Skateboard Trick Recognition through an AI-based Approach

Kris Saliba

Supervisor: Prof. Joseph Bonello
Co-Supervisor: Dr John Abela

June 2024

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

Abstract

Acknowledgements

Contents

List of Figures

List of Tables

1 Introduction

1.1   Background

1.1.1   Skateboarding

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 Olympics [**SkatebaordingOlimpics**]. Skateboarding comprises the dynamic activities of riding a skateboard and skillfully performing a repertoire of tricks, manifesting as a popular and exhilarating "extreme sport".

This dynamic and multifaceted sport encompasses various disciplines and styles of riding, each of them offering unique challenges for skateboarders to delve into. Two of the most prominent styles are "vert" and "Street", vert skateboarding revolves around riding on specialised structures known as half-pipes or ramps, emphasising areal manoeuvres. Whilst, street skateboarding transpires in urban environments, utilising a diverse array of obstacles, including stairs, rails, ledges, gaps or flat ground for skaters to showcase their creativity and versatility through various techniques [**skateStyles**].

1.1.2   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.

1.1.3   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.

1.1.4   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.

1.1.5   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 as 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. In this research paper, images are passed through an object detection model as one of the pre-processing steps before passing through the main AI model. This strategy is employed with the primary objective of utilising the bounding box output to facilitate the cropping of individual skateboarder frames, thus contributing to a notable reduction in computational overhead.

1.1.6 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. Similarly 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 activi-
ties[**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 1.1. If we were to feed a single frame of this trick into an AI model, as shown in Figure 1.1a, it may misinterpret the manoeuvre as another trick. Whereas, by providing the model with a sequence of frames, as shown in Figure 1.1b, 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.
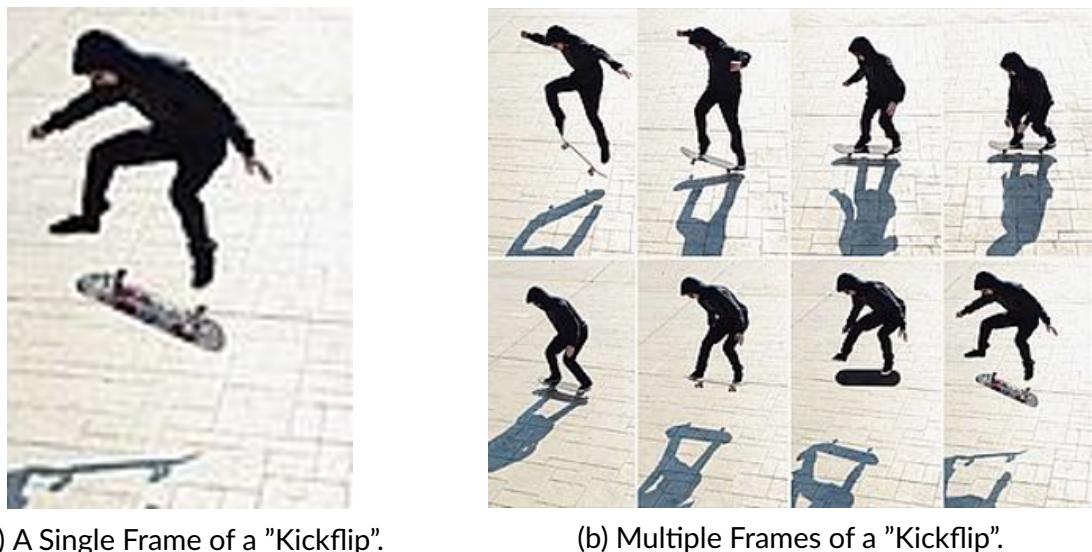
1.1.7 Neural Networks

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**]. Figure 1.2 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**].

(a) A Single Frame of a "Kickflip".   (b) Multiple Frames of a "Kickflip".

Figure 1.1 Comparison of a Single Frame and Multiple Sequential Frames.

