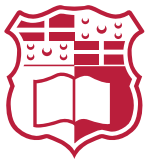
Skateboard Trick Recognition through an AI-based Approach

Kris Saliba

Supervisor: Dr. 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

In progress — incorporate these paragraphs

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**].

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 Background

2.1 Skateboard Tricks

Skateboard tricks can be described as dynamic manoeuvres that involve complex orchestration of the skateboard and the skateboarder's body. Skateboard tricks can be categorised into various types, including rotations (flips, rails), grinds (on ledges or rails) and manuals (balancing on two wheels). Successful execution of these tricks relies heavily on precise foot placement, which determines the boards velocity and direction, enabling a skateboarder to manipulate the board to replicate specific tricks.

- **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.
- **Pop-Shuvit:** A trick where the skateboarder scoops the board with their back foot causing a 180°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 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 interpretable, and as the volume of available datasets continues to surge so does the demand for machine learning [**ML_Algorithms**].

Morris (2019) [**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 it becomes at executing tasks. Numerous critical factors shape a model's performance within this phase, as exmplified by Budach et al. (2022) [**TheEffectsofDataQualityonMachineLearningPerformance**]. Such factors include dataset quality and diversity, data preprocessing, the selection of a suitable model architecture, training time and the fine-tuning of hyper-parameters.

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.

2.3 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 by Watson et al. (2016) [**NeuralScience**], it is an essential feature that enables our performance and communication. Numerous researchers have shown a deep interest in this technology, focusing on various applications where object detection may play a major role, such as surveillance systems, face detection and autonomous driving [**RecentAdvancesObjectDetection**].

The output of an object detection model returns the location of the instance, as the object's centre or in the form of a bounding box. The research paper by Agarwal et al. (2018) [**RecentAdvancesObjectDetection**] defines object detection as the following equation where an image is denoted as $\mathcal{I}$, and $O(I)$ represents the collection of object descriptions for objects within the image.

$$O(I) = \{(Y_1^*, Z_1^*), \ldots, (Y_i^*, Z_i^*), \ldots, (Y_{N^*i}^*, Z_{N^*i}^*)\} \tag{2.1}$$

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 $(x_{min}, y_{min}, x_{max}, y_{max}) \in \mathcal{R}^4$.

2.4 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**].

2.5 Neural Networks

2.5.1 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 capable of learning from their environment and adapting to complex patterns in data [**FundamentalsOfNeuralNetworks**]. Figure 2.1 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 non-linear 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.

2.5.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs), as described by Gu et al. (2018) [**RecentAadvancesInConvolutionalNeuralNetworks**] are a category of Deep learning

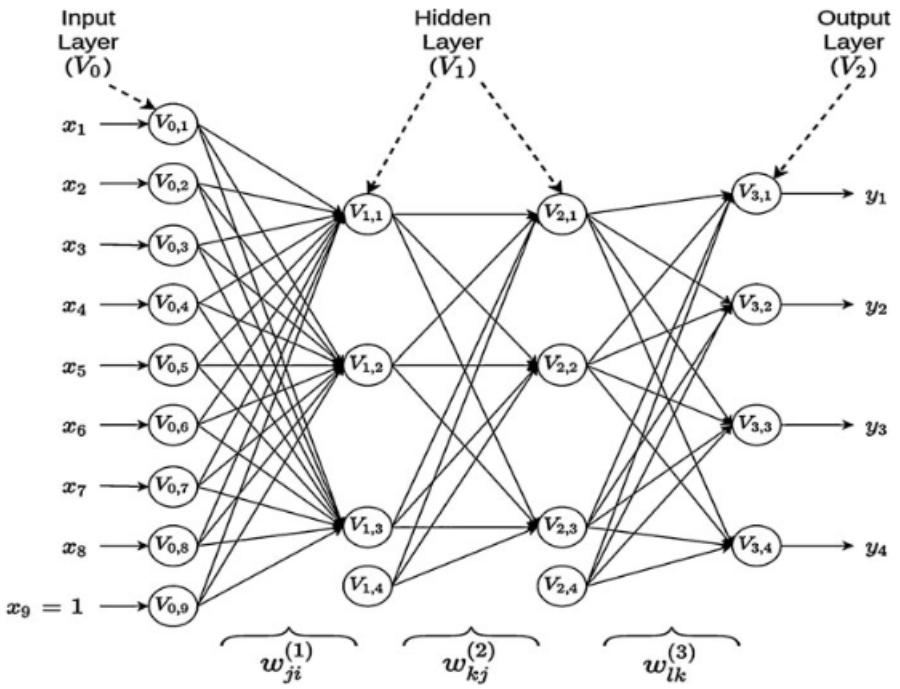


Figure 2.1 Schematic Representation of an Artificial Neural Network. Reproduced from López et al. (2022) [**FundamentalsOfArtificialNeuralNetworksAndDeepLearning**]

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

– continue explanation, explain how feature maps are created through convolutions

2.5.3 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a subset of Neural networks that are designed for sequential data processing. RNNs are capable of modelling dynamic relationships in sequential data by feeding signals from past time steps back into the network. However, they are limited by their inability to access long-term data, limited to approximately ten sequential time steps. This constraint arises from the challenge of dealing with vanishing or exploding gradients in the backpropagation procedure while processing 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 sports 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 point 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 personal development, 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 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 feeding it to Machine Learning algorithms. —continue or leave out

As a result of the limited research on 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 as highlighted in prior works, [**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**]. Min-max normalisation is particularly prevalent in normalising pixel intensities to a 0-1 scale as follows:

$$\text{Normalized Value} = \frac{\text{Pixel Value} - \text{Min Value}}{\text{Max Value} - \text{Min Value}} \tag{3.1}$$

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

$$\text{Z-score} = \frac{x - \mu}{\sigma} \tag{3.2}$$

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 both studies suggest 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 through the elimination of redundancies, resulting in a set of numerical features ideal for classification. CNNs excel at this due to their capabilities in detecting complex patterns and enhancing features. While they are primarily used in image recognition, as an initial step before deploying classification algorithms, their effectiveness in feature extraction is demonstrated by studies like Manjunath Jogin et al. (2018) [**FeatureExtractionUsingCNNandDeepLearning**] where they achieved 86% accuracy using CNNs 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. Such features can then be passed through a sequence analysis model like an LSTM, proven to be efficient at understanding temporal information.

Following the discussion on CNNs 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 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 these models for feature extraction, followed by an SVM classifier for final classification. 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 field of skateboard trick classification. Two other studies [**skatePaper1**], [**SkateboardAIPaper**] have utilised this approach and achieved high accuracies, however, a more in-depth analysis of these papers can be found in the 'Advancements in Skateboard Trick Classification' section. These studies strongly suggest the exploration of various pre-trained models for feature extraction in the development of a skateboard trick classifier.

3.1.3 Activity Recognition Techniques

The accurate classification of skateboard tricks poses a unique challenge for computer vision due to the sport's dynamic and complex nature, characterised by rapid movements, potential occlusions and camera angles. This section explores various computer vision techniques and architectures employed in previous literature, exploring their potential impact on this specific task.

Deep learning (DL) techniques have become increasingly popular over traditional ML methods, for their ability to learn feature representations automatically from raw data, significantly improving performance [**AReviewOnComputerVisionBasedMethodsForHARRecognition**]. One particularly effective DL architecture is the CNN-LSTM. This approach leverages the CNNs strength in extracting spatial features from individual frames and LSTMs ability to capture temporal information across frames, leading to a deeper understanding of video content.

The study by Orozco et al. (2020)

[**HARRecognitionInVideosUsingARobustCNNLSTMApproach**] adopted this approach to test its effectiveness against three activity recognition datasets: KTH, UCF-11, HMDB-51, particularly focusing on how the number of LSTM units impacted performance. The authors employed transfer learning, utilising the VGG16 model [**VGG**] for feature extraction from videos, followed by an LSTM network for classification. Their findings show that 360 LSTM units achieved an accuracy of 93.86% on the KTH dataset, while 320 units led to an accuracy of 91.93%. However, the performance on the HMDB-51 dataset dropped, with 400 LSTM units resulting in a lower accuracy of 47.36%. These findings show the potential of the CNN-LSTM approach, particularly for simpler datasets, highlighting the need for further investigation and optimisation for more complex datasets like HMDB-51.

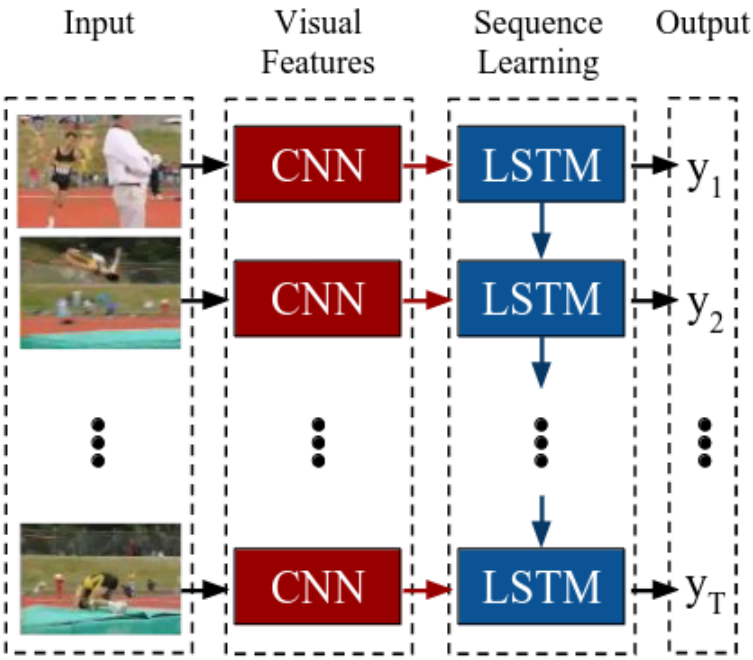


Figure 3.1 CNN-LSTM Architecture. Reproduced from Donahue et al. (2015) [**LongTermRecurrentConvolutionalNetworksForVisualRecognitionAndDescription**]

Building upon the foundational CNN-LSTM architecture, a recent study by Saoudi et al. (2023) [**Advancinghumanactionrecognition:ahybridapproachusingattention-basedLSTMand3DCNN**] enhances this model by incorporating three key advancements:

1. **3D Convolutional Neural Networks (3D CNNs):** Unlike regular CNNs, data is processed as 3D volumes, enabling them to capture both spatial and temporal information by performing convolution operations on all three dimensions: width, height and time. The authors opted to use the I3D model, leveraging transfer learning to bypass the resources required to train one from scratch. The I3D model was selected for its proven efficiency and its ability to be fine-tuned for activity recognition tasks.

2. **Bi-directional Long-Short-Term Memory (BiLSTM) network:** The authors chose to use a variant of LSTM called Bi-directional Long Short-Term Memory (BiLSTM). Whereas LSTMs only process data in one direction, BiLSTMs can analyse data in both directions, allowing them to understand relationships between preceding and subsequent actions.

3. **Attention Mechanisms:** Saoudi et al. incorporated attention mechanisms after their BiLSTM, enabling the model to prioritise particular parts of the input data. This technique allowed the model to learn which temporal features were most

applicable to the task, providing a more detailed representation of the input and ultimately, improved performance.

With the integration of these advancements, Saoudi et al. achieved model accuracies of 97.98% and 96.83% on the HMDB51 and UFC101 datasets respectively, demonstrating the potential of 3D CNNs, BiLSTMSs and attention mechanisms for activity recognition tasks.

3.2 Advancements in Skateboard Trick Classification

In the emergent field of skateboard trick classification, leveraging activity recognition techniques from a video have 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 by Abdullah et al. (2021) [**skateboardClassificationTransferLearningPipelinesAccelermetry**] and Corrêa et al. (2017) [**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 Shapiee et al. (2020) [**skatePaper1**] and Hanciao Chen (2023) [**SkateboardAIPaper**].

3.2.1 Accelerometer-based approaches

The study by Abdullah et al. (2021) [**skateboardClassificationTransferLearningPipelinesAccelermetry**], makes use of a custom dataset comprising 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 consider these findings 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 94.8%, 96.7% and 98.7%, respectively, while the combined ANN XYZ achieved an accuracy of 92.8%.

3.2.2 Computer Vision-based Approaches

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 with the application of three rotations to the images: 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%.

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 Methodology

4.1 Class Establishment

This study pursued a multi-class classification strategy, targeting three fundamental skateboard tricks: ollie, pop shuvit and kickflip. These tricks were selected based on two primary criteria. Firstly, they are often associated with the first tricks learnt by beginners, highlighting their role in foundational skateboarding skills. Secondly, their popularity within the skateboarding community, often performed in competitions emphasises their relevance, making them highly relevant for analysing and evaluating competitive performance.

4.2 Data Preparation

4.2.1 Dataset

This study utilised video recordings of skateboarders performing tricks as its primary data source. To ensure model robustness, the final dataset consisted of videos across diverse environmental conditions and varying skateboarder skill levels.

The initial dataset was sourced from the publicly available ”SkateboardML” repository on GitHub [**lightningdrop2020skateboardml**], comprising 200 video clips corresponding to two common tricks: the ollie and the kickflip. To expand the dataset's diversity, additional data was obtained through direct communication with Hanxiao Chen, the author of the SkateboardAI paper [**SkateboardAIPaper**]. This communication yielded a dataset containing 750 videos covering 15 distinct tricks. Given the wide range of tricks included in this dataset, many were beyond the scope of this study, therefore only a subset of these videos were selected and included in the final dataset.

4.2.2 Labelling techniques

This study investigated two primary labelling techniques: the folder-based and the text-based approach as depicted by Figures 4.1 and 4.2. In the folder-based method, videos were categorised into folders named after their corresponding class label offering a simple organisation method. On the other hand, in the text-based approach, each video's path and corresponding label were listed on a text file, providing more flexibility. Given the limited number of classes and manageable dataset size, this study chose to utilise the folder-based approach, as the extra complexity from the text-based method wasn't necessary for this project.

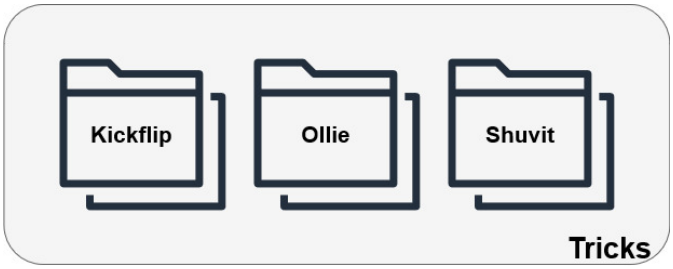


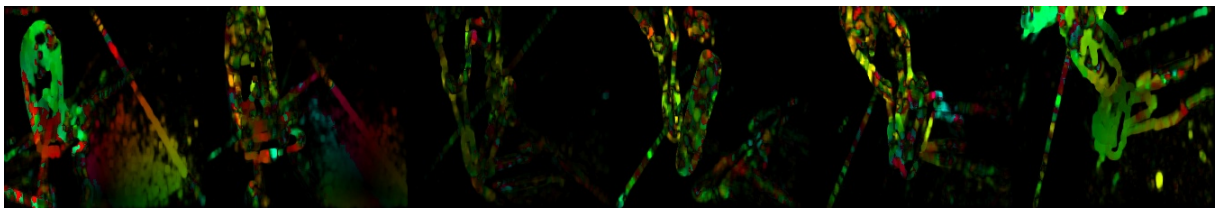
Figure 4.1 Folder-based labelling.

Figure 4.2 Text-based labelling

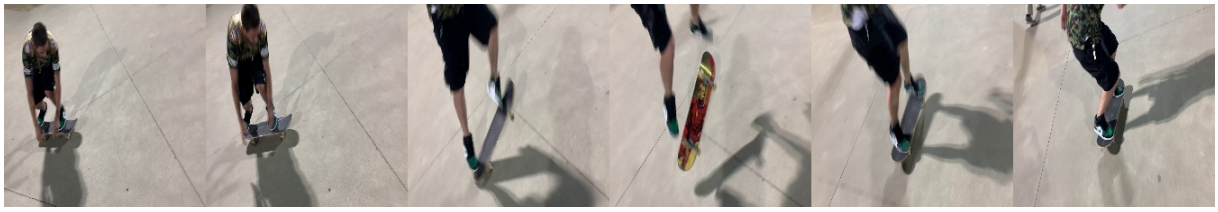
4.3 Preprocessing

4.3.1 Frame Extraction

Given that the nature of this research is an activity recognition task, this study employed frame extraction to convert videos into sequences of frames suitable for processing by ML models. This technique aims to extract a subset of frames that encapsulate the essential dynamics of the activity. Thus, in the context of skateboard trick recognition, the frames should capture the entire sequence, including the wind-up, the trick and the landing.



(a) Optical flow visualisation of a kickflip sequence, highlighting regions with significant motion.



(b) Extracted frames from the sequence based on the greatest weightings from the optical flow representation.

Figure 4.3 Visualisation of optical flow: (a) Optical flow representation (b) Individual frames extracted.

Two primary frame extraction techniques were explored, uniform sampling and optical flow method. The former technique evenly splits a video into its corresponding

frames by using a pre-calculated step size determined by the number of frames of a video. The second technique leveraged optical flow, a method that assigns weights to frames determined by the motion of pixels between them. With this approach, frames with the greatest weightings were chosen for extraction, resulting in capturing frames with the most movement. An example of this technique can be observed in Figure 4.3.

4.3.2 Data Augmentation

Due to the limited number of data used in this research, data augmentation techniques were implemented to increase the dataset's size before model training. These techniques including rotating, flipping and adding noise to images also help reduce overfitting, which was a risk evident due to the over-parametrised nature of the chosen models, meaning that their number of trainable parameters surpass the size of the dataset, making them vulnerable to overfitting.

4.3.3 Normalisation

During the preprocessing stage, normalisation played a crucial role in preparing the video frame data before further processing. In particular this study utilised min-max normalisation to scale the pixel intensities between 0 and 1. Normalising data in this manner ensured that the model's input values data was within a standardised range.

4.3.4 Feature Extraction with Transfer Learning

This research leveraged two well-established CNNs for this task: VGG16 and ResNet50. Both models were pre-trained on the large ImageNet dataset [**imageNetDataset**], leveraging extensive image recognition knowledge that were then fine-tuned to the domain of skateboard trick recognition.

VGG16: VGG16, introduced in 2014 by Simonyan and Zisserman [**VGG**], is a CNN architecture known for its success in image classification and feature extraction for action recognition tasks. This model is relatively simple made up of 16 convolutional and fully connected (FC) with repeated use of 3x3 convolutional filters to preserve spatial information within frames.

Resnet50: Resnet50, introduced in 2016 by He et al. [**Resnet50**], is another powerful CNN architecture that is well known for its advancements in image classification and frame extraction capabilities. Unlike VGG16, this model has a deeper architecture employing 50 convolutional layers allowing it to learn more intricate feature representations of the data. Furthermore, it incorporates residual connections to counteract the vanishing gradient problem that is often caused by many layers.

Table **??** summarises the key characteristics of VGG16 and ResNet50.

Model	VGG16	ResNet50
Published	2014	2016
Layers	16 (Convolutional and FC layers)	50 (Convolutional layers)
Parameters	∼138 million	∼25 million
Input Image Size	224x224	224x224

Table 4.1 Characteristics of Transfer Learning Models VGG16 and ResNet50

The final four fully-connected (FC) layers of these models which are typically used for classification were removed. This was crucial to adapt these models for feature extraction, as the aim was not to classify each frame of the video, but to extract a collection of features that could help sequence models gain a better understanding. These features in the form of vectors not only encapsulate the visual information present in each frame but also more abstract ideas such as low-level information (like edges/shapes) to higher-level features like objects and even actions themselves. By removing the final classification layers, both models resulted in a feature vector shape of $(7 \times 7 \times 512)$.

4.3.5 Dimensionality Reduction

The use of pre-trained models used in this study generate high-dimensional feature vectors, while these features capture valuable information, a large number of dimensions can lead to challenges. Firstly, it increases the computational costs of training models. Secondly, it could potentially lead to the ”curse of dimensionality” as explained by Venkat (2018) [**curseofdimensionality**], where models may struggle to learn effectively with high dimensional data. To address these issues, dimensionality reduction is employed after feature extraction to reduce the number of features while still retaining valuable information for classification.

4.4 Adapted Models

This research investigates the performance of four different deep learning architectures for skateboard trick classification. these models leverage pre-trained CNNs for feature extraction followed by sequence modelling techniques to capture the temporal dynamics of skateboard tricks. The architectures investigated are as follows.

1. **VGG16-LSTM:** This architecture leverages the pre-trained VGG16 model to extract features, which are then fed to an LSTM for sequence modelling. LSTMs

are powerful for capturing long-term dependencies within sequential data such as skateboarding tricks.

2. **ResNet50-LSTM:** This architecture utilises the deeper ResNet50 pre-trained CNN for feature extraction, followed by an LSTM for sequence modelling. The increased depth of ResNet50 potentially allows it to learn more complex feature representations than VGG16.
3. **VGG16-BiLSTM:** This architecture employs the VGG16 model for feature extraction and a BiLSTM network for sequence modelling. BiLSTMs are known to be able to learn dependencies in both directions of a sequence, which could be useful for capturing subtle details in skateboard tricks.
4. **ResNet50-BiLSTM:** This architecture combines ResNet50 for feature extraction with a BiLSTM network. This combination leverages the potential advantages of deeper feature learning and the bi-directional capabilities of BiLSTMs.

The selection of these models is motivated by their effectiveness in prior research for video-based action recognition tasks. For instance, Orozco et al. (2020) achieved an outstanding 91.93% accuracy using a VGG-LSTM architecture [**HARRecognitionInVideosUsingARobustCNNLSTMApproach**]. Additionally Chen's work [**SkateboardAIPaper**] explored all four of these models in the context of skateboard trick classification and demonstrated potential for competitive results. While this study also explored attention mechanisms within these models, the main aim is to build upon existing knowledge and potentially achieve better performance.

4.5 Evaluation Methods

To thoroughly evaluate the performance of the proposed models, this study employed various evaluation metrics, including accuracy, precision, recall, F1-score and confusion matrix. These metrics provided valuable insights into the model's ability to correctly classify different tricks.

The above mentioned metrics depend on the following definitions, described in the context of the "kickflip" class.

- **True Positive (TP)**: The model correctly identified a video as containing a kickflip.
- **True Negative (TN)**: The model incorrectly identified a video as containing a kickflip.
- **False Positive (FP)**: The model correctly identified a video as not containing a kickflip.

- **False Negative (FN)**: The model incorrectly identified a video as not containing a kickflip.

Accuracy: Measures the proportion of correctly classified instances, over the total number of predictions made by the model. This metric provides a generic indicator of the model's reliability, however this specific metric can be sensitive in scenarios with class imbalance [**classificationAssessmentMethodstharwat2020**].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Preci-
sion: Measures the proportion of predicted positive instances that are truly real positives [**Evaluation:fromprecisionrecallandF-measuretoROCinformednessmarkednessandcorrelation**]. In the context of kickflip detection, it represents the proportion of true kickflips against all predicted kickflips.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

Recall: Measures the proportion of positive instances that are truly predicted positive [**Evaluation:fromprecisionrecallandF-measuretoROCinformednessmarkednessandcorrelation**].In the context of kickflip detection, it represents the proportion of true kickflips that are identified correctly.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.3}$$

F1-score: This metric provides a balanced measure of performance by combining both precision and recall. It is calculated using the harmonic mean on both values and outputs a value ranging from zero to one, with values closer to one, indicating better performance.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

Confusion Matrix:

5 Implementation

6 Sample A

7 Sample B