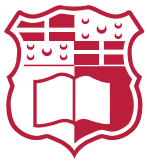
Skateboard Trick Recognition through an AI-based Approach

Kris Saliba

Supervisor: Prof. Joseph Bonello

June 2024

Submitted in partial fulfilment of the requirements
for the degree of Bachelor of Science in Information Technology (Honours)
(Software Development).



L-Università ta' Malta
Faculty of Information &
Communication Technology

Abstract

Acknowledgements

Contents

List of Figures

List of Tables

1 Introduction

Skateboarding dates back to the 1940s when handmade skateboards first appeared [**SkateboardingEncyclobeadia**]. It has since developed into a worldwide phenomenon, with its popularity skyrocketing, after gaining recognition as an official sport in the 2020 Tokyo Olympic Games [**SkatebaordingOlimpics**]. Skateboarding comprises the dynamic activities of riding a skateboard and skilfully performing a repertoire of tricks, manifesting as a popular and exhilarating “extreme sport”.

This dynamic sport encompasses various disciplines and riding styles, each offering unique challenges for skateboarders to explore. Two of the most prominent styles are “vert” and “street.” Vert skateboarding revolves around riding on specialised obstacles, namely, half-pipes and ramps, emphasising aerial manoeuvres. Street skateboarding transpires in urban environments, utilising various obstacles that can be found outdoors, including stairs, rails, ledges, gaps or flat ground for skaters to showcase their creativity [**skateStyles**].

2 Background

2.1 Skateboard Tricks

Skateboard tricks are the heart and soul of skateboarding. These tricks originate from the dynamic orchestration of rotations and revolutions of a skateboard along various axes emphasising the significance of precise placement of a skateboarder's feet to initiate these rotations. These tricks serve as excellent examples of how the skateboarder's body and skateboard work in perfect harmony. Some common skateboard tricks include:

- **Ollie:** One of the first tricks beginners learn. Where the skateboarder pops the tail of the board while sliding their foot across the board, causing the board to level out in the air, used to jump over obstacles.
- **Kickflip:** A trick where the skateboarder flips the board under their feet while jumping, making it spin 360°around the x-axis.
- **360 kickflip:** A combination of a kickflip and a 360°board rotation around the y-axis.

Skateboarders continually innovate and come up with new trick combinations, contributing to the dynamic nature of the sport.

2.2 Problem Definition

The problem at hand revolves around the lack of an efficient and objective method for recognising skateboard tricks in competitions and practice sessions. Currently, the skateboarding community relies on using subjective methods for evaluations of these tricks, thus, this can lead to inconsistencies and disputes in scoring. This lack of objectivity not only disrupts the fairness in competitions but also inhibits skaters' ability to receive real-time feedback for skill improvement.

The integration of Artificial Intelligence (AI) into skateboarding provides great potential for the standardisation of tricks, real-time feedback and injury prevention. The development of an AI model that can accurately recognise skateboard tricks is a tool that the skateboarding community would greatly benefit from. Consequently, there is a need for the development of an AI model that can accurately recognise and standardise skateboard tricks, thereby enhancing the sport's objectivity, fairness and quality.

2.3 Machine Learning

Machine Learning (ML) can be defined as a field of study that explores algorithms and statistical models employed by computer systems to execute tasks without the need to be explicitly programmed. It is particularly applicable in situations where the information we seek from a dataset is not immediately evident or interpretable, and as the volume of available datasets continues to surge so does the demand for machine learning [**ML_Algorithms**].

Morris [**UnderstandingLSTM**] characterises ML as the advancement of algorithms that progressively enhance their performance through practice, suggesting that the more training the learning algorithm undergoes, the better the algorithm becomes at executing tasks. Training is a multifaceted process that directly influences the overall accuracy of the model. Within this phase, numerous critical factors come into play, shaping the model's performance. These factors encompass dataset quality and diversity, the meticulous preprocessing of data, the selection of an appropriate model architecture, the optimal duration of training, and the fine-tuning of essential hyper-parameters [**TheEffectsofDataQualityonMachineLearningPerformance**].

