Novel data augmentation schemes for pose classification using a convolutional neural network

JS du Toit

orcid.org/0000-0001-5655-3083

Dissertation accepted in fulfilment of the requirements for the degree *Master of Science in Computer Science* at the North-West University

Supervisor:	Prof JV du Toit
Co-supervisor:	Prof HA Kruger

Graduation May 2024

25088246

ACKNOWLEDGEMENTS

I am grateful to all those who have supported and guided me throughout the completion of this dissertation. Your unwavering encouragement made this pursuit possible.

First and foremost, I would like to express my sincere appreciation to both my supervisor, Prof. Tiny du Toit, and co-supervisor, Prof. Hennie Kruger, for providing me with valuable insights, constructive feedback, and continuous motivation. I am grateful for the time, effort, and expertise you have invested in me and my research, and for your patience and understanding during the extended timeline of this project. Thank you both for being such exemplary mentors, and for helping me achieve my goals.

Secondly, I would like to extend my appreciation to my employer, Dr. Martin Puttkammer, for sympathising with me on the challenges that come with writing a dissertation and for accommodating my research schedule. Your willingness to provide me with the time and flexibility needed to balance both my work and academic commitments was instrumental in my ability to achieve this milestone.

Thirdly, I would also like to thank my family, friends, and colleagues, who offered their perspectives and insights, and provided endless support throughout the entire research process. I am grateful for their willingness to listen, discuss, and share their own knowledge and experiences, which enriched my own perspectives on the subject matter and helped me to develop new insights and ideas.

Finally, I dedicate this dissertation to my late father whose unwavering belief in my abilities was always a constant source of encouragement and my incentive for pursuing a higher education. Although he was not able to witness the completion of this work, his belief in me will continue to inspire and motivate me. This dissertation is a tribute to his memory and his enduring influence on my life.

"There you saw how the Lord your God carried you, as a father carries his son,
all the way you went until you reached this place."
~ Deuteronomy 1:31

ABSTRACT

Population aging is a global trend that can be attributed to declining fertility rates, longer life expectancies, and aging cohorts in many developed countries. The number of individuals over the age of 60 is expected to rise significantly in the coming decades and will likely place additional pressure on healthcare systems worldwide. Currently, falls are one of the leading causes of hospitalisation among the elderly and can result in life-threating injuries and long-term disability. Promptly administering aid to a fall victim can significantly improve their chances of recovery and reduce the need for specialised care. To address these problems, fall detection systems have been developed which facilitate early intervention and reduce the negative consequences of falls. This allows seniors to maintain their independence even after incurring a fall while also alleviating the anticipated pressures on healthcare since injuries are readily treated. However, there is no universal approach or solution to fall detection due to the complexity of the problem domain. New avenues of research are continually stimulated and explored as technological advances are made and new sensor technology becomes available.

Vision-based approaches for fall detection are more favourable than wearable and ambient sensor-based approaches since they are less obtrusive and can provide more information about the context of a fall. However, fall recognition requires quantifying the recorded human body from footage data, which can be achieved through pose estimation and approximating the location of different body parts. This study investigates the effectiveness of data augmentation techniques on such pose estimated mappings to improve pose classification when conducted using a convolutional neural network (CNN). A novel pose descriptor is designed that can be superimposed onto the imaged human body to encode the kinematic arrangement of limbs and relevant positional cues. This method emphasises the postural differences between poses in the abstracted feature space of a CNN classifier, thereby improving the classifier's ability to reliably differentiate between poses. The approach is demonstrated on a fall dataset that consists of multiple pose classes that are typical of everyday activities.

The study results indicate that the proposed visual augmentation schemes are effective in improving CNN-based pose classification. An improvement of up to 11 percentage points was achieved over a baseline pose recognition accuracy that is free of any augmentations. The simplicity of the approach makes it useful for real-time applications that warrant timely action and response. These findings are expected to pave the way for more accurate fall detection systems that minimise the risk of fall-related injuries while reducing the intensity of their medical care.

Key terms: convolutional neural network, data augmentation, deep learning, fall detection, human pose classification, human pose estimation, neural network

CONFERENCE CONTRIBUTIONS

Excerpts from this dissertation have been presented at conferences as follows:

Improved human pose differentiation in Convolutional neural network classification using colour-based data augmentation

J.S. du Toit*, J.V. du Toit, H.A. Kruger

(Abstract presented at the 50th ORSSA Annual Conference, Virtual/Online, 12-15 September 2021)

Colour-based encoding schemes for improved human pose recognition using a Convolutional neural network

J.S. du Toit*, J.V. du Toit, H.A. Kruger

(Full paper presented at the SATNAC 2021 Conference, Central Drakensberg, South Africa, 21-23 November 2021)
See Annexure A or the full paper.

Improving human pose classification using a Convolutional neural network through data augmentation

J.S. du Toit*, J.V. du Toit, H.A. Kruger

(Abstract presented at the 48th ORSSA Annual Conference, Cape Town, South Africa, 12-19 September 2019)

Heuristic data representation augmentation for improved human activity recognition

J.S. du Toit*, J.V. du Toit, H.A. Kruger

(Full paper presented at the SATNAC 2019 Conference, Ballito, South Africa, 1-4 September 2019)
See Annexure B for the full paper.
Recognised as first runner-up for the outstanding student paper award.

*Presenting author

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

2D	Two dimensional
3D	Three dimensional
CIFAR	Canadian Institute for Advanced Research
CNN	Convolutional neural network
FADE	Acoustic fall detection system
FN	False positive
FP	False positive
GPS	Global Positioning System
HPE	Human pose estimation
IDT	Improved dense trajectory
IMU	Inertial measurement unit
MFCC	Mel frequency cepstral coefficients
NFI	Near-field imaging
NN	Neural network
NWU	North-West University
PAF	Part affinity fields
PIR	Passive infrared
PNG	Portable Network Graphics
ReLU	Rectified Linear Unit
RGB	Red Green Blue
RMS	Root mean square
SDK	Software development kit
SVM	Support Vector Machine
TN	True negative
TP	True positive
WHO	World Health Organization

CHAPTER 1 GENERAL INTRODUCTION AND PROBLEM CONTEXTUALISATION

1.1 Introduction

Individuals older than 65 are susceptible to falling and are vulnerable to fall-related injuries. According to the World Health Organization (WHO, 2008), it is estimated that 28–35% of seniors over the age of 65 incur a fall at least once a year, increasing to around 32–42% when over the age of 75. The estimates rise to roughly 50% for the portion of the population aged over 80 (Yacchirema *et al.*, 2019). In the mid-2000s, the WHO identified falling as the second leading cause of accidental death among the elderly. At the same time, close to 50% of hospitalisations were due to fall-induced injuries, of which 40% accounted for non-natural mortalities among individuals over 65 (Kalula *et al.*, 2016; WHO, 2008). Recent research in Finland confirms these statistics, showing a consistent proportion of fall-related deaths among the elderly from 2003 to 2015 (Kannus *et al.*, 2018). In southern Sweden, data on the elderly population indicates a steady increase in the mean age of fall-related deaths, from 77.5 years in 1998–2002 to 82.9 years in 2010–2014. Interestingly, the average age for other causes of death in this age group shows only a marginal increase from 78.5 to 79.8 years. Related research studies ascribe the mean age shift to improvements in the quality of life in the modern day. However, the proportion of elderly individuals experiencing falls remains consistent with the WHO predictions, indicating that the risks remain unmitigated. The relationship between age groups and fall susceptibility was researched by Lan *et al.* (2020) who statistically demonstrated that senior frailty, which naturally increases with advancing age, serves as a significant predictor of falls.

The probability of falling and the associated risks restrict the elderly in their freedom to live alone or without assistive care. Modern trends indicate that most seniors choose to live independently for as long as possible, even after the death of a spouse. However, approximately one in three Europeans over 65 resides alone. This living arrangement can leave them vulnerable in the event of an accident or medical emergency unless they opt for care homes, cohabitation, or hiring personal caregivers (De Miguel *et al.*, 2017; Bloom & LeeLuca, 2016). Cohabiting or assisted living helps prevent elderly individuals avoid what is referred to as a *long lie*, where a fall victim remains on the floor for an extended period (often longer than an hour) due to muscle weakness or a debilitating injury. If not aided in time, complications such as pneumonia, pressure sores, and infection can result from their incapacitation. Providing prompt assistance to fall victims increases their chances of recovery, especially when administered within the first 12 minutes after a fall. This is considered the critical window of opportunity to prevent long-term injuries (Solbach & Tsotsos, 2017; Baraka *et al.*, 2012). Advancements in assistive technology have resulted in

specialised sensor systems that automatically detect falls and notify relevant authorities when triggered. Continuous monitoring by such systems ensures a swift reaction to a fall and reduces intervention time by securing immediate medical attention. However, the intrusive nature of assistive technology infringes on the privacy of individuals, especially when subjected to continuous surveillance. For this reason, the greatest challenge of implementing a fall detection solution lies in selecting less invasive sensors while maintaining high detection accuracy.

Wearable, ambient, and vision sensor technologies are commonly used for fall detection (Ramachandran & Karuppiah, 2020). Each sensor type has its associated advantages and disadvantages, which should be considered based on the application environment and the individuals being monitored for accidental falls. Wearable devices predominately use inertial sensors that are accurate in recording motion, but it requires that the sensor be worn by the user. Active cooperation from the user is necessary to keep the device charged and on their person, which may hinder its effectiveness if the intended user is forgetful or considers the device uncomfortable. Ambient sensors circumvent these challenges through vibration and pressure sensors that operate independently. However, these sensors are only effective if deployed throughout an entire home to ensure complete coverage. Vision sensors such as cameras also operate independently and can provide accurate fall signals through video and image data. Although the visual nature of vision sensor recordings is considered highly intrusive and can be met with privacy concerns, vision sensors can capture the human silhouette, enabling reliable motion and pose tracking compared to other sensors.

To detect fall events using visual sensors, it is first necessary to synthesise potential fall cues from the data which relate to the body's posture and the location of specific body parts. Standard calculations in human activity recognition involve measuring fluctuations in stance, rate of change, and body orientation. However, vision-based approaches are frequently met with technical hindrances resulting from dim lighting, viewpoint variations, and visual occlusions (Vallabh & Malekian, 2018). Image processing techniques such as data augmentation have been used to mitigate some of these obstacles and suppress their influence in machine-learned classifiers. Past research studies show that data augmentation can influence the accuracy of neural networks (NNs) by injecting and removing visual cues that serve as features for classification. Applying data augmentation to human pose recognition can potentially enhance fall detection. The possibility is evident in the elective learning format of convolutional neural networks (CNNs), which selectively extract features that best distinguish class instances during training (Goodfellow *et al.*, 2016). However, little existing research has focused on utilising data augmentation to improve pose classification for the purpose of fall detection.

The potential to enhance pose-estimated key points of the human body using novel data augmentation techniques is investigated as part of this dissertation. Multiple CNNs are trained on augmented datasets to evaluate the introduced augmentation schemes that are inspired by established data augmentation techniques. Given the visual nature of the task, a CNN is considered highly proficient at abstracting features from image data and has achieved state-of-the-art performance in many related research areas (Sze *et al.*, 2017; Goodfellow *et al.*, 2016). In the end, the most successful augmentations will be combined and applied to a real-world video dataset to evaluate its feasibility for fall detection.

The remainder of this chapter is structured as follows. The research question is stated in Section 1.2, followed by the primary objective and secondary objectives of this dissertation in Section 1.3. Section 1.4 provides an overview of the research design regarding the research paradigm and methodology. The ethical implications of this study are briefly discussed in Section 1.5, along with evidence of ethical clearance in undertaking this study. The purpose of each chapter in this dissertation is detailed in Section 1.6 before concluding in Section 1.7.

1.2 Research question

This study entails performing pose recognition for the purpose of fall detection using enhanced pose-estimated mappings of the human body. Novel data augmentation schemes are applied to a dataset of human poses that are then comparatively evaluated based on differences in pose recognition accuracies from separately trained classifiers. Consequently, the research question can be stated as: "*Can pose estimation data be enhanced through heuristic data augmentation to obtain an improvement in CNN image classification accuracy when performing pose recognition?*"

1.3 Research goals

The primary objective of this study is to improve pose classification using a CNN by establishing a set of data augmentation techniques that enhance pose-estimated mappings of the human body in an image or video. The following secondary objectives contribute to the primary objective:

1. Conduct a literature study of domain-related topics, namely fall detection, pose estimation, CNNs, and data augmentation.
2. Collect and assemble a human pose dataset that consists of two classes, namely sitting and standing poses.
3. Identify a pose estimation implementation that can pre-process pose data to derive points of interest on the human body.

4. Propose data augmentation techniques that enhance the feature space of a CNN.

5. Determine how data augmentation can be applied to pose data for improved pose classification.

6. Establish a suitable CNN architecture and hyperparameter configuration that supports pose classification.

7. Evaluate the effectiveness of the data augmentation techniques when applied to a real-world dataset.

1.4 Research design

Research design is the framework of methods and techniques for planning, implementing, and evaluating a research study. According to Trochim and Donnelly (2006), research design is the strategy that determines how the various components of a study are integrated and structured in a cohesive and logical manner to effectively answer the research question. Essentially, it dictates how the study will be conducted and ensures that informed decisions can be made regarding the methods for data collection, analysis, and interpretation. The research paradigm and methodology are central to the research design since it is informed by the underlying assumptions and principles of each. In this section, the chosen research paradigm and methodology are described regarding how they shape the research approach of this dissertation.

1.4.1 Research paradigm

Research can be conducted along one of three general dimensions. Each dimension consists of ontological and epistemological assumptions that collectively constitute how knowledge can be defined within "a system of interrelated practices of thinking that define the nature of enquiry" (Thomas, 2010:292). This study follows an experimental design approach and therefore fits within the conceptual framework of positivism, which is fundamentally concerned with experimentation through empirical means. Moreover, positivism states that knowledge is both objective and quantifiable and can be enacted by remaining unbiased during data acquisition and when drawing insights from the conducted research (Oates *et al.*, 2022; Thomas, 2010).

The objective approach of positivism is also reflected in the scientific method, consisting of three techniques: reductionism, repeatability, and refutation (Oates *et al.*, 2022). These techniques are centred on an epistemological worldview which states that relationships between phenomena can be identified and tested through reasoning and analysis. This view is supported by the two basic assumptions fundamental to the scientific method. First, the world is considered logical and cannot be regarded as random, which dictates that the world can be objectively observed (Oates

et al., 2022). Second, the world can be modelled by observing phenomena and proving the discoveries through repeatability as part of a reductionist process. The research conducted as part of this dissertation is approached under positivism to ensure that observations remain impartial and that findings drawn from experiments support objective outcomes for pose recognition and are not influenced by personal bias.

1.4.2 Research methods

A CNN image classifier for human pose recognition is constructed using an experimental design methodology. Novel data augmentation techniques are applied to the datasets used in training and evaluating the pose classifier models. The pose data is derived from public-domain images that depict people in sitting and standing poses and is pre-processed using pose estimation to delineate the human body. As part of the study, a suitable CNN architecture is first established that achieves sufficient classification accuracy on the non-augmented dataset. Then, a preliminary set of experiments is performed to identify which augmentation techniques benefit pose recognition accuracy. The findings from these experiments inform the second set of primary experiments, where pose recognition is performed on a fall dataset to assess the effectiveness of the proposed augmentation techniques in addressing a real-world problem (Adhikari *et al.*, 2017).

Augmented data samples are generated using the results of pose estimation. First, a pose estimator is employed to derive a skeletal mapping of 18 key points representing body joints and other significant body parts. These key point coordinates are then augmented with varying marker sizes, colours, and blending where markers overlap. A set of preliminary experiments are then conducted to establish a favourable augmentation scheme by evaluating the individual and combined effect of these factors on classification performance. In the primary experiments, the augmentation scheme is refined, and the new key point mapping is projected directly onto pose images (video frames). In both sets of experiments, a CNN-based pose classifier is trained using their respective datasets. The augmentations are evaluated based on the classification results and how well it emphasises underlying feature-based similarities among the same sets of poses. Measures such as classification accuracy, sensitivity, and recall are utilised to quantify the improvements that result from applying the augmentations.

1.5 Ethical considerations

The proposal for this study was presented to the Faculty of Natural and Agricultural Sciences' ethics committee of North-West University, where ethical clearance was granted with minimum to no risk under the reference number NWU-01162-20-A9. The following ethical considerations aimed at safeguarding the integrity of this study were observed throughout its course:

- All institutional, national, and international laws, regulations, and policies applicable to the research are adhered to in this study.
- The research performed is conducted in a manner that will allow verification and replication by others.
- Appropriate research methods are employed at all times; conclusions are based on critical analysis of the evidence, and findings and interpretations are thoroughly and objectively reported.

It should be noted that both the compiled and the real-world datasets allow for unrestricted use. The compiled pose dataset is derived from images in the public domain or under a creative commons licence agreement that provides for unrestricted use. The real-world dataset provided by Adhikari *et al.* (2017) is available online, and its licencing agreement allows for free use towards research ends, provided the original creator and author are acknowledged.

1.6 Outline of chapters

The remainder of this thesis is organised as follows:

Chapter 2 provides the necessary background information on fall detection. The study is introduced with a motivation for fall detection in the context of an ageing world demographic, its expected impact on society, and its consequent demand on healthcare. A literature study of the most prominent approaches to fall detection is also provided, focusing on the variety of sensor technologies and their existing challenges, shortcomings, and advantages.

Chapter 3 offers a brief literature study of pose estimation, where the different approaches to achieving this task are discussed and demonstrated in the context of past research attempts. The chapter continues with an explanation of pose estimation as a feature synthesis method and its use in quantifying the human body from raw visual data. The chapter concludes with a discussion of *OpenPose*, a pose estimation implementation selected as the preferred solution to perform data pre-processing in this dissertation.

Chapter 4 contains an introduction to NNs regarding their inception, internal functioning, and utility in machine learning to produce reliable classifiers. The chapter also contains a discussion of CNNs and how their unique properties that benefit image classification can also favour pose classification. To capitalise on the favourable strengths of a CNN, data augmentation is explained and credited as a strategy to enhance pose classification performance by manipulating significant features that are relied upon in image classification.

Chapter 5 contains the results and discussions of two sets of experiments. The preliminary experiments are presented as an initial investigation into how data augmentation can be purposed for pose classification by engineering a pose descriptor using body key points. A set of primary experiments are then presented, which build on the initial experiment results to yield a pose descriptor that encodes spatial information through colour. The chapter concludes with a discussion of the valuable insights gained from these experiments and their potential value in pose-dependent applications.

Chapter 6 concludes the dissertation by reflecting on the research question and intended objectives of the study and assessing how they were addressed throughout the dissertation. This chapter summarises the research outcomes while reflecting on which data augmentation schemes and techniques were found to best support pose classification. Identified limitations and possible improvements regarding the research study are also presented, and potential future research opportunities are discussed.

1.7 Conclusion

This chapter served as an introduction to the dissertation by outlining its structure, purpose, and objectives. The research domain was described, which emphasised the importance of fall detection among the elderly population and motivated the research question regarding data augmentation and the enhancement of pose signals to achieve improved CNN-based pose classification. An overview of the research paradigm, methodology, and ethical considerations was provided in the preceding sections before concluding with an outline and description of each chapter presented in the dissertation. This chapter helped establish a foundation for the study and the research process required to investigate how pose classification can be enhanced through the strategic application of data augmentation techniques.

CHAPTER 2 FALL DETECTION

2.1 Introduction

The goal of this chapter is to provide an overview of fall detection concepts, techniques, and approaches. Because human-orientated technologies require situational awareness of their immediate environment, a variety of sensors are necessary to capture the movement and actions of people. A literature study of notable research from recent years is presented in this chapter that highlights the diverse range of sensor technologies while also showcasing the evolution of fall detection systems. Furthermore, relevant concepts are explained regarding falls and their identification, the degrees of automation in fall detection, and the standard evaluation metrics of such systems. Fall detection is defined in this chapter as the problem domain by contextualising the experimental design of this dissertation in terms of a real-world problem, thereby substantiating its practicality and applicability.

The remainder of this chapter is organised in the following manner. Section 2.2 details the motivation for fall detection in terms of an ageing world demographic, its likely impact on society, and the expected demand on healthcare. Section 2.3 provides an overview of the technical concepts related to fall detection, such as the indicators of an accidental fall, the generational categories of detection solutions, and the typical evaluation metrics for these solutions. In Section 2.4, both the structural framework and the sensor-based approaches to fall detection are discussed. Finally, a survey of prominent technological solutions in each of the three sensor-based approaches to fall detection is presented in Section 2.5 and focuses on the advantages and disadvantages of each before the chapter is concluded in Section 2.6.

2.2 Ageing populations as a motivation for fall detection

The global demographic trend towards an ageing populace brings new causes for concern related to geriatric healthcare and the rise in life-threatening propensities like accidental falls. Many parts of the world currently possess a low birth rate that no longer supplements past levels of population growth, and fertility rates in most developed countries have fallen below replacement levels. According to the Organisation for Economic Cooperation and Development (OECD, 2017), the number of individuals aged over 65 is projected to nearly triple by 2060, resulting in 58 retirees per 100 working individuals, up from the current ratio of 20 to 100. The net effect of this population shift will be apparent by as early as 2050, at which time the WHO predicts that elderly individuals will outnumber children and adolescents for the first time in known human history (Uddin *et al.*, 2018; WHO, 2008). This change will lead to new economic uncertainties and unprecedented pressure on healthcare systems.

The last known systematic review of costs associated with fall-related injuries in major countries recommended that fall prevention programmes be implemented to help curb the total healthcare expenditures (Heinrich *et al.*, 2010). Such measures are meant to relieve the increasing demands surrounding geriatric care, which often require specialised treatments and lengthy hospital stays when seniors are admitted (Bloom & LeeLuca, 2016). In addition, the anticipated population shift will likely lead to a diminished active workforce, giving way to a shortage of skilled care workers and medical personnel (Campbell *et al.*, 2013; Fu *et al.*, 2008). Prevention programmes and technological innovation are among the cited means that could help alleviate the expected pressure on healthcare. Some countries, such as Singapore, have begun taking pre-emptive action by heavily investing in their healthcare infrastructure and hosting conferences dedicated to assistive technology for the elderly (Yu, 2008).

Fall detection is primarily aimed at preventing a long lie, which occurs when an individual remains immobilised on the floor for an extended period of time after sustaining a fall. A long lie is associated with a 50% mortality rate among seniors, even in the absence of any physical injuries (Kalula *et al.*, 2016). Prompt assistance is essential for fall victims to improve their chances of recovery. In severe cases, timely treatment within the first 12 minutes after a fall can increase the survival rate by up to 75% (Baraka *et al.*, 2012). Falling can also induce a fear of falling, which can restrict an individual's participation in physical activities, potentially resulting in social isolation and reduced productivity (Vallabh & Malekian, 2018). The elderly population is particularly vulnerable to these consequences, especially those who live alone. Providing these individuals with an automated fall detection solution can ensure timely aid and prompt medical attention without requiring them to be supervised or relocating to a monitored facility such as a nursing home.

2.3 Fall detection concepts

Defining the nature and characteristics of accidental falls is necessary to understand how such an event can be automatically recognised. Subsection 2.3.1 provides insights into the characteristics of a fall and how these indicators might be relied upon for fall detection. Additionally, the modes of fall reporting are described in Subsection 2.3.2, which range from manual solutions to pre-emptive warning systems, and demonstrates the trade-offs between simplicity, accuracy, and real-time detection capabilities. Finally, Subsection 2.3.3 provides an overview of the evaluation metrics that measure the efficacy and reliability of a standard fall detection system. These topics establish a foundation for the subsequent discussions and analyses presented in this dissertation.

2.3.1 The nature and indicators of accidental falls

According to the authoritative work of Hyndman *et al.* (2002), a fall is defined as an unintentional loss of balance resulting in an involuntary impact with a lower surface. Following a fall, the posture assumed by the victim is likened to that of a lying pose. A person who is unexpectedly found in a lying position demonstrates a sign of abnormal behaviour, indicating that a fall was likely incurred. The probability is higher when a lying position is assumed in the proximity of lounging furniture, such as a bed or a couch. Additionally, falls can be classified based on their originating posture, as described in the work of Yu (2008). This classification distinguishes falls based on their originating pose, namely from upright positions (e.g., standing or walking), sitting positions, lying positions, as well as transitional and inter-transitional postures like bending or crawling. Similarly, Zhang *et al.* (2015) argue that falls should be categorised not only based on the final resting position but also by their direction namely, forward falls, backward falls, and side falls.

When observing the human body in video and image data, its visual appearance remains consistent across frames, while variations are apparent in the position of limbs, orientation, and posture. Identifying and focusing on indicative cues from these factors of variability can help to differentiate poses, especially those with similar characteristics. A fall is commonly associated with an abrupt change in velocity and can be used to distinguish between purposefully lying down and the accidental loss of balance. In other cases, an indicative cue can inadvertently increase ambiguity. For example, an abrupt change in velocity can be mistaken for a fall when purposely sitting down swiftly. Other characteristics of fall detection include measured changes in the height and width dimensions of the human body, an abnormal inclination angle, and sudden head displacement (Zhang *et al.*, 2015). Automated fall detection systems often rely on one or more of these indicators to identify a fall event.

2.3.2 Degrees of automation in accidental fall recognition

Fall detectors can be classified into three generations based on their level of user interaction and embedded intelligence (Martin *et al.*, 2006). First-generation devices are typically rudimentary remote-controlled signalling devices that are worn by the user and activated by pressing a button. These devices, often referred to as personal alarm systems, have limitations because they depend on the user's capacity to activate them, making them ineffective in situations where the individual is incapacitated and unable to report a fall. These shortcomings extend to situations where the user experiences a medical emergency such as a heart attack or stroke and the device happens to not be at hand or within reach (Vallabh & Malekian, 2018; Hazelhoff & Han, 2008).

Second-generation devices automate fall detection through passive monitoring and sensor readings, eliminating delays in alerting responders to a fall. Any readings that deviate from the

expected norm can be analysed to determine if they indicate a probable fall. Most automated devices can be advanced to third-generation devices if a pre-emptive (rather than reactive) approach is taken by recognising early warning signals from incoming signals. Third-generation devices monitor factors that contribute to the risk of falling, such as dim lighting or abnormal vital signs, and produce a warning when the potential of a fall is high. The approach taken in this study targets improving second-generation fall detectors that operate based on a reactive approach and which use visual sensors to perceive fall indicators.

2.3.3 Evaluation metrics for fall detection systems

Fall detection can be approached as a binary classification and implemented using supervised machine learning. The goal of a fall detection classifier is to determine whether or not a sequence or selection of video frames contains a fall. Its ability to reliably recognise such events is commonly expressed through sensitivity (Equation 2-1) and specificity (Equation 2-2) (Mubashir *et al.*, 2013). The former denotes to its capacity to correctly classify a fall instance as a fall and is also referred to as *recall* or the *true positive rate*. The latter denotes the classifier's ability to correctly identify a non-fall instance as not a fall and is referred to as the *true negative rate*. Essentially, sensitivity measures how effectively a system predicts falls, while specificity measures how well it differentiates falls from other events. Because falls are naturally infrequent in everyday life, a truly representative dataset would reflect this imbalance with fewer fall instances compared to other classes. Sensitivity and specificity measures are not influenced by such class imbalances and maintain a true reflection of the classifier's performance (Nunez-Marcos *et al.*, 2017). Additionally, accuracy (Equation 2-3) is employed in this study to allow for comparison with other approaches in the literature since it is a standard evaluation metric in classification tasks. In summary, the evaluation and comparison of fall detection classifiers in this study are based on the following three metrics:

$$Sensitivity/Recall = \frac{TP}{TP + FN} \tag{2-1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2-2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2-3}$$

where TN refers to true negatives, TP to true positives, FN to false negatives, and FP to false positives.

2.4 Approaches to fall detection

Fall detection solutions share a common operational framework regarding its design and implementation. The framework is described in Subsection 2.4.1 and highlights the fundamental steps of data acquisition, processing, and fall reporting. In addition, the sensor technology frequently integrated into these systems is introduced in Subsection 2.4.2. Examining these underlying principles and key components facilitates a better understanding of the core elements behind the development of effective and reliable fall detection systems. This information serves as a conceptual foundation for the subsequent discussions and evaluations of various fall detection approaches and methodologies in this chapter.