- **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.

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.

CNN's, incorporate multiple layers and are capable of extracting effective representations
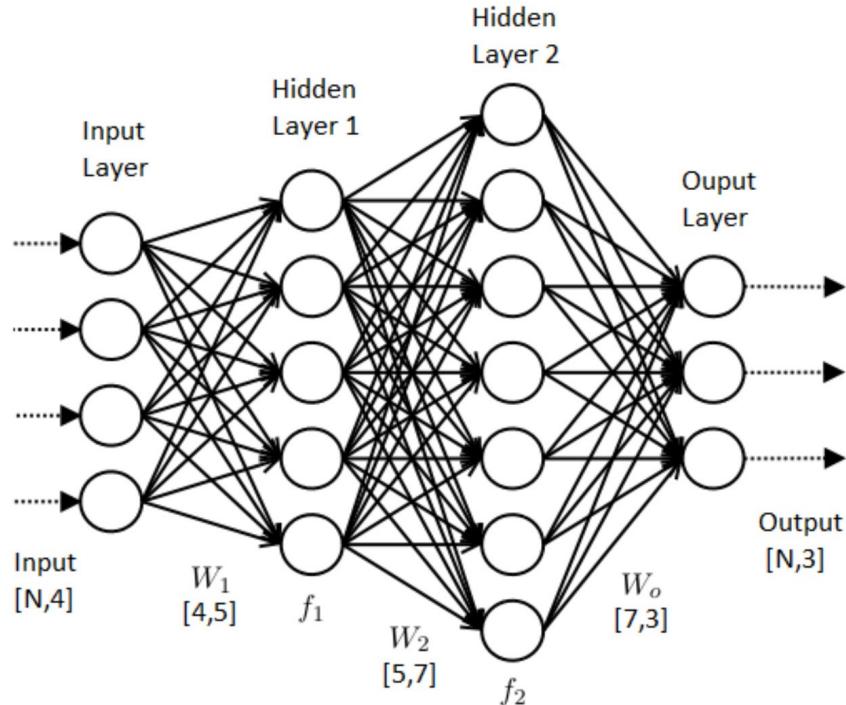
Figure 1.2 Schematic Representation of an Artificial Neural Network with four input variables, three output variables and two hidden layers.

1.1.8 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a subset of Neural networks

1.2 Literature Review

1.2.1 Data Collection

The assortment of data stands out as the foundation of this study, it lends significance to the results as the quality and quantity of data have a direct relation with the capacity of AI algorithms to learn and generalise effectively.

The study [**skatePaper1**] leverages a custom data set, comprising videos capturing the execution of five distinct skateboard tricks, each attempted five times. Each video spans a duration of two to three seconds, yielding a total of 750 images by extracting 30 frames per video. Notably, each skateboard manoeuvre was recorded against a uniform white backdrop, maintaining a consistent distance of 1.26 meters between the skateboarder and the camera. In this study, the researchers applied data augmentation techniques to expand their initial data set while simultaneously reducing the potential for model over-fitting [**DietterichOverfittingAI**] Consequently, they introduced an additional 2,250 images, achieving a total of 3,000 images in their data set. The researchers applied three distinct augmentation techniques, including horizontal rotation, positive 90°rotation and negative 90°rotation.

On the other hand, Chen [**SkateboardAIPaper**] compiled a comprehensive data set by collecting videos from multiple platforms including YouTube, Twitter and Instagram. Furthermore, the researcher trained the model using 15 fundamental tricks commonly observed in competitive settings. They 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 5 were reserved for validation.

1.2.2 Data Preprocessing

An AI model cannot rely solely on raw data to achieve a favourable accuracy. Instead, it benefits from the utilisation of preprocessing techniques that enhance features and guarantee that data is formatted consistently [**DataPreprocessingForImageClassificationByCNN**].

Appendix A   Sample A

Appendix B Sample B