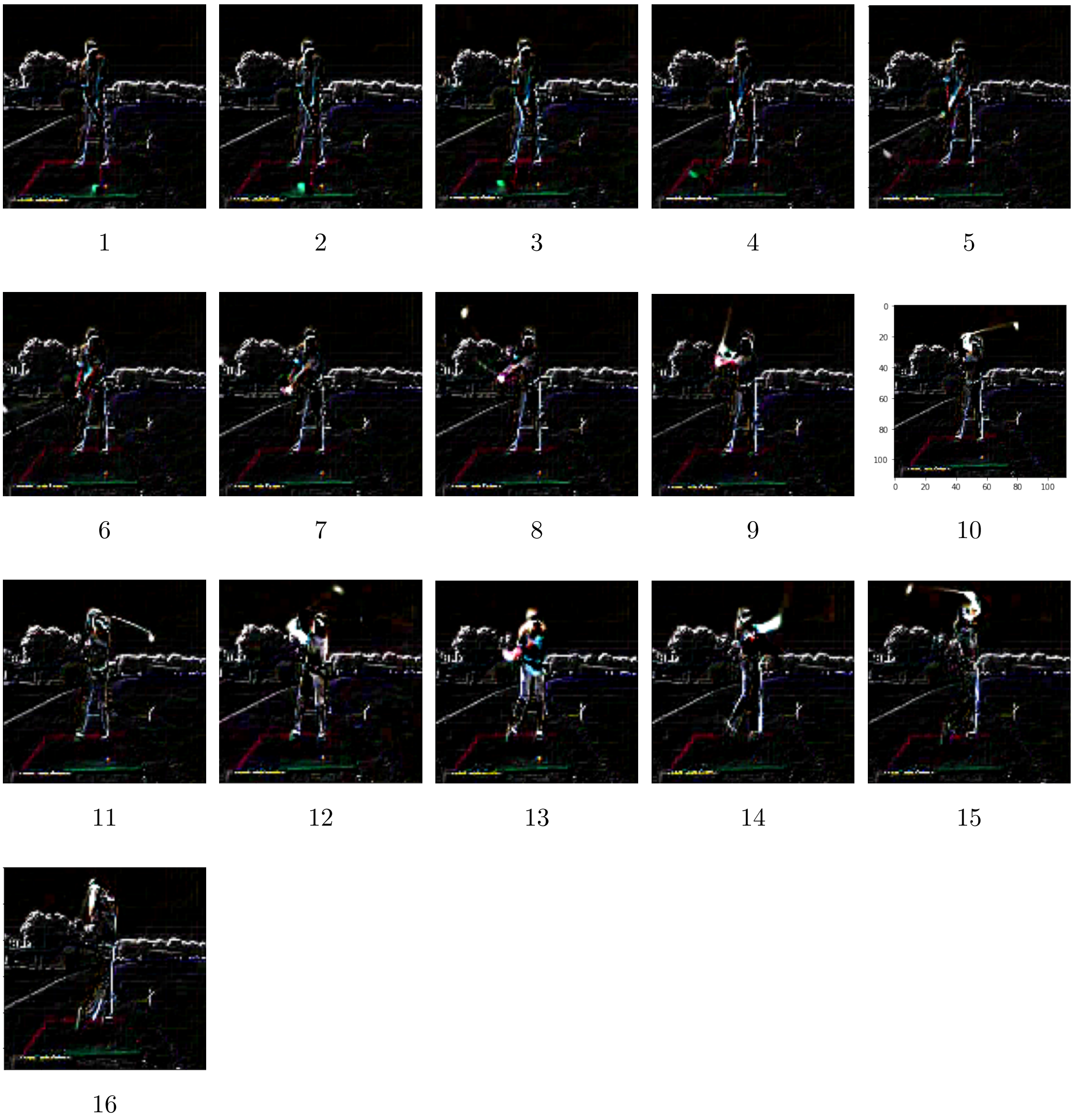


Figure 4.3: A sub-sampled 32-frame cube for a golf video clip from UCF11[45]



1 2 3 4 5

6 7 8 9 10

11 12 13 14 15

16

Figure 4.4: A 3D-DWT representation generated corresponding to the golf video clip of UCF11[45] in figure 4.3

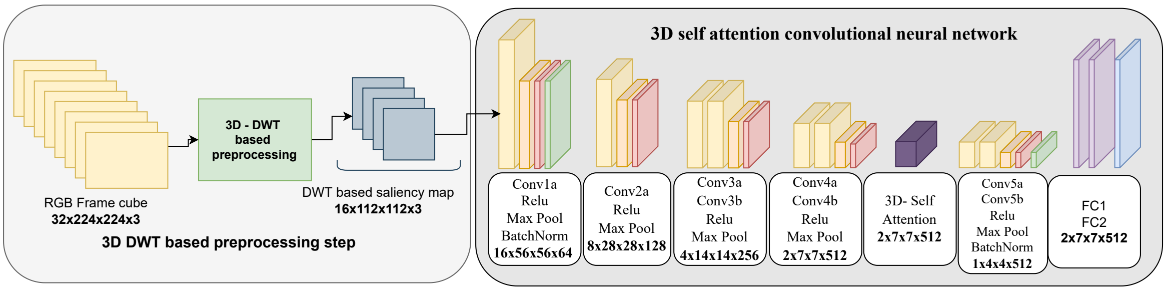


Figure 4.5: Block diagram of our proposed action recognition framework - SAC3D with 3D DWT preprocessing on RGB video clip

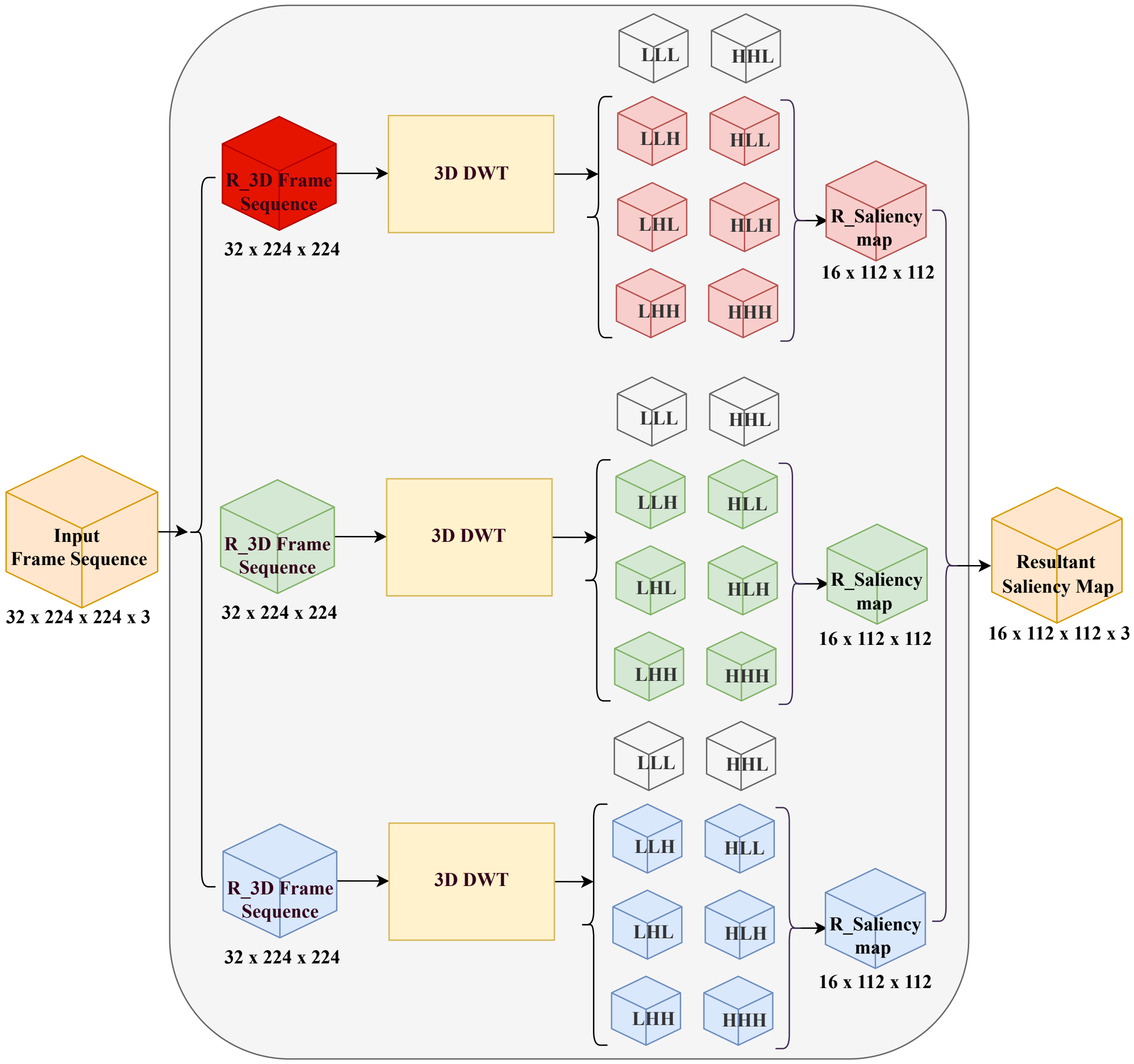


Figure 4.6: Proposed 3D Discrete Wavelet Transform preprocessing step.