2.4.1 Structural framework of fall detection solutions

Automatic fall detection has become a recurring topic in elderly-centred healthcare. With the ongoing development and state-of-the-art achievements in computer vision and machine learning, new approaches and sensor technologies are continually being adopted to improve fall detection (Birku & Agrawal, 2018). Each proposed solution aims to satisfy the following four design requirements: high accuracy, low cost, automation, and the capacity to perform real-time computing (Lu & Chu, 2018). Diverging approaches to fall detection have emerged that often rely on a single sensor technology that best aligns with the design requirements (Mubashir *et al.*, 2013). The choice of sensor technology is also based on its advantages, such as accuracy and time efficiency, as well as its potential drawbacks, including privacy concerns (Daher *et al.*, 2017). Regardless of the technology implemented, the same general operational framework is maintained for most fall detection systems, as is illustrated in Figure 2-1.

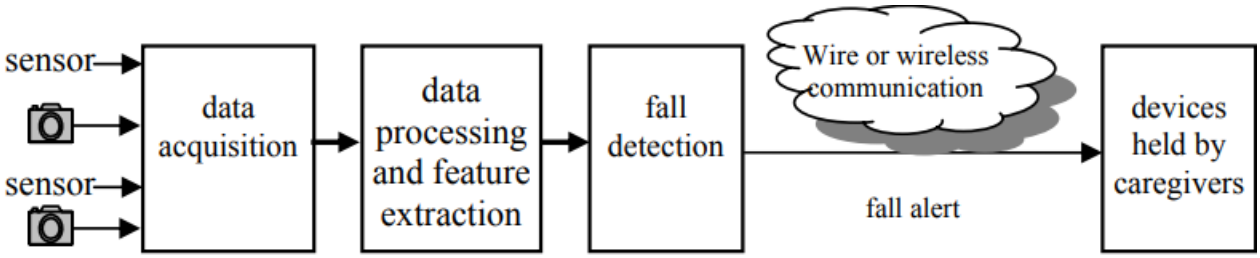


Figure 2-1: General framework for fall detection systems (Yu, 2008)

The framework illustrates how a deployed sensor is used to monitor the environment and evaluate readings as part of a fall detection algorithm. If a positive signal is registered, the system triggers an alert to notify the relevant caretakers using either a wired or wireless communication method. The next subsection provides a description of each sensor-based category of fall detection solutions that typically follow this framework.

2.4.2 Sensor-based categories of fall detection solutions

Different approaches and technologies have been experimentally applied to fall detection over the years, resulting in three broad sensor-based categories being established (Ramachandran & Karuppiah, 2020; Vallabh & Malekian, 2018). These categories are based on *wearable sensors*, *ambient sensors*, and *camera sensors*:

1. Wearable sensors

   Wearable technology encompasses sensors such as accelerometers, gyroscopes, tiltmeters, and oscillometers, which are typically integrated into wearable devices to monitor movement (Ramachandran & Karuppiah, 2020; Mohamed *et al.*, 2014). However, wearable technology is highly prone to false alarms in fall detection, given the similarities in readings generated by activities of daily life and falls (Rougier *et al.*, 2011). Additional drawbacks of this technology relate to the limited battery life of a wearable device and its dependency on the user regarding their accountability to continually wear the device.

2. Ambient sensors

   An ambient sensor monitors its immediate environment for changes in motion, light, pressure, sound, and vibration to detect the presence of a person (Birku & Agrawal, 2018; Uddin *et al.*, 2018). Specific actions can be derived from the sensor readings by identifying patterns in the data. However, ambient readings are prone to false fall detections because the source of the sensor reading cannot be verified as either originating from a person or an object in the environment. Moreover, privacy is a common concern for individuals monitored by this technology, particularly when sensors can reveal their location, like in the case of underfloor pressure sensors.

3. Camera sensors

   Visual approaches rely on camera technology and image processing techniques to identify falls in various settings, such as public spaces or personal homes (Birku & Agrawal, 2018; Zhang *et al.*, 2015). Cameras are often favoured over other sensors for fall detection since they can operate independently without relying on any action from the individuals being monitored. It is also cost-efficient since camera technology is readily available, and a single sensor is typically sufficient to monitor a large area. However, cameras are regularly met with privacy concerns, given the visual nature of their readings.

A literature study on each sensor-based approach to fall detection is provided in the next section.

2.5 Literature study of approaches to fall detection

Exemplary solutions toward fall detection in each sensor-based approach are presented in this section that help to demonstrate the operational concerns and successes of each technology. Moreover, these solutions exemplify how fall indicators are derived from various sensor readings. To provide a holistic perspective of these insights, this section concludes with a summary outlining the key advantages and disadvantages of each approach.

2.5.1 Wearable-based fall detection

This approach to fall detection relies on sensors embedded within a wearable device that record signals related to body movement, actions, and often location. These devices monitor various physiological parameters such as electrocardiogram, heart rate variability, and pulse oximetry, in addition to motion data from gyroscopes, accelerometers, and magnetometers. These values are typically passed as features to either a machine-learned or threshold-based system to detect falls. Wearable devices are considered a disruptive method of fall detection because they depend on the user's accountability to wear the device. However, this approach helps preserve the freedom of movement for the individuals being monitored, unlike ambient- and vision-based sensors that are limited to an area of operation. Furthermore, the cost of implementing a wearable approach is comparatively lower since no additional sensors are required to support the coverage of additional or more expansive areas (Ramachandran & Karuppiah, 2020).

Wearable devices are practical solutions when designed to operate wirelessly, but this comes with functional restrictions related to battery life and the mode of signal processing. Onboard signal processing ensures that the device can register a fall event and notify a relevant caregiver by incorporating the complete fall detection framework into the device itself. On the other hand, remote processing involves transmitting sensor data to an external processing unit, such as a phone, for analysis. The trade-off between the two modes is apparent in the size and subsequent usability of the device. A fully equipped device is often bulky and uncomfortable and may require frequent recharging, which is cumbersome for the user. Developing reliable and robust wearable fall detection solutions necessitates balancing size, functionality, and user experience (Ramachandran & Karuppiah, 2020; Yu, 2008).

2.5.1.1 Fall detection solutions that implement wearable sensors

Wearable devices typically consist of three components (González-Cañete & Casilari, 2020):

1. The *sensing sub-system* incorporates various inertial sensors and sometimes biosensors to monitor the movement and physical condition of the user. Additionally, a global positioning sensor (GPS) is sometimes incorporated to track the location of the user.

2. The *decision core* produces a decision by applying a detection algorithm to the incoming sensor signals. The decision can be binary, as in fall detection, or multi-classed when performing other forms of activity recognition.

3. Wearables are often equipped with a *communication module* that can produce a local or remote alert when a specific action, such as a fall, is registered by the decision core.

The above-mentioned components are present in many modern smartphones and have therefore emerged as an easily accessible solution to implement fall detection. They offer adequate computing power to accommodate a detection algorithm and are natively equipped with an inertial measurement unit (IMU) and telecommunication technology. Consequently, subcategories for wearable fall detection systems have emerged: *dedicated devices*, *smartphone-adopting architectures,* and *hybrid architectures*. In addition, solutions are broadly categorised in terms of the detection algorithm used, typically threshold-based or machine learning-based.

2.5.1.1.1 Dedicated devices

Dedicated fall recognition devices are specifically designed single-purpose devices worn ideally on the waist, wrist, or head (Kangas *et al.*, 2009). These devices integrate both the decision core and the communication module components into their construction, and their sensors are either embedded in the device itself or strategically placed on the user's body. Most commercial fall detection systems employ this design and are sold as bracelets, pendants, or waist belts. Dedicated devices primarily rely on an accelerometer and often incorporate additional sensors to reduce false detections (Tong *et al.*, 2013). The detection of falls is based on monitoring variations in gravitational acceleration (expressed in *g*) and is measured using a standard 3-axis accelerometer. These values are analysed to isolate the four phases of an accidental fall which are illustrated in Figure 2-2 and discussed below (Pierleoni *et al.*, 2015; Kangas *et al.*, 2009):

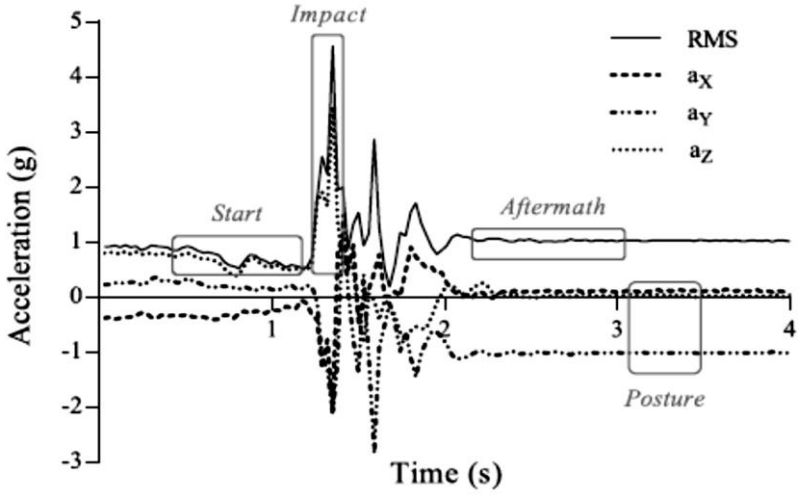


Figure 2-2: Changes in acceleration during an accidental fall (Pierleoni *et al.*, 2015)

1. The *start phase* commences when the subject loses contact with the floor and approaches the ground. During this phase, the body accelerates as it is pulled down by gravity and will achieve a maximum speed just before impact. This free-fall phase is identified by a root mean square (RMS) value that drops below 1*g* and approaches 0*g*.

2. The *impact phase* initiates once contact is made with the ground and the free-fall is abruptly halted, resulting in a sharp and rapid spike in the RMS reading. The RMS often peaks at just above 2*g* in this phase, depending on the speed of the fall and the intensity of the impact.

3. The *aftermath phase* denotes the immediate period of inactivity following the impact phase where the person remains motionless on the floor (Kangas *et al.*, 2009). This phase is characterised by a flat trend in the RMS reading and is only affected when a still-conscious person attempts to move or get up.

4. The *recovery (posture) phase* occurs approximately two seconds after the initial impact, denoting the end of a fall event. This phase is characterised by low acceleration readings across all three accelerometer axes. At this point, a person will typically adopt a resting supine, prone, or lateral recumbent posture as a result of the fall (Kangas *et al.*, 2009).

Most accelerometer-based devices for fall detection utilise at least two of the previously mentioned signal phases, namely the start, impact, or posture phases, as the basis for identifying a fall event (Kangas *et al.*, 2009). However, Pierleoni *et al.* (2015) has argued for the inclusion of the aftermath phase in fall detection, as it provides valuable insights into whether a person was able to recover from a fall. In their study, Pierleoni *et al.* (2015) proposed an alternative signal calculation approach using an inertial unit fitted to a belt, which incorporated an accelerometer, a gyroscope, and a magnetometer. Their unit could measure the velocity, angle, and pitch of a person, thereby facilitating a more comprehensive characterisation of the four phases of a fall. Furthermore, it provided additional information on the severity of a fall, the orientation of the person before and after the fall, and whether the individual managed to regain an upright posture.

One limitation of the proposed approach was the absence of a communication module, which relied on a wireless connection to a smartphone to deliver an alert when a fall was registered. The system was found to be unreliable when the device was out of range of the connected phone. Pierleoni *et al.* (2015) also observed that accelerometers were easily affected by external vibrations and accelerations caused by trembling of the device or its rapid movement, particularly when placed on an extremity (Pierleoni *et al.*, 2015; Fisher, 2010). Moreover, accelerometer data alone did not provide information about the posture of a fall victim, which the researchers argued is a critical indicator for reliable fall detection. In a subsequent study, Pierleoni *et al.* (2016) addressed these limitations by adding a barometer sensor into their device to measure the altitude

during a fall. Although this addition did not directly capture the person's posture, it significantly improved the identification of the aftermath phase, resulting in improved performance compared to other existing fall detection systems at the time.

2.5.1.1.2 Smartphone-adopting architectures

Smartphone-based solutions are not yet as widely adopted as dedicated devices, with the earliest study employing a smartphone only being conducted in 2009 by Sposaro and Tyson (2009). The IMU in most smartphones is equipped with the necessary set of sensors to perform fall detection, namely a gyroscope, a magnetometer, and an accelerometer. While the specialised sensor technology in dedicated devices may be more advanced than smartphone IMUs, the high cost associated with dedicated devices poses a challenge for their commercial viability. In contrast, smartphone adoption helps circumvent the cost and expertise needed to develop dedicated devices. This approach takes advantage of a smartphone's data collection capacity, computing power, storage, dedicated IMU, and built-in communication module.

iFall, designed by Sposaro and Tyson (2009), is one of the first smartphone applications to be developed for fall detection. It uses the built-in smartphone accelerometer to identify falls by analysing acceleration magnitudes, timeouts, and any periods of inactivity following a suspected fall. The application alerts nominated individuals of a fall event once a period of inactivity is detected, signalling the onset of a long lie. iFall was mimicked in later smartphone-based solutions with improvements to the detection algorithm. One such improvement was proposed by He *et al.* (2012), who opted to use a smartphone's motion sensor to monitor the gravitational vector against the longitude axis of the device. If the angle surpassed 40°, it signified a significant change in posture or orientation, suggesting a fall position. However, this approach required users to keep the phone securely positioned on their waist to ensure accurate readings. Similarly, Thammasat and Chaicharn (2012) required that the phone in their experiments be kept in the front shirt pocket to maintain consistency in the orientation of the sensors. Their system also improved the detection algorithm by incorporating tri-axial acceleration readings and identifying any actions that surpass predefined threshold speeds of normal everyday activities.

Shen *et al.* (2016) proposed a third-generation solution to fall detection that focused on intrinsic factors that lead to falls among the elderly. Their system made use of gait information sampled through a smartphone accelerometer to predict an impending fall and provide a preventive alert to the user. By training the system on an individual's regular gait patterns, any deviations in their expected walking pattern could be used to recognise muscle weakness or balance instability. While third-generation solutions cannot guarantee fall prevention, they enable the earliest possible alert before an actual fall occurs. Using a smartphone in such a solution is not ideal since

its reliability is subject to a limited battery life, which is heavily strained by continuous monitoring and signal processing (Yacchirema *et al.*, 2019).

Additionally, smartphone sensors generally lag behind the capabilities of dedicated devices. For example, the last generation of smartphones included accelerometers with a measurement range of only 2*g*, while dedicated devices at that time offered ranges from 3*g* to 16*g* (Medrano *et al.*, 2016; Casilari *et al.*, 2015). Furthermore, variations in the sensor reading range among different manufacturers make it difficult to establish reliable fall detection thresholds across all smartphones (Vallabh & Malekian, 2018). As sensor technology improves and becomes more cost-effective, the advances will cascade down into smartphones, and widespread adoption of smartphone-based fall detection will likely soon follow (Habib *et al.*, 2014).

2.5.1.1.3 Hybrid architectures

Smartphone-based solutions regularly require that phones be placed in specific locations, such as on the waist or chest, to obtain accurate sensor readings (Abbate *et al.*, 2010). However, this requirement is unnatural for people who habitually keep their phone in their pocket or elsewhere. Moreover, sensors onboard a phone are of lower quality than dedicated devices since they support the functionality of a phone and not that of a dedicated fall detection system. Hybrid solutions aim to resolve these limitations by using a dedicated device to record sensor data and a smartphone to process the data. Some advanced hybrid approaches even rely on the monitoring device to pre-process the sensor data and extract features prior to its processing as a means to conserve the battery life of the connected smartphone (Habib *et al.*, 2014).

One example of a hybrid solution is SensorTag which was developed by Terroso *et al.* (2013). In their approach, a wearable device was integrated into an underwear garment and connected to a smartphone through Bluetooth. All incoming data was processed by the smartphone to identify falls, and an alert was generated when a positive detection was made. The notification also included location information derived from its onboard GPS. Similarly, Santoyo-Ramón *et al.* (2018) proposed a hybrid approach using a variety of sensors and different device placements to evaluate their effectiveness. The sensors and smartphone application were organised into a piconet topology where the smartphone acted as the master by receiving multiple sensor readings through Bluetooth Low Energy. Santoyo-Ramón *et al.* (2018) concluded that sensors near the body's centre of gravity, such as the chest and waist, offered the most reliable readings. However, their approach still suffered from the same constraints as other hybrid solutions regarding limited battery life and the need to keep sensors in close proximity to the connected smartphone for continuous monitoring.

2.5.1.2 Summary of wearable sensor-based fall detection solutions

The literature presented in Subsection 2.5.1.1 reveals a common drawback of wearable fall detection solutions, namely the technology's dependency on the user to consistently wear the device. Innovative solutions that instead make use of devices that a person may already be wearing, such as a smartphone, seek to circumvent user compliance. Still, researchers argue that no one is guaranteed to always wear a monitoring device, especially overnight, where trips to the lavatory and kitchen can leave them vulnerable in the event of a fall. Other limitations related to the battery life of devices and the need to optimally place sensors on a person bring the practicality of the approach into question. In light of this, other approaches to fall detection operate without relying on user interaction and battery power, such as those centred on ambient sensors (Ramachandran & Karuppiah, 2020).

2.5.2 Ambient-based fall detection

Ambient sensors include proximity and floor sensors that detect motion, sound, light, and vibration and are best employed when distributed throughout a home or living space (Birku & Agrawal, 2018). Human movement and actions produce distinct readings when a person is in range of an ambient sensor. The unique data signatures can help distinguish different actions and can be used to recognise a fall based on abnormal weight displacement or sudden impact with the floor. Implementing this approach is often expensive since multiple sensors are necessary to monitor activity throughout any chosen space. On the other hand, this approach is less intrusive in terms of privacy and does not require the active participation of the monitored individual. However, because an ambient sensor monitors all activity within its range, it can be prone to false alarms and, depending on sensor sensitivity, can also miss detections (Mubashir *et al.*, 2013).

2.5.2.1 Fall detection solutions that implement ambient sensors

The use of ambient sensors in the monitoring of human environments began to emerge in the late 1900s with projects like the ORL active floor, where office flooring was equipped with pressure-sensitive cells to measure load distribution throughout an office space (Addlesee *et al.*, 1997). Similarly, the Magic Carpet was created to monitor the position and pressure of feet so as to map physical gestures to musical expression (Paradiso *et al.*, 1997). An evolution of the Magic Carpet – the Smart Floor – performed user identification based on weight and gait signatures (Orr & Abowd, 2000). The variety and capacity of ambient sensors have evolved over the years, leading to new applications such as fall detection. The following subsections describe prominent solutions and advances in ambient sensor-based fall detection.

2.5.2.1.1 Vibration sensors

One of the first automated fall detection systems using floor sensors was proposed by Alwan *et al.* (2006). Their system focused on detecting two of the most common types of falls among the elderly: falls that occur while walking and falls that occur when getting seated in or rising from a chair (Lu & Chu, 2018). Common fall scenarios were emulated using an anthropomorphic dummy with a mass distribution similar to that of an average person. The dummy was repeatedly toppled over onto a concrete floor. A piezoelectric sensor measured the floor's vibration upon impact and recorded vibration signatures based on vibration frequency, duration, and amplitude. Their readings provided clear distinctions between simulated fall events, everyday activities, and objects hitting the ground. In the end, the study achieved a fall detection accuracy of 100% based on emulated falls, provided the sensor was relatively close to the fall location. Alwan *et al.* (2006) demonstrated that their approach was inexpensive and did not infringe on privacy. However, its success is dependent on a type of flooring that can carry vibration, and its detection range is restricted to that of the sensor (Vallabh & Malekian, 2018).

2.5.2.1.2 Pressure sensors

Pressure sensors have been used in elderly care to monitor physical health over time. In particular, assistive railings equipped with sensors have been used to measure the pressure asserted during sit-to-stand and stand-to-sit transfers in both the bedroom and lavatory to detect early signs of degradation in physical strength (Arcelus *et al.*, 2010; Arcelus, Herry*, et al.*, 2009; Arcelus, Holtzman*, et al.*, 2009). Although the size of the contact area limits the range of pressure sensors, the low cost and non-obtrusive nature of these sensors make them ideal for passive monitoring. In the context of fall detection, the limited range of the sensors and their inherent inability to differentiate between a fall and a person lying down make them inadequate for recognising fall events. To address these limitations, Daher *et al.* (2017) added three-axis accelerometers to a pressure sensor-equipped tiled flooring to facilitate human activity detection. The accelerometer readings allowed them to distinguish between the identical pressure signatures of a fall and the action of lying down. Their approach yielded an accuracy (sensitivity) of 94.1%, but the number of sensors required to monitor a home makes this option expensive to deploy, and the overall accuracy is dependent on the sensitivity of the sensors (Vallabh & Malekian, 2018).

2.5.2.1.3 Acoustic sensors

The audible impact of a person making sudden contact with the ground can be used to recognise a fall. Popescu *et al.* (2008) proposed the acoustic fall detection system (FADE), which relied on a linear array of microphones mounted on vertical pre-amplifier boards to record and classify sounds based on their originating height. Training data for the approach was generated by a stunt

actor who was instructed to act out a variety of everyday tasks and simulated falling to the ground. The recorded audio was processed to extract mel frequency cepstral coefficients (MFCC) as signal features, and a classifier model was trained to differentiate sounds based on the originating location. Classification accuracy was suboptimal, primarily due to background noise, reverberation, and other everyday acoustic interference, which resulted in noisy data.

Popescu and Mahnot (2009) later improved the acoustic-FADE system by replacing the vertical microphone array with a circular configuration. They also used beamforming to isolate the source of a sound more accurately, a technique used in video conferencing applications and bird monitoring that enhances sound quality (Li *et al.*, 2012). The second iteration of acoustic-FADE could successfully isolate ambient sounds, locate their source, improve the signal, and classify the event as a fall or non-fall using a machine-learning algorithm. In the end, the system was capable of accurate fall detection with a sensitivity of 100% and a specificity of 97%. However, both studies raised concerns about the practicality of the approach. The recording range of a microphone array was limited to that of a single room, and it was assumed that no more than two people would be present in a room. Deploying the solution would also require that a microphone array be set up in the centre of each room, which is impractical and costly. In addition, the system was trained on recordings of simulated falls, which could likely undermine its accuracy in a real-world setting.

2.5.2.1.4 Electric near-field sensors

Near-field imaging (NFI) is used in many large nursing homes for person-tracking because of its discreet nature (Oksanen *et al.*, 2009). The sensors monitor the whereabouts and movement patterns of people based on impedances within a matrix of electrodes fitted beneath the floor, making them ideal for passive monitoring (Henry *et al.*, 2008). The sensors are visually undetectable, immune to furniture shading, and have a sensor range of 5 cm above the ground. Rimminen *et al.* (2010) were the first to utilise an NFI-equipped floor for fall detection by evaluating the probability of a fall event based on the near-field image of a human silhouette on the floor. The system could accurately recognise falls with an equal sensitivity and specificity measure of 91%. Similar implementations soon emerged, such as the CapFloor project, which recorded the influence of a person passing through low-intensity electric fields over a carpet embedded with capacitive sensors similar to those used in touchscreen devices (Braun *et al.*, 2012). Ropponen *et al.* (2011) proposed using low-frequency radio frequency identification technology to identify individuals based on tags they would carry on their person. Medical centres successfully used this technology to pull up relevant medical records as patients enter a consulting room.

2.5.2.1.5 Doppler radar sensors

Doppler radars, commonly used in weather forecasting and aviation, provide velocity readings on objects from a distance. They operate by emitting a continuous electrometric wave signal and estimating the speed of a moving object based on the reflected wave, a technique referred to as the Doppler effect (Tomii & Ohtsuki, 2012). As part of their innovative solution, Liu *et al.* (2011) creatively applied Doppler radar readings by manipulating reflected wave signatures of recorded falls using sound signal processing techniques. First, MFCC features were extracted from the reflected waves of various daily activities. Then, the features were classified into fall and non-fall actions by a support vector machine (SVM) and a *k*-nearest neighbour classifier. Despite having a limited dataset, the system demonstrated promising results, achieving sensitivity scores in the range of 91% to 97%. Doppler-based fall detection has advantages over most other approaches since it can operate in environments with dim or no lighting, and its readings preserve user privacy (Su *et al.*, 2014). Radar has the added benefit of penetrating obstacles, such as furniture, that would otherwise obstruct the readings of many other sensors. On the other hand, Doppler sensors are only sensitive to motion and cannot detect a long lie. Furthermore, because the radar can penetrate walls, it invalidates any option to restrict coverage to a particular room or area (Tomii & Ohtsuki, 2012).

2.5.2.1.6 Wireless technology sensors

Wireless technology can be exploited to perform indoor person-tracking and localisation and it was previously featured in a Microsoft indoor positioning competition (Huang *et al.*, 2019). Indoor positioning systems have evolved to use different wireless technologies such as Bluetooth low energy, wireless area networks, radio frequency, optical light, and ultra-wideband. In the context of fall detection, Wang *et al.* (2016) successfully used Wi-Fi signals to detect human activity as part of their WiFall solution. Wi-Fi connections, which are shared by multiple devices, create a network of reflected rays that undergo changes as people move through them, leading to measurable variations in the amplitude and phase of the Wi-Fi signals. It is these fluctuations that WiFall exploits to evaluate variations in the reflection, diffraction, and scattering of signals to recognise when a person suffers a fall. Wang *et al.* (2016) extracted signal features to learn and predict human actions using a Random Forest algorithm and an SVM, achieving precision scores of 94% and 90%, respectively. As advanced and easily deployable as this system is, its reliability is constrained to measuring signal fluctuations within a space where no more than one person is present at a time. This limitation may not be practical in a real-world setting.

2.5.2.1.7 Passive infrared motion sensors

Passive infrared (PIR) motion sensors are primarily used in security alarms and automatic lighting. These sensors detect heat and respond to sudden changes in temperature, making them useful for various applications such as monitoring stove usage, tap and water use, opening cabinets, and detecting the presence of people (Uddin *et al.*, 2018). However, PIR motion sensor readings alone are not suited to fall detection since both falls and non-fall activities produce identical infrared signatures. Nonetheless, Lee *et al.* (2007) successfully used motion readings to help recognise potential falls by monitoring older adults living alone and recording their daily movement patterns as they move about their homes. Any deviation from an established pattern would signal abnormal behaviour, and caregivers would be alerted to a possible fall or other concern. PIR motion sensors have also been used by Yazar *et al.* (2013) to reduce false detections in their fall detection solution. They employed PIR sensors to verify the presence of a person in the room where their vibration sensor-based solution made a positive detection. A common challenge related to PIR motion sensors lies in optimising their line of sight and maximising the coverage area within the layout of a home or living space (Vallabh & Malekian, 2018).

2.5.2.2 Summary of ambient sensor-based fall detection solutions

The literature presented in Subsection 2.5.2 demonstrates that many innovative fall detection solutions implement a range of ambient sensors. This approach is not subject to the same user accountability related to wearable devices but is more prone to interference and false detections. Several of the examined solutions also operate under the unlikely assumption that no more than one person will be present in a room or space when monitoring activity. In other systems, the limitations of the ambient sensors prevent it from being used as a primary sensor. Instead, the ambient sensors are often used as secondary sensors to verify fall detections. These factors bring the practicality of the approach into question, especially when considering the costs of solutions that necessitate multiple sensors to provide complete coverage of a space. Still, certain scenarios favour ambient sensor-based approaches regardless of the shortcomings. For instance, underfloor sensors offer an unobtrusive and discreet solution. Additionally, there are instances where ambient technologies, such as PIR alarm systems or Wi-Fi systems, can be adapted to support fall detection. Although these approaches may have shortcomings, they can still contribute to fall detection in specific contexts.

2.5.3 Vision-based fall detection

Camera footage is the most interpretable sensor data of the three approaches. However, it necessitates that action-related cues be synthesised from the recorded footage to achieve human action recognition. Unlike wearable and ambient sensor readings that are interpreted directly,

discrete features need to be generated from the visual data to indicate relevant actions or poses (Birku & Agrawal, 2018). Although camera-based fall detection is less intrusive and user-reliant than other sensor-based approaches, its adoption is hindered by privacy concerns because of the visual nature of its readings. The cost of implementation mainly depends on the number and quality of camera sensors, but this becomes easily feasible when a camera system is already in place. Moreover, unlike other solutions, a single well-positioned camera sensor is sufficient to monitor a large area and the approach is not restricted to monitoring only a single person.