There exist three main categories for ML models[**ML_Algorithms**]:

- **Supervised:** This is a machine learning concept that centres around the development of algorithms to make predictions or classifications using labelled data. The model is trained on a dataset comprising of example input-output pairs.
- **Unsupervised:** This ML concept concentrates on discovering relationships within data when there are no predefined "correct" answers or labelled examples to guide the learning process. These algorithms are left to autonomously explore and divulge structures in the data.
- **Reinforcement:** This type of learning consists of an agent that interacts with the environment and learns from the continuous feedback it receives in the form of rewards or punishment.

2.4 Object Detection

Object detection is a computer vision task that detects instances of objects in images and videos and maps them to a predefined class. For humans, the act of recognising and responding to objects is a trivial task as described in [**NeuralScience**], it

is an essential feature that enables our performance and communication. Numerous academic and industry researchers have shown a deep interest in the technology, focusing on various applications where object detection plays a major role. These applications include but are not limited to autonomous driving, surveillance systems and face detection [**RecentAdvancesObjectDetection**].

The output of an object detection model yields the instance's location, as the object's centre, a bounding box or even a list of pixels containing the object. The research paper [**RecentAdvancesObjectDetection**] further implies that object detection is consistently defined within the context of a data set that consists of images mapped to a list of relevant object properties, such as their locations and scales, that are specified within each image. This definition makes references to the equation below, where an image is denoted as $\mathcal{I}$, and $O(I)$ represents the collection of object descriptions for objects within the image.

$$O(I) = \{(Y_1^*, Z_1^*), \ldots, (Y_i^*, Z_i^*), \ldots, (Y_{N^*i}^*, Z_{N^*i}^*)\}$$

In the above equation, each description encompasses two parts, $Y_i^* \in \mathcal{Y}$ characterises the category or type of an object, and $Z_{N^*i}^* \in \mathcal{Z}$ represents information about its location, size or shape within the image. $\mathcal{Z}$ represents the different ways to describe an object, this is typically done by specifying the object's centre $(x_c, y_c) \in \mathcal{R}^2$ or as a bounding box $(x_{min}, y_{min}, x_{max}, y_{max}) \in \mathcal{R}^4$. By utilising these notations, according to [**RecentAdvancesObjectDetection**], object detection can be defined as the operation of combining an image with a set of detections.

2.5 Activity Recognition

Activity recognition is the process of identifying and categorizing human activities from video sequences. Human activity involves a wide range of motions and interactions with objects, varying from simple isolated actions like dancing to more complex activities that engage multiple body parts and external objects. Similar to object detection, the human ability to perceive these behaviours is a trivial task; yet, it is a challenging problem for computers due to the sequential nature and the resemblance of visual content in such activities[**ActionRecognitionDeepBi-DirectionalLSTM**][**AReviewOfHumanActivityRecognitionMethods**

Recognising complex human actions demands the examination of sequential data as opposed to relying on single frames or images

[**ActionRecognitionDeepBi-DirectionalLSTM**]. To illustrate this point, consider the example of a skateboarder executing a challenging trick like the ”Kickflip”, as depicted in Figure 2.1. If we were to feed a single frame of this trick into an AI model, as shown in Figure 2.1a, it may misinterpret the manoeuvre as another trick. Whereas, by providing the model with a sequence of frames, as shown in Figure 2.1b, it captures the entire essence of the trick portraying the dynamic progression of actions such as foot placement, board rotation and landing which collectively define the skateboard trick.

(a) A Single Frame of a ”Kickflip”.

(b) Multiple Frames of a ”Kickflip”.

Figure 2.1 Comparison of a Single Frame and Multiple Sequential Frames.

2.6 Neural Networks

2.6.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) are a class of machine learning models that are inspired by the interconnected systems of neurons found in the nervous system of living organisms. They consist of connected nodes organised in layers capable of learning from their environment and adapting to complex patterns in data [**FundamentalsOfNeuralNetworks**]. One common type of ANN is the Feed-Forward Neural Network (FNN), which serves as a foundational architecture in neural network design. Figure 2.2 depicts a schematic representation of an ANN. The diagram is organised into three fundamental layers: the Input Layer, the Hidden Layer(s) and the Output Layer [**FundamentalsOfArtificialNeuralNetworksAndDeepLearning**].

- **Input Layer:** This is the set of neurons that serve as the initial entry point for external data or features. Each input neuron in this layer corresponds to a specific feature or variable used in the Neural Network model.