is implemented by combining three 1D DWTs applied in three directions. It convolves the signal with a low-pass filter L and high-pass filter H along each dimension. Let V(x, y, t) represent the volume corresponding to the 3D signal. Implementing a one-level 3D DWT decomposes the signal into eight sub-volumes LLL, LLH, LHL, LHH, HLL, HLH, HHL, HHH, reducing the volume size by 8 with each dimension downsampled by a factor of 2. These eight subbands have specific spatiotemporal orientation properties. LLL subband captures the slowly moving average signal while LLH highlights the signal's quick average changes. LHL and HLL capture slowly changing vertical and horizontal features, respectively, whereas HHH captures the quickly changing diagonal features. Subbands LLH, LHH, HLH, and HHH relate to fast-changing foreground actions, while LHL, HLL, and HHL relate to background changes. Different combinations of these subbands capture distinct saliency regions within the input volume. We incorporate 3D DWT as a preprocessing step for the input frame sequences in a video. As outlined earlier, applying 3D DWT to each color channel yields eight wavelet subbands. These subbands, representing different spatiotemporal information, are combined across the color channels through average fusion, resulting in eight subbands for the entire RGB frame sequence. Experimentally, we fuse six wavelet coefficient bands, excluding LLL and HHL, to eliminate redundant spatiotemporal information. Figures 4.2 and 4.4 illustrate two preprocessed 16 frame sequences from the input videos, serving as input to the model.

4.1.2 3D self-attention module

Self-Attention is a powerful attention mechanism that effectively captures dependencies between two input positions. It computes a correlation matrix along different dimensions of the input sequence to capture the pairwise relationship between feature positions. The correlation value at each position is calculated through the weighted sum of all other positions. Considering all the positions while computing the similarity between pairwise positions of the input, the self-attention mechanism can identify and represent the strong and weak relationships in the input feature map. Based on the strength of the similarity relationships among pixels, it can infer the strong

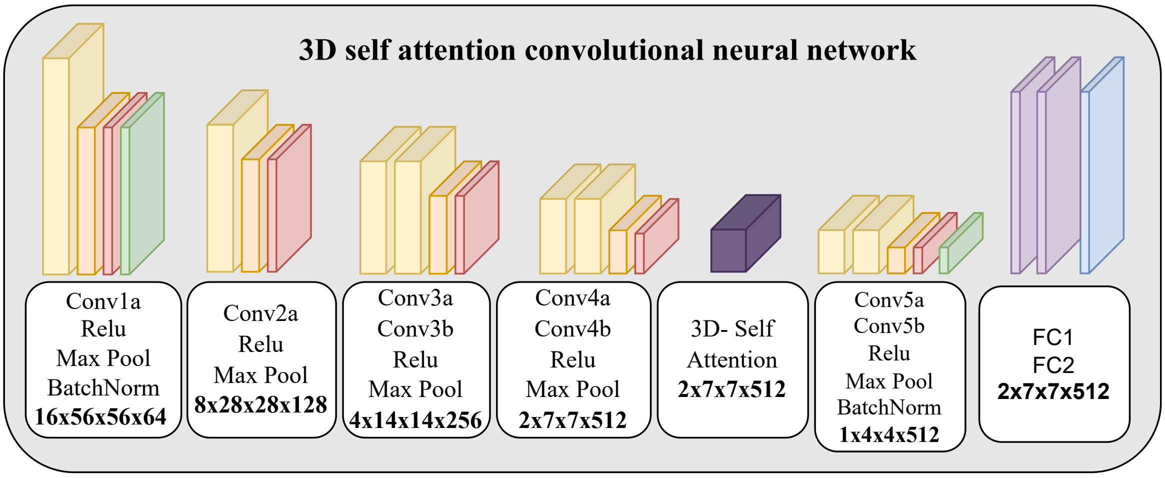


Figure 4.7: Block Diagram of proposed 3D self attention convolutional neural network (SAC3D)

global dependencies and weak local dependencies in the input sequence. Therefore, a self-attention mechanism aids in assigning the necessary relative importance to different parts of the video. Mathematically, a self-attention mechanism involves mapping a query, a key, and a value to the input feature map, where the query, key, and value are all derived from the input feature map. We categorize dependencies present in the spatiotemporal volume of the hidden feature map into intra-frame and inter-frame dependencies. Intra-frame dependencies are relationships in the pixels in spatial dimensions within each frame. In contrast, inter-frame dependencies are the relationships in the features across different frames along the temporal dimension. In this work, we capture the long-range spatiotemporal feature dependencies using a 3D Self-Attention mechanism. The proposed 3D self-attention mechanism gives specific attention to the intra-frame dependencies of the feature maps by the plane attention module; likewise, it provides specific attention to inter-frame dependencies by the temporal attention module. Let $X \in R^{D \times H \times W \times C}$ denote the hidden convolutional feature map that is input to the 3D self-attention module, where D is frame depth, H and W are the height, and width and C corresponds to the number of channels in the input feature map. We first apply $1 \times 1 \times 1$ convolution to the input feature map to form embedding vectors query $X_q \in R^{C \times D \times H \times W}$, *key* $X_k \in R^{C \times D \times H \times W}$ and *value*

$X_v \in R^{C \times D \times H \times W}$. These embedded vectors further served as input to the defined sub-modules below.

4.1.2.1 Plane attention module

To capture intra-frame dependencies within the feature map, we initially flatten the 2D spatial dimensions of the query X_q, key X_k, and value X_v vectors into 1D vectors. We transform each query X_q, key X_k, and value X_v vectors from $(C \times D \times H \times W)$ to flatten 1D vectors $\in (C \times D \times N)$ feature space, where N is the total number of pixels. We next perform the inner dot product of the transposed query vector ${X_q}^T(i)$ and key vector $X_k(j)$ to compute the correlation matrix s_{ij}.

$$s_{ij} = X_q(i)^T \times X_k(j) \tag{4.1}$$

' The correlation matrix describes the similarity-relationship between the features within each frame of the input feature map. The softmax function is employed to derive the self-attention map plane, denoted as a_{ij}, by normalizing the correlation matrix s_{ij}.

$$a_{ij} = SOFTMAX(s_{ij}) = \frac{\exp^{s_{ij}}}{\sum_{i=1}^{k} \exp^{s_{ij}}} \tag{4.2}$$

The final output of the plane attention module is calculated by the dot product of the value vector X_v to the weight matrix of self-attention map a_{ij}.

$$P_{att} = \sum_{i=1}^{N} X_v(i) \times a_{ij} \tag{4.3}$$

The output of the plane attention module P_{att} signifies the weighted strength of the pairwise feature dependencies within each frame of the input feature map.

4.1.2.2 Temporal attention module

It captures the intra-frame dependencies present in the features across the different frames of the input feature map over the temporal dimension. Similar to the plane attention module, we begin by transforming each of the queries X_q, key X_k, and value X_v vectors from $(C \times D \times H \times W)$ to $((H \times W) \times D \times C)$. We next perform the inner dot product of the transposed query vector ${X_q}^T(i)$ and key vector $X_k(j)$ to compute the correlation matrix s_{ij}.

$$s_{ij} = X_q(i)^T \times X_k(j) \tag{4.4}$$

The correlation matrix s_{ij} describes the similarity-relationship between the features that exist across the different frames of the input feature map. The softmax function obtains the temporal self-attention map $a_(ij)$ upon normalizing the correlation matrix s_{ij}.

$$a_{ij} = SOFTMAX(s_{ij}) = \frac{\exp^{s_{ij}}}{\sum_{i=1}^{k} \exp^{s_{ij}}} \tag{4.5}$$

The final output of the temporal attention module is calculated by the dot product of the value vector X_v to the weight matrix of self-attention map a_{ij}.

$$D_{att} = \sum_{i=1}^{N} X_v(i) \times a_{ij} \tag{4.6}$$

The output of the temporal attention module signifies the strength of the pairwise feature dependencies that exist across different frames of the input feature map.

4.1.2.3 Attention fusion module

The output of the plane attention module and temporal attention module presents the strength of the feature dependencies within each frame and across the frame sequence of the input feature map. We perform a weighted fusion of the outputs of both the plane and temporal attention modules to obtain the overall strength of the spatio-temporal feature dependencies of the input feature map. We use a trainable weight parameter γ to provide the best fusion of the attention to input feature map based on the performance. We add the results of the plane and depth attention module output to the input feature map through a residual connection to obtain the final output Y of the 3D self-attention module.

$$Y = \left[P_{att} + D_{att} \right] * \gamma + X \tag{4.7}$$

An overview of our proposed 3D Self-attention convolutional neural network model (SAC3D) is shown in figure 3.2. We have extended the popular 3D CNN architecture-C3D by incorporating a 3D self-attention mechanism. An overview of C3D architecture[70] is shown in figure 3.1. We followed the modular design of the C3D architecture. Our SAC3D model comprises five convolutional blocks with eight 3D-convolutional layers with 64, 128, 256, 256, 512, 512, 512, and 512 feature channels, respectively. Each convolutional block has a 3D max-pooling layer followed by ReLU activation. Our proposed model further consists of two batch normalization layers, two fully connected FC layers, a 3D self-attention layer, and a dense softmax output layer. We have augmented the C3D baseline by adding two batch normalization layers after the first conv-block and another before starting FC layers. After experimental analysis, we have encapsulated a 3D self-attention layer between the fourth and fifth conv-block. We randomly initialized the weights of new layers: the 3D Self-attention layer, dense output layer, and batch normalization layers. We have adopted the weights of the 3D convolutional layers and FC layers from pretrained C3D on the sport1M [31] dataset.

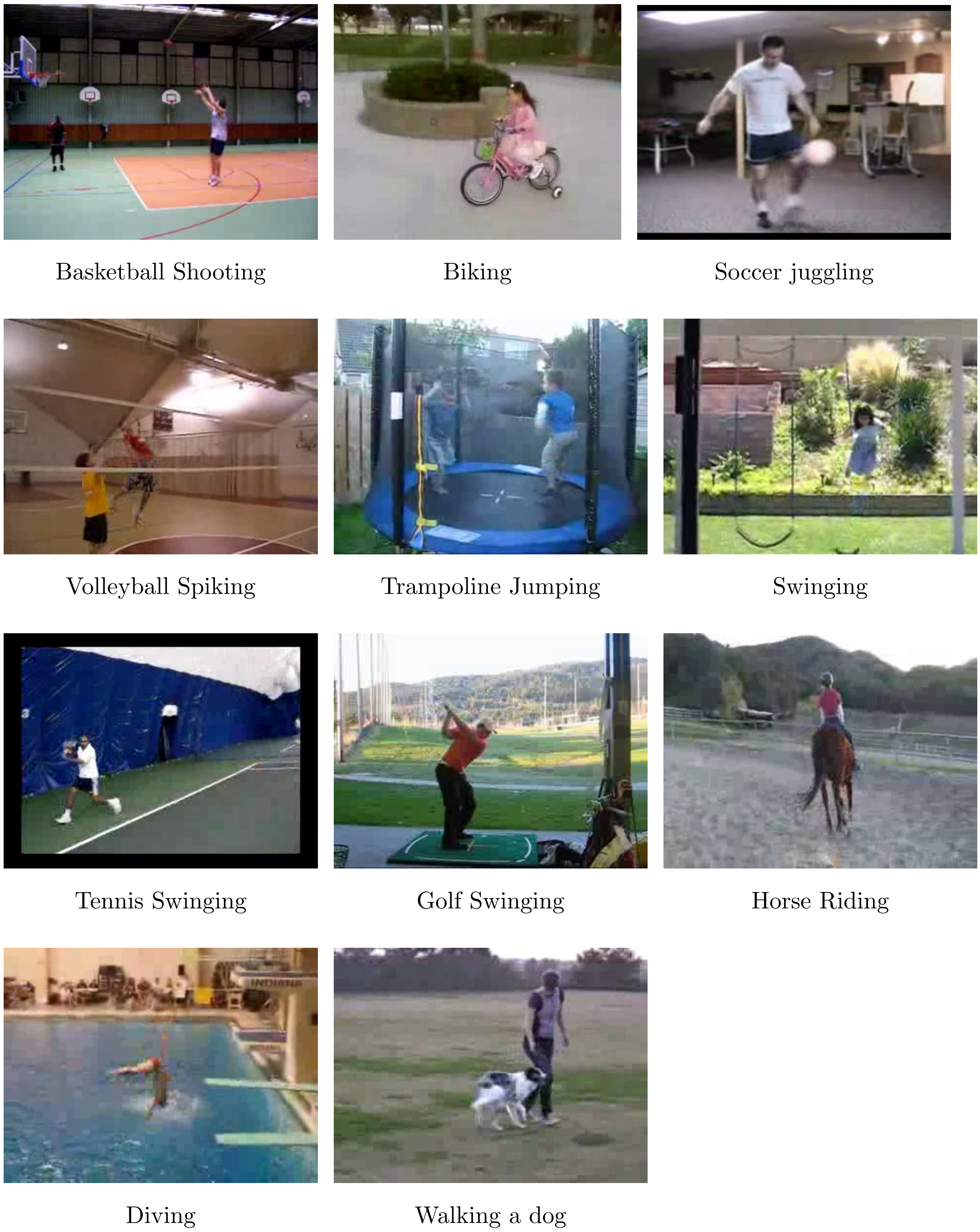
4.2 Experiments

In this section, we have evaluated our proposed action recognition framework – a 3D self-attention convolutional neural network (SAC3D) with 3D DWT preprocessed

RGB frame sequences as input. We conducted extensive experiments in determining the effectiveness of the proposed augmentation to C3D architecture upon encapsulating the proposed 3D self-attention module to improve recognition accuracy as the performance measure. We have evaluated the point of adding a 3D DWT-based preprocessing step to improve recognition accuracy further. We showcase the results of our proposed model, a 3D self-attention convolutional neural network (SAC3D) on a 3D DWT preprocessed RGB frame sequence of videos for action recognition on the UCF11[45] and UCF Sports action[60] datasets. We further provide our proposed model's performance in evaluation metrics such as confusion matrix, precision, recall, f1-score, and AUC score. The remainder of the section contains details of the UCF11 dataset, experimental details and results, and a comparison of our proposed approach with state-of-the-art methods on the UCF11 dataset. Further, this section contains details of the UCF Sports action dataset, experimental details and results, and a comparison of our proposed approach with state-of-the-art methods on the UCF sport-actions dataset.

4.2.1 UCF11 Dataset

UCF11 (YouTube actions) [45] is a small benchmark video dataset for action recognition. The task of performing action recognition on the UCF11 dataset becomes challenging due to significant variations in camera motion, viewpoint, illumination conditions, the inconsistent appearance of variant scale objects, cluttered backgrounds, etc. The dataset has 11 sports action categories, including basketball shooting, soccer juggling, volleyball spiking, trampoline jumping, swinging, tennis swinging, golf swinging, biking/cycling, horseback riding, trampoline jumping, and walking with a dog. The dataset contains 25 groups with more than four video clips for each group in each category. The video clips within the same group share similar characteristics, such as having the same actor, similar backgrounds, similar viewpoints, etc.