Real-time fall detection is necessary to facilitate the timely response of caregivers and authorities. However, a fall motion typically has a short duration of only a few hundred milliseconds, which can be partially or even completely omitted from a video sequence when recording in a low frame rate (Bian *et al.*, 2015). Increasing the frame rate requires more computational power for processing of the incoming video feed, which may offset the potential for real-time detection. Striking a balance between the frame rate and computational load has often been a challenge in visual fall detection but increasing computational power over the years has since overcome these concerns. Nonetheless, environmental influences such as dim lighting and an obstructed line of sight can hinder capturing the fall action and observing the fallen position of a person, an essential indicator in its detection. Various attempts to overcome these influences are proposed as part of the different solutions described in this section.

2.5.3.1 Fall detection solutions that implement camera sensors

Visual fall detection is becoming increasingly feasible in assistive care due to advances in image processing and the accessibility of low-cost cameras (Xu *et al.*, 2018; Vaidehi *et al.*, 2011). The approach relies on visual cues to identify a person in a video frame and to determine whether their actions or posture indicate the possibility of a fall. Because the human shape is diverse in its appearance, many methods have been explored to accurately diagnose a fall and limit the influence of non-indicative or contradictory visual information. The relevant information that describes a human pose can be extracted from recorded footage using image processing techniques (Zhang *et al.*, 2015). The footage itself can take the form of 2D video that is recorded using a single RGB or depth camera, or it can take the form of a 3D representation that is compiled from multiple such cameras (Zhang *et al.*, 2015). The following subsections contain discussions on the different approaches to processing visual data and the limitations and advantages of camera-based fall detection.

2.5.3.1.1 RGB cameras

Camera-based solutions offer an affordable and easily implementable approach to fall detection since the sensor itself is generally accessible and simple to install (Vallabh & Malekian, 2018).

Conventional single-camera solutions record full-colour footage (RGB) and rely on cues such as shape-related features, human motion analysis, and inactivity detection that are synthesised from raw footage. For example, Rougier *et al.* (2011) analysed the deformation of the human silhouette across multiple video frames by evaluating the incremental pose changes against the characteristics of a fall motion. More advanced methods derive cues based on human motion and orientation dynamics, like Kreković *et al.* (2012) who quantified motion from video data to identify sudden movements or actions that resemble an accidental fall. Inactivity detection and the recognition of horizontal poses are often also used to confirm suspected fall actions and further increase the robustness and reliability of a system.

Camera sensors produce 2D video imaging which is a challenging data format from which to reliably identify a person and any visual cues that signal a fall. Other factors, such as low-quality video, variable lighting, occluding objects or people, and texture-rich or cluttered backgrounds, can exacerbate the ability to differentiate elements within a scene (Zhang *et al.*, 2015). These challenges can be addressed through the use of multiple-camera setups, where synchronised video feeds are processed to create a 3D representation (Xu *et al.*, 2018). It is a complex process that requires precise calibration of the cameras to accurately observe the same scene (Ren & Peng, 2019). Reconstructing a subject in 3D space involves back-projecting and unifying multiple video feeds to allow for features to be extracted against a third dimension. For example, Auvinet *et al.* (2010) analysed the distribution of the body's volume along the vertical axis of a 3D avatar and would register a fall when the distribution was concentrated near the floor. The advantage of this approach is evident in the accuracy of such systems, but its complexity makes it challenging to implement without the necessary expertise.

Occlusion is a significant challenge to visual fall detection, and it occurs when objects obstruct the line of sight between a camera and the person being monitored. It is especially common in densely packed living spaces, and it is often a result of seniors downsizing to smaller homes, leading to rooms being generously populated with belongings (Vallabh & Malekian, 2018). Large furniture items and obscure camera placement can cause partial and complete occlusion of people as they move about their environment. Occlusion can be mitigated by strategically placing cameras at high vantage points for a broader field of view. Alternatively, multiple cameras can be installed to capture different perspectives, ensuring that at least one camera will provide for an unobstructed view (Vallabh & Malekian, 2018). Privacy concerns related to camera surveillance, particularly in private areas such as homes, complicate its deployment. Research shows that areas associated with a high risk of falling include the staircase, kitchen, and bathroom, which are often frequented overnight with trips to the lavatory and should therefore be monitored (Su *et al.*, 2014).

2.5.3.1.2 Depth cameras

Much research interest has emerged since the release of affordable and accessible low-cost depth sensors like the Microsoft Kinect. This particular sensor is equipped with an RGB camera, an infrared camera, and an infrared laser-based emitter that records the degrees of depth in a scene, albeit within a limited range. Because the Kinect provides both RGB and depth information, the pixels and depth values can be overlayed and associated to denote important areas or objects in its field of view. For example, the floor can be estimated using a v-disparity map where any incrementing depth values along a linear path signify a ground plane. Supplementing image data with depth information can produce a richer dataset and help reduce ambiguity between fall actions and daily activities (Vallabh & Malekian, 2018). Additionally, depth sensors can capture footage in low light, enabling overnight activity monitoring while preserving privacy due to the lack of discernible details in the depth readings.

The Kinect software development kit (SDK) provides skeletal joint tracking functions and body analysis tools that support human action recognition (Bian *et al.*, 2015). Planinc and Kampel (2013) developed a fall detection solution that calculated the major axis of the human body in real time by tracking the head, shoulder, spine, and knee joints. A fall was registered when the orientation of the major axis was parallel to the floor and the spine coordinate was below an expected height threshold. Similarly, Rougier *et al.* (2011) measured the rate of tracked movement across multiple frames to detect any sudden change in velocity that might indicate a fall. Zhang *et al.* (2012) recorded depth information in low-light conditions overnight to monitor falls in the dark (Vallabh & Malekian, 2018). The capabilities of the Microsoft Kinect sensor make it a valuable sensor in the development of fall detection applications.

2.5.3.2 Summary of camera sensor-based fall detection solutions

Camera-based approaches are more convenient to implement than wearable sensors and are easier to install than ambient sensors. In many instances, a camera-based solution may also be the most cost-effective, depending on the coverage area and the number of people to monitor. Moreover, a camera may be met with the least number of challenges to operate reliably since it does not depend on user interaction and is not limited by battery life. Depth sensors complement RGB data and offer the added benefit of capturing footage in low-light environments.

However, occlusions and low-quality footage pose challenges to accurate fall detection. The sophistication of the techniques used to extract pose related cues and features from a video feed also determines the overall accuracy of the approach. In addition, ethical concerns related to confidentiality and privacy are important to consider in deploying such a fall detection system, given the visual nature of the sensor readings. Privacy concerns are especially stressed when

monitoring private homes and intimate spaces such as bedrooms or bathrooms. However, the potential to implement a real-time fall detection algorithm on a live feed without recording and storing the video data may be a reasonable compromise that preserves personal privacy.

2.5.4 Advantages and disadvantages of each sensor-based approach

A summary of all the known advantages and disadvantages of each sensor type regarding its influence to fall detection is presented in Table 2-1. Significant characteristics regarding the implementation and processing of sensor readings are also highlighted in the table. Comparing these attributes helps to define the most appropriate strategy in terms of ease of use, privacy preservation, reliability, accuracy, and cost.

Table 2-1: Summary of advantages and disadvantages of each sensor in the context of fall detection (Gutiérrez *et al.*, 2021; Vallabh & Malekian, 2018)

Advantages and disadvantages of each sensor type in the context of fall detection	
Wearable sensors	
Advantages	*Disadvantages*
• Preserves freedom of movement by not restricting a person to a predefined area • A single device monitors only the target individual and is not influenced by people close to the device or its sensors • A smartphone can be appropriated as a detection device using its built-in suite of sensors	• Complex sensor readings that require expertise • Relies on user accountability for the device to be worn • It has limited battery life and requires frequent recharging • The device should remain within range of its receiver when dependent on remote signal processing • The device needs to be continually worn and sometimes in a fixed position and orientation that is not natural to the user • A user could be reluctant to wear the device if they consider it heavy or uncomfortable • Different actions can produce similar sensor readings, which may require additional sensors to aid in differentiating between confusable actions for the sake of accuracy • Smartphone accelerometer sensors do not necessarily record readings in as wide a range as those associated with dedicated devices

Table 2-1: (continued)

Ambient sensors	
Advantages	*Disadvantages*
• Uncomplicated sensor readings • A device does not have to be worn by the person being monitored • It is a non-intrusive approach with little to no user accountability and its sensors are discreet • Some sensor readings, such as vibration, can be considered privacy-preserving since they are not directly interpretable • A variety of ambient sensors are available; therefore, a tailored selection of sensors can be chosen to suit different environments or living arrangements	• It is considered one of the least reliable approaches to fall detection • The area of detection is limited to the range of the chosen sensor • Numerous sensors are often required to cover areas for monitoring • Installing multiple sensors can be costly when providing coverage of large areas • Each type of sensor has its own set of associated limitations that are to be taken into consideration • Certain sensors are only operational when no more than one person is present within their range • Some sensor readings may intrude on privacy, most notably audio sensors and sensors that express a user's location
Camera sensors	
Advantages	*Disadvantages*
• Uncomplicated sensor readings • A device does not have to be worn by the person being monitored • No user accountability is relied upon for the sensors to be operational • Camera sensors are easily accessible and affordable • A camera system can be retrofitted to process video feeds without recording footage, preserving user privacy without recording footage. • Camera sensors are cost-effective since fewer sensors are needed for large areas compared to ambient or wearable sensors. • Camera sensors serve more than a single, dedicated purpose since they also provide surveillance	• It is considered the most intrusive approach in terms of privacy, given the interoperability of video data • The area of detection is limited to the camera's field of view • A high frame rate is required to capture swift motions, such as a fall • A high frame rate requires more computational power to process the incoming frames in real time • Camera sensors are easily affected by occlusions • People are wary of installing cameras in private areas such as the bathroom, which is where seniors are often most susceptible to slipping and falling

Camera sensors are a compelling choice since they offer uncomplicated sensor readings, they operate independently, and provide for a large coverage area using a single sensor. Existing surveillance systems can be retrofitted to support fall detection without compromising their primary purpose of facilitating safety and security. Additionally, camera systems can be adapted to preserve privacy by opting to process video feeds without recording footage. Despite the challenge posed by occlusions, camera sensors remain a reliable and versatile solution for fall detection. In terms of ease of use, privacy preservation, reliability, accuracy, and cost, camera sensors demonstrate the most advantageous attributes among the available sensor types.

2.6 Conclusion

Worldwide population aging has prompted countries to invest in innovative solutions to address the anticipated increased demand for geriatric healthcare. Technological advancements hold the opportunity to manage and mitigate the influx of elderly patients by reducing the intensity of care required. To this end, fall detection was examined in this chapter as a means to alleviate the need for prolonged hospital stays due to a fall by ensuring that timely access to assistance and medical attention is made possible. Different approaches rely on different sensor technologies to capture pose information to identify actions and behaviours that might resemble or suggest that a fall event has taken place. The selection of an appropriate sensor type depends on factors such as the environment, living arrangements, and the number of individuals being monitored.

Visual sensors are considered the most convenient and reliable option when compared to wearable and ambient sensors. Camera technology also provides interpretable sensor readings and can operate independently without relying on user interaction. Additionally, the challenges associated with camera-based approaches, such as occlusion and dim lighting, can be more easily addressed than those that encumber other sensor types. Furthermore, the rich visual nature of camera readings offers the opportunity to use recent advances in image processing that may benefit pose-dependent applications such as fall detection. The next chapter provides an overview of vision-based pose estimation and how the human body is quantified in video and image data for fall detection.

CHAPTER 3 VISUAL POSE DETECTION

3.1 Introduction

The previous chapter showcased how fall detection is achieved by processing and interpreting the signals provided by sensor technologies. Regarding camera-based approaches, cues are regularly derived from the recorded footage by synthesising information that describes the human body. Pose estimation is a technique for quantifying human poses in video and image data. This chapter includes an overview of prominent research in human pose estimation (HPE) and the variety of methods that can be used to describe the human body from 2D imaging. Additionally, the use of *OpenPose* as the chosen HPE method used in this dissertation is substantiated by the advances that deep learning has afforded both fall detection and HPE.

Three common modes of pose estimation within the context of fall detection are described in Section 3.2 with reference to past research attempts and successful solutions. Pose estimation is then defined in Section 3.3, and key point localisation as a method of HPE is introduced by providing an overview of research focused on the task. Section 3.3 continues with a discussion of the different approaches in performing key point localisation and how each addresses the complexity and natural variability of how the human body is expressed in visual data. The benefits of machine learning highlighted in HPE methods and approaches culminate in Section 3.4, where *OpenPose* is introduced and selected as the preferred method for HPE. Section 3.5 concludes the chapter with a summary of the literature findings presented in this chapter.

3.2 Literature study of methods for pose estimation in fall detection

Three pose estimation methods are presented in this section to illustrate how the derived visual cues offer insight into potential fall occurrences. The examined implementations showcase prominent approaches to visual fall detection while also highlighting the strengths and shortcomings of each method. Surveys on visual fall detection, such as those compiled by Xu *et al.* (2018) and Dhiman and Vishwakarma (2019), reveal the extent of novel approaches and innovative technologies that have been applied to the task. The rich diversity of approaches underscores the ongoing pursuit of a practical and universally applicable solution to this challenging problem, consequently stimulating new avenues of research (Xu *et al.*, 2018).

One major challenge that still hinders research efforts in this domain is the lack of authentic data. Fall actions that accurately resemble real-life scenarios are challenging to capture and even harder to recreate. Nevertheless, using visual signals to detect falls is still considered more reliable than other methods, prompting continued research efforts in this direction. Over the years, various successful forms of visual pose estimation have emerged, including body shape analysis,

head motion analysis, and machine learning-based approaches. These methods offer indicators for pose and activity recognition that are applicable to fall detection and are discussed as part of the literature of review in the following subsections.

3.2.1 Body shape analysis

Fall detection using body shape analysis first entails silhouette extraction through foreground segmentation and background subtraction from image or video data. The segmented silhouette can then be used to quantify shape deformation to infer posture as an indicator of human activity. This approach is demonstrated in the work of Vaidehi *et al.* (2011) and is illustrated in Figure 3-1, where an isolated human silhouette is encapsulated with a bounding box around its contour to represent the pose as a shape. A bounding box allows for its aspect ratio and inclination angle to be observed as signals for fall events by measuring their values against predefined thresholds. Vaidehi *et al.* (2011) opted to use these static features over dynamic features such as velocity since they are less computationally intensive to compute and can therefore support real-time fall detection.

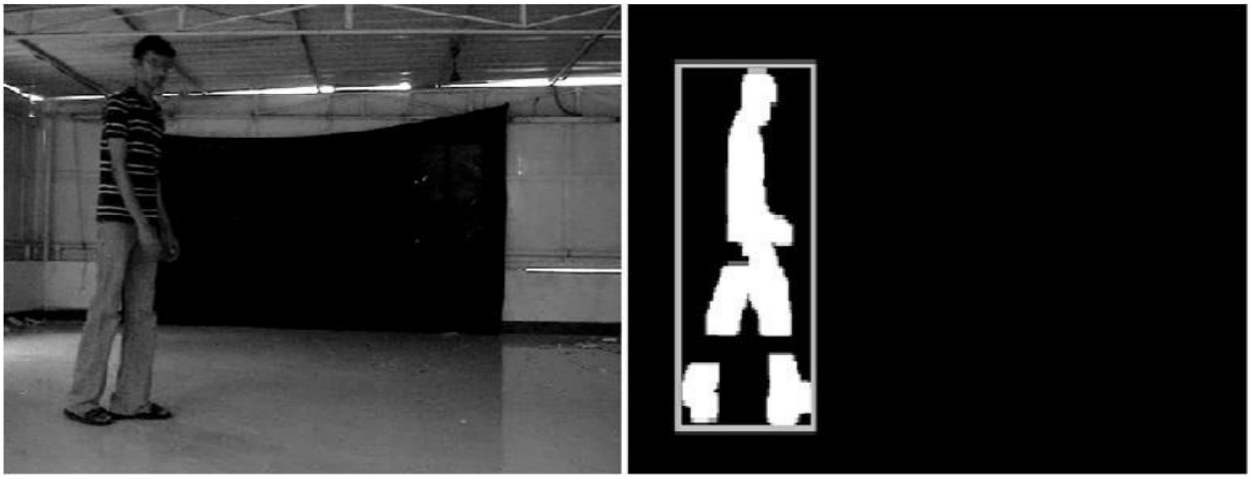


Figure 3-1: Bounding box placed around a segmented human silhouette to represent the body and its pose as a shape (Vaidehi *et al.*, 2011)

Even if established from a large and representative dataset, the thresholds used to evaluate fall movements can fail because of the natural variability in real-life scenarios. Vaidehi *et al.* (2011) attempted to counter this variability in their approach by calculating an inclination angle from the bounding box, which helped to differentiate between actual falls and ambiguous poses (*e.g.*, where a person is kneeling or bending over). However, intentional lying poses were easily misclassified using this method. Without additional information, a person lying down in a horizontal position could trigger a false detection since this action resembles the final resting position of a fall. Interestingly, this concern was addressed in an earlier study by Nait-Charif and McKenna (2004), which proposed using inactivity detection alongside body shape analysis to

confirm whether a fall had occurred. By monitoring a living space using an overhead camera, Nait-Charif and McKenna (2004) identified areas of routine inactivity (such as a bed or sofa). Any inactivity detected outside these learned zones was regarded as abnormal behaviour and, by extension, probable falls.

As an alternative to the bounding box approach, Chua *et al.* (2015) proposed expressing the human shape as three points across the length of the body. These points represented the head, body, and legs and were computed after foreground segmentation of a person in a video, as illustrated in Figure 3-2. Features such as orientation, the sum of sub-region heights, and the ratio of heights between sub-regions could be easily calculated from the localised key points. Any actions that exceeded a change in orientation of 30 degrees and 40% in height were considered likely fall events. Regardless of the simplicity of the approach, it achieved a reliable accuracy of 90.4%. Still, this method remains susceptible to the potential distortion of a camera's perspective when not optimally positioned and angled within a room.

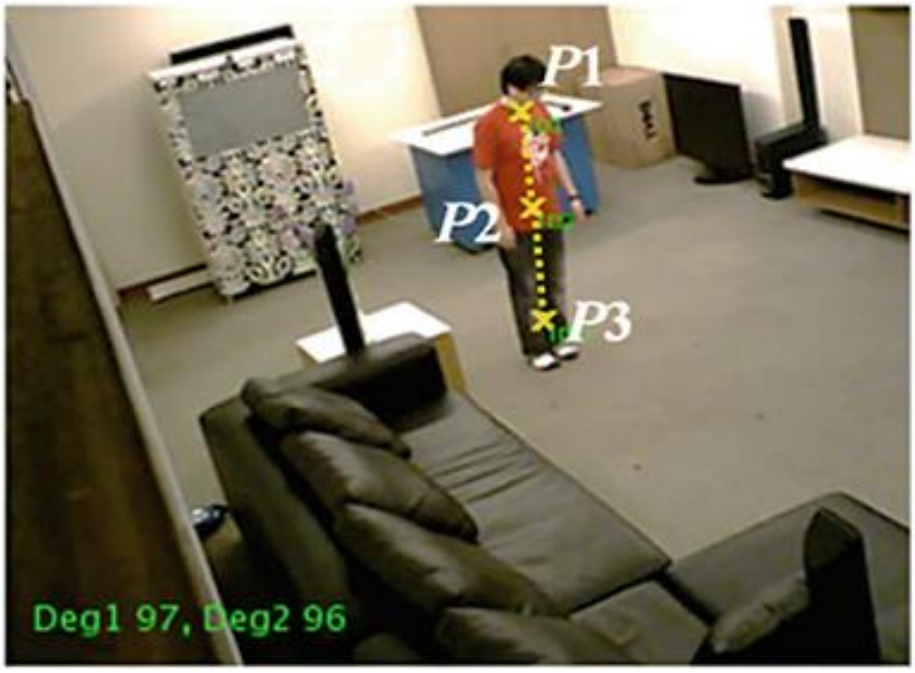


Figure 3-2: Three-point representation of the human body to detect fall events (Chua *et al.*, 2015)

Zhang *et al.* (2012) attempted a similar form of pose estimation by representing the human body as a collection of key points identified by a Microsoft Kinect sensor. By tracking eight body key points across multiple frames, Zhang *et al.* measured the deformation in the key point structure as the person's pose changed. When person tracking would fail due to the limited range of the depth sensor, the ratio between the width and height of a projected bounding box in the RGB footage was instead monitored for fall cues. A support vector machine was trained on the pose-

related cues to distinguish between fall and non-fall events. Their system achieved an accuracy of 98% when differentiating between the two sets of activities on the test dataset.

3.2.2 Head motion analysis

Aside from body tracking, the head has also been found to correlate well with the motion of a fall. For this reason, head tracking is a generally well-established approach for analysing human action within video footage (Yu, 2008). The method is also less susceptible to occlusions compared to other forms of motion tracking, making it a valuable form of pose estimation. However, owing to its smaller size and simpler shape, the head is less conspicuous than the human body, making it more challenging to identify and track across the frames of a video. This inherent challenge is cited by Yu (2008) as the reason why only a limited number of fall detection applications employ this approach.

Rougier and Meunier (2006) were among the first to capture fall signals based on the motion of the head. They proposed monitoring its trajectory and speed and founded the now-established principle of faster vertical motion than horizontal motion as a measure for fall detection. By representing the head as an ellipsoid in 3D space derived from a 2D video feed, their system could project its position in real-world space using the coordinate system of a specialized video camera calibrated to the monitored environment. Subsequently, the head was tracked across multiple frames to measure its trajectory and speed. A fall action was then registered when the vertical and horizontal velocity of the head surpassed pre-established thresholds. Although successful, it was later realised in subsequent research that the velocity measured in 2D space yielded higher values for a person near the camera than when further away. This inconsistency challenged their ability to define a fixed velocity threshold that would be universally applicable (Rougier & Meunier, 2006).

Hazelhoff and Han (2008) attempted rudimentary head tracking based on silhouette segmentation followed by searches for skin-coloured regions at the body extremities using a Gaussian skin-colour model, as illustrated in Figure 3-3 on the next page. Head motion was monitored across multiple frames by repeatedly searching for these regions near the head's last known position in the previous frame. However, Hazelhoff and Han (2008) did not employ motion cues for fall detection because they argued that falls were highly dynamic events with varying duration, speed, and direction characteristics. Therefore, they focused on detecting falls based on a consistent feature: the lying posture. When a significant change was registered for the head's position between frames, a fall was confirmed if the body silhouette was angled horizontally and the person was not in an area of expected inactivity (*e.g.*, on a bed).

Figure 3-3: Head identification using a Gaussian skin-colour model on a human silhouette (Hazelhoff & Han, 2008)

Hazelhoff and Han (2008) achieved a 100% detection rate for falls using only pose orientation and zones of inactivity as signals in video sequences free of occlusion. Their method maintained a high accuracy of 91% on video sequences that included partial occlusion. However, due to flawed silhouette segmentation, accuracy declined to 44% when video sequences featured significant occlusion. Although such a simplistic approach seems to provide reliable results, its effectiveness may likely be undone in a real-world setting where occlusions are common, and the misidentification of a human silhouette and its head is more probable (Mubashir *et al.*, 2013).

3.2.3 Machine-learned analyses

Machine learning is advantageous in performing fall detection over other methods because of its capacity to deal with complex data in large volumes, such as video footage. Attempting manual feature construction from such a data source is unrealistic. Yet, deep learning can automate this process through example-based training, where it learns to extract fall-related features from raw data into a high-level multi-dimensional feature space. Quantifying fall-related motion in this manner establishes feature relationships across separate frames by considering the series of poses in a fall action, all without having to quantify any motion cues outright. The described learning capacity and proficient feature selection of machine-learned techniques are demonstrated in the research studies mentioned in this section.

CNNs are machine-learning algorithms that are proficient in processing 2D image data. The systems proposed by Ji *et al.* (2013) and Simonyan and Zisserman (2014) were the first to extend a CNN's permissible input from only 2D images to multi-dimensional video data for human action

recognition. This made it possible for the models to extract features from both the spatial dimension and an experimental temporal dimension that expressed motion across adjacent frames. Motion was represented in between frames using optical flow displacement fields, as illustrated in Figure 3-4. Each displacement field was placed as an additional frame between consecutive video frames, thereby allowing the network to learn motion explicitly rather than implicitly.

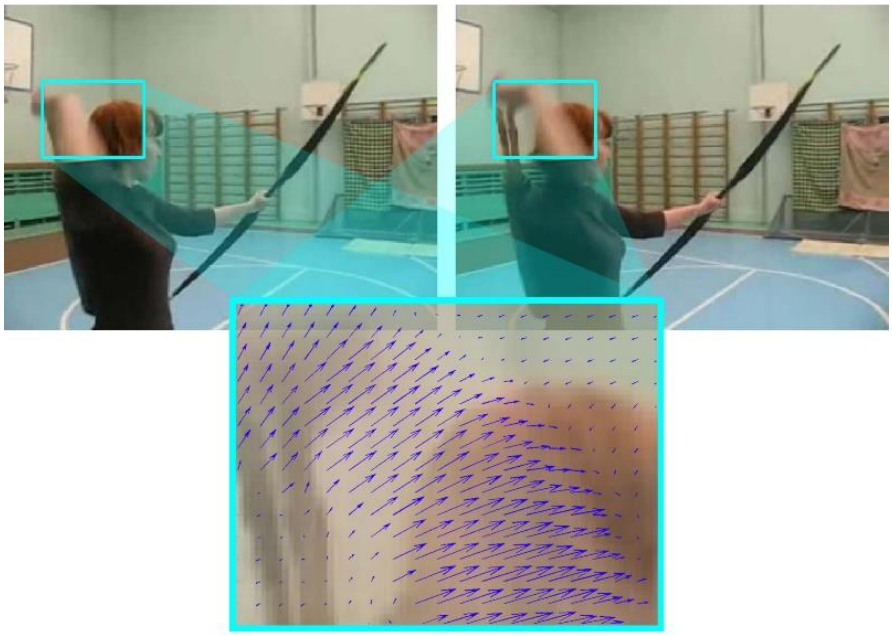


Figure 3-4: Difference between frames captured and represented as optical flow displacement fields to explicitly record motion (Simonyan & Zisserman, 2014)

Panahi and Ghods (2018) attempted to leverage deep learning for well-established fall detection measures. Their approach involved first separating depth-imaging video footage into individual images and pre-processing the data using background subtraction, silhouette segmentation, and image noise filtering. Rectangular and elliptical bounding shapes were fixed to the human silhouette, and the distance between the body centroid and the floor was calculated. Additionally, the inclination angle was calculated as a property of the elliptical bounding box and the aspect ratio as a property of the rectangular bounding box. These features were used to train a support vector machine to distinguish between fall and non-fall events. The system performance realised sensitivity and specificity scores of 97.05% and 97.20%, respectively. However, Panahi and Ghods (2018) did not address the limited range of depth imaging and its practicality in a real-world scenario. Additionally, their approach operated on the assumption that all falls would end with a person taking a resting position on the floor.

3.2.4 Summary of methods for pose estimation

The diversity of research attempts toward pose estimation for fall detection discussed in Section 3.2 helped highlight the processes and strengths of successful approaches. Additionally, the discussed research demonstrates the significance of quantifying the human body to derive pose-related prognostic values. Subsections of the chapter were divided according to three general analyses used to quantify human movement: body shape analysis, head motion analysis, and holistic machine-learned approaches for fall recognition. Most methods that rely on body shape analysis entail measuring silhouette deformations against threshold values that are typical of a fall. Similarly, head motion analysis methods establish velocity and height measurement thresholds to recognise fall-related motion. Whereas machine learning solutions rely on abstracted visual information, such as optical flow, dynamic speed measures, and inclination angle, to learn indicative fall cues.

3.3 Human pose estimation

A diverse range of methods are applicable to fall detection; common to them all is the need to quantify the pose or motion of a person. Silhouette segmentation and key point estimation are among the most regularly employed methods for pose estimation. The same techniques are used in this dissertation by adopting a pre-existing pose estimation solution to support pose recognition. Section 3.3 describes what pose estimation entails and how it is achieved by providing a brief overview of research focused on the task.