- **Hidden Layer(s):** This is the set of neurons that are situated between the Input and Output Layers where the network captures complete nonlinear behaviours of data and feature transformations.
- **Output Layer:** This is the set of neurons that provide the final predictions produced by the neural network. Depending on how the ANN is configured, the final output can be continuous, binary, ordinal, or count.

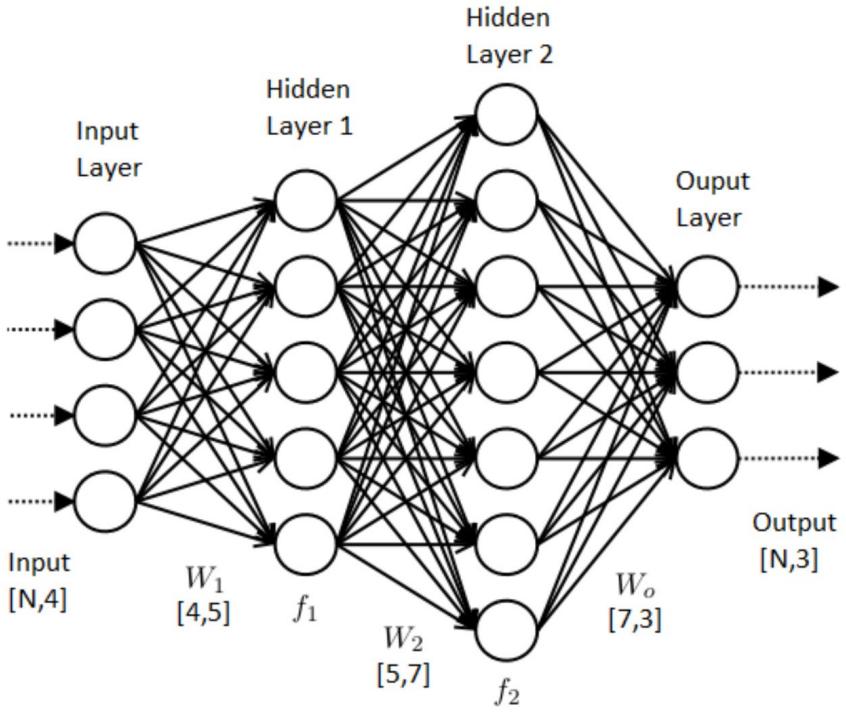


Figure 2.2 Schematic Representation of an Artificial Neural Network with four input variables, three output variables and two hidden layers.

2.6.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs), as depicted by [**RecentAadvancesInConvolutionalNeuralNetworks**] are a category of Deep learning architectures with roots in the biological visual perception mechanisms of living organisms. These networks have gained widespread attention for their incredible performance in various fields such as visual recognition, speech recognition and natural language processing.

CNNs, incorporate multiple layers and are capable of extracting effective representations

2.6.3 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a subset of Neural networks that are designed for sequential data processing. While feed-forward neural networks provide static mappings from input to output, RNNs are capable of modelling dynamic relationships in sequential data by feeding signals from past time steps back into the network. However, as we delve deeper into the workings of RNNs, it becomes apparent that these networks are not without their constraints. One prominent limitation lies in their ability to effectively capture information from the past, often constrained to approximately ten sequential time steps. This constraint arises from the challenge of dealing with vanishing or exploding gradients during the backpropagation procedure during the processing of extended sequences, as discussed in prior works [**UnderstandingLSTM**],[**Long-termConvolutionalNeuralNetworksforVisualRecogntion**].

To address the limitation that RNNs encounter in capturing long-term dependencies, Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) were introduced [**learningPriceseTimingWithLSTM**]. LSTM networks unlike RNNs are capable of capturing dependencies across more than 1,000 time steps depending on the network's complexity and are more biologically credible [**UnderstandingLSTM**].

3 Literature Review

3.1 Activity Recognition

Building on the foundational understanding of activity recognition defined in the background, it's important to acknowledge the impact it has on various sectors. The recent advancements in recognising human actions in videos have have not only revolutionalised sectors such as healthcare and security as demonstrated in the studies [**3DhumanActionDetectionForHealthCareSystems**], [**HumanActivityRecognitionSecurityAndMonitoring**], but also hold extensive potential in the realm of sports.

The research conducted by K. Host and M. Ivašić-Kos in [**HARinSportsComputerVision**] along with Wu et al. in [**AsurveyonvideoactionrecognitioninsportsDatasetsmethodsandapplications**] further elaborate on this potential, such as its numerous applications in categorising complex sport actions, injury prevention methods and refinement of game strategies through video analysis. These studies also propose various methodological advancements, including the utilisation of Deep Learning models to enhance accuracy, the exploration of multimodal data sources for sports activities and an emphasis on real-time analysis capabilities. These insights provide valuable techniques that can be leveraged for the development of activity recognition models in the field of sports.

The study by Beddiar et al. [**VisionBasedHARASurvey**],identifies two main streams of human computer interaction (HCI) technologies: Contact-based and Vision-based systems. The authors categorise contact-based HCI as those technologies that require physical user interaction through mediums such as accelerometers, wearable sensors and multi-touch interfaces. Alternatively, the authors describe vision-based methods as the simplification of HCI due to more natural human communication, eliminating the need for physical contact or equipment. These methods use image and video data to recognise human activities offering an advantage in terms of societal acceptance and usability.

While both contact and vision-based systems have their merits, this study will specifically focus on vision-based techniques and their application in the development of a skateboard trick classifier. These methods, which utilise image and video data to recognise human activities are selected due to their societal acceptability and their applicability in sports broadcasts.

3.1.1 Challenges in this field

The domain of Activity recognition comes with many challenges as depicted by Zhang et al. (2017) [**ReviewOfHumanActivityRecognitionUsingVisionBasedMethod**]. The researchers suggest that while certain scenarios utilise static cameras such as surveillance systems, most situations that benefit from Activity Recognition adopt dynamic recording devices such as mobile phones and sports event broadcasts. These dynamic devices introduce a significant level of complexity as a result of their tricky dynamic backgrounds present in video footage. Zhang et al. also points out specific difficulties posed by long-distance and low-quality videos, often encountered in environments like crowded public spaces and sports events. The camera distance, results in smaller subjects, making detailed analysis of human movements more challenging while lower quality videos further complicate the task for Human Activity Recognition (HAR) systems.

The challenges highlighted by Zhang et al. [**ReviewOfHumanActivityRecognitionUsingVisionBasedMethod**] in the domain of activity recognition are directly relevant to the development of a skateboard trick classifier. In scenarios like televised skateboarding events, skaters may appear relatively small to accommodate the entire skatepark, This factor along with the complexity of the background as a result of dynamic recording poses a challenge for trick recognition in live broadcasts. Furthermore, if a skateboard trick classifier is intended for home use, then it may encounter videos of lower quality further complicating the task of trick recognition. Addressing these challenges is crucial for the development of a skateboard trick classifier that performs well in real-world conditions.

3.1.2 Activity Recognition Techniques

Notably the CNN-LSTM architecture is an effective model in accurately recognising human activities, this is outlined in the study by Ronald Mutegeki [**CNN-LSTMApproachToHAR**]. The author reports this architecture achieving 99% and 92% on the iSPL and UCI HAR datasets respectively, showcasing the strength of CNN's for extracting spatial features from data, especially in the field of activity recognition.

CNN-LSTM is a popular approach in HAR use cite [**HumanActivityRecognitionSystem:StateOfTheArt**]

3.2 Preprocessing techniques

Preprocessing is a crucial step in the development of ML models, especially in the field of activity recognition. It involves the application of various techniques to enhance raw data before

As a result of the limited research in the emergence of a skateboard trick classifier, there is a significant lack of open-source datasets featuring skateboard tricks. This lack of data presents an opportunity to employ data augmentation techniques, especially valuable in studies with limited data. Methods such as flipping, rotating, scaling and colour manipulation not only artificially enhance the size of the original dataset but also lower the likelihood of overfitting [**DataAugmentationCanImproveRobustness**], [**AnOverviewOfOverfittingAndItsSolutions**]. In the realm of Skateboard trick classifiers, Shapiee et al (2020) [**skatePaper1**] effectively employed data augmentation techniques to expand their dataset, demonstrating the application of these methods in improving model performance for trick classification.

After establishing the role of data augmentation in addressing the lack of data available, another important step in computer vision is data normalisation. This technique is essential for standardising the range and distribution of pixel values in an image, and has shown an increase in model performance, [**RobustnessInMLNormalisation**], [**RealWorldMicroGraphDataQualityNORMALIZATIONCITE**]. Among methods like Z-score, min-max normalisation is particularly prevalent in normalising pixel intensities to a 0-1 scale as follows:

- Normalized Value $= \frac{\text{Pixel Value} - \text{Min Value}}{\text{Max Value} - \text{Min Value}}$

While less common in computer vision due to its normality assumption, Z-score normalisation is defined as:

- Z-score $= \frac{x - \mu}{\sigma}$

Here, x represents the individual pixel value, μ is the mean of all pixel values, and σ is their standard deviation.

The research by Pei et al. (2023) [**RobustnessInMLNormalisation**], explored the impact of normalisation on classification accuracy. They demonstrated that, for 8-bit images, min-max normalisation outperformed Z-score, significantly enhancing classification accuracy. On the other hand, the study by de Raad et al.

[**EffectOnPreProcessingOnCNNForMedicalImageSegmentation**] suggests that the impact of normalisation on model performance varies depending on certain dataset characteristics. These contrasting conclusions presented by the studies [**RobustnessInMLNormalisation**] and [**EffectOnPreProcessingOnCNNForMedicalImageSegmentation**] suggests that while normalisation is an important preprocessing step, its application should be carefully adapted to the characteristics of the dataset.

Having optimised the distribution of pixel values through the normalisation process, the next crucial step is feature extraction. Xudong Jiang [**FeatureExtractionForImageRecognitionAndComputerVision**], describes this technique as the process of capturing the core attributes of an object by eliminating redundancies, resulting in a set of numerical features ideal for classification. CNN's excel at this due their capabilities in detecting complex patterns and enhancing features. Used primarily in image recognition, as an initial for classification algorithms, their effectiveness is demonstrated by studies like Manjunath Jogin et al. (2018) [**FeatureExtractionUsingCNNandDeepLearning**] where they achieved 86% accuracy rate through the use of CNN's for feature extraction alongside several classifiers. Beyond this primary function, CNN's ability to extract complex features from images makes them very effective when coupled with other models for more complex tasks such as activity recognition. In the context of skateboard trick classification, CNNs can efficiently extract detailed spatial features like board rotations, foot positioning and limb movements. This involves CNNs operating on individual frames to extract spatial features, which can then be passed through a sequence analysis model like an LSTM, proven to be efficient at understanding temporal information.

Following the discussion on CNN's for feature extraction, it is important to highlight the role of transfer learning in enhancing ML models, especially in fields with limited data like skateboard trick classification. The concept of transfer learning as detailed by the studies [**ASurveyOfTransferLearning**] and [**ApplicationAndAnalysisOfTransferLearningSurvey**] involves using a pre-trained ML model to leverage its past experience for a new, but related task. The study by Sargano et al. (2017) [**HARUsingTransferLearningWithDeepRepresentations**], employed transfer learning with pre-trained deep CNNs like AlexNet [**AlexNetCite**] and GoogleNet [**GoogleNetCite**], for human activity recognition. Notably, they utilised the pre-trained models for feature extraction, followed by an SVM classifier for final recognition. Their approach showcases the resource-efficient and time-saving nature of transfer learning, evident in their impressive accuracies of 98.15% and 91.47%. Moreover, research on transfer learning is not unique within the realm of skateboard

trick classification. Two other studies [**skatePaper1**], [**SkateboardAIPaper**] have utilised this approach and achieved good accuracies, a more in-depth analysis on these papers can be found in the 'Advancements in Skateboard Trick Classification' section. These studies strongly suggests the exploration of various pre-trained models for feature extraction in the development of a skateboard trick classifier.

3.3 Advancements in Skateboard Trick Classification

In the emergent field of skateboard trick classification, leveraging activity recognition techniques from a video has led to two primary methodologies among researchers. The first technique involves utilising signals obtained from skateboard-mounted accelerometers or signals artificially generated based on the findings of prior studies. These signals are then fed into a study-dependent model for classification, as outlined in [**skateboardClassificationTransferLearningPipelinesAccelermetry**] and [**skateboardTrickClassifierUsingAccelerometryAndML**]. The second approach employs computer vision techniques, leveraging video footage of skateboard tricks to train and refine models for accurate trick identification, as depicted by the studies [**skatePaper1**] and [**SkateboardAIPaper**].

3.3.1 Accelerometry approach

The study by Abdullah et al (2021) [**skateboardClassificationTransferLearningPipelinesAccelermetry**], makes use of a custom dataset comprising of six skateboard tricks most commonly executed in competitive events. Amateur skateboarders performed each trick five times on a modified skateboard equipped with an Inertial Measurement Unit (IMU) to record the signals produced. The researchers capture six signals for each trick, including linear accelerations along the x, y, and z axes (aX, aY, aZ) and angular accelerations along the same axes (gX, gY, gZ). They then opt for the unique approach of concatenating all six signals onto a single image corresponding to one trick, employing two input image transformations: raw data (RAW) and Continuous Wavelet Transform (CWT).

With the application of six transfer learning models on this data, Abdullah et al. [**skateboardClassificationTransferLearningPipelinesAccelermetry**] reports exceptionally high accuracies, achieving a 100% test accuracy over multiple models. While these results are remarkable, very high levels of accuracy are rare in ML applications and are typically associated with models that may be overfitting the data. Recognising the rarity of such high accuracies, this study will take these findings into

consideration and efforts will be made to ensure a robust model by employing techniques to avoid overfitting such as early-stopping and the use of a diverse dataset [**AnOverviewOfOverfittingAndItsSolutions**].

The study by Corrêa et al (2017) [**skateboardTrickClassifierUsingAccelerometryAndML**], obtained their sample data by artificially generating 543 signals based on prior research, utilising tools such as MATLAB 2015 and Signal Processing Toolbox. These signals were then categorised into five distinct classes representing different skateboard tricks, each with various samples ranging from 30 to 50 per class, across three axes (X, Y and Z). This study developed and validated individual Artificial Neural Networks (ANNs) for each axis, as well as the combination of the three: ANN XYZ, displaying the potential of Neural Networks to categorise multidimensional skateboard tricks. The ANNs are all multilayer feed-forward neural networks (MFFNNs), structured into three distinct layers. They feature an input layer with 82 neurons, a hidden layer, comprised of 23 neurons utilising a tan-sigmoid transfer function and an output layer consisting of 5 neurons with a softmax function. Finally, the study achieved high accuracies, with ANNs X, Y and Z achieving accuracies of 94.8%, 96.7% and 98.7%, respectively, while the combined ANN XYZ achieved an accuracy of 92.8%.

3.3.2 Computer Vision Approach

The paper by Shapiee et al (2020) [**skatePaper1**] leverages a custom data set comprising videos capturing the execution of five distinct skateboard tricks, each attempted five times. Each video spans two to three seconds, yielding a total of 750 images by extracting 30 frames per video. This study made use of data augmentation techniques to expand their dataset further. Consequently, they introduced an additional 2,250 images, achieving 3,000 images in their data set. On the other hand, Chen (2023) [**SkateboardAIPaper**] compiled a comprehensive data set by collecting videos from multiple platforms, including YouTube, Twitter and Instagram. Furthermore, Chen trained the model using 15 fundamental tricks commonly observed in competitive settings. The researcher collected 50 videos per trick, summing up to a total number of 750 videos. Of these, 45 videos per trick were allocated for training, and the remaining 5 were reserved for validation.

The paper by Shapiee et al. [**skatePaper1**] utilises data augmentation techniques and applies three rotation augmentation techniques: horizontal rotation, positive 90°rotation and negative 90°rotation. The researchers experimented on three Transfer learning models: MobileNet, NASNetMobile and NASNetLarge, each evaluated using a

k-Nearest Neighbor (k-NN) classifier. As a result, the models demonstrated impressive classification accuracies, with MobileNet achieving 95%, NASNetMobile 92% and NASNetLarge 90%.

Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs) are popular architectures due to their capabilities in modelling the dynamic relationships in sequential data [**UnderstandingLSTM**]. In the student abstract by Hanciao Chen [**SkateboardAIPaper**], extensive experimentation is conducted using diverse models, exploring various combinations of CNN-LSTM and CNN-BiLSTM architectures. The study also incorporated attention mechanisms and explored transfer-based methods for activity recognition. This study further documents and analyses important metrics such as training time, training accuracy and validation accuracy for each model experimented on. Among these, the top three models that stood out in terms of validation accuracy were the ResNet50 with Attention and BiLSTM (84%), ResNet50 with BiLSTM (81%) and ResNet50 with LSTM (80%). Chen's study provides valuable insight into the application of diverse models in activity recognition in skateboarding.

4 Sample A

5 Sample B