Basketball Shooting

Biking

Soccer juggling

Volleyball Spiking

Trampoline Jumping

Swinging

Tennis Swinging

Golf Swinging

Horse Riding

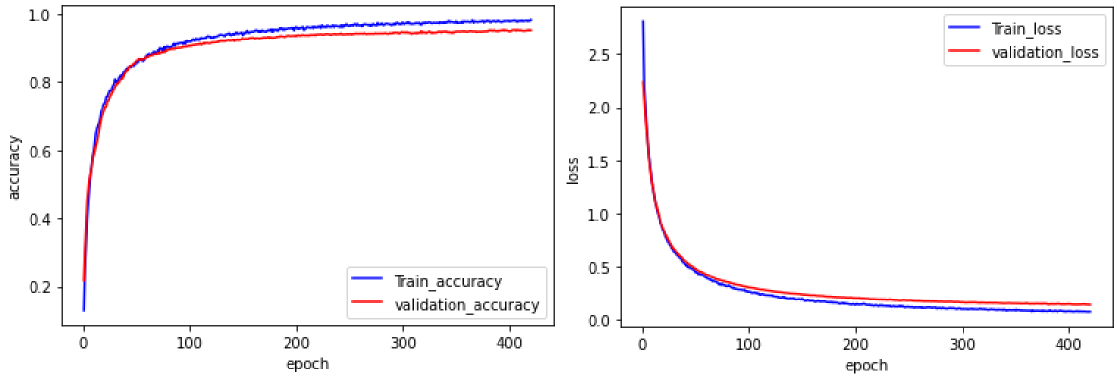
Diving

Walking a dog

Figure 4.8: UCF11 Dataset[45]

4.2.1.1 Training and Testing details

In this work, we have used the Leave-One-Group-Out cross-validation scheme as in [45] as a train and test dataset split for the experiments. The proposed methodologies begin by extracting all the frames from each video. The next step is the formulation of continuous frame sequences from the extracted frames. We have used 32 frames as depth for each frame sequence of a video; thus, we divide the video into continuous segments with 32 frames. For each video segment, We first resize each frame to 256×342 spatial dimension, and then a center cropping is performed to limit spatial dimension 224×224. All the pixels in the cropped frames are scaled to limit the values $0 - 1$. We next stack all of the preprocessed 32 consecutive frames to form an RGB frame sequence with dimension $32 \times 224 \times 224 \times 3$. We have proposed a novel representation of the input video using 3D DWT as described in 4.1.1 as an additional preprocessing step on the above extracted RGB-frame sequence. The final representation of each input video segment obtained is a fixed size $4D$ tensor of dimension $16 \times 112 \times 112 \times 3$. The batch size of 8 frame sequences has been for the experiment. Finally, 8 of these 3D DWT preprocessed frame sequence representations are batched together to obtain the $5D$ tensor of dimension $8 \times 16 \times 112 \times 112 \times 3$ as the final input to the model. The next step in the experiment is to input the final representation of the video segment to the proposed model for feature learning. We have optimized our proposed model with an ADAM optimizer with an initial learning rate of 10_{-4} and using Sparse Categorical Cross Entropy as the loss function. We have proposed a transfer learning- a fine-tuning scheme by adopting the weights of pretrained C3D on the Sports1M dataset [31] for initializing the weight of 3D convolutional layers and fully connected layers. We First freeze all model layers except the last dense output layer and train the model for 20 epochs. We unfreeze the FC layers of the model and train the model for another 30 epochs. We finally make all layers trainable and train the model for 370 epochs. We have also experimented with other strategies of freezing and unfreezing layers with different epochs but have experimentally found this combination strategy to result in optimal performance accuracy for the model. Figure 4.9



a Training and Validation accuracy

b Training and Validation loss

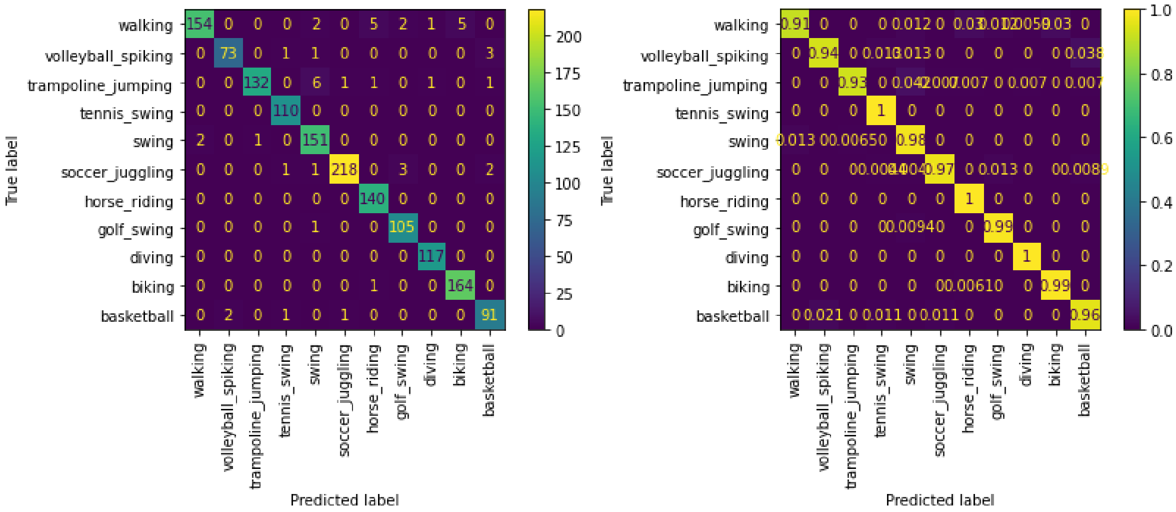
Figure 4.9: (a) Training accuracy (b) Training loss for our SAC3D model with DWT representation on UCF11 dataset.

presents the model's train and validation accuracy and loss plots during training.

We adopt the same steps for the input test video as for the training video. We first segment the input test video into continuous video segments. We next obtain the final input representation for each test video segment using the same preprocessing steps as during training. Finally, we predict action categories for each test video segment and evaluate our model performance based on the results obtained by comparing the predictions with the ground truth label of each video segment.

4.2.1.2 Results

Our proposed model SAC3D with 3D-DWT preprocessing achieves an accuracy of 96.93% during testing. Our SAC3D model with 3D-DWT preprocessing excels in the performance with the baseline C3D model's accuracy of 89.07% by roughly 8%. Tables 4.1 present the performance of the our proposed SAC3D model with 3D-DWT preprocessing. Our proposed model, SAC3D with 3D-DWT, beats the baseline C3D model and our previous model SAC3D on all of the performance measures of accuracy, precision, recall, f1-score, and AUC score. We further present in the figure 4.10 the confusion matrix and normalized confusion matrix the classwise performance of our SAC3D model with 3D-DWT preprocessing, respectively. We present the comparison