3.3.1 Defining human pose estimation

As considered within the field of computer vision, HPE is typically defined as estimating “the configuration of a person’s body parts - or ’pose’ - in an image” (Johnson & Everingham, 2011:1). The goal of HPE is to automatically infer the configuration of the human body by approximating a set of key points that correspond to joints and features such as the hands or eyes. On the next page, Figure 3-5 (a) depicts instances of single-person pose estimation where specific body parts are identified, such as the head, neck, elbows, and shoulders (Martinez *et al.*, 2017). This localisation process is considered a constrained regression problem in that a known and limited number of body parts are to be identified for a single person. More complex instances include multiple people, as shown in Figure 3-5 (b). These instances are addressed by either performing single-person pose estimation for each individual or using a multi-person pose estimation algorithm that regresses an unknown number of key points (Cao *et al.*, 2017; Pishchulin *et al.*, 2016).

(a) (b)

Figure 3-5: Single-person pose estimation (a) and multi-person pose estimation (b) illustrated by a skeletal mapping of key points (Cao *et al.*, 2017; Toshev & Szegedy, 2014)

Pose estimation is relevant in various domains with different but specific tasks. The following are a few example applications:

- **Sports and exercise performance**: Athletes' motions are tracked to determine posture and technique to reproduce such actions or identify areas of improvement in an athlete's training (Baraka *et al.*, 2012).
- **Human–computer interaction**: This form of interaction aims to replace equipment such as a keyboard and mouse to allow users to interact with technology through gestures and physical actions (Wang *et al.*, 2019).
- **Smart surveillance systems**: Video monitoring identifies or tracks pedestrians for security and commercial purposes. Systems such as these have been employed in airports and supermarkets; the most notable of which is the chain of Amazon Go convenience stores, a smart store that tracks item purchases of its customers through video surveillance (Wingfield *et al.*, 2018). Another example is the deployment of gait recognition in China to identify individuals in public spaces based on their walking mannerisms (Dake, 2018).
- **Medical care**: Advances in medical care allow for new approaches to patient monitoring and automated diagnoses. Examples include fall detection, where an individual can be monitored in a care facility or in the home to deploy immediate aid for their assistance. Similarly, epilepsy and the onset of such an event can be observed through posture tracking (Ntinou, 2018; Abbate *et al.*, 2010).

3.3.2 Approaches to human pose estimation

HPE is considered an important task in computer vision, given its extensive applicability in significant domains such as those listed in the previous section. For this reason, much research has been invested in accurately achieving HPE. However, the task remains challenging owing to the degrees of freedom inherent in a complex articulated object such as the human body (Gong *et al.*, 2016; Moeslund & Granum, 2001). This complexity is compounded by its variable appearance in terms of differences in clothing, body shapes, inconsistent lighting, image noise, changing backgrounds, varying viewing angles, and the occurrence of occlusions (Moeslund & Granum, 2001). Deriving value from such data remains an ongoing area of research due to the natural variability. Existing approaches to pose estimation are grouped into *generative*, *discriminative*, and *extended machine learning* methods, each of which is described in the following section.

3.3.2.1 Generative approach

Generative methods are also referred to as top-down or model-based approaches to HPE and entail representing the human body using a mathematical representation referred to as a body model or template that is defined by a set of parameters. The parameters capture characteristics of the body, such as the joints, limbs, and body proportions, which can be adjusted to model a wide range of possible pose configurations to provide a flexible and realistic representation of the human body. In this manner, the model can be used to estimate the pose of a person from an image by finding the set of parameter values that best matches the observed body features, such as joint locations and limb orientations. A pose is therefore interpolated based on comparability with the body template to provide estimations that are representative of real-world poses. However, the approach is highly dependent on accurate person detection to ensure that the parameters are correctly aligned and maintained against an input instance (observed individual) which can be hampered and disrupted by people overlapping in crowds and groups (Kreiss *et al.*, 2019).

Generative methods operate by attempting to minimise the error between the given input instance and the parametrised model according to the known kinematic limitations of the human body (Yang *et al.*, 2012; Salzmann & Urtasun, 2010). Because the model parameters encode this a priori understanding of the human body, a brute-force local search can be executed against the parameters to identify potential poses that best represent the current observation. However, conducting a comprehensive search across the highly dimensional pose space would be computationally expensive. Instead, a more efficient cost function is optimised through the continuous prediction and updating of the model's variables. This optimisation can be achieved

using either stochastic search, which involves random sampling from the parameter space, or local optimisation, which entails iterative updates to the model parameters in the direction of the steepest descent of the cost function. Updating and maintaining the model against the current observation ensures continuous access to pose information for all key points, offering real-time insights (Moeslund & Granum, 2001). When generative methods are employed for 2D data, the models are typically represented as either a rigged kinematic model or a planar model, both illustrated in Figure 3-6

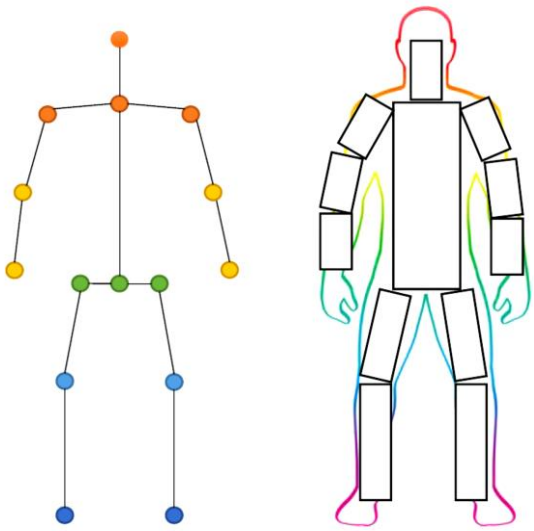


Figure 3-6: Kinematic model (left) and planar model (right) that represent human pose in generative approaches (Zheng *et al.*, 2020)

Kinematic models are also referred to as tree-structured models, where localised key points are grouped into limb-like connections. Planar models involve decomposing the search space into body parts to estimate a pose hierarchically, typically first localising the head and torso and then other limbs. Inaccurate estimations for any initial key point in either of the models can disrupt and propagate errors to other key points further down the kinematic chain, highlighting a vulnerability of generative modelling (Poppe, 2007). Another drawback of top-down approaches is that the model's parameters require initialisation by manually identifying the body key points or parts in the first frame of a video sequence since the pose is continuously maintained against the estimation produced in the preceding frame (Salzmann & Urtasun, 2010; Poppe, 2007). In addition, generative methods are prone to becoming trapped in local minima, which means the algorithm converges to suboptimal solutions instead of the global optimum, resulting in imprecise key point estimates. The difference calculation required in these methods can also be computationally expensive, as it involves continuously updating and maintaining the human body model against an incoming video feed (Salzmann & Urtasun, 2010; Poppe, 2007).

Early and notable work that employed part-based models was first proposed to represent general objects as planar structures and was also adopted for the human pose (Felzenszwalb & Huttenlocher, 2005). Poses were modelled as collections of parts arranged in a deformable configuration, and local connections among neighbouring parts provided qualitative descriptions in pose recognition, as depicted in Figure 3-7. Similarly, Yang and Ramanan (2011) adopted this representation to model the human body for more detailed pose estimation. Their part-based models articulated limbs and limb orientation through a combination of limb templates and the encoding of contextual co-occurrence among parts. They showcased how these relations capture notions of rigidity since the model learned to constrain two parts located on the same limb by assuming that they have similar orientations. This is illustrated in the collections of colour-coded key points in Figure 3-8.

Figure 3-7: Parts-based model of the human body (Felzenszwalb & Huttenlocher, 2005)

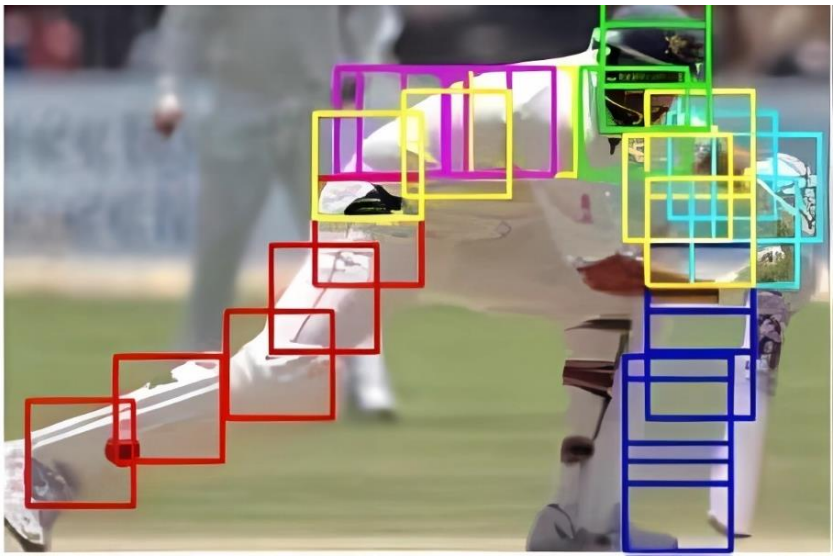


Figure 3-8: Parts-based model where kinematic limitations of the human body are used to associate key points as depicted by coloured regions (Yang & Ramanan, 2011)

Parts association is a critical component of generative approaches to HPE and is best demonstrated in the work of Epshtein and Ullman (2007). They modelled the spatial relations among parts in a facial model and extended them with semantic classes. By analysing various eye and mouth models, Epshtein and Ullman (2007) found that open eyes often occur with a smiling mouth, helping to indicate a specific expression. This understanding of semantic correlation and co-occurrence among parts was also evident in the work of Yang and Ramanan (2011). It has since become an established principle in HPE where body key points and their correlation are based on underlying limb structures and the fixed arrangement of the human body.

3.3.2.2 Discriminative approach

Unlike generative methods that involve person detection followed by key point estimation, discriminative methods directly learn a mapping between an image and the corresponding pose. These methods take a bottom-up approach, initially localising all observable key points in an image and then grouping them using an association formula to form the human structure (Dang *et al.*, 2019; Kreiss *et al.*, 2019). As a result, discriminative methods do not rely on a parameterised body model or its initialisation, making them more responsive to rapid movements and uncommon poses. They also do not typically employ a cost function and instead infer a pose directly from the input image, resulting in faster performance (Van den Bergh *et al.*, 2009). Therefore, discriminative pose estimation is considered a constrained search problem within the 2D image space rather than an optimisation problem in a high dimensional space (Gong *et al.*, 2016).

Key point estimation is performed by a function that renders the pose representation directly from the input based on salient aspects of the image or video frame (*e.g.*, corners or edges). The function is learned from annotated training data either through classification or regression. Regression estimates the best location candidates for each key point using a mapping function based on image features, while classification employs a search-based mechanism to estimate each joint (Gong *et al.*, 2016). Given the logical association between connecting limbs, these localised key points are then grouped into two-dimensional templates that designate different body parts (Poppe, 2007). However, false associations can occur due to limb-like regions in images and overlapping individuals in crowded scenes. Pioneering work like DeepCut (Pishchulin *et al.*, 2016) and DeeperCut (Insafutdinov *et al.*, 2016) addressed this issue by first differentiating among people in an image and by treating pose estimation as an integer linear program which intelligently clusters body parts by person. Later works further accelerated parts association by using greedy decoders while refining association operations through methods such as associative embedding (Newell *et al.*, 2017) and part affinity fields (Cao *et al.*, 2021; Cao *et al.*, 2017) like in *OpenPose*.

3.3.2.3 Extended machine-learning-based approach

Machine learning approaches have undergone significant advancements, particularly with the emergence of deep network structures that utilise features and metrics identified by the model itself. The advent of deep learning and deep neural network structures has resulted in new state-of-the-art achievements in many classification and recognition tasks. Their proliferation can be attributed to the ever-increasing availability of data and the emergence of efficient computing techniques, such as CNNs (Munea *et al.*, 2020). CNNs, as hierarchically extended NNs, are characteristic of deep learning methodologies. The initial layers of the deep network act as powerful feature extractors, propagating the most salient features to the deepest layers for classification. These deep structures are regularly used in a wide range of computer vision tasks, including image classification, object recognition, facial recognition, and even pose estimation (Taylor & Nitschke, 2018).

Deep learning-based pose estimation methods can be classified as discriminative HPE approaches since a feature representation is learned directly from the training examples annotated with key points. However, the remarkable performance improvements have prompted researchers to differentiate machine learning-based approaches from traditional discriminative methods. One of the earliest advances in pose estimation using deep learning is the DeepPose solution developed by Toshev and Szegedy (2014). DeepPose employed cascading CNNs to regress the coordinates of body joints directly from raw input images without employing part detection or introducing any explicit feature representations. At the time, this approach used novel deep-learning techniques and outperformed contending methods on recognised HPE benchmarks (Tian *et al.*, 2012; Yang & Ramanan, 2011; Andriluka *et al.*, 2009).

Although DeepPose performed well when it was introduced in 2014, some inefficiencies were noticed by other researchers in the field related to its method of key point regression. Tompson *et al.* (2014) proposed combining a CNN-based part detector with a separate part-based spatial model that encoded the relational association between parts. The part detector used a confidence representation which depicted the likelihood of key points occurring at a given position based on calculated per-pixel confidence scores. This indirect form of key point regression produced sets of confidence maps depicting 2D Gaussians centred at each key point's estimated location in the input image. Because confidence maps represent joint locations more loosely than direct regression, a more lenient representation was produced that helped to better refine the error between the input image and output key point locations.

Carreira *et al.* (2016) also proposed an alternative approach to error minimisation through their iterative error feedback approach. Rather than directly localising key point positions on the body

in a single feed-forward pass, their system relied on a self-correcting model. The initial key point predictions were progressively refined by feeding weak estimates back into the system. This refinement process involved iteratively passing the raw input image stacked with a confidence map of the predicted joint positions from a previous pass back through the network. The network then calculated a correction displacement, which was represented in the produced confidence map. This iterative feedback loop helped to refine the network's estimates in each iteration until convergence was achieved for the most probable key point locations.

These advancements laid the foundation for the *OpenPose* solution proposed by Cao *et al.* (2017). *OpenPose* utilises a multi-stage and two-branching CNN to predict the confidence of potential key point locations as well as the connections and orientation of limbs in parallel. The predictions from both branches are combined and help refine predictions in the subsequent stages of its predictive process. *OpenPose* is used as part of the experiments in this dissertation, and its functioning is explained in the following section.

3.4 OpenPose

OpenPose, proposed by Wei *et al.* (2016), is a vision-based approach for pose estimation that maps the human skeleton by detecting body joints in video and image data. Initially, Wei *et al.* introduced a CNN-based pose machine framework to learn implicit spatial models from images for pose recovery. Cao *et al.* (2017) then further refined the approach, culminating in the release of *OpenPose*. This deep-learning-based solution achieved state-of-the-art accuracy and ranks among the top 10 performing pose estimation implementations, evaluated on the multi-person HPII Human Pose Dataset (MPII, 2023; Andriluka *et al.*, 2014). *OpenPose* operates using a multi-stage CNN that implements two complementary techniques, namely parts detection and parts association, as illustrated in Figure 3-9.

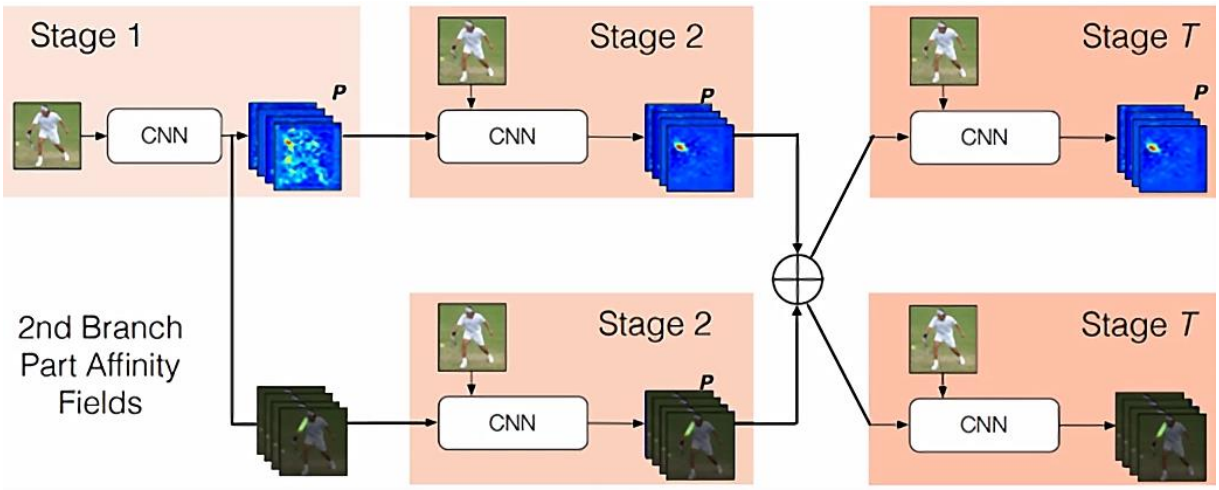


Figure 3-9: The two-branch, multi-stage CNN architecture of *OpenPose* (Cao *et al.*, 2017)

OpenPose adopts a discriminative approach to HPE by utilising a multi-stage CNN for key point prediction. The CNN pipeline consists of consecutive prediction tasks that generate confidence maps for individual body key points at each stage. The first branch predicts key points as confidence maps, while the second branch establishes logical associations between them. These processes are performed concurrently and in stages where, prior to the next iteration, the outputs of both branches are combined and passed as input to the next stage. This sequential prediction framework and the incrementally expanding receptive field of the CNN enable the model to learn local associations and long-range spatial relationships among the key points, which aids in the kinematic modelling of the human body (Munea *et al.*, 2020).

The first branch of *OpenPose* generates a confidence map from an input image depicting signature zones that reflect the confidence of its prediction. Each layer of the confidence map corresponds to a specific joint or body part. The prediction accuracy of the first branch is favourable to parts that are consistent in their appearances, such as the head and shoulders. The prediction is less favourable toward parts that often vary in their position in relation to other parts, such as the hands and legs. However, the generated confidence maps tolerate such variations through their multi-stage refinement, where previous predictions are taken into account when predicting new key points in succeeding stages. Figure 3-10 illustrates the prediction process. In the first stage, the right elbow is initially predicted in the location of the right knee. Once all the key points are predicted in the first stage, their positions serve as a guide to refine the location of the right elbow in stages two and three. The second branch, which is responsible for associating the various key points, also aids in correcting ill-detected key points in later stages based on logical constraints associated with the body's natural configuration.

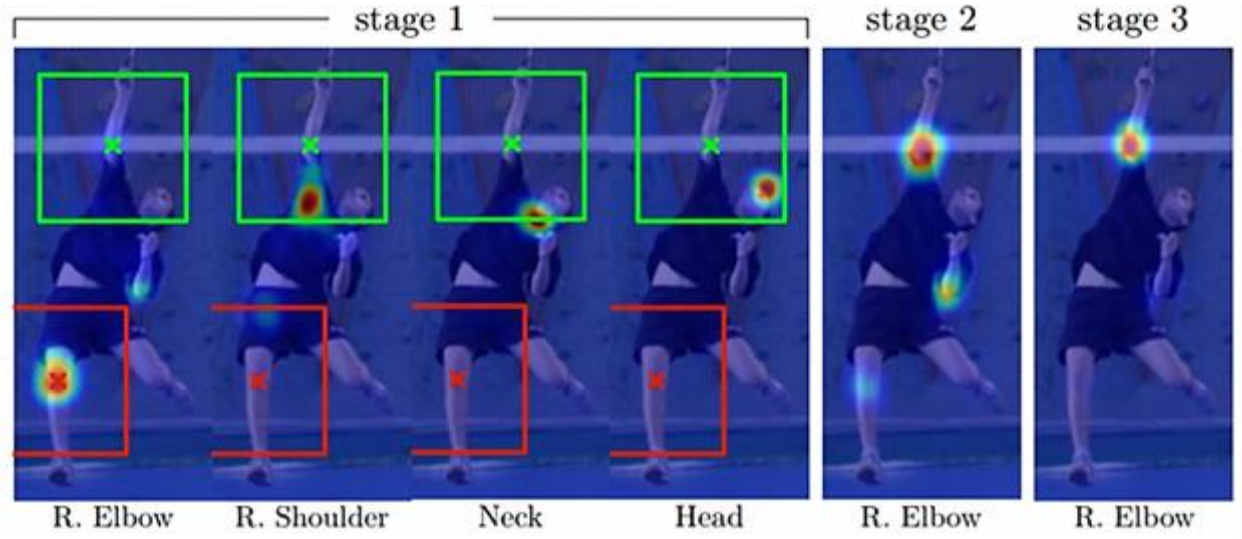


Figure 3-10: Multi-stage sequential refinement of part predictions in *OpenPose* based on preceding predictions (Wei *et al.*, 2016)

The association between predicted key points is a specialised and novel solution referred to as part affinity fields (PAFs), and it is represented in the second branch of Figure 3-11. By way of 2D vectors, information related to key point locations and their connections is combined and represented after each stage. To illustrate this, consider the first image in Figure 3-12 where red and blue dots represent the estimated locations of two key points for multiple people. Various candidate associations can be drawn between all the key points, as represented by the grey lines in the figure. However, after predicting a third key point, illustrated in yellow in the second image of Figure 3-12, these associations can be refined accordingly by eliminating implausible associations based on the kinematic constraints of the human body. The remaining associations depicted in the third image of Figure 3-12 demonstrate how key points are finally associated across a supporting limb.

Figure 3-11: Two branches of *OpenPose* conceptually illustrated alongside their respective outputs (Cao *et al.*, 2021)

Figure 3-12: Process of parts association in part affinity fields (Cao *et al.*, 2021).

As illustrated in Figure 3-12, PAF encodes pairwise relationships between parts by estimating a midpoint representing an underlying and connecting limb. The candidate associations are given a confidence score for all paths that cross the projected midpoint, which represents the likelihood of the PAF association. However, in crowded scenarios, false connections can occur, as illustrated by the green paths in the second image of Figure 3-12. This limitation arises from the path representation, which only captures information about the limb's position while neglecting its orientation. PAFs address this limitation by capturing the position and orientation across the region of support between the midpoint and its anchoring joints within the 2D vector. This information is conceptually illustrated with colour in Figure 3-11, where the orientation of the limb coincides with a colour in the depicted colour wheel.

3.5 Conclusion

The literature reviewed in Section 3.2 illustrated the diverse range of techniques used in vision-based fall detection. These methods revealed some of the challenges associated with pose recognition and pose estimation related to occlusion and the influence of camera perspective and placement. Some researchers also acknowledged the difficulty of distinguishing between visually similar poses. Different methods of pose estimation seek to address or negate these influences, but regardless of the approach, the proposed solutions share the need to quantify the human body in order to realise human activity recognition. HPE is a complex task, and the different approaches discussed in Section 3.3 reflect this in the variety of methodologies and algorithms employed. Despite the complexity, significant progress has been made in HPE, with advancements in computer vision and deep learning techniques enabling more accurate and robust detection.

Traditionally, fall detection research relied on calculating thresholds to recognise actions from body movement and pose. In contrast, this dissertation proposes a novel approach that avoids manually crafting features and instead leverages the efficiency and effectiveness of deep neural networks to learn fall-indicative cues. Such an approach acknowledges the proven benefits of deep learning and how it can effectively manage feature design through supervised learning and feature extraction. Furthermore, the research presented in this chapter demonstrates the effectiveness of pose estimation using body key points in quantifying the human body from images and videos. *OpenPose* was proposed as a suitable non-invasive solution for pose estimation, which is employed as part of the experiment-focused investigation described in Chapter 5. The subsequent chapter provides further motivation for the chosen machine learning-based approach and how data augmentation can be utilised to improve classification performance.

CHAPTER 4 CONVOLUTIONAL NEURAL NETWORKS AND DATA AUGMENTATION

4.1 Introduction

The goal of this chapter is to introduce and explain NNs, CNNs, and the concept of data augmentation. NNs are powerful learning algorithms that are efficient at classification tasks, while data augmentation is a technique used to enhance the accuracy and performance of these tasks. Over the years, significant advancements have been made in computer vision and HPE, primarily driven by the application of these learning approaches and techniques. By examining how NNs operate and process visual information, the applicability and value of a CNN-based approach to pose classification are illustrated in this chapter.

Neural networks are introduced in Section 4.2 by examining the artificial neuron, its inception, internal functioning, and its utility in machine learning to produce reliable classifiers. A specific type of NN is then described in Section 4.3, namely the CNN. The unique properties of a CNN that make it a powerful classifier are explained in Section 4.4, and their influence and benefit to pose classification are discussed. Research evidence is then presented in Section 4.5, which provides insight into the training process of a CNN and the discriminative feature identification regarding colour information and other class-related cues in images. From this evidence, a proposal is made to augment data and improve classification performance by stimulating the saliency of colour information in pose images. The viability of such an approach is supported in Section 4.6, where colour-based data augmentation from existing research is presented, albeit in different but related classification tasks. Finally, the chapter and proposed arguments are summarised in Section 4.7.

4.2 Artificial neural networks

The NN is a powerful computational model inspired by the interconnections and processing capabilities of the human brain. This branch of machine learning has led to notable breakthroughs in diverse domains, including computer vision (Sze *et al.*, 2017). The NN is designed to simulate the computational faculty and functioning of the biological neuron. An artificial neuron acts as a basic computational unit that emulates the biological mechanisms involved in accepting and integrating incoming signals, conducting computational operations on the input, and generating output signals (Walczak & Cerpa, 2003). The various components of the biological neuron facilitate these operations: the dendrites receive input signals, the nucleus performs calculations on the input values, and the axon passes output values to the next neuron.

The biological neuron is depicted in Figure 4-1. Information transmission occurs along the dendrites and is influenced by the synaptic strength, which regulates the flow of information. Likewise, the information passed along the axon is also influenced by the synaptic strength of the receiving neurons (SU, 2023; Walczak & Cerpa, 2003). A mathematical abstraction of this process, depicted in Figure 4-2, emulates the process by multiplying an input value (x_i) with an associated weight (w_i) that represents the synapse strength. The connections determine the influence of one neuron on another by acting as either excitatory (positive weight value) or inhibitory (negative weight value) in the neuron chain. Negative weights inhibit a neuron's potential to fire, creating weak connections that do not contribute to the decision process of the NN as a whole. The network operates by passing multiple signals down these connections and, once weighted, the values are summed in the cell body ($\sum_i w_i\, x_i$), along with the given neuron's own bias value (b), resulting in a pre-activation value, namely z. An activation operation, f, is then performed on the pre-activation value, and if this value, $f(z)$, exceeds a threshold, the neuron can fire and pass the resulting output to the next neuron in the chain.