a Confusion Matrix

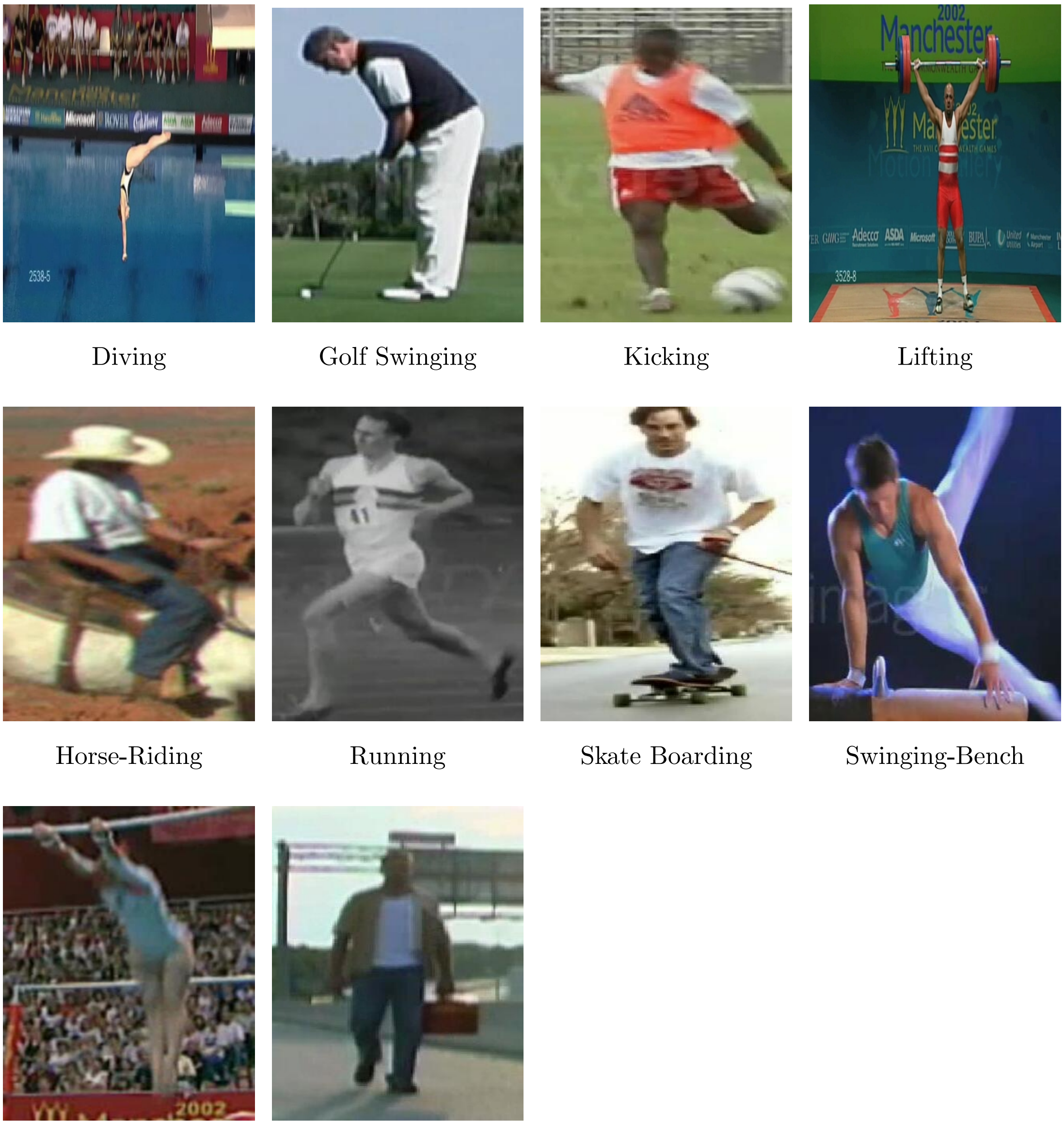
b Normalized Confusion Matrix

Figure 4.10: (a) Confusion Matrix (b) Normalized Confusion Matrix on test UCF11[45] dataset for proposed 3d self attention convolutional neural network on RGB Dataset.

of our SAC3D model with 3D-DWT with other state-of-the-art methods on the UCF11 dataset in table 4.2.

Performance Measure	**Score**
Accuracy	0.9693
Precision	0.9710
Recall	0.9693
F1 - score	0.9690
AUC- score(one vs one)	0.9997
AUC - score(one vs rest)	0.9996

Table 4.1: Performance of SAC3D with 3D DWT on UCF11[45]



Diving

Golf Swinging

Kicking

Lifting

Horse-Riding

Running

Skate Boarding

Swinging-Bench

Swinging-Side

Walking-front

Figure 4.11: UCF SPORT ACTION Dataset[60]

Method	*Accuracy*
Liu et al. [45]	71.20
Ravanbakhsh et al. [59]	77.10
Sharma et al. [67]	84.96
Liu et al. [43]	89.70
Pan et al. [53]	86.90
Patel et al. [55]	89.43
Meng et al. [49]	89.70
Gharaee et al. [21]	89.50
Gammulle et al. [20]	89.20
Pan and Li [54]	89.24
Ullah et al. [74]	92.84
Wang et al. [76]	93.7
Javidani and Mahmoudi-Aznaveh [28]	95.73
Dai et al. [13]	96.90
Zebhi et al. [86]	93.4
Abdelbaky and Aly [1]	81.4
Muhammad et al. [51]	96.60
Akbar et al. [2]	100
Zebhi et al. [87]	97.13
Xiao et al. [82]	98.91
Saif et al. [62]	95.44
C3D*	89.07
SAC3D*	93.20
SAC3D + 3D DWT*	96.93

Table 4.2: Comparison with the state of the art methods on UCF11[45]

4.2.2 UCF Sports Action Dataset

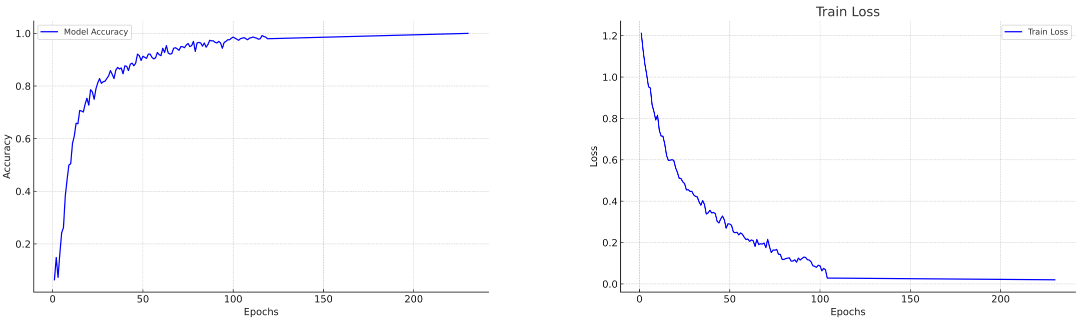
UCF Sports action dataset[60] consists of 150 sequences with a resolution of 720 × 480 collected from broadcast television channels such as BBC and ESPN. The dataset includes 10 actions like diving, golf swing, kicking, lifting, riding-horse, running, skateboarding, swing-bench, swing-side, and walking. Video frames from the UCF Sports action dataset represent natural and genuine actions from different perspectives in a wide variety of scenes.

4.2.2.1 Training and Testing details

In this work, to decrease background correlation between the training and test sets, we split the dataset into 107 training samples and 43 test samples, as suggested by [37], ensuring a 70:30 balance for each action class. Additionally, clips in the training and testing sets could not come from the same video file. These sets were built to ensure that clips from the same video were not used for both training and testing. The proposed methodologies begin by extracting all the frames from each video. The next step is the formulation of continuous frame sequences from the extracted frames. We have used 32 frames as depth for each frame sequence of a video; thus, we divide the video into continuous segments with 32 frames. For each video segment, We first resize each frame to 256×342 spatial dimension, and then a center cropping is performed to limit spatial dimension 224×224. All the pixels in the cropped frames are scaled to limit the values $0 - 1$. We next stack all of the preprocessed 32 consecutive frames to form an RGB frame sequence with dimension $32 \times 224 \times 224 \times 3$. We have proposed a novel representation of the input video using 3D DWT as described in 4.1.1 as an additional preprocessing step on the above extracted RGB-frame sequence. The final representation of each input video segment obtained is a fixed size $4D$ tensor of dimension $16 \times 112 \times 112 \times 3$. The batch size of 8 frame sequences has been used for the experiment. Finally, 8 of these 3D DWT preprocessed frame sequence representations are batched together to obtain the $5D$ tensor of dimension $8 \times 16 \times 112 \times 112 \times 3$ as the final input to the model. The next step in the experiment is to input the final representation of the video segment to the proposed model for feature learning. We have optimized our proposed model with an ADAM optimizer with an initial learning rate of 10_{-4} and using Sparse Categorical Cross Entropy as the loss function. We have proposed a transfer learning- a fine-tuning scheme by adopting the weights of pretrained C3D on the Sports1M dataset [31] for initializing the weight of 3D convolutional layers and fully connected layers. We First freeze all model layers except the last dense output layer and train the model for 20 epochs. We unfreeze the FC layers of the model and train the model for another 30 epochs. We finally make all layers trainable and train the model for 250 epochs. We have also experimented with other strategies of freezing and unfreezing layers with different epochs but have experimentally found this combination strategy to result in optimal performance accuracy for the model. Figure 4.12 presents the model's train and validation accuracy and loss plots. We adopt the same steps for the input test video as for the training video. We first segment the input test video into continuous video segments. We next obtain the final input representation for each test video segment using the same preprocessing steps as during training. Finally, we predict action categories for each test video segment and evaluate our model performance based on the results obtained by comparing the predictions with the ground truth label of each video segment.

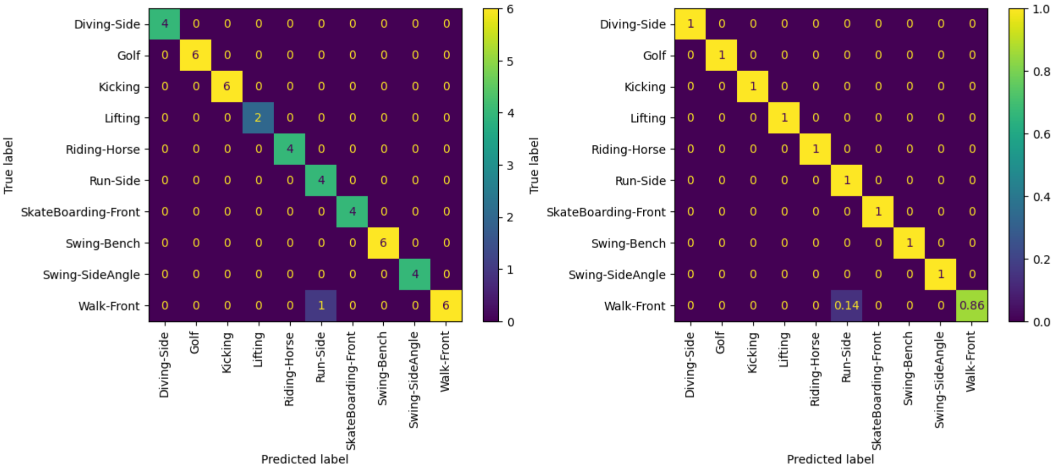
4.2.2.2 Results

Our proposed model achieves an accuracy of 97.87% during testing. Our model excels in the performance with the baseline C3D model's accuracy of 89.36% by roughly 8%. Tables 3.4 and 4.3 present the performance of the C3D and our proposed SAC3D model with 3D-DWT. Our proposed model, SAC3D, beats the baseline C3D model on all of the performance measures of accuracy, precision, recall, f1-score, and AUC score. We further present in the figures 3.12 and 4.13 the confusion matrix and normalized confusion matrix the classwise performance of the C3D and our SAC3D model with 3D-DWT, respectively. We present the comparison of our SAC3D model with other state-of-the-art methods on the UCF Sports action dataset in table4.4.



a Training accuracy **b** Training loss

Figure 4.12: (a) Training accuracy (b) Training loss for SAC3D model on UCF sports action[60] dataset with 3D DWT preprocessed representation.



a Confusion Matrix **b** Normal Confusion Matrix

Figure 4.13: (a) Confusion Matrix (b) Normalized Confusion Matrix of proposed SAC3D on test UCF sports action [60] dataset with 3D DWT preprocessed representation.

4.3 Summary

In this chapter, we proposed a novel action recognition framework, a 3D Self-Attention Convolutional Neural Network with Discrete Wavelet Transform on RGB video clips of the UCF11 and UCF Sports action dataset. We chose a 3D CNN architecture C3D as our baseline model in this work. We attempt to alleviate various

Performance Measure	**Score**
Accuracy	0.9787
Precision	0.9829
Recall	0.9787
F1 - score	0.9791
AUC- score(one vs one)	0.9787
AUC - score(one vs rest)	0.9789

Table 4.3: Performance of 3D self convolutional neural network with 3D DWT on UCF Sports action[60]

limitations of 3D CNN architectures. We tackle the incompetence of C3D to model long-range dependencies by incorporating a 3D Self-attention mechanism in the C3D model baseline. The encapsulation of the 3D self-attention mechanism in the C3D model also helps the model improve its feature learning by allocating appropriate weightage to the information content. 3D self-attention mechanism allows the model to focus on important information and ignore redundant irrelevant information. We tackle the overfitting issues caused by large training data requirements by adopting a transfer learning approach - fine-tuning scheme. We adopt the weights of the 3D convolutional and fully connected layers of the pretrained C3D model on the Sport1M dataset to initialize our model layer. Instead of using RGB representation of videos as input to our model, In this chapter, we proposed to use 3D Discrete Wavelet Transform as preprocessing step. Additional 3D DWT preprocessing of RGB video clips produces a motion saliency representation that localizes action in the space and frequency domain. This motion saliency representation of video is a more sparse and discriminative representation that enhances the model's learning capability. It also mitigates the significant computational training time of 3D CNN required, owing to sparseness in data representation. Overall, in this work, we effectively and efficiently performed action recognition on the UCF11 and UCF Sports action video datasets

Method	*Accuracy*
Rodriguez et al. [60]	69.20
Lan et al. [37]	83.70
Wang and Schmid [77]	88.20
Ravanbakhsh et al. [59]	88.11
Wang et al. [78]	91.89
Meng et al. [49]	93.20
Gharaee et al. [21]	97.80
Gammulle et al. [20]	92.20
Liu et al. [43]	95.00
Tu et al. [72]	97.50
Nazir et al. [52]	97.30
Dai et al. [13]	98.60
Zebhi et al. [86]	92.6
Abdelbaky and Aly [1]	92.67
Kumar et al. [36]	96.8
Muhammad et al. [51]	99.10
Russel and Selvaraj [61]	99.26
Akbar et al. [2]	99.8
Xiao et al. [82]	97.84
Saif et al. [62]	95.74
C3D*	89.37
SAC3D*	93.62
SAC3D + 3D DWT*	97.87

Table 4.4: Comparison with the state of the art methods on UCF Sports action[60]

and achieved an accuracy of 96.93% and 97.87% comparable to the available state-of-the-art results.

Chapter 5

Conclusion and Future Work

This thesis predominantly delved into deep learning for enhancing action recognition in videos. Our dedicated efforts were directed towards refining every facet of the action recognition framework, from preprocessing techniques to the deployment of advanced neural network architectures. By addressing key challenges and incorporating innovative methodologies, we aimed to elevate the overall efficacy of the action recognition pipeline. In this thesis, we produced two end-to-end trainable architectures in chapters 3 and 4. As previously discussed, our approach to action recognition encompasses two integral subtasks. The first involves crafting a robust representation of the input video data, a pivotal step that lays the foundation for the model's feature learning capability. The second subtask revolves around the meticulous design and development of model architecture, ensuring efficiency and effectiveness in the task of action recognition. Together, these subtasks form a comprehensive strategy to tackle the complexities of action recognition. In chapter 3, We designed and developed a 3D self-attention convolutional neural network (SAC3D) for action recognition on RGB video data. In this work, we focused on improving the feature learning capabilities of 3D CNN, the prevalent choice of baseline architecture for video action recognition. We used a simple standard 3D CNN architecture, C3D, for our model baseline and improved its action recognition accuracy by overcoming its drawbacks. We incorporated a 3D self-attention mechanism into the C3D model baseline to enhance its feature representation learning. In Chapter 4, we extended the SAC3D model architecture introduced in our previous work to enhance its input representation for video data. This extension involved the incorporation of a preprocessing step based on the 3D Discrete Wavelet Transform (3D DWT). By leveraging the 3D DWT, we obtained a motion-salient representation of the input video data. The introduction of this preprocessing step contributed significantly to refining the action recognition accuracy of our SAC3D model, further solidifying the effectiveness of our proposed approach. Further, We used transfer learning – a fine-tuning scheme to counteract the excessive need for training data. We developed and tested our architectures on the benchmark UCF11 and UCF sports action datasets. The results obtained in chapters 3 and 4 exhibited the efficacy of our works in improving the action recognition of C3D architecture. Our approaches outperform many state-of-the-art methods, showing their precedence in video action recognition. Further, we summarize the contributions achieved in this thesis in section 5.1 followed by the possible future directions in section 5.2.

5.1 Summary of Contributions

We have achieved the objectives specified in Section 1.4 in this thesis by making the following main contributions:

1. **3D Self-Attention Convolutional Neural Network for action recognition in videos**
   We introduced a pioneering action recognition framework, the 3D Self-Attention Convolutional Neural Network, designed to process RGB video clips from the UCF11 and UCF sports action datasets. Adopting the C3D architecture as our baseline model, we aimed to address several limitations inherent in 3D CNN architectures. To address the challenge of C3D's inability to model long-range dependencies, we introduced a 3D self-attention mechanism into the C3D model baseline. This enhancement not only overcame the incompetence in capturing long-range dependencies but also facilitated improved feature learning. The incorporation of the 3D self-attention mechanism enabled the model to allocate appropriate weightage to information, focusing on crucial details while ignoring

redundant and irrelevant data. Additionally, to tackle overfitting issues arising from the substantial training data requirements of the 3D CNN model, we adopted a transfer learning approach—a fine-tuning scheme. Leveraging the weights of the 3D convolutional and fully connected layers from the pre-trained C3D model on the Sport1M dataset, we initialized the corresponding layers of our model. Overall, This work successfully executed action recognition on the UCF11 and UCF sports action video datasets, attaining remarkable accuracies of 93.20% and 93.62%, respectively. These achievements stand on par with or even surpass the current state-of-the-art results, underscoring the effectiveness and efficiency of our proposed approaches.

2. **3D Self Attention Convolutional Neural Network with 3D Discrete Wavelet Transform preprocessing for Action Recognition in videos** we developed a novel action recognition framework, a 3D Self-Attention Convolutional Neural Network with Discrete Wavelet Transform preprocessing on RGB video clips of the UCF11 and UCF sports action dataset. In this work, we deployed an additional 3D Discrete Wavelet Transform-based preprocessing step on the input RGB frame sequences. Applying the 3D DWT to the input RGB frame sequences generates different wavelet subbands, which provide localized information in the frequency and space domain. We formulated a representation that recombined specifically selected subbands that concentrate on fast-changing information. The formulated representation had a good separation of the foreground performed human action to the background scenes. The 3D DWT processed representation facilitated the feature learning and thus improved the performance of our SAC3D model. Overall, This work elevated the feature representation quality of SAC3D by training the model on the motion saliency representation instead of the raw RGB frame sequence. This strategic modification resulted in the successful execution of action recognition on the UCF11 and UCF sports action video datasets, yielding impressive accuracies of 96.90% and 97.87%, respectively. These results align with or surpass exist-

ing state-of-the-art benchmarks, underscoring the effectiveness of our proposed methodology.

5.2 Future Research Directions

As discussed above, our work in this thesis focuses on determining a better representation for the input video and developing a 3D CNN-based model for learning the feature representation. We understand that our current work can be explored in the following future directions:

1. **Two stream SAC3D architecture for action recognition** Two-stream network architecture designs have proven to improve action recognition accuracy in videos. As our next research step, we adopt the two-stream network architecture design. We plan to use our proposed SAC3D model with an RGB frame cube and 3D DWT preprocessed RGB frame cube as two input streams.

2. **Variant of 3D CNN as the baseline** In this thesis, we adopted the C3D model as the baseline architecture. Our proposed work schemes helped improve the action recognition accuracy of the C3D model baseline. As our baseline model, we plan to experiment with more refined 3D CNN models such as ResNet3D[24], I3D[9], etc.

3. **3D DWT-based representation as an internal layer in the model architecture** In this thesis work, The proposed salient representation of the RGB frame cube obtained from the 3D DWT preprocessing step by including various subband coefficients is decided by experimental analysis. In our future work, we will include the 3D DWT preprocessing step as an internal layer to model architecture. Our assumption is a more refined representation of the RGB frame cube through trainable weighted fusion architecture.

Bibliography

[1] A. Abdelbaky and S. Aly, "Human action recognition using three orthogonal planes with unsupervised deep convolutional neural network," Multimedia Tools and Applications, vol. 80, no. 13, pp. 20 019–20 043, 2021.

[2] M. N. Akbar, F. Riaz, A. B. Awan, M. A. Khan, U. Tariq, S. Rehman et al., "A hybrid duo-deep learning and best features based framework for action recognition," Computers, Materials & Continua, vol. 73, no. 2, pp. 2555–2576, 2022.

[3] M. N. Al-Berry, H. M. Ebied, A. S. Hussein, and M. F. Tolba, "Human action recognition via multi-scale 3d stationary wavelet analysis," in 2014 14th International Conference on Hybrid Intelligent Systems. IEEE, 2014, pp. 254–259.

[4] M. Al-Faris, J. Chiverton, D. Ndzi, and A. I. Ahmed, "A review on computer vision-based methods for human action recognition," Journal of imaging, vol. 6, no. 6, p. 46, 2020.

[5] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6836–6846.

[6] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding," arXiv preprint arXiv:2102.05095, vol. 2, no. 3, p. 4, 2021.

[7] R. A. Bhuiyan, S. Tarek, and H. Tian, "Enhanced bag-of-words representation for

human activity recognition using mobile sensor data," Signal, Image and Video Processing, vol. 15, no. 8, pp. 1739–1746, 2021.

[8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in European conference on computer vision. Springer, 2020, pp. 213–229.

[9] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[10] H. Chang, J. Chen, Y. Li, J. Chen, and X. Zhang, "Wavelet-decoupling contrastive enhancement network for fine-grained skeleton-based action recognition," arXiv preprint arXiv:2402.02210, 2024.

[11] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12 299–12 310.

[12] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," arXiv preprint arXiv:1911.03584, 2019.

[13] C. Dai, X. Liu, and J. Lai, "Human action recognition using two-stream attention based lstm networks," Applied soft computing, vol. 86, p. 105820, 2020.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

[15] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "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.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv preprint arXiv:2010.11929, 2020.

[17] E. Essa and I. R. Abdelmaksoud, “Temporal-channel convolution with self-attention network for human activity recognition using wearable sensors,” Knowledge-Based Systems, vol. 278, p. 110867, 2023.

[18] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, “Multiscale vision transformers,” in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 6824–6835.

[19] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1933–1941.

[20] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, “Two stream lstm: A deep fusion framework for human action recognition,” in 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2017, pp. 177–186.

[21] Z. Gharaee, P. Gärdenfors, and M. Johnsson, “First and second order dynamics in a hierarchical som system for action recognition,” Applied Soft Computing, vol. 59, pp. 574–585, 2017.

[22] X. Guo, X. Guo, and Y. Lu, “Ssan: Separable self-attention network for video representation learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12 618–12 627.

[23] K. Hara, H. Kataoka, and Y. Satoh, “Learning spatio-temporal features with 3d residual networks for action recognition,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 3154–3160.

[24] ——, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6546–6555.

[25] ——, "Towards good practice for action recognition with spatiotemporal 3d convolutions," in 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018, pp. 2516–2521.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[27] H. Imtiaz, U. Mahbub, G. Schaefer, and M. A. R. Ahad, "A multi-resolution action recognition algorithm using wavelet domain features," in 2013 2nd IAPR Asian Conference on Pattern Recognition. IEEE, 2013, pp. 537–541.

[28] A. Javidani and A. Mahmoudi-Aznaveh, "A unified method for first and third person action recognition," in Electrical Engineering (ICEE), Iranian Conference on. IEEE, 2018, pp. 1629–1633.

[29] 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.

[30] G. Jiang, X. Jiang, Z. Fang, and S. Chen, "An efficient attention module for 3d convolutional neural networks in action recognition," Applied Intelligence, vol. 51, no. 10, pp. 7043–7057, 2021.

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

[32] M. Khare and M. Jeon, “Towards discrete wavelet transform-based human activity recognition,” in Second International Workshop on Pattern Recognition, vol. 10443. SPIE, 2017, pp. 35–39.

[33] M. Kim, H. Kwon, C. Wang, S. Kwak, and M. Cho, “Relational self-attention: What’s missing in attention for video understanding,” Advances in Neural Information Processing Systems, vol. 34, pp. 8046–8059, 2021.

[34] A. Klaser, M. Marszałek, and C. Schmid, “A spatio-temporal descriptor based on 3d-gradients,” in BMVC 2008-19th British Machine Vision Conference. British Machine Vision Association, 2008, pp. 275–1.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in neural information processing systems, vol. 25.

[36] B. S. Kumar, S. V. Raju, and H. V. Reddy, “Human action recognition using a novel deep learning approach,” in IOP Conference Series: Materials Science and Engineering, vol. 1042, no. 1. IOP Publishing, 2021, p. 012031.

[37] T. Lan, Y. Wang, and G. Mori, “Discriminative figure-centric models for joint action localization and recognition,” in 2011 International conference on computer vision. IEEE, 2011, pp. 2003–2010.

[38] K. Li, Y. Wang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, “Uniformer: Unified transformer for efficient spatiotemporal representation learning,” arXiv preprint arXiv:2201.04676, 2022.

[39] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, L. Wang, and Y. Qiao, “Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer,” arXiv preprint arXiv:2211.09552, 2022.

[40] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, “Uniformer: Unifying convolution and self-attention for visual recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[41] M. Li, W. Hsu, X. Xie, J. Cong, and W. Gao, "Sacnn: Self-attention convolutional neural network for low-dose ct denoising with self-supervised perceptual loss network," IEEE transactions on medical imaging, vol. 39, no. 7, pp. 2289–2301, 2020.

[42] Z. Li and G. Liu, "Video scene analysis in 3d wavelet transform domain," Multimedia Tools and Applications, vol. 56, no. 3, pp. 419–437, 2012.

[43] A.-A. Liu, Y.-T. Su, W.-Z. Nie, and M. Kankanhalli, "Hierarchical clustering multi-task learning for joint human action grouping and recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 1, pp. 102–114, 2016.

[44] C. Liu, Y. Jin, K. Xu, G. Gong, and Y. Mu, "Beyond short-term snippet: Video relation detection with spatio-temporal global context," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10 840–10 849.

[45] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in 2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009, pp. 1996–2003.

[46] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10 012–10 022.

[47] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 3202–3211.

[48] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, and M. Chiaberge, "Action transformer: A self-attention model for short-time pose-based human action recognition," Pattern Recognition, vol. 124, p. 108487, 2022.

[49] B. Meng, X. Liu, and X. Wang, "Human action recognition based on quaternion spatial-temporal convolutional neural network and lstm in rgb videos," Multimedia Tools and Applications, vol. 77, no. 20, pp. 26 901–26 918, 2018.

[50] E. Mohammadi, Q. J. Wu, Y. Yang, and M. Saif, "Effect of wavelet and hybrid classification on action recognition," in 2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017, pp. 1787–1791.

[51] K. Muhammad, A. Ullah, A. S. Imran, M. Sajjad, M. S. Kiran, G. Sannino, V. H. C. de Albuquerque et al., "Human action recognition using attention based lstm network with dilated cnn features," Future Generation Computer Systems, vol. 125, pp. 820–830, 2021.

[52] S. Nazir, M. H. Yousaf, J.-C. Nebel, and S. A. Velastin, "A bag of expression framework for improved human action recognition," Pattern Recognition Letters, vol. 103, pp. 39–45, 2018.

[53] Y. Pan, J. Xu, M. Wang, J. Ye, F. Wang, K. Bai, and Z. Xu, "Compressing recurrent neural networks with tensor ring for action recognition," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, 2019, pp. 4683–4690.

[54] Z. Pan and C. Li, "Robust basketball sports recognition by leveraging motion block estimation," Signal Processing: Image Communication, vol. 83, p. 115784, 2020.

[55] C. I. Patel, S. Garg, T. Zaveri, A. Banerjee, and R. Patel, "Human action recognition using fusion of features for unconstrained video sequences," Computers & Electrical Engineering, vol. 70, pp. 284–301, 2018.

[56] M. Patrick, D. Campbell, Y. Asano, I. Misra, F. Metze, C. Feichtenhofer, A. Vedaldi, and J. F. Henriques, "Keeping your eye on the ball: Trajectory attention in video transformers," Advances in neural information processing systems, vol. 34, pp. 12 493–12 506, 2021.

[57] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," Advances in Neural Information Processing Systems, vol. 32, 2019.

[58] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Spatiotemporal saliency for event detection and representation in the 3d wavelet domain: potential in human action recognition," in Proceedings of the 6th ACM international conference on Image and video retrieval, 2007, pp. 294–301.

[59] M. Ravanbakhsh, H. Mousavi, M. Rastegari, V. Murino, and L. S. Davis, "Action recognition with image based cnn features," arXiv preprint arXiv:1512.03980, 2015.

[60] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in 2008 IEEE conference on computer vision and pattern recognition. IEEE, 2008, pp. 1–8.

[61] N. S. Russel and A. Selvaraj, "Fusion of spatial and dynamic cnn streams for action recognition," Multimedia Systems, vol. 27, no. 5, pp. 969–984, 2021.

[62] A. S. Saif, E. D. Wollega, and S. A. Kalevela, "Spatio-temporal features based human action recognition using convolutional long short-term deep neural network," International Journal of Advanced Computer Science and Applications, vol. 14, no. 5, 2023.

[63] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in Proceedings of the 15th ACM international conference on Multimedia, 2007, pp. 357–360.

[64] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," arXiv preprint arXiv:1312.6229.

[65] L. Shao and R. Gao, "A wavelet based local descriptor for human action recognition." in BMVC, 2010, pp. 1–10.

[66] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 806–813.

[67] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," arXiv preprint arXiv:1511.04119, 2015.

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

[69] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7464–7473.

[70] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.

[71] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6450–6459.

[72] Z. Tu, W. Xie, Q. Qin, R. Poppe, R. C. Veltkamp, B. Li, and J. Yuan, "Multi-stream cnn: Learning representations based on human-related regions for action recognition," Pattern Recognition, vol. 79, pp. 32–43, 2018.

[73] S. Tuli, I. Dasgupta, E. Grant, and T. L. Griffiths, "Are convolutional neural networks or transformers more like human vision?" arXiv preprint arXiv:2105.07197, 2021.

[74] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional lstm with cnn features," IEEE access, vol. 6, pp. 1155–1166, 2017.

[75] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[76] D. Wang, G. Zhao, G. Li, L. Deng, and Y. Wu, "Compressing 3dcnns based on tensor train decomposition," Neural Networks, vol. 131, pp. 215–230, 2020.

[77] H. Wang and C. Schmid, "Action recognition with improved trajectories," in Proceedings of the IEEE international conference on computer vision, 2013, pp. 3551–3558.

[78] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, "Human action recognition by learning spatio-temporal features with deep neural networks," IEEE access, vol. 6, pp. 17 913–17 922, 2018.

[79] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, "Towards good practices for very deep two-stream convnets," arXiv preprint arXiv:1507.02159, 2015.

[80] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.

[81] J. Wei, H. Wang, Y. Yi, Q. Li, and D. Huang, "P3d-ctn: Pseudo-3d convolutional tube network for spatio-temporal action detection in videos," in 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019, pp. 300–304.

[82] L. Xiao, Y. Cao, Y. Gai, E. Khezri, J. Liu, and M. Yang, "Recognizing sports activities from video frames using deformable convolution and adaptive multiscale features," Journal of Cloud Computing, vol. 12, no. 1, p. 167, 2023.

[83] S. Yan, X. Xiong, A. Arnab, Z. Lu, M. Zhang, C. Sun, and C. Schmid, "Multiview transformers for video recognition," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 3333–3343.

[84] B. Yang, L. Wang, D. Wong, L. S. Chao, and Z. Tu, “Convolutional self-attention networks,” arXiv preprint arXiv:1904.03107, 2019.

[85] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, “Learning texture transformer network for image super-resolution,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5791–5800.

[86] S. Zebhi, S. AlModarresi, and V. Abootalebi, “Action recognition in videos using global descriptors and pre-trained deep learning architecture,” in 2020 28th Iranian Conference on Electrical Engineering (ICEE). IEEE, 2020, pp. 1–4.

[87] S. Zebhi, S. M. T. AlModarresi, and V. Abootalebi, “Human activity recognition using pre-trained network with informative templates,” International Journal of Machine Learning and Cybernetics, vol. 12, no. 12, pp. 3449–3461, 2021.

[88] W. Zeng and M. Li, “Crop leaf disease recognition based on self-attention convolutional neural network,” Computers and Electronics in Agriculture, vol. 172, p. 105341, 2020.

[89] C. Zhang, Y. Xu, Z. Xu, J. Huang, and J. Lu, “Hybrid handcrafted and learned feature framework for human action recognition,” Applied Intelligence, vol. 52, no. 11, pp. 12 771–12 787, 2022.

[90] X. Zhang, L. Han, W. Zhu, L. Sun, and D. Zhang, “An explainable 3d residual self-attention deep neural network for joint atrophy localization and alzheimer’s disease diagnosis using structural mri,” IEEE journal of biomedical and health informatics, 2021.

[91] Y. Zhang, X. Li, C. Liu, B. Shuai, Y. Zhu, B. Brattoli, H. Chen, I. Marsic, and J. Tighe, “Vidtr: Video transformer without convolutions,” in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 13 577–13 587.

[92] Y. Zhang, J. Li, N. Jiang, G. Wu, H. Zhang, Z. Shi, Z. Liu, Z. Wu, and X. Liu, “Temporal transformer networks with self-supervision for action recognition,” IEEE Internet of Things Journal, vol. 10, no. 14, pp. 12 999–13 011, 2023.

[93] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable detr: Deformable transformers for end-to-end object detection,” arXiv preprint arXiv:2010.04159, 2020.

[94] Y. Zhu, X. Li, C. Liu, M. Zolfaghari, Y. Xiong, C. Wu, Z. Zhang, J. Tighe, R. Manmatha, and M. Li, “A comprehensive study of deep video action recognition,” arXiv preprint arXiv:2012.06567.