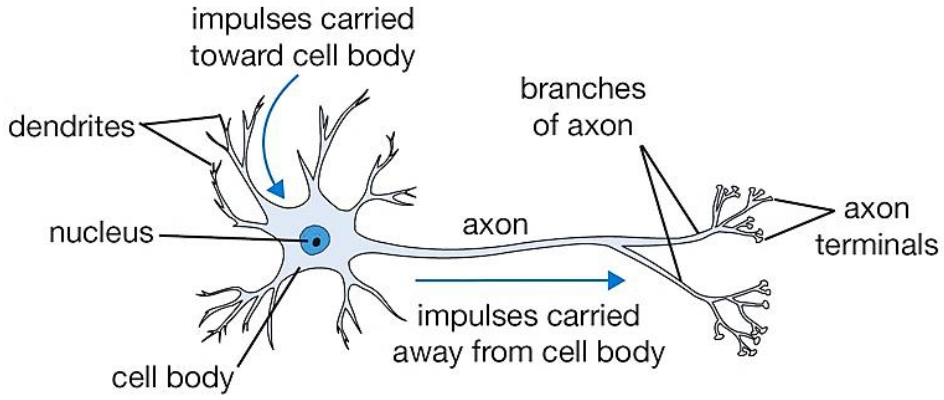


Figure 4-1: Illustration of the biological neuron (SU, 2023)

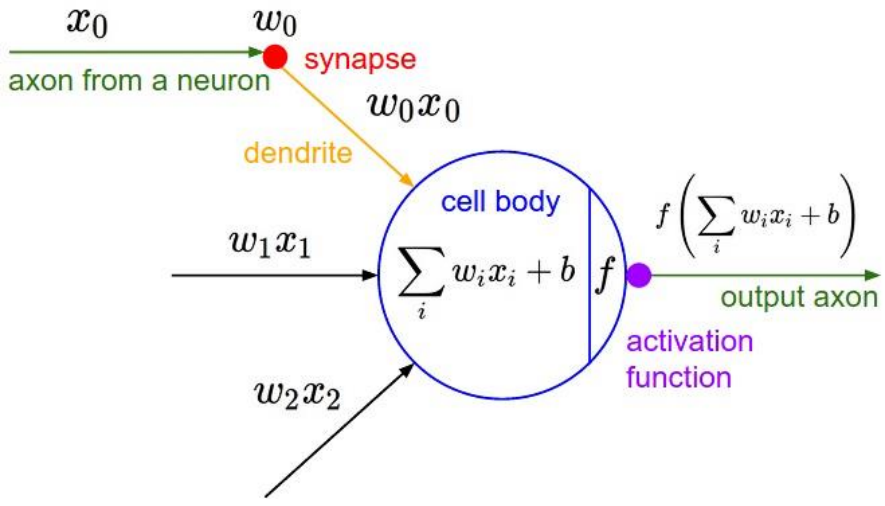


Figure 4-2: Mathematical model derived from a biological neuron (SU, 2023)

Artificial neurons are generally organised into acyclic graphs, the typical topology of a feed-forward NN, as illustrated in Figure 4-3. Neurons are conceptually structured into layers or vectors, where the output of one layer serves as the input to the next layer (Walczak & Cerpa, 2003). Information is passed along connections that simulate the brain's synapses. None of the neurons in a single layer are connected to each other, but two adjacent layers regularly share pair-wise connections that represent the dendrites and axons in the biological brain. The first NN layer facilitates accepting and receiving input data, and the last layer associates classification probabilities with the given input instance. Any layers between these are referred to as the hidden layers and are predominately responsible for feature extraction.

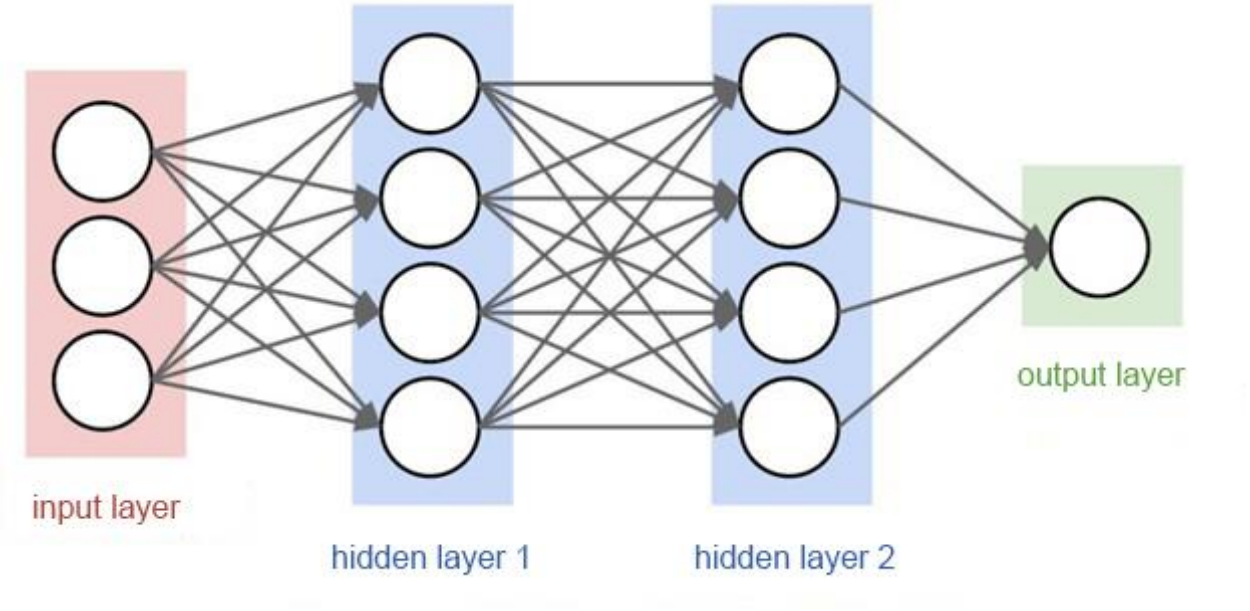


Figure 4-3: Illustration of a four-layer neural network (SU, 2023)

The densely connected structure of processing units within an NN makes it adept at processing complex data and abstracting knowledge to solve complex problems (Sze *et al.*, 2017). During training, the NN learns a mapping between the raw input data and the intended classification by adjusting weights in the inter-layer connections through backpropagation. These connections filter or maintain values passing through the network to yield significant features related to the classes under consideration. Extending a network with multiple hidden layers results in a deep network structure, lending its name to the concept of deep learning. Deeper networks can handle more complex data, such as images and video, since the process of knowledge abstraction is extended and features are further refined. A CNN is a type of deep NN containing characteristic convolutional layers that allow the model to learn and identify image features using local connectivity patterns referred to as filters or kernels (Lee *et al.*, 2017). The operations and proficiencies of a CNN are discussed further in Section 4.3.

4.3 Convolutional neural networks

In recent years, CNNs have proven to be highly successful in sophisticated tasks such as computer vision (Krizhevsky *et al.*, 2012), speech recognition (Abdel-Hamid *et al.*, 2014), and natural language processing (Sze *et al.*, 2017). CNNs utilise a distinct hierarchical architecture that facilitates the high-level abstraction of features from unprocessed, raw data to represent the input space (Sze *et al.*, 2017). Through a series of convolution transformations, the spatial resolution of an input image is sequentially down-sampled to produce higher-dimensional, useful representations of the image. These NNs use parameterised, sparsely connected kernels that are trained to identify and extract relevant image information through iterative learning over large amounts of example data (Shorten & Khoshgoftaar, 2019).

CNNs are structured as a feed-forward artificial NN consisting of three characteristic neural layers: convolutional, pooling, and fully connected layers. The typical deep CNN structure is depicted in Figure 4-4, which illustrates the process of accepting input images, performing feature extraction, and producing a classification. First, the input layer receives an image in its raw state that consists of a matrix of pixel values potentially divided across colour channels when the image is not in greyscale. The convolutional layers then apply several learned filters to the incoming data using a receptive field smaller than the dimensions to extract features. Each kernel convolves across the entire height and width of the image by computing the dot product between the matrix of pixel values and the kernel weights to produce multiple feature maps (convolved features) that collectively yield a feature vector. The filters are trained to only activate on class-related patterns found in the image, which relate to low-level features such as edges and shapes in early layers of the network and evolve to complex patterns in deeper layers of the network.

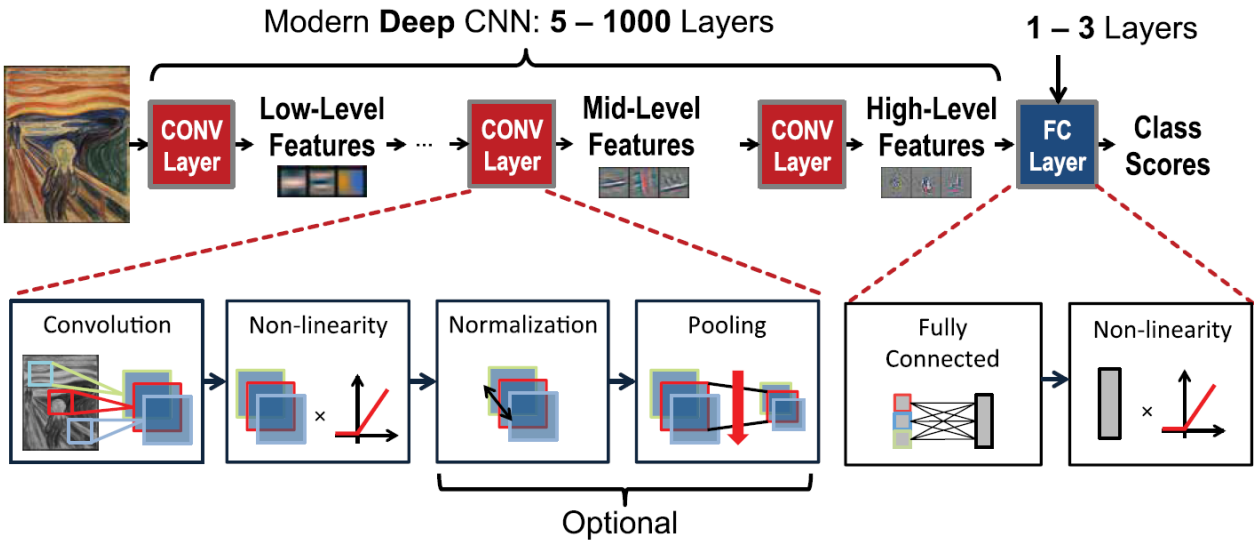
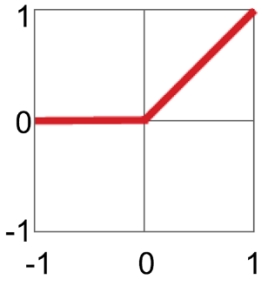


Figure 4-4: Typical CNN structure and operations (Sze *et al.*, 2017)

As a complement to convolution, normalisation is optionally applied to the outputs of convolution to prevent weight imbalances within the network. In doing so, the network can converge faster by controlling the input distribution across layers, thereby preventing outlying large weights from over-influencing the training process (Sze *et al.*, 2017). Following convolution, pooling is commonly employed as a non-linear down-sampling technique and is depicted in Figure 4-4. It reduces the dimensionality of the feature vector while retaining key elements indicative of the associated class (Sze *et al.*, 2017). Convolution and pooling are often stacked as complementary layers in a CNN, thereby encouraging the iterative sampling of an image to extract only the most discriminative elements. The final layers, shown in Figure 4-4, consist of fully connected NN layers that collectively act as a general classifier over the extracted features to assign confidence values that represent each possible classification. Both convolutional and fully connected layers employ non-linear activation functions, such as Softmax and Rectified Linear Unit (*ReLU*) operations (Table 4-1). These functions perform a non-linear transformation to linearly separate the data for classification. In the case of *ReLU*, the input value is passed to the next neuron when it is positive; otherwise, zero is passed as the output value. Softmax is mainly used as an activation function in the final layer of an NN where the number of nodes coincides with the number of classes. The extracted feature values are converted into probabilities that collectively sum to one, where each probability represents the network's confidence in assigning the corresponding class label.

Table 4-1: Activation functions (Sze *et al.*, 2017)

Activation function	Mathematical formula	Functional plot
Softmax	$y_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_i}} for\ i = 1, \dots, n$	*No functional plot*
Rectified Linear Unit (*ReLU*)	$y = \max(0, x) = \begin{cases} 0, if\ x \leq 0 \\ x, if\ x > 0 \end{cases}$	

The CNN neural layers enable characteristic concepts such as shared weights, sub-sampling, and local receptive fields. These concepts instil the CNN with some degree of shift (translation), scale, and distortion invariance, all of which contribute to its robust performance (Zhang & Suganthan, 2016). The convolution operation also aids in the NN's learning capacity through three key properties discussed in the remainder of this section, namely *sparse interactions*, *parameter sharing*, and *equivariant representations*.

4.3.1 Convolutional layer

A key component of CNNs is the convolutional layer, which conducts a linear operation between a set of learnable filters and the input data. These filters, also known as kernels, are represented as two-dimensional arrays of adjustable parameters and serve as feature detectors during the convolution process (Goodfellow *et al.*, 2016). Convolution produces two-dimensional feature maps that are then stacked to form a three-dimensional output volume, with the number of feature maps corresponding to the number of filters employed. Ultimately, the spatial dimensions of the output volume are determined by factors such as the kernel size (K), the stride (S), the use of padding (P), and the size of the input image or vector (W_{in}), as expressed by the following formula (SU, 2023):

$$W_{out} = \frac{W_{in} - K + 2P}{S} + 1 \tag{4-1}$$

The receptive field, determined by the kernel size, specifies the scope of inputs available to a neuron from the previous layer. In other words, the receptive field is the size of the region observed within the input image. As the network undergoes successive convolutions, the receptive field progressively condenses the input, extracting local features. In deeper layers, convolution combines these local features into global features, enhancing the network's capacity to generalise to unfamiliar input samples. Training further refines this process, as filter parameters adapt to elements most relevant to the classification task at hand.

Convolution in neural networks offers the advantage of **sparse interactions** or sparse weights. Unlike standard neural networks that utilise matrix multiplication involving all input and output values, convolutional networks establish fewer connections between these values. This sparsity arises from using a kernel smaller than the size of the image, yielding fewer connections. Sparse connections require fewer parameters and memory, thereby enhancing the network's statistical efficiency. **Parameter sharing** also benefits the computational efficiency of the network since it is a result of the kernel being reused at every position of the input, allowing features to be detected regardless of their position within the image. In traditional neural networks, each weight is used only once and never revisited, whereas convolutional layers reuse the same set of weights for each activation. This approach enables convolutional layers to learn a single set of parameters for each kernel rather than separate sets for each location (Goodfellow *et al.*, 2016). Because this effectively alleviates memory requirements, it is possible to stack multiple neural layers in a CNN to yield a deep network that can undertake increasingly complex classification tasks.

Due to parameter sharing, CNNs exhibit the property of **translational equivariance**. This means that variations in the input are reflected similarly in the transformed CNN output. In the context of image processing and object detection, this property allows the filters to learn and recognise features regardless of where they occur within an image. For example, if a filter is trained to detect the contours of a face, it will be applied across the entire input volume, enabling activation regardless of where a face might be positioned in the image. However, convolution is not innately equivariant to other transformations, such as scale and orientation (Goodfellow *et al.*, 2016). Still, equivariance toward these factors can be approximated by varying the kernel size across multiple layers to capture features at different scales. Similarly, data augmentation methods, such as rotating training samples, can generate plausible variations of the original input data and further enhance the network's property of equivariance.

4.3.2 Pooling layer

Pooling layers are commonly used in conjunction with convolutional layers to further reduce the spatial resolution of convolved features. The pooling function operates by replacing groups of feature values in a defined region with a summary statistic based on hyperparameters like the pooling function, kernel size, and stride (Goodfellow *et al.*, 2016). Pooling is applied iteratively across the input volume, pooling values within the receptive field. As illustrated in Figure 4-5, max-pooling selects the maximum value within the kernel's scope (2 x 2), discarding the remaining values as it passes across the input in strides of 2. Subsequently, this process condenses the input's dimensionality based on the highest values (*i.e.,* the most significant features), thereby diminishing the potential for a model to overfit since the input representation becomes further abstracted (SU, 2023; Sze *et al.*, 2017).

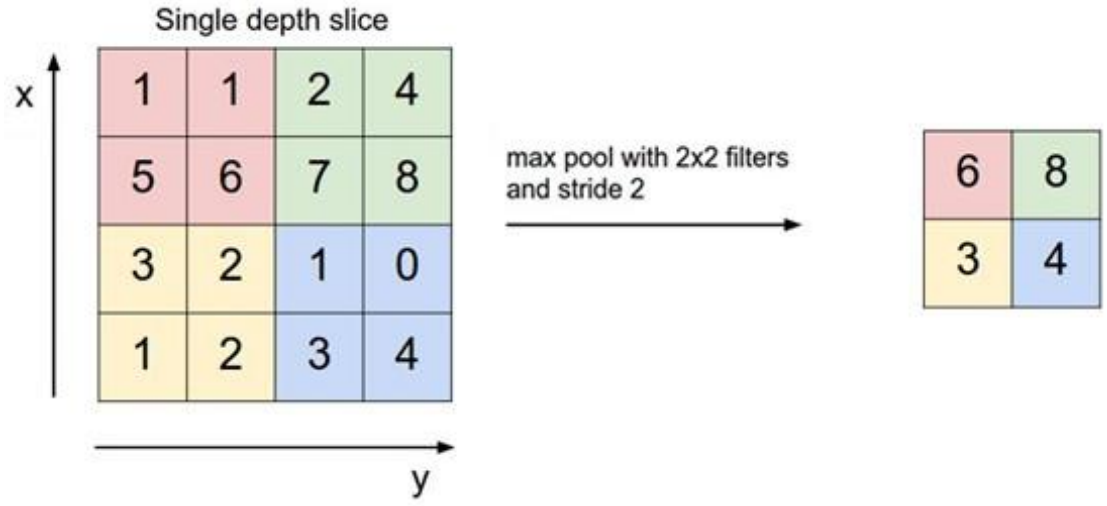


Figure 4-5: Results of max-pooling a region with a kernel of size 2x2 and a stride of 2 (SU, 2023)

Pooling also helps to approximate translation invariance toward small and insignificant translations that remain within the receptive field of the kernel (Goodfellow *et al.*, 2016). Local translation invariance favours the recognition of a feature based on its presence in the input rather than its location. In addition, pooling reduces the size of the convolved input, alleviating the computational burden on the next layer of the network and aiding in the efficiency of the network (Goodfellow *et al.*, 2016).

4.3.3 Fully connected layer

The final layer in a CNN is typically a fully connected layer, also known as a dense layer, and serves as the network's output layer. It determines the classification of an input sample by the value associated with each of its neurons that refer to a specific class, as illustrated by Figure 4-6. Because convolution is a linear operation, a non-linearity layer such as a dense is added directly after convolution to introduce non-linearity to the activation map. It is possible to stack multiple dense layers, thereby extending the network's depth and improving its approximation quality for complex tasks like image recognition. Apart from the final dense layer, it is important to have a sufficient number of neurons in the preceding dense layers to accommodate enough of the extracted features during the classification process.

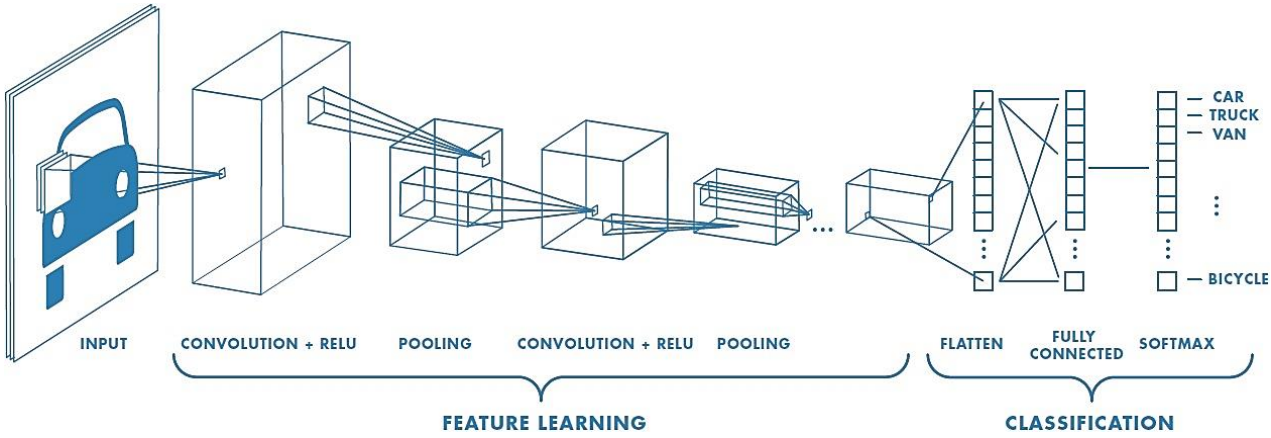


Figure 4-6: Typical CNN network structure conceptually grouped by action, namely feature learning and classification (MathWorks, 2017)

The final fully connected layer is referred to as the output layer, and its number of neurons equals the number of classes. Typically, the Softmax function is used as the activation function in this layer, allowing each neuron to express a fractional probability value that sums to 1. The highest probability signifies the most likely class based on the observed feature set. All the layers in a CNN contribute to the efficiency and accuracy of the network, which promotes specific properties and benefits explained further in Section 4.4 regarding their influence on pose classification.

4.4 Leveraging a CNN for pose classification

The unique operations of a CNN endow it with properties that greatly enhance its performance in classification tasks. Properties such as translation equivariance can be relied upon in image classification, while other properties can be exploited for the benefit of improved classification accuracy using data augmentation. This section explores this notion and examines the way in which CNNs can serve pose classification.

4.4.1 Hyperparameter configuration and layer organisation

The architecture of a CNN can be tailored to effectively extract features and classify specific tasks by adjusting the arrangement of layers and tuning network hyperparameters. Interestingly, research suggests that the depth and organization of intermediate CNN layers play a more decisive role in its classification capacity than the specifications and values of its kernels and weights. For instance, the studies conducted by Saxe *et al.* (2011) and Jarrett *et al.* (2009) demonstrate that a CNN initialised with random kernel weights can achieve satisfactory results when suitable pooling layers and non-linear activation functions are employed. The insights are further evidenced in the work of Pinto *et al.* (2009), who explored numerous CNN configurations for object recognition and found that networks with pre-trained weights only marginally outperformed those with random kernel weights.

CNN architectures exhibit inherent translation invariance as a result of convolution and pooling operations, making them effective in handling spatial variations. Additionally, CNNs are frequency-selective, as they prioritise the most frequent and significant features during the feature extraction process (Saxe *et al.*, 2011). Furthermore, Pinto *et al.* (2009) also concluded that adjusting network hyperparameters can significantly enhance classification performance. For these reasons, this dissertation includes a series of preliminary experiments aimed at identifying an optimal CNN architecture and hyperparameter configuration that best serve to classify pose data, as described further in Chapter 5.

4.4.2 Translation equivariance

Translation equivariance is realised by weight-sharing among kernels in the convolutional layers. The learned kernels adjust to recognise distinguishing features in images of the same class regardless of their position, resulting in kernels being invariant to different forms of translation that the object may exhibit in real-world instances (Goodfellow *et al.*, 2016). In addition, pooling (particularly max-pooling) further encourages translation invariance because the feature vector is condensed to only the most significant features, thereby disregarding the positional origin of the features in the original image (Goodfellow *et al.*, 2016). Regarding pose recognition, translation

equivariance is beneficial to the objective of this dissertation. The property can be capitalised on to ensure that a pose remains recognisable, regardless of where a person is located within a camera's field of view. This allows a subject to be positioned anywhere within a frame of a surveillance video and still have their posture and body cues discerned for fall detection.

Although convolution and pooling reduce the dimensionality of the input during feature extraction, the spatial information of local pixel arrangements is still preserved. Evidence of this is present in the work of Zeiler and Fergus (2014), where the learned kernels of convolutional layers were projected, revealing that when a rudimentary feature detected in lower levels of the network contributes to the classification of an image, kernels in deeper layers of the network tend to capture that information. As an example, the classification of a dog was shown to rely heavily on its facial features as opposed to its body since different breeds and sizes of dogs instead share a common arrangement of facial features rather than a similar silhouette. Lower-level kernels were responsible for the detection of a dog's individual features (*e.g.*, eyes, snout, and ears), while the deepest layers learned the arrangement of those features, for instance that the eyes are located below the ears and the snout between the eyes. This characteristic is beneficial to pose recognition since the human body's kinematic limitations (*e.g.*, the fixed points at which joints are located), and the positional organisation of its limbs (such as legs being located below the torso) can be learned in deeper layers to help indicate a person's actual pose or stance.

This section helped to explain how translation equivariance and the architecture of a CNN are likely to support pose recognition. Nevertheless, other properties such as the colour sensitivity of a network, can still be capitalised on to improve the accuracy of pose recognition and are discussed further in the following section.

4.5 Enhancing the feature space of a CNN for pose classification

Image classification using a CNN relies on feature extraction to obtain abstractions of the image content. The process involves identifying salient features that provide a meaningful distinction between classes. Yet these features may not strictly represent the class, such as, for example, when a feature related to the blue hue of the ocean is associated with a class representing boats. Coincidental class and feature associations in a CNN can occur when differentiating between classes that are similar in nature and appearance but are distinguished by slight variations, as is the case for different poses of the human silhouette. Features based on the colour of an object can often manifest as a salient feature in these instances, especially when they allow for a higher correlation with a class from among the available training samples. This phenomenon is further examined in Sections 4.5.1 and 4.5.2 and explained using evidence from supporting research.

4.5.1 The significance of colour in CNN-based classification

CNNs interpret raw sensory input through representation learning instead of feature engineering, making them adept at image classification. Mapping the identity of an object from a collection of image pixels is a complex task and is considered humanly insurmountable if confronted directly (Goodfellow *et al.*, 2016). CNNs resolve this complexity to a series of simple, nested mappings where each function mapping is defined in a separate layer of the network. During training, the network learns to recognise common image features related to the different classes. Depicted in Figure 4-7 are projected mappings of the increasingly abstracted features learned in the hidden layers of a CNN. Shallow layers typically learn to identify rudimentary image elements such as colours, edges, corners, contours, and shapes, while deeper layers tend to identify recurring combinations of these elements. Ultimately, the deepest layers of the network aim to describe the image in terms of its most characteristic elements that reflect the class of the image as determined by their recurrence (Goodfellow *et al.*, 2016; Zeiler & Fergus, 2014).

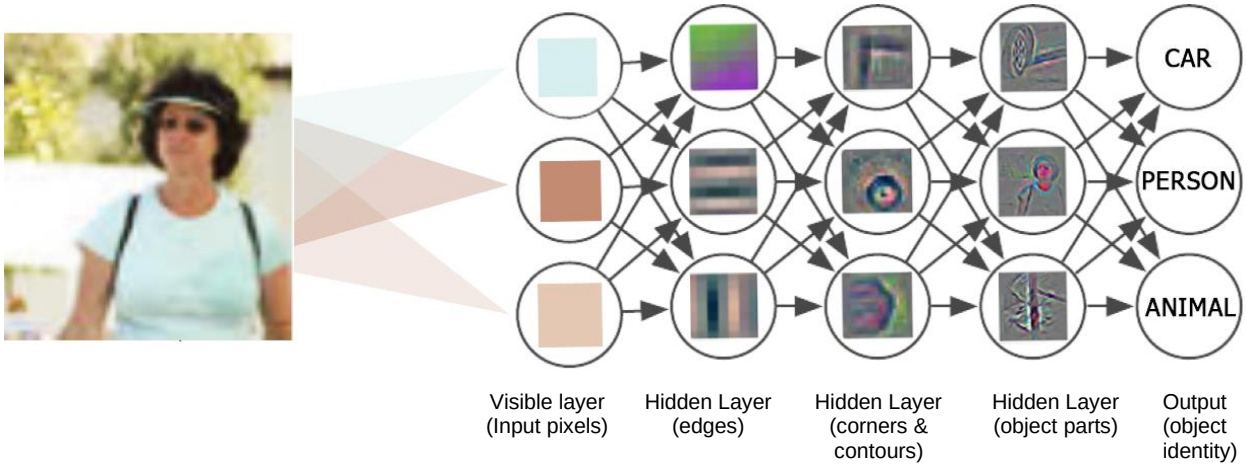


Figure 4-7: Process of representation learning in a deep convolutional neural network (Goodfellow *et al.*, 2016)

Colour information has been recognised as a distinctive feature that CNNs come to learn in instances where it helps the classifier distinguish between similar object classes. Zeiler and Fergus (2014) were among the first to allude to the importance of colour in image classification when they performed deconvolution and un-pooling on a trained CNN, projecting feature activations back to the pixel space. They observed that lower layers focus on edge and shape detection, while deeper layers seek out more complex and class-specific patterns. Surprisingly, colour information, previously associated with shallow layers, emerged as a feature sought in intermediate convolutional layers. These findings demonstrate the significance of colour as a salient feature in CNN-based image classification since the abstracted features in deeper layers of the network are highly correlated with its classification.

Krizhevsky *et al.* (2012) also explored the significance of colour in image classification when they performed image classification on 1 000 object classes. To increase the number of training samples, Krizhevsky *et al.* (2012) augmented their data with label-preserving transformations and adjustments to RGB colour channels. Their success in performing reliable classification despite changes in illumination intensity and colour showcased an important property of object recognition, namely that “object identity is invariant to changes in the intensity and colour of the illumination” (Krizhevsky *et al.*, 2012:6). This finding demonstrates that colour is preserved as an identifying attribute even after applying augmentations, furthering underscoring the value of colour as a salient feature in object identity and image classification. The notion is supported by the work of Buhrmester *et al.* (2019), which is discussed in the following subsection.

4.5.2 The influence of colour in CNN-based classification

Buhrmester *et al.* (2019) were among the first to provide comprehensive evidence for the influence of colour on the classification performance of CNNs. Through their experiments, they obtained insights into the most significant features in classification by manipulating several benchmark and custom-made datasets using augmentation techniques related to colour, noise, and blur. By studying neuron activations across different layers of the CNN, they examined the stability of confidence scores between models trained on augmented and non-augmented image sets. The findings revealed that certain object classes exhibited colour, texture, and shape dependencies. Notably, colour dependency was observed in the FlickrScene dataset depicted in Figure 4-8 (consisting of four categories: desert, snow, urban, and forest scenes) when its samples were transformed into greyscale. Buhrmester *et al.* (2019) attributed this to the similarity in texture and appearance of each landscape, which likely encouraged the classifier to learn characteristic hues of either warm or cool tones to distinguish samples from the desert and snow classes. Images in the urban category were least influenced by the omission of colour, suggesting that colour either varied too often between samples or because the geometric shapes of cityscapes played a more significant role in class identification.

Figure 4-8: Four sample categories of scenery images from FlickrScene dataset: deserts, mountains, forestry, cities (Buhrmester *et al.*, 2019)

Buhrmester *et al.* (2019) included the CIFAR-10 and CIFAR-100 datasets in their experiments, which comprise 10 and 100 object classes, respectively. Classification results on the CIFAR-10 dataset exhibited less reliance on colour cues compared to the CIFAR-100 dataset. This difference can be attributed to the CIFAR-10 dataset containing more distinct and easily distinguishable classes, such as birds and dogs. In contrast, the CIFAR-100 dataset consisted of numerous subcategories, like oranges and apples, that would otherwise belong to a broader class like fruit. Subcategory instances often share many similarities in their appearance, requiring the classifier to identify unique and characteristic features during training to differentiate among them. Experiments using greyscale augmentation indicated that distinguishing factors among samples of classes like oranges and apples, as well as roses and sunflowers, were primarily based on colour associations.

In contrast, the cat class comprising different cat species and shades of fur, achieved improved classification accuracy when colour information was omitted, indicating that colour added ambiguity rather than being a relevant attribute. Similarly, increased misclassification occurred for samples of the ship class where its characteristic blue colour, typical of its ocean backdrop, was omitted when applying greyscale augmentation. Noise and blur further degraded the shape information of ship images, leading to additional misclassifications and highlighting the importance of the boat's distinct bow shape and silhouette. In the case of crustacean images, degrading shape information increased misclassification for lobster and cockroach samples due to their unique silhouettes. However, omitting colour from crustacean images did not impact classification accuracy, indicating the non-dependence of colour-based features in the classifier. Similar findings were observed in a CNN model trained on the PersonFinder dataset depicted in Figure 4-9. Augmentations applied to the dataset showed that removing colour information did not negatively affect the classifier's classification performance. This result supports the logical assumption that factors like image background, colour variations, and clothing are irrelevant in person detection. Instead, the maintained performance suggests that the classifier relies on cues related to the shape and silhouette of the human body as primary indicators of the class.

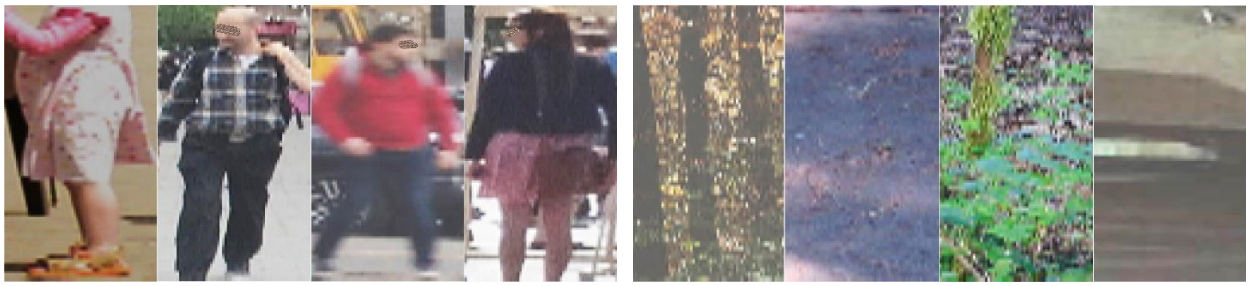


Figure 4-9: PersonFinder dataset samples of two image classes: including (left) and excluding (right) a person (Buhrmester *et al.*, 2019)

The findings of Buhrmester *et al.* (2019) demonstrate how CNNs adaptively learn the features that provide the most distinction between the classes of a dataset. In the cases where classes were found to be less dependent on colour information for their classification, unique shape and texture information was instead relied upon as significant and characteristic features. By applying data augmentation techniques, Buhrmester *et al.* (2019) were able to uncover these feature dependencies within the learned classes. By extension, it should be theoretically possible to use similar augmentation techniques to impress any desired feature dependencies on a classifier. The potential of this approach is further explored in Chapter 5, where the experimental design for pose classification and fall detection is detailed. To further motivate the viability of such an approach, evidence regarding the successful application of data augmentation in other tasks is provided in Section 4.6.

4.6 Data augmentation

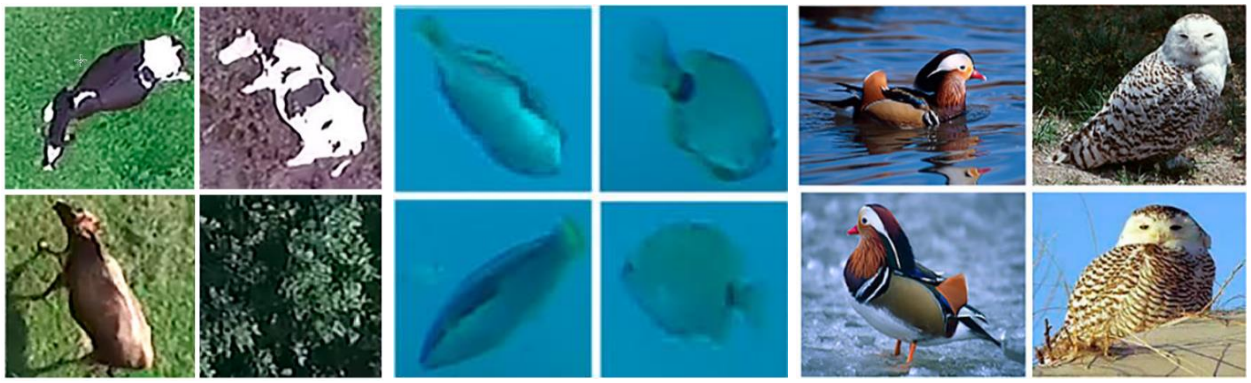
Variations naturally occur in the presentation and perception of objects. Yet, the representational property of the biological visual system enables humans to recognise objects accurately despite variations in size, position, or illumination. This invariance is a key mechanism for accurate object recognition and has been a prized objective since the inception of artificial NNs (Kheradpisheh *et al.*, 2016). View-invariant object recognition is well-established in neuroscience and is an inherent property of human vision. It is argued that invariance can be emulated using purposeful data augmentation and related techniques to enhance a NN's tolerance of different types of variations (Goodfellow *et al.*, 2016). Growing evidence suggests that this can be credited to deep networks exploiting highly discriminative features that do not necessarily correlate with human perception, given their large capacity and unconstrained learning format (Hernández-García *et al.*, 2019). This section examines approaches to data augmentation and reasoning is provided for the way in which augmentations influence network performance when training a CNN.

4.6.1 Improving performance through general invariance

Deep learning methods heavily rely on high-quality and representative training data to achieve optimal performance since machine-learned models become more effective when generalising from diverse data (Dang *et al.*, 2019). To ensure that a model can generalise its classification capacity to unseen instances, a large dataset with representative and variable class samples is essential. However, data is finite, and a limited training set may contain fewer class samples than is necessary to represent the population well. In such cases, label-preserving data augmentation techniques can artificially expand the training set by generating additional instances from the existing data (Shorten & Khoshgoftaar, 2019). Common linear transformations, such as cropping, shifting, scaling, flipping, and rotating, can be applied to the image geometry to create new

samples that retain their natural plausibility. By introducing variations and generating semantically similar samples, the model is encouraged to disregard the inter-sample differences and instead learn the common and recurrent features of each class (Taylor & Nitschke, 2018).

Okafor *et al.* (2018) achieved high classification accuracies for their CNN models by applying augmentations related to colour constancy, affine transformations, and matrix-based transformations across various datasets. Colour constancy, a well-studied area in image processing and computer vision, aims to preserve the perceived colours of objects captured under different lighting conditions by mitigating environmental sources of colour influence (Okafor *et al.*, 2018). Essentially, the goal of colour constancy is to maintain the perceived colours of objects captured in variable lighting by identifying and counteracting environmental sources of colour influence (*e.g.*, household lighting) by using techniques like contrast and lightness enhancement or colour rendition. Okafor *et al.* (2018) explored the application of colour constancy and rotation matrix algorithms as data augmentation methods. They evaluated the impact of these techniques on classification performance using different image datasets depicted in Figure 4-10, which include Croatian fish species, multiple bird species, and aerial shots of grazing livestock.

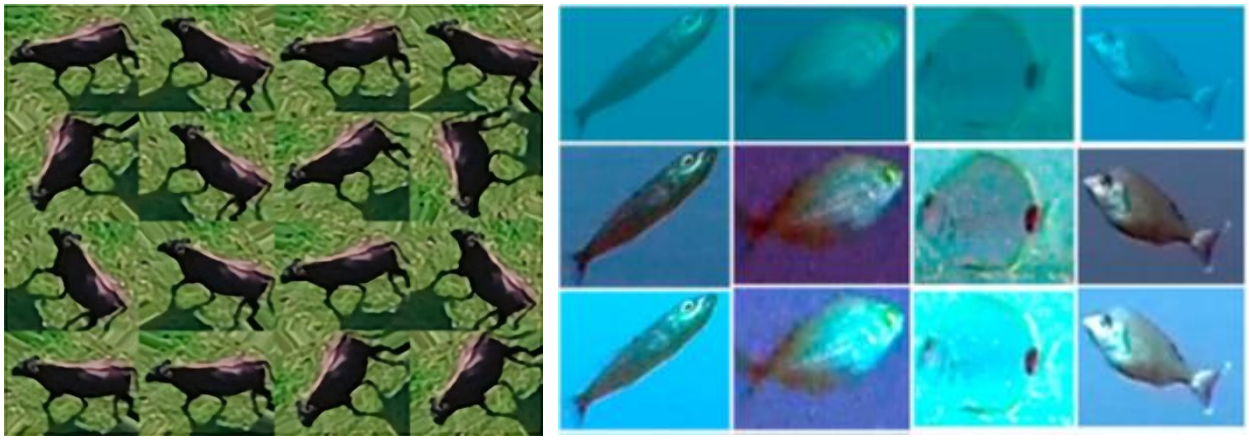


(a) (b) (c)

Figure 4-10: Sample images of the Aerial UAV dataset (a), the Croatia Fish dataset (b), and the Bird-600 dataset (c) (Okafor *et al.*, 2018)

Okafor *et al.* (2018) experimented with various data augmentation techniques to expand their available training dataset by producing new image samples from their original counterparts. Their findings indicated that rotational transforms were significant to the Aerial UAV dataset, where the natural orientational variations of drone-captured images could be simulated in the generated samples, as depicted in Figure 4-11 (a) on the next page. By including multiple directional positions for the grazing livestock, the classifier became invariant to this variation. On the other hand, colour constancy was most significant in the Croatia Fish and Bird-600 datasets. Due to the diverse colour variations in the bird species images due to differences in scenery and lighting conditions, the classifier benefitted from additional samples modified through specialised retinex augmentation techniques that simulate variations in illumination through tinting and shading. The

variation in the appearance of the birds allowed the classifier to learn the characteristic, coloured plumage of the different species while disregarding any environmental influences. The fish species classification demonstrated even greater dependence on colour, likely attributed to the consistent angular shape of the fish and the blue ocean backdrop. By providing the classifier with lighter and darker shades of the same images that are depicted in Figure 4-11 (b), invariance to different lighting conditions was achieved, allowing the classifier to differentiate between classes based on the coloured scales that are unique to each species.



(a) (b)

Figure 4-11: Generated variations of original images produced through augmentation for aerial images of grazing livestock (a) and Croatian fish (b) (Okafor *et al.*, 2018)

Colour-based augmentation techniques help to expand and improve the representative nature of the dataset. By introducing artificial diversity, machine-learned models become more robust and effective in classification by disregarding irrelevant differences among class samples (Dang *et al.*, 2019). The work of Okafor *et al.* (2018) showcases the discriminative nature of colour and how data augmentation can be used to either diminish or amplify its influence.

4.6.2 Improving performance through selective invariance

In the same manner that achieving general invariance supports reliable classification for instances whose features vary naturally, selective invariance can be instilled for a chosen variable. Encoding selective invariance is often accompanied by domain-specific knowledge relevant to the classification problem and dictates the most applicable augmentations for training an accurate model. Instead of applying a wide range of augmentations to the entire dataset, selective invariance focuses on identifying the distinguishing features of a class and introducing variability specifically for these features. This approach enhances model robustness and improves classification performance (Shorten & Khoshgoftaar, 2019).

Galdran *et al.* (2017) researched the early detection of melanoma, a highly aggressive form of skin cancer, using CNN-based analysis. Effective detection relied on identifying the known visual indicators of the disease related to the colour, size, shape, and texture of skin lesions (*e.g.*, moles). Existing computerised image analysis systems deduce possible cases of melanoma using hand-crafted measurements evaluated against a set of feature representations drawn from an image of a skin lesion. Deep learning, specifically CNNs, automates this feature design process but requires a substantial database of labelled, high-quality image samples. However, in medical scenarios, building such a database poses challenges due to the rarity of the disease, the need for expert labelling, and patient privacy restrictions (Shorten & Khoshgoftaar, 2019).

Furthermore, the image samples used in their experiments were obtained from various hospitals that use different acquisition systems to capture dermoscopic skin images. The inconsistent acquisition conditions lead to variations in the appearance of the skin lesions, further adding to the sparsity of the data. To address this challenge, Galdran *et al.* (2017) applied the shades of grey colour constancy technique to colour-normalise the training set by estimating the colour of the illumination source and subtracting it from the image. In addition, they retained the colour of the estimated illuminates and used it to generate new samples by applying the illuminates to a white-balanced version of the training set, thereby simulating diverse yet plausible illumination conditions, as depicted in Figure 4-12.

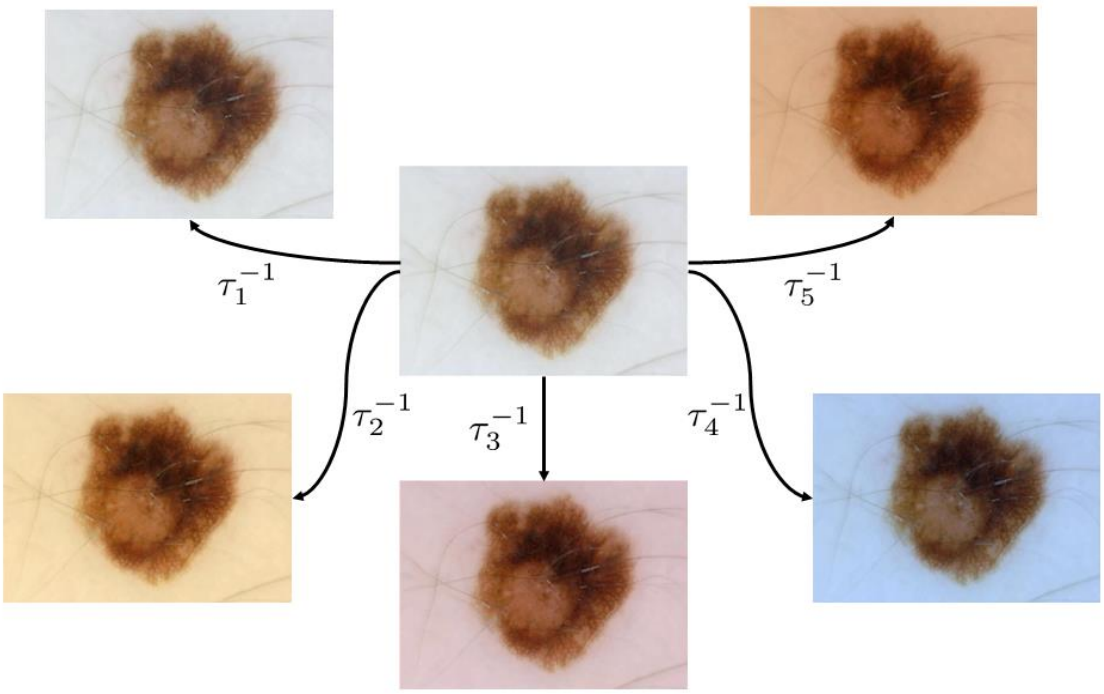


Figure 4-12: New samples of skin lesion images generated through colour-casting white-balanced samples with different illuminants (Galdran *et al.*, 2017)

Selective invariance takes advantage of the expressive power of NNs by training them on induced variability to learn to identify features not influenced by the known factors of variation. Colour augmentation facilitates robust training by transforming and expanding a limited dataset with more diverse samples to aid the network in better generalising to unseen instances. Through this method, Galdran *et al.* successfully trained a sophisticated image classifier capable of accurately identifying melanocytic skin lesions, regardless of the imaging device or acquisition conditions. by instilling the classifier with colour-invariance enabled it to disregard the colour-casting effects caused by an illumination source (2017).

The research results produced by Okafor *et al.* (2018) and Galdran *et al.* (2017) illustrate an important property of colour in image classification in that an object's identity (or classification) is independent of changes in the intensity and colour of its illumination (Krizhevsky *et al.*, 2012). This highlights the value of colour-based augmentation, particularly in classification tasks where objects exhibit similar appearances across instances (*e.g.*, fish) or are commonly captured under diverse lighting conditions (*e.g.,* throughout different times of day). Exposing a CNN-based classifier to the natural variability related to the imaged objects promotes the learning of more informative features based on the relevant factors underlying the associated class, rather than relying solely on its predictive capability given a limited dataset (Shorten & Khoshgoftaar, 2019).

4.6.3 Improving performance through feature engineering

During supervised training, the goal of a machine learning algorithm is to accurately classify a problem based on a set of candidate features. However, the training data may not adequately represent the underlying problem and may require restructuring to reveal any inherent saliency to the algorithm. The training data may also contain irrelevant features that do not support the classification task or are redundant in its characterisation (Kotsiantis, 2011). Data restructuring can be done through either feature engineering or feature selection to address these issues. Feature engineering involves decomposing or aggregating data to better highlight the relationship between relevant features and the target class, while feature selection involves removing redundant features that do not contribute to generalisation. Both techniques are commonplace in machine learning data preparation and allow applicable features to be identified and retained to improve the predictive performance of a classification model (Brownlee, 2017).

Classification of a data sample is performed based on a set of available features and should therefore be relevant to the classification task (Zheng & Casari, 2018). Transformation functions such as arithmetic and aggregate operators are often used to better present data to the learning algorithm. These functions generate samples that emphasise any significant correlations between the features and the associated class labels. For instance, restructuring dates to represent yearly

seasons or fiscal quarters can be beneficial when predicting class-related events that exhibit seasonal or periodic patterns. Restructuring features does not follow a predefined method, as the best approach varies depending on the data and the specific task at hand. Moreover, determining the best feature representation in complex tasks is likely unknowable and, at best, intractable a priori (Brownlee, 2017). Therefore, feature engineering requires human creativity and domain expertise. Still, the task remains mainly relevant to comprehensible data where the insights necessary to best model a task are understood beforehand and can be engineered from the available data.

Media data such as images and video can be too voluminous and complex to model directly, and is not easily comprehensible for manual feature engineering. Feature extraction is instead relied upon to reduce the dimensionality of these observations to a subset of abstracted features when modelled by an NN, a process referred to as representation learning (Goodfellow *et al.*, 2016). Still, the derived feature-rich abstractions cannot be (directly or easily) interpreted to understand which features do well for the purpose of restructuring or engineering new features. In addition, because the NN is a discriminative classifier, it seeks only to learn which features (however crudely) delineate the training data into the prescribed classes (Liu *et al.*, 2017; Jurafsky & Martin, 2000). For example, if a classifier were tasked with differentiating between images of dogs and cats – and only the dogs were coincidently all wearing collars – that feature alone would satisfy the model. In such a model, its only knowledge of cats is that they do not wear collars which is likely not a true reflection of reality and would not generalise well to unseen data. Representation learning entails an NN automatically electing salient features from its training data and is most useful for complex data such as images and video. Still, the freedom afforded to a classifier by representation learning may result in unintended or unrelated underlying structures in the data being reflected in its extracted feature set (Jurafsky & Martin, 2000). The effects can, however, be countered through feature engineering by emphasising or generating features that are relevant to the classification task.

The work of Sung *et al.* (2010) demonstrates the use of feature-engineered data for terrain classification in autonomous off-road navigation. Differentiating between various terrain surfaces is necessary to optimise the navigation strategy of mobile robots. For instance, maximum speeds must be reduced when traversing sand to prevent entrenchment. Sung *et al.* employed a pre-processing step using footage from a roof-mounted camera and trained a maximum likelihood classifier. The classifier performs region localisation by leveraging colour and texture information which explicitly encodes the correlation between raw pixel regions and their corresponding region classes, as shown in Figure 4-13 on the next page. Texture information was feature engineered using a discrete wavelet transform by representing image information as wavelets (a wave-like signal where its amplitude oscillates around the value zero and is influenced by changes in pixel

information when applied to images) to accurately extract localisations of colour and texture from the image set. Furthermore, spatial coordinates from the camera aided in refining the class-labelling task based on the positional information of terrain regions within an image, such as the sky typically occupying the upper field of view. The augmented and feature-extracted data were used to train a multilayer perceptron neural network, enabling the classification of soil, gravel, pavement, grass, foliage, and sky classes as regions within a video frame.

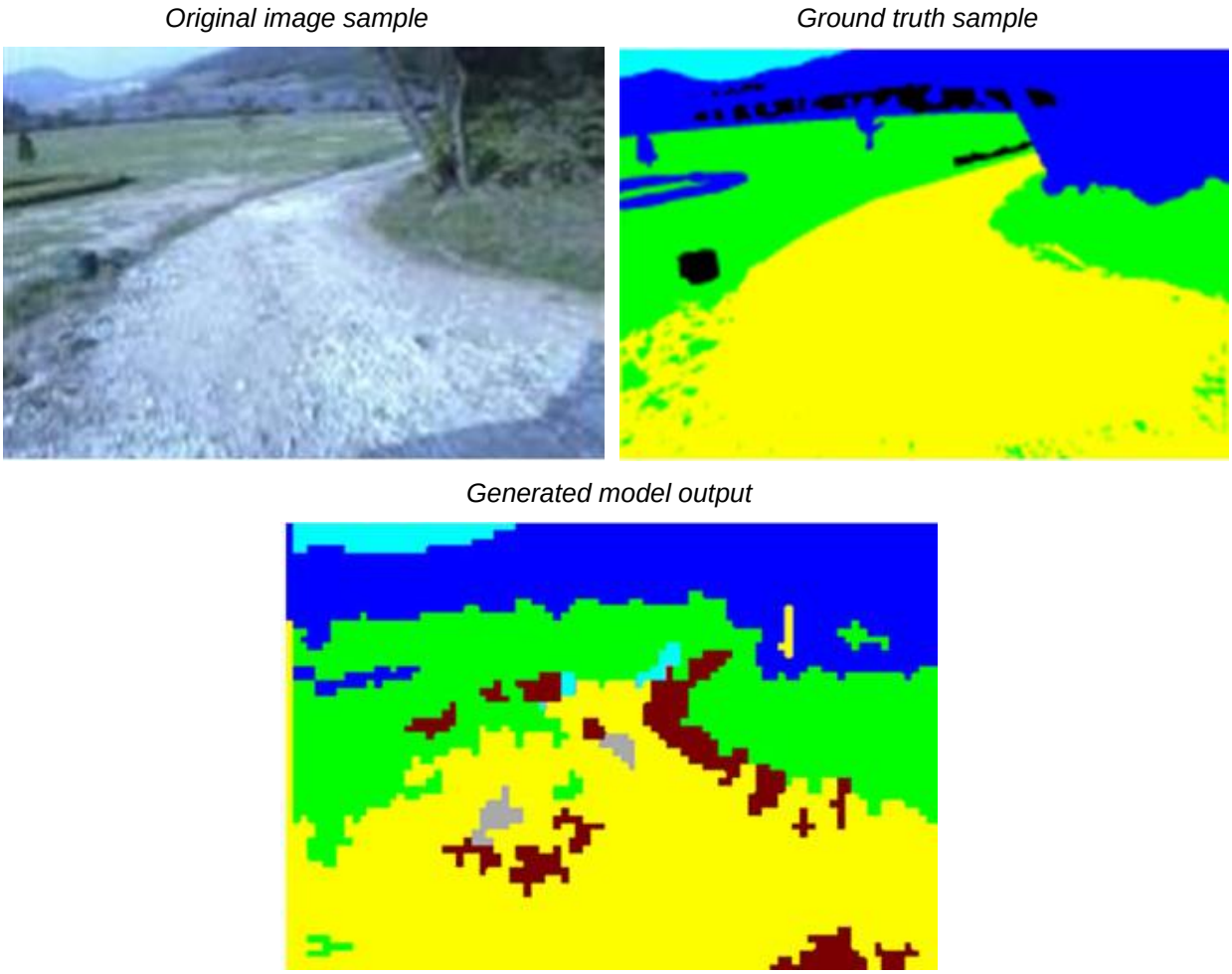


Figure 4-13: Original photo image (left) alongside its labelled ground truth (right) and generated region-labelled sample (below) (Galdran *et al.*, 2017)

Similarly, feature-engineered image data for human activity recognition was performed by Roy *et al.* (2015). Interpreting actions from video footage is challenging due to the complexity of the data and the high degree of freedom in how the body expresses movement. While representation learning can extract features efficiently, it does so without considering the relevance of the feature set to the classification task. According to Roy *et al.*, unambiguous classification tasks often rely only on a small subset of features (2015). For instance, distinguishing between walking and running activities is primarily inferred from leg movement. Relying on fewer features is preferable and can improve the generalisation capacity of a model, especially since high-dimensional

abstractions are more prone to overfitting than low-dimensional representations or raw features (Guyon & Elisseeff, 2003). Roy *et al.* therefore performed task-aware feature selection on features engineered from their footage data that express body movement for multiple video frames. In their experiments, improved dense trajectory (IDT) feature vectors were derived by tracking the movement of pixels across sequential frames. Figure 4-14 illustrates three dimensions of the projected IDT features, each expressing changes in tracked pixels using different descriptors.

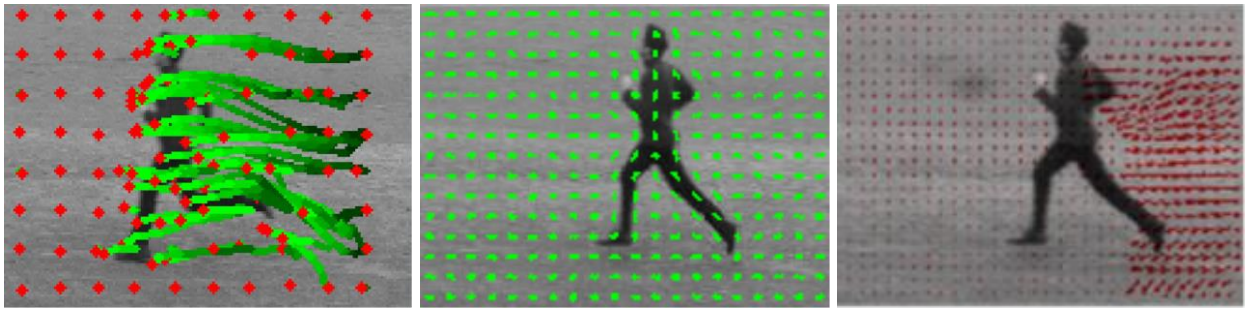


Figure 4-14: Three sub-dimensions of the projected IDT feature vector expressing pixel movement across multiple frames for the same video clip (Roy *et al.*, 2015)

The IDT feature vector represents human motion by tracking the movement of pixels within a neighbourhood and expressing the change according to 426 dimensions, encompassing heavy overlap to ensure complete coverage of all motion. However, Roy *et al.* (2015) argued that such a high-dimensional feature vector contained redundant motion information that was not relevant to their task classes, namely boxing, clapping, waving, jogging, running, and walking. They performed purposeful dimensionality reduction to refine the feature vector, isolating and retaining only the motion features that effectively describe each task class. Using an NN model trained on the IDT feature vectors, Roy *et al.* could identify 30 relevant IDT vector dimensions by monitoring the neuron activations in the initial NN layer and observing their overall contribution to classification. Limiting the training set to these 30 feature dimensions improved classification and runtime performance. However, misclassifications became more prevalent in class instances that looked alike, such as running, walking, and jogging actions. The research performed by Roy *et al.* (2015) demonstrates how representation learning can be capitalised upon when dealing with complex video data while guiding the NN training process using features engineered to be relevant to the classification task.

4.7 Conclusion

This chapter included a comprehensive overview of NNs, CNNs, and the concept of data augmentation. NNs were acknowledged as powerful learning algorithms that approximate complex functions which are otherwise challenging to achieve through conventional approaches. The unique operations of a CNN were also credited as the reason for its ability to extract features

from complex data such as images and video. Additionally, its translation equivariance property was considered advantageous for human pose classification since it allows for reliable recognition regardless of the camera perspective or the person's position within the image or video frame. Moreover, the chapter highlighted the potential of using feature engineering and data augmentation to emphasise feature saliency for improving pose classification. The influence of colour in CNN-based classification was demonstrated, revealing that augmenting pose data samples with colour to express indicative cues and emphasise variations may have the potential to help distinguish pose classes.

The literature findings in this chapter provided a theoretical background and rationale for the experimental approach described in the next chapter. Chapter 5 demonstrates the effectiveness of feature engineering in the context of pose recognition. The experimental results provide empirical evidence to support the theoretical claims made in this chapter and establish the viability of using data augmentation to encode relevant and underlying patterns that distinguish human poses visually.

CHAPTER 5 EXPERIMENTAL DESIGN AND RESULTS

5.1 Introduction

Based on the evidence presented in the previous chapter, the importance of colour as a significant feature in CNN-based image classification was illustrated. Building upon this knowledge, the present chapter explores the use of colour-focused data augmentation techniques to improve pose classification. To this end, a pose descriptor is introduced that delineates the human frame and exposes relevant information about the kinematic arrangement of the body to a CNN classifier. The influence of colour is investigated by assigning hues to body key points through a dynamic and a fixed association, and the impact on classification performance is evaluated.

To investigate the effectiveness of data augmentations on pose data, a preliminary experiment is conducted and presented in Section 5.2. These experiments evaluate different pose descriptor formats based on adjustments to the hue, saturation, blending, and coverage area of key point body markers when applied to a two-class pose dataset. The insights gained from the preliminary experiments are used to inform the augmentation schemes proposed in the primary experiment, detailed in Section 5.3 and evaluated on a real-world fall dataset. The findings and conclusions of both experiments are summarised in Section 5.4. Additionally, the section includes a discussion on the applicability of the proposed augmentation techniques for fall detection and other pose-dependent implementations. Finally, the chapter concludes in Section 5.5 by summarising the key contributions and insights presented throughout this chapter.

5.2 Preliminary experiments

The preliminary experiments are an initial investigation into how data augmentation as a form of feature engineering can be applied to pose classification. These experiments are intended to act as a proof of concept to establish the viability of novel augmentation techniques for pose expression. The experimental process and results discussed in this section are also published in a peer-reviewed conference paper presented at the 2019 Southern Africa Telecommunication Networks and Applications Conference (Du Toit *et al.*, 2019b) and is provided in Annexure B. The next section provides an overview of the experimental design regarding the construction of a pose descriptor, the applied augmentations, and the motivation for their use.

5.2.1 Overview of experimental design

Human pose recognition can enable safety and risk monitoring for human-centred applications that depend on pose information. In this study, data augmentation techniques are explored with the goal of improving pose recognition in real-time human activity recognition tasks such as fall

detection. The novel augmentation techniques described in this section are inspired by the literary evidence regarding feature saliency in the extraction process of a CNN. Besides texture and shape features, colour is considered a discriminative feature in image classification for tasks where it functions as a unique class identifier. Given this insight, multiple augmentation techniques are used through experimentation to assess how colour could be applied as a unique identifier for human pose. The proposed augmentation takes the form of a pose descriptor that resembles a skeletal outline of a person, which can be superimposed onto images or videos of people. Different hues, body marker sizes, and saturation intensities are experimentally applied to the descriptor and their impact on pose classification is observed to determine their effectiveness. The expectation is that the CNN will heuristically learn to leverage the encoded colour information to supplement its feature space and improve its capacity to distinguish between different human poses.

Ultimately, the goal is to capitalise on the adaptive and comprehensive learning capability of a CNN by supplying it with ancillary pose-related information as part of its training and prediction process. The approach also leverages its elective learning process to account for the subtle cues, recurring patterns, and natural variations in body posture that would otherwise go unobserved, or would introduce ambiguity in conventional fall detection solutions that rely on complex metrics and motion-related thresholds. Exploratory experimentation of colour-focused augmentations can yield the necessary insight to establish a favourable and effective collection of techniques that are congruent with neural learning and pose expression. If successful, purposeful data augmentation could be used as a simple technique to enhance the accuracy of a vision-based machine-learned fall detection system that relies on a single camera feed, precluding the need for any additional sensor technology to improve its detection accuracy.

The remainder of this section details the experimentation process regarding the dataset, the different augmentation techniques, and the resulting influence on classification performance. Subsection 5.2.2 introduces a two-class dataset that was collected to serve in the experiments; its origin and necessary pre-processing are explained. Individual augmentation techniques and generated datasets are presented in Subsection 5.2.3 and represent the different formats in which colour can be encoded into an image. The experimentally derived CNN architecture used to facilitate the experiments is described in Subsection 5.2.4, followed by a discussion of the model performance results in Subsection 5.2.5. Finally, all insights observed in the experiments are summarised in Subsection 5.2.6, which establishes the necessary basis for further experimentation in Section 5.3.

5.2.2 Image dataset

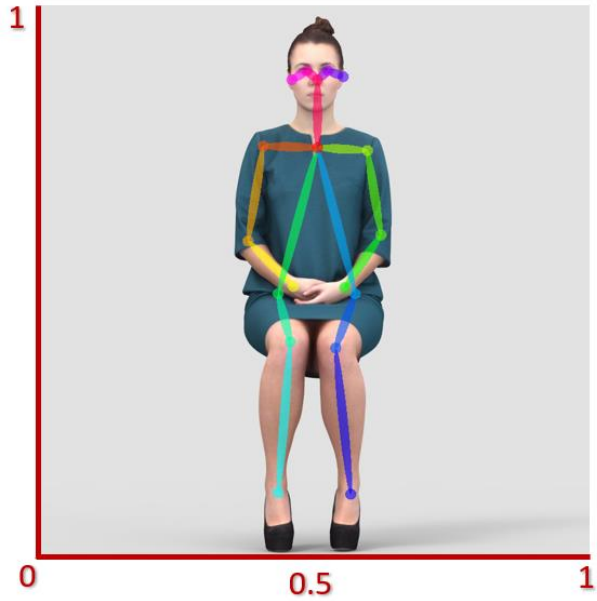
The preliminary experiments are conducted on a curated dataset[1] of pose images collected from both the public domain using image searches and batch downloading and websites that freely offer the use of photos under a creative commons licence agreement. The dataset comprises 25 000 unique images that depict people in diverse environments in either of two poses, namely sitting or standing. Each pose is equally represented in the dataset with 12 500 images, samples of which are depicted in Figure 5-1, while an additional excerpt of the data can be viewed in Annexure C. The dimensions and resolutions vary across all the images and are therefore incompatible with the fixed height and width input format of a CNN. All datasets are therefore pre-processed by *OpenPose* to derive body key points and are illustrated by the corresponding skeletal mappings in Figure 5-2.

Figure 5-1: Image dataset samples of sitting and standing poses

Figure 5-2: Image dataset samples of superimposed OpenPose skeletal mappings

[1] https://doi.org/10.25388/nwu.23290937

The key points are made up of 18 localised joints and body parts that are illustrated in Figure 5-3 (a). *OpenPose* outputs the identified key points as normalised XY-plane coordinates which are listed as attribute values in the adjoining table of Figure 5-3 (b). Each localised coordinate is associated with a degree of confidence by *OpenPose* that defines the estimated probability of the specific joint or body part's predicted location. When occlusion occurs and the body is partially obstructed from the view of the camera, it can cause difficulty in accurately estimating the corresponding key point, resulting in a weak confidence score for the occluded body parts. Since *OpenPose* iteratively refines key point estimations, it uses the confidence scores to reject low-confidence predictions or to weigh the contributions of different key points in further processing.



(a)

Key point	X-Value	Y-Value	Confidence (%)
R_Ear	0.54	0.87	84.36
L_Ear	0.44	0.87	93.05
R_Eye	0.51	0.88	93.40
L_Eye	0.47	0.88	90.97
Nose	0.49	0.87	86.24
Neck	0.49	0.74	85.02
R_Shoulder	0.59	0.75	78.90
L_Shoulder	0.40	0.74	80.47
R_Elbow	0.62	0.57	83.25
L_Elbow	0.38	0.58	77.15
R_Wrist	0.53	0.49	77.32
L_Wrist	0.44	0.49	81.03
R_Hip	0.56	0.48	58.15
L_Hip	0.42	0.48	63.32
R_Knee	0.54	0.39	79.99
L_Knee	0.46	0.38	84.46
R_Foot	0.56	0.12	68.07
L_Foot	0.43	0.11	72.44

(b)

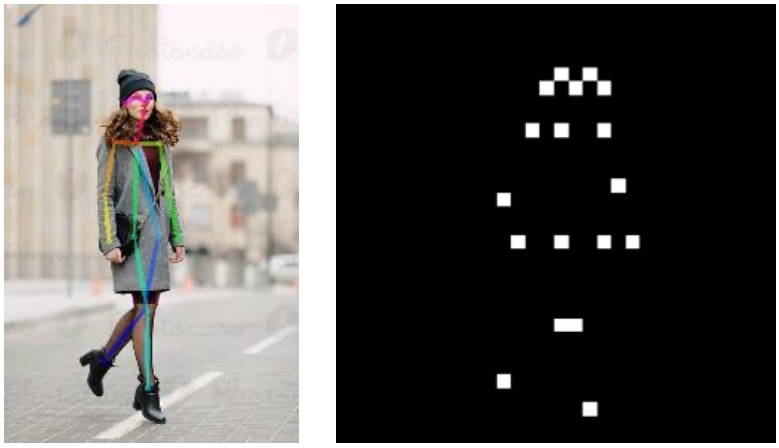
Figure 5-3: Superimposed estimated key points of a sitting pose (a), alongside the XY-coordinates of the corresponding 18 body key points (b) produced by *OpenPose*

The *OpenPose* output provides a high-level understanding of the human pose in an image or video by offering a key point delimitated outline of the body that implies the pose and posture of the person. In terms of this study, the proposed augmentation techniques seek to extend the synthesised pose information to a CNN pose classifier. The XY-plane coordinates and confidence values can be encoded into an input image through data augmentation to embed discriminative cues that relate to its classification. Subsection 5.2.3 explores various augmentation techniques to derive a pose descriptor from the *OpenPose* estimations.

5.2.3 Dataset augmentation

Data augmentation is a technique commonly employed to enhance the performance of machine learning models by artificially increasing the number of samples in a dataset. This is often achieved through label-preserving transformations, which introduce variations into the data and encourage the classifier to seek out salient structures underlying the data. Essentially, the model learns to identify important features shared by samples of the same class and, by extension, characteristic properties of the imaged object itself, thereby helping it to generalize to unseen data (Taylor & Nitschke, 2018). In addition to enhancing feature saliency, data augmentation can also serve as a means of feature engineering. In this chapter, preliminary experiments focus on the creation of a pose descriptor that can be overlaid onto images to effectively convey the underlying pose. Various aspects, such as key point sizes, colours, and blending, are explored to determine the optimal configuration in terms of encoding heuristic information.

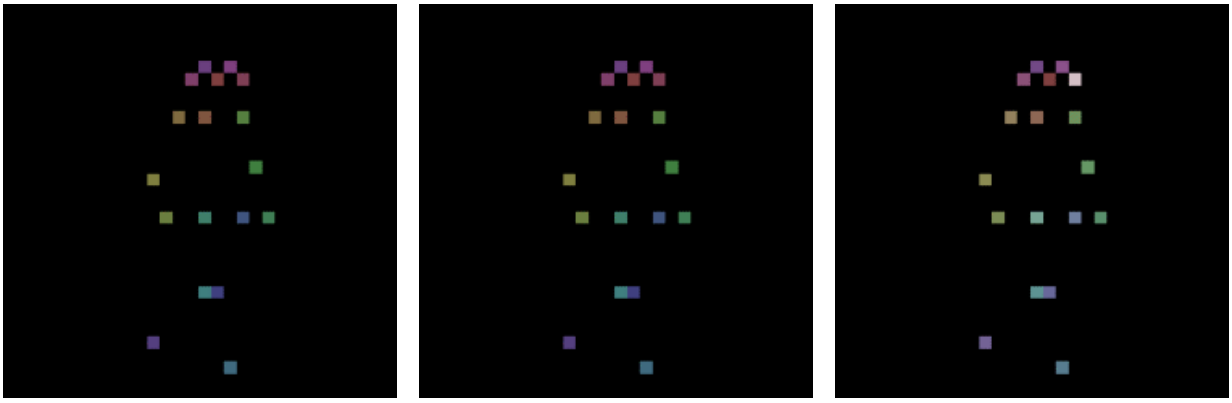
Figure 5-4 (a) depicts a sample pre-processed image from the dataset that is overlayed with an illustrative key point mapping generated by *OpenPose*. In Figure 5-4 (b), a baseline pose descriptor is constructed by representing the estimated *OpenPose* key point locations as individual white pixels within a standardized 32 x 32 pixel image frame against a black background. The generated key point mapping maintains the person's position within the original image and is not centred in the frame, since positional information is not considered by a CNN due to its properties of equivariance. This pose representation is encoded with the least amount of heuristic information to allow the influence of other augmentations to be comparatively evaluated against the baseline regarding classification performance.



(a) (b)

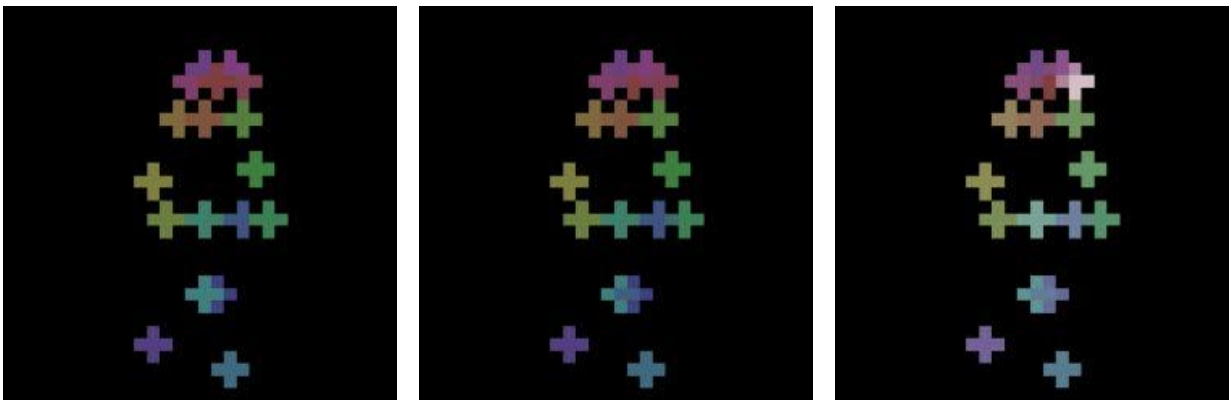
Figure 5-4: Original data sample (a) alongside a generated pose descriptor (b) that depicts key points as white pixels

The baseline dataset consists of generated images that represent the human poses present in the pose image dataset. Similarly, six additional datasets are created using the same key point estimations but using different pose descriptors, resulting in varying levels of encoded heuristic information. The first collection of generated images is represented in Figure 5-5 (a), where each *OpenPose* coordinate is depicted by a single-pixel marker with a unique colour corresponding to specific body parts. Each colour is sampled from a uniform distribution of RGB colour gradients to ensure that distinct hues are utilised. Likewise, Figure 5-5 (b) presents the same pose mapping, but overlapping key point colours are blended to ensure obscured body parts remain observable in the pose descriptor. Figure 5-5 (c) again depicts the same amount of heuristic information but encodes the *OpenPose* localisation confidence through tinting, where lighter shades indicate lower degrees of confidence and white signifies zero confidence. These datasets are reproduced in Figure 5-6 but instead uses a larger, crosshair-style key point marker to represent the body key points. An additional excerpt of the generated dataset can be viewed in Annexure C.



(a) (b) (c)

Figure 5-5: Representative samples of generated datasets that encode varying degrees of heuristic information using coloured pixels that represent body key points



(a) (b) (c)

Figure 5-6: Representative samples of generated datasets that encode varying degrees of heuristic information using coloured crosshairs that represent body key points

In summary, seven sets of images are generated using different augmentation techniques to depict varying levels of heuristic information. The baseline dataset serves as a neutral pose descriptor, representing key points with white pixels and no additional augmentations. The six remaining datasets use coloured key points and are incrementally augmented by blending and tinting marker colours. These datasets are distinguished by the type of key point marker used: either a single-pixel marker or a crosshair marker. Ultimately, each dataset serves in training pose classifier models using a favourable CNN architecture that is described in the subsequent section.

5.2.4 Convolutional neural network architecture

A non-exhaustive search is conducted to identify an optimal CNN architecture conducive to classifying the generated pose data. The aim is to ensure a fair evaluation by selecting a CNN layer and hyperparameter configuration that does not favour a particular augmentation over another. This approach allows for a neutral CNN architecture that is suitable for comparing the classification performance of models trained using the augmented data described in the previous subsection. The search space consists of 32 different architectures derived from all possible combinations of standard hyperparameter specifications listed in Table 5-1. These composite architectures are trained on two datasets that differ in the amount of encoded information. The first dataset is the baseline data depicted in Figure 5-4 (b), representing the minimum heuristic information. The second dataset is the most informative pose representation depicted in Figure 5-6 (c), featuring colour-coded, crosshair key points, blending, and confidence-based tinting. By deducing a beneficial CNN architecture based on these two dissimilar pose descriptors, the resulting classifier is optimised for the classification task rather than a specific augmentation.

Table 5-1: Experimental CNN architecture hyperparameter specifications

CNN hyperparameter configuration values	
Network component	Configuration
Number of convolutional layers	[1, 2, 3, 4] layers
Number of trainable convolutional filters	[32, 64, 128, 256] filters
Number of fully connected layers	[1, 2] layers
Number of nodes in hidden fully connected layer	[32, 64, 128, 256] nodes
Number of nodes in output fully connected layer	2 nodes
Batch size	32
Normalisation technique	Batch normalisation
Activation function	Rectified linear unit (*ReLU*)

The CNN architecture search space, as defined in Table 5-1, encompasses a varying range of configurations. Convolutional layers are adjusted from a minimum of one to a maximum of four, and the filters within each layer are collectively adjusted by powers of two, starting at 32 and reaching a maximum of 256. The fully connected layers (or dense layers) positioned at the end of the network consist of one or two hidden layers and are responsible for condensing the feature vector before the final dense layer assigns a classification probability to the input instance. The number of nodes in the final dense layer is fixed at 2 nodes – which is equal to the number of pose classes – whereas the preceding dense layers each have a number of nodes equal to the selected number of filters.

A total of 32 CNN classifiers are trained on two datasets, with an 80% and 20% split, resulting in a training set of 20 000 images and a validation set of 5 000 images. Samples of sitting and standing poses are equally represented in both these subsets. During training, the CNN processes the dataset in batches of 32 samples which allows the network to adjust to the data patterns over time by incrementally updating the weights after each batch. Using a small batch size is advantageous when training with limited data because it facilitates more frequent parameter updates and helps to avoid local optima by promoting randomness and exploration during training (Goodfellow *et al.*, 2016). Batch normalisation is applied after each convolution to help stabilise the distribution of the layer outputs by preventing layer activations from becoming too large or too small, thereby preserving the generalisation capacity of the network. The average classification accuracy of architecturally identical models trained on the two distinct datasets is evaluated at 10 and 15 epochs, with accuracy dropping off after the 15th epoch owing to anticipated overfitting on the training set. Figure 5-7 depicts the reported accuracies between the model pairs at these epochs.

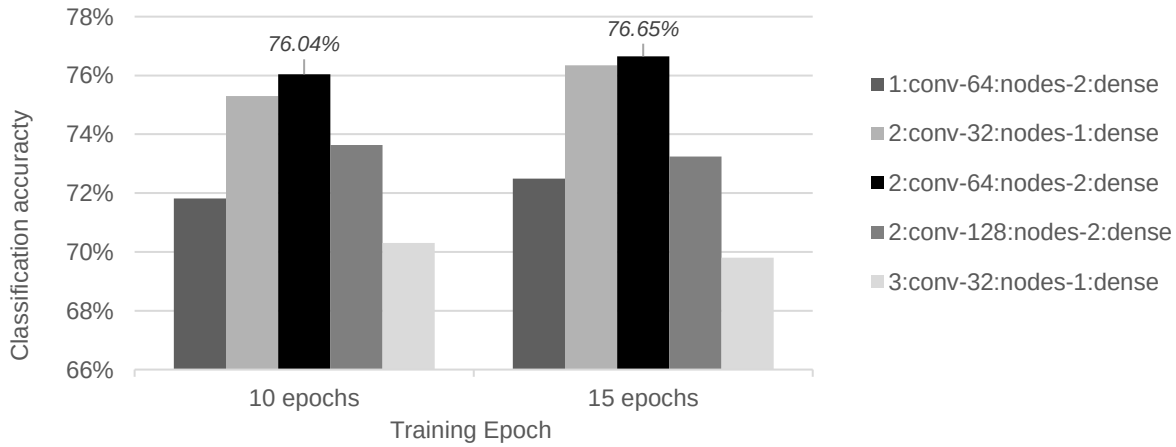


Figure 5-7: Top five mean classification accuracy scores for models trained on the least augmented (baseline) and highly augmented (crosshair + blend + confidence) datasets

Figure 5-7 indicates that the 2:conv-64:nodes-2:dense CNN architecture achieves the highest mean classification accuracy at both the 10th and 15th epoch, indicating its proficiency in the classification task. This CNN architecture consists of two convolutional neural layers (2:conv), 64 trainable convolutional filters (64:nodes), and two additional fully connected layers (2:dense) preceding the final dense layer. This architecture is implemented in Python, and the code can be obtained by referring to Annexure E. A visual representation of the CNN architecture is depicted in Figure 5-8. The initial layer accepts a colour image of size 32 x 32 pixels, and the subsequent convolutional layers employ 64 trainable filters of size 3 x 3 and a stride of 1. Each convolutional layer is paired with a max pooling operation that uses a 2 x 2 window size. Following the convolutional and pooling layers, the two fully connected layers condense the feature vector and assign a classification to the input pose in the third layer. *ReLU* serves as the activation function for all convolutional and fully connected layers in the network.

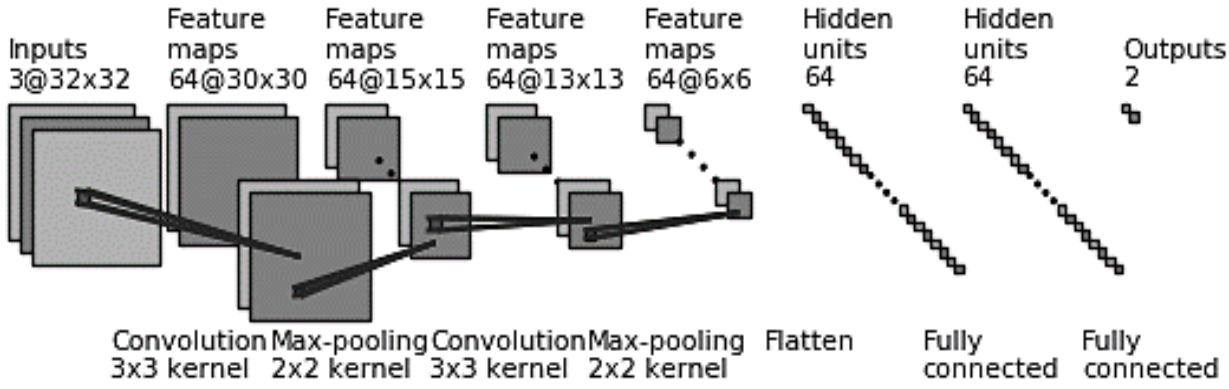


Figure 5-8: The best performing CNN architecture across two of the augmented datasets

The composition of a CNN, in terms of its depth and the organisation of intermediate layers, facilitates discriminative classification capability more effectively than does the selective assignment of filter hyperparameters, namely the size and number of filters. This notion is substantiated by the findings of Jarrett *et al.* (2009) and is again verified in the empirical approach of Pinto *et al.* (2009), where the comparative evaluation of numerous CNN architectures for object detection was performed. In line with these findings, a limited set of configurations are considered in the preliminary experiments to identify an architecture that adequately processes and classifies the pose descriptor representation without needing to explore a comprehensive selection of adaptions to its per-layer hyperparameters. The proposed network architecture serves as a neutral CNN composition for the fair comparison of the augmentation techniques, the results of which are presented in the following section.

5.2.5 Experiment results

Using the optimal CNN configuration identified in Subsection 5.2.4, models are trained on each of the seven generated image datasets described in Subsection 5.2.3. The validation accuracy reported in Figure 5-9 is measured during training, and the individual lines on the graph represent each of these models. The results indicate that encoding additional information in the pose representation by way of colour-focused augmentation techniques evidently benefits classification accuracy. Improvements are most prominent when compared to the baseline model, where an average increase of approximately five percentage points is achieved across all the models in the depicted epoch range. The improvement in classification performance was anticipated and is substantiated by the findings of Zeiler and Fergus (2014) and Buhrmester *et al.* (2019) who showcased colour as a discriminative feature in CNN-based image classification.

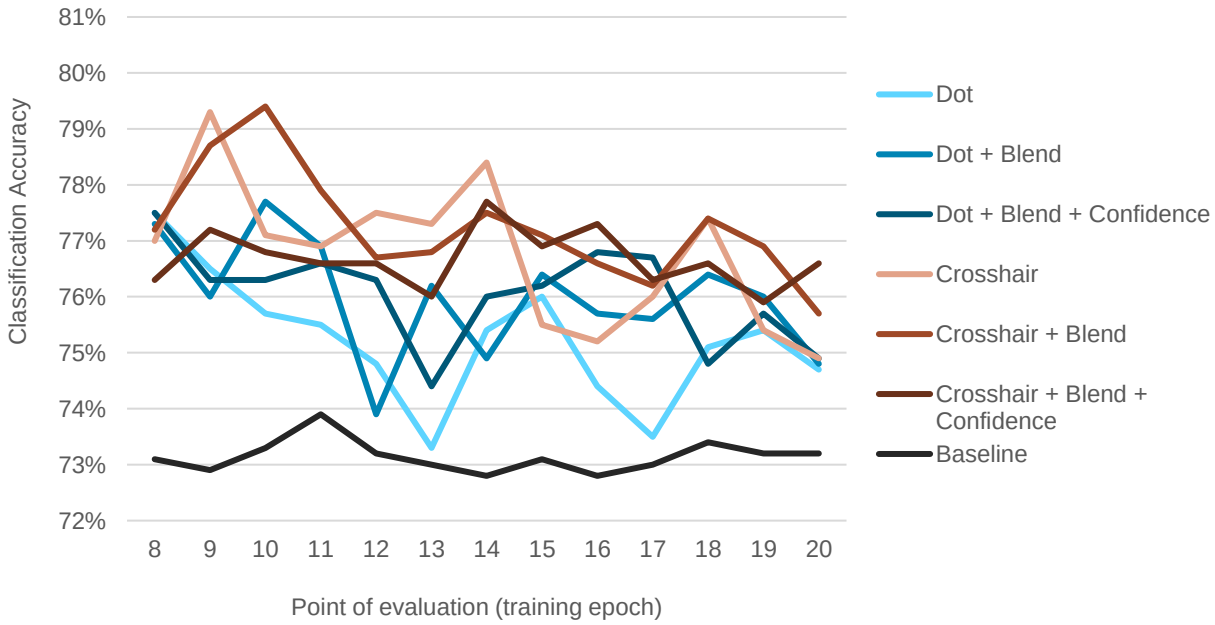


Figure 5-9: Validation accuracy during training across all models of respective datasets

The classification results in Figure 5-9 indicate that assigning colour to the key point formats, namely dot (single pixel) and crosshair, provided a measurable improvement in classification performance over the baseline as indicated by the sets of red and blue graph lines. A crosshair-style marker proves to be more beneficial to performance than a dot key point format as indicated by the set of overarching red graph lines that surpass the blue. The larger surface area covered by each joint marker in the crosshair format likely allows for its retention in the CNN feature space during convolution and pooling, thereby contributing to enhanced performance.

When considering the performance of models that incorporate blending, the results do not indicate a significant change over its less informative dot or crosshair model counterparts. The outcome

is logical since fully occluded body parts are not reported by *OpenPose* and are therefore not represented in the pose descriptor. Consequently, blending is concentrated to partially occluded key points in laterally observed poses where self-occlusion is more common. These poses are naturally infrequent, resulting in few occurrences of overlapping key points that would express the augmentation. The effect of tinting key point colours to represent the localisation confidence of *OpenPose* is similarly indiscernible and does not significantly influence or improve the classification accuracy of either the dot or crosshair models. However, this effect is unsurprising since it aligns with the arguments of Krizhevsky *et al.* (2012), who stated that an object's identity remains invariant to variations in RGB intensities that are typically caused by differences in illumination conditions. According to their work, saliency is related to the hue of a colour rather than its intensity. Therefore, the results indicate that tinting key point colours is only advantageous for visual interpretability but does not contribute to the model's performance.

5.2.6 Conclusion of preliminary experiments

The preliminary experiments establish the feasibility of the proposed data augmentation approach and provide insight into how the different techniques enhance pose classification when using a CNN classifier. The results demonstrate that a skeletal pose descriptor derived from localised body key points can potentially be used to encode valuable information indicative of a pose. A notable improvement in classification performance is attainable through colour-based augmentations, where the most advantageous augmentation entails assigning a fixed hue to the key point of each body part. When compared to the baseline pose descriptor, an average increase in classification accuracy of approximately five percentage points was attained. These improvements align with existing literature on CNN learning regarding the discriminative nature of shape, texture, and colour information in the feature extraction process.

The crosshair-style marker, employed in the pose descriptor, outperformed the dot representation due to its larger surface area since it provides more opportunity for the key points to be retained in the CNN feature space. Opting to blend overlapping key points was intended to mitigate occlusion-related issues. Yet, its effectiveness was limited since *OpenPose* only localises key points of visible and partially occluded body parts, rendering the augmentation sparse and not readily observable in the results. Tinting the colour of a key point to reflect its associated localisation confidence yielded no significant influence on classification performance since the intensity of RGB values do not inform a classifier of the image class. These insights guided the design of a novel augmentation scheme proposed in the next set of experiments discussed in Section 5.3, which serve as the primary investigation in this study.

5.3 Primary experiments

In this section, a second set of experiments is presented that builds on the insights of the preliminary experiments regarding the influence of colour-focused augmentations. These experiments form the core of this study and entail the experimental augmentation of a video dataset designed for fall detection that best demonstrates the feasibility of data augmentation in a real-time, pose-dependant application. The experimental process and results discussed in this section are also published in a peer-reviewed conference paper presented at the 2021 Southern Africa Telecommunication Networks and Applications Conference (Du Toit *et al.*, 2021b) and is provided in Annexure A. The next section provides an overview of the experimental design regarding the inspiration for alternative augmentation schemes and how their design is guided by the findings of the preliminary experiments.

5.3.1 Overview of experimental design

In the previous section, the preliminary experiments demonstrated the potential to improve pose classification when enriching image data using augmentation. The use of key point markers helped to delineate the human body in the image data and subsequently induced discriminative cues that relate to the pose class. The experiment results indicate that such a pose descriptor is best defined using a crosshair key point marker and by assigning colour based on the associated body part or joint. Its benefit is attributed to the fact that the pose descriptor is leveraged during training to supplement the model feature space and thereby enhance its capacity to differentiate between pose classes. The augmentations introduced in the current section build upon the insights gained from the preliminary experiments and are informed by the literary evidence reviewed in previous chapters.

The position of a key point and knowledge of the body part to which it belongs are both highly relevant in deducing a pose. In the preliminary experiments, coloured key points aided the pose classifier by demarcating the important body parts. However, literary evidence suggests that the effectiveness of coloured key points is reduced by translation invariance since positional information is not maintained in the abstracted feature space of a CNN. To counteract these effects, an alternative augmentation technique is proposed in the primary experiments where key point colours are dynamically encoded rather than using a fixed association. Each marker colour is derived from either a radial or a ringed colour wheel which is notionally intended to encode spatial information. The effect is a consequence of the direct correlation between the key point colour and its location within the colour wheel. Essentially, the augmentation represents the kinematic arrangement of the body and its limbs through colour, thereby helping to retain relevant positional cues in the abstracted feature space of the pose classifier model. In the preliminary

experiments, tinting and shading of the key points did not provide any notable improvements and are therefore not included in this augmentation. Similarly, overlapping key points are not blended in this set of experiments since the occluded key points are not reported by *OpenPose*. Furthermore, key points that overlap are assigned the same colour when derived from a colour wheel, making any attempt to blend such key point colours ineffective. The augmentation method is evaluated on a fall dataset to illustrate its applicability to pose-dependent classification tasks and to demonstrate how a simple augmentation technique might be readily applied to improve classification accuracy.

The remainder of this section details the dataset used in the experimentation process, the different augmentation schemes, and the classification results. Subsection 5.3.2 contains a description of the fall dataset regarding its origin, class distribution, and the steps necessary to prepare it for augmentation. Four augmentation schemes are proposed and described in Subsection 5.3.3, each providing varying degrees of supplemental information that encourage class similarity among related pose samples with the aim of promoting accurate pose classification. The adopted CNN architecture that is used to facilitate the experiments is described in Subsection 5.3.4, followed by an analysis of the performance results which are detailed in Subsection 5.3.5 in terms of general and class-specific classification accuracy. These experimental results ultimately provide evidence supporting the effectiveness of the proposed augmentations for a real-world problem. These notions and other general conclusions drawn from both sets of experiments are discussed in Section 5.4.

5.3.2 Video dataset

Subsection 5.3.2.1 provides a description of the video dataset used in the experiments regarding its origin and format. The pre-processing of the data, which is necessary to perform the augmentation, is described in Subsection 5.3.2.2.

5.3.2.1 Data acquisition

The video dataset[2] used in the experiments was developed by Adhikari *et al.* (2017) to facilitate their own research towards vision-based fall detection which focused on differentiating human poses to identify fall occurrences. Adhikari *et al.* recorded the actions of participants as they moved about a room to capture and represent five different pose classes, namely sitting, standing, lying, bending, and crawling. The recordings were manually annotated by frame to yield an image dataset with a distribution that naturally fluctuates for each of the five poses. Variability within the

[2] http://www.falldataset.com/

dataset was minimised by limiting real-world complications, such as severe occlusions, background changes, and the presence of multiple individuals in a scene. Each recording maintains unobstructed viewing angles and features only one participant per scene, thereby ensuring clear observations of the human poses. Out-of-frame observations were retained in the dataset and occur when an individual walks into and out of the camera's field of view. These instances constitute a sixth class in the dataset and are labelled as empty frames since they do not contain a human pose. All footage was captured using a Microsoft Kinect sensor which records both depth footage and traditional RGB footage simultaneously. Samples of each type of footage is depicted in Figure 5-10 and Figure 5-11, respectively. An additional excerpt of the dataset can be viewed in Annexure D.

Figure 5-10: Sample RGB video frames from the video dataset (Adhikari *et al.*, 2017)

Figure 5-11: Sample depth video frames from the video dataset (Adhikari *et al.*, 2017)

All footage was recorded from a high vantage point by positioning the Kinect sensor at a ceiling height of approximately 2.4 m above ground in each room where participants were monitored. A total of five different participants were recorded from eight distinct viewing angles within these rooms. The training set, totalling 16 794 frames, consists of recordings from two participants in two separate rooms. The validation set, comprising 3 299 frames, features a single participant from the training set but the footage was recorded from a different viewing angle of the same room. The remaining three participants make up the test set of 2 543 frames and are recorded in rooms not featured in either the training or validation sets. Footage recorded using such a diverse combination of viewing angles, rooms, and participants ensures that the dataset expresses natural variability and perspectives of the aforementioned human poses.

5.3.2.2 Data preparation

The first step in pose classification involves person detection using background subtraction. This technique entails discounting the changeless elements in a scene to identify a moving object, such as a person. It is most effective in controlled environments where only human movement is expected, avoiding confusion from pets and other displaceable objects. Subtracting the static background scene from the current video frame yields a silhouette of the person under observation; however, noise can be introduced if the background undergoes any environmental changes. Adaptive background subtraction is used to mitigate such noise, which weights incoming observations of a video feed more heavily than preceding observations, thereby counteracting irrelevant scene changes such as shifting shadows and lighting variations (Zivkovic & Van Der Heijden, 2006). Regarding the preliminary experiments, the data preparation entails adaptive background subtraction to extract the silhouette of each participant in the fall detection dataset for both RGB and depth footage. Figure 5-12 and Figure 5-13 depict the original video frames alongside the background subtracted counterparts that contain the isolated human silhouette which is utilised as part of the augmentation process described in the next subsection.

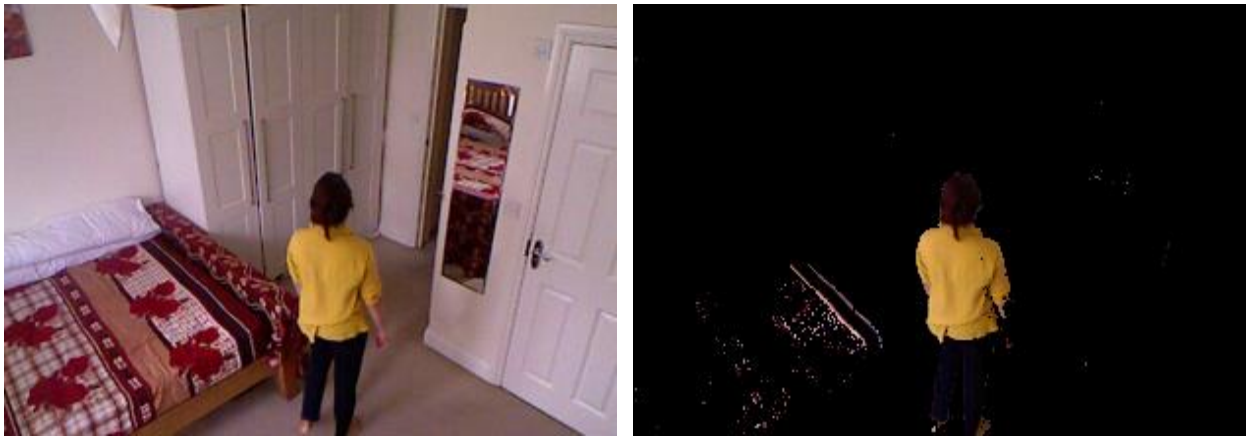


Figure 5-12: RGB video frame (left) and its background-subtracted counterpart (right)

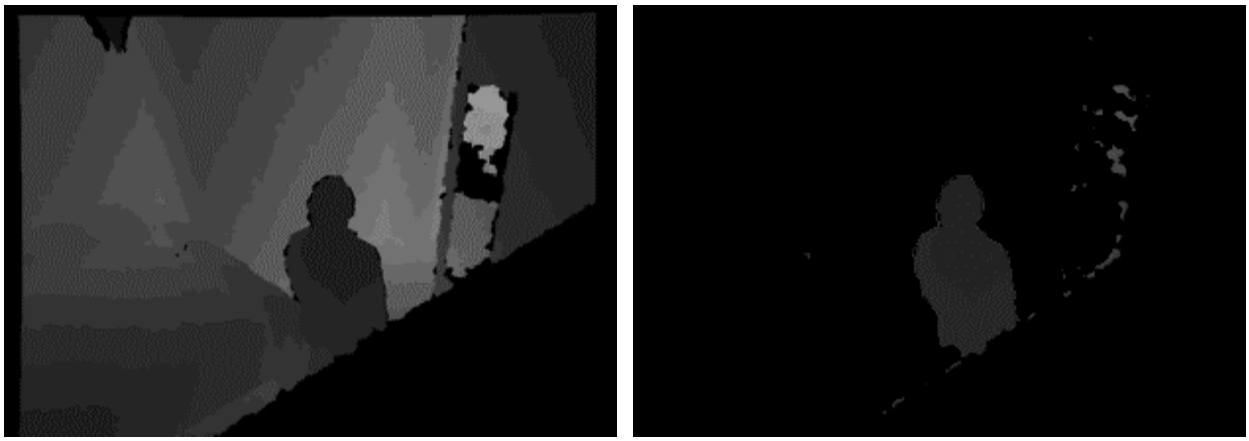


Figure 5-13: Depth video frame (left) and its background-subtracted counterpart (right)

In similar fashion to the preliminary experiments, *OpenPose* is used to derive up to 18 skeletal key points from the original video frames. *OpenPose* estimates these key points by iteratively refining a heat map of probable locations for each body part and joint based on their relationship with other key points. The localised body key points are output to a set of XY-plane coordinates that relate to the image dimensions and are mapped onto the extracted human silhouettes as part of the data augmentation described later in Subsection 5.3.3. To ensure the integrity of the evaluation results, video frame instances where OpenPose fails to localise at least 14 of the available 18 key points are removed from the dataset. Typically, these instances occur when a participant enters or leaves the camera's field of view, resulting in a silhouette only being partially visible. Excluding these instances ensures that the evaluation results only reflect the influence of the proposed data augmentation schemes and are not tainted by the negative influence of occlusion. The adapted video frame counts for the Adhikari *et al.* (2017) dataset are listed in Table 5-2, and the resulting class distribution is listed in Table 5-3.

Table 5-2: Video frame counts of the original and the adapted dataset

Dataset video frame counts			
Dataset	Original frame count	Adapted frame count	Frame count change (%)
Training set	16 794	14 938	-11.05
Validation set	3 299	3 063	-7.15
Test set	2 543	2 344	-7.83

Table 5-3: Class distribution across each of the data subsets

Dataset class distribution						
Pose class	Training set		Validation set		Test set	
	Count	*Distribution (%)*	*Count*	*Distribution (%)*	*Count*	*Distribution (%)*
Bending	760	5.09	148	4.83	83	3.54
Crawling	542	3.63	199	6.50	37	1.58
Empty	419	2.80	192	6.27	205	8.75
Lying	3 512	23.51	880	28.73	577	24.62
Standing	6 595	44.15	982	32.06	657	28.03
Sitting	3 110	20.82	662	21.61	785	33.49
Total	*14 938*	*100.00*	*3 063*	*100.00*	*2 344*	*100.00*

The video dataset was originally recorded in the native Microsoft Kinect resolution of 640 x 480 but was resized and published in a resolution of 320 x 240. In their experiments, Adhikari *et al.* (2017) further resized the dataset to a resolution of 180 x 120 to reduce the computational load and model training time. Additionally, Adhikari *et al.* applied a random crop to all their image samples, ultimately reducing the resolution to only 156 x 108. This crop is replicated as part of this study but is instead informed by the pose estimated key points, ensuring that pixels are only

removed from the frame edges where no human silhouette is present. By adopting a similar data format to Adhikari *et al.*, existing expertise of their research can be capitalised upon, and their favourable CNN architecture can be replicated to train pose classifier models in these experiments. The following subsection details each novel data augmentation scheme, which involves projecting body key points onto a colour wheel to dynamically assign their colours before mapping them onto the video frames.

5.3.3 Dataset augmentation

The approach presented in the present section builds on the findings of the preliminary experiments presented in Section 5.2 by using colour-based augmentations that embed positional information. Four colour wheel arrangements are proposed in Figure 5-14 that each lend themselves to different poses based either on body orientation or silhouette deformation. When projected onto a human pose, each body key point is assigned a colour related to its location within the colour wheel. Similar poses are expected to exhibit the same spectrum of colours in their corresponding key points, thereby introducing an added dimension of patterned body behaviour that a machine-learned model can learn during training. The effectiveness of each augmentation scheme in enhancing pose classification is comparatively evaluated by training separate models on four generated datasets, each containing augmentations based on one of the colour wheels. Additionally, a fifth non-augmented dataset containing the raw pre-processed image samples is also employed for training a model that serves as a comparable baseline against which the improvements can be measured.

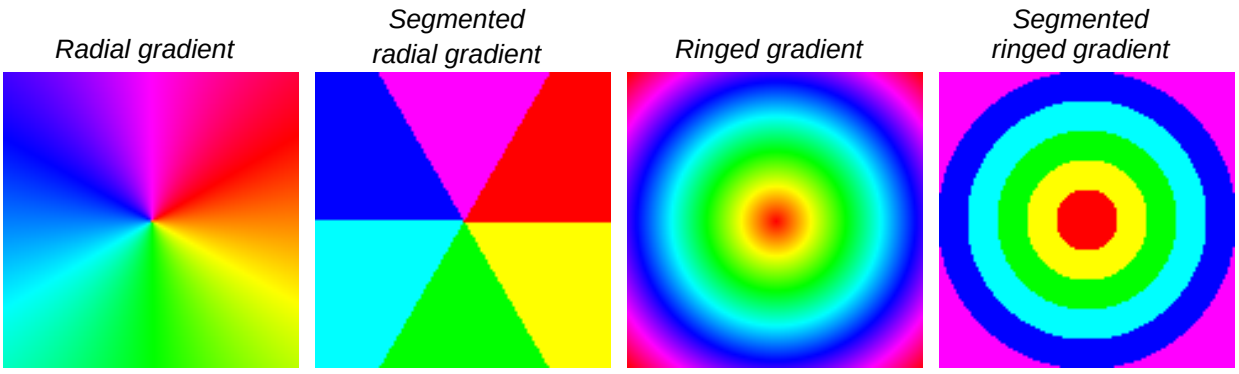


Figure 5-14: Four colour wheel arrangements for key point colour assignment as part of the experimental data augmentation approach

The pre-processed video dataset is augmented by projecting the centre of the colour wheel onto the body torso and assigning colours to key points based on their location within the wheel. Two example augmentations are depicted in Figure 5-15 and Figure 5-16, where the right-most images illustrate the positional key point colour assignment alongside the augmented video frame. The

information encoded by the colour wheels is determined by the available spectrum of colours and their structural arrangement. Gradient colour wheels, namely Figure 5-14 (a) and (c), provide greater granularity through 360 distinct hues, which allows for slight pose variations to be captured and emphasised by multiple colours. Segmented colour wheels, namely Figure 5-14 (b) and (d), instead account for only six distinct colours – which helps to disregard any insignificant pose variations, thereby encouraging CNN generalisation for poses of the same class. The two types of granularities are illustrated and contrasted in the figures below for the same standing pose. Figure 5-15 demonstrates the use of less granularity in the skeletal mapping to achieve a relatively consistent key point colour distribution among the same sets of joints for identical poses. In contrast, Figure 5-16 depicts how varying gradients of similar colours will instead be expressed in the same set of key points across identical poses when encoded with a gradient colour wheel.

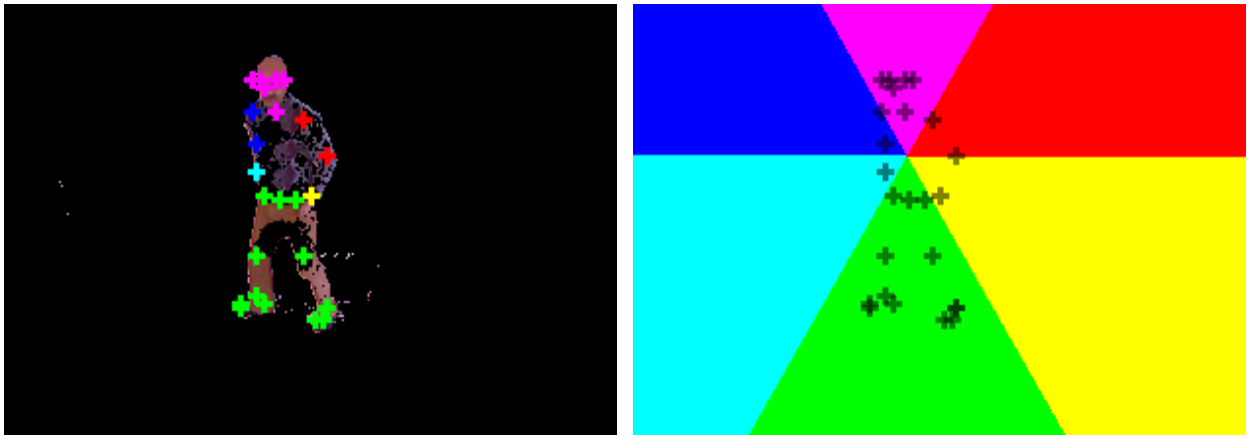


Figure 5-15: Augmented pose sample (left) populated with key point colours based on their position within a segmented colour wheel (right)

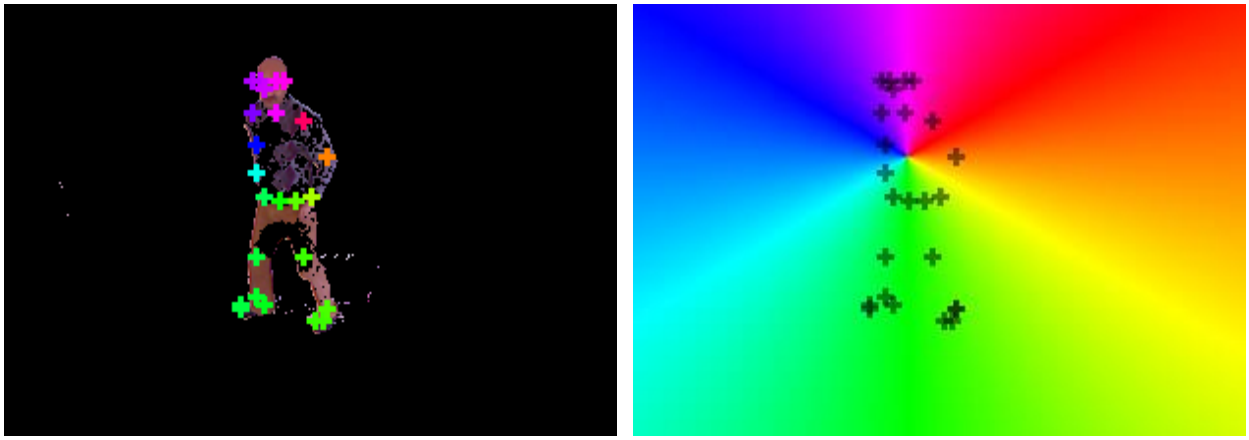


Figure 5-16: Augmented pose sample (left) populated with key point colours based on their position within a gradient colour wheel (right)

Besides the granularity offered by the range of colours, the ringed and radial structural arrangements replicate spatial cues, highlighting postural distinctions often overlooked in a CNN's abstracted feature space due to its properties of invariance. The radial colour wheels depicted in

Figure 5-14 (a) and (b) help to encode the orientation of a pose through key point colours, whereas the ringed colour wheels in Figure 5-14 (c) and (d) help encode silhouette deformations. These arrangements allow for identical poses to be distinguished based on either body orientation or deformation. For example, a crawling pose and sitting (on a chair) share the same silhouette but are opposed in their orientation. This postural difference is captured in Figure 5-17 which shows how orientation is encoded using radial-based key point colours. The crawling pose aligns horizontally with the colour wheel thereby saturating key points with mostly cyan and yellows. In contrast, Figure 5-18 key points are saturated with magenta and green to signify a vertical orientation when encoding a sitting pose.

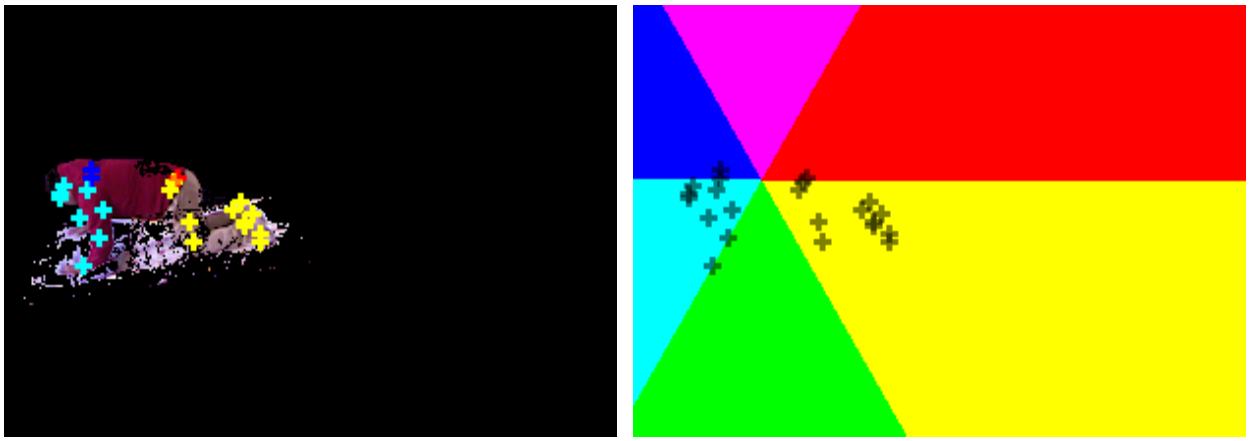


Figure 5-17: Data augmented crawling pose sample (left) populated with key point colours from a radial colour wheel (right)

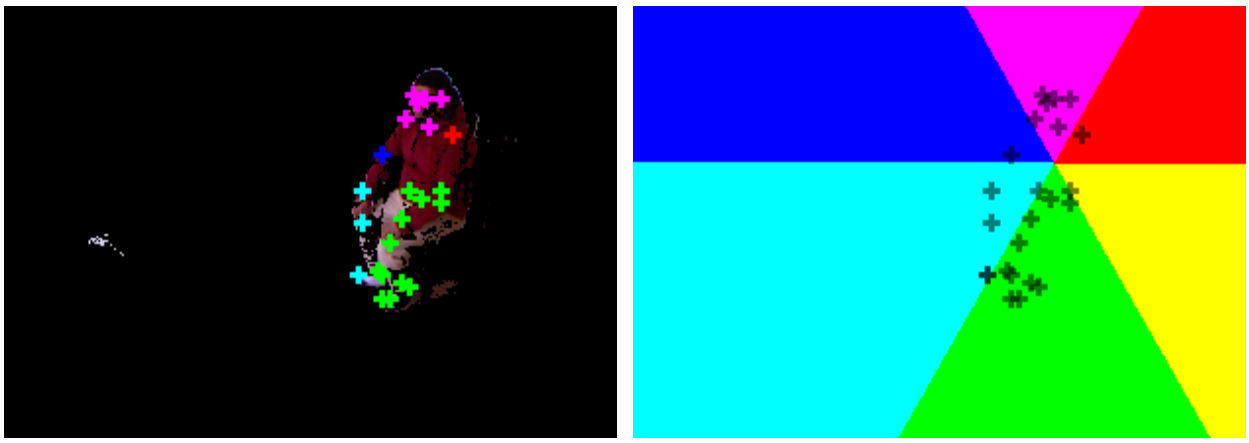


Figure 5-18: Data augmented sitting pose sample (left) populated with key point colours from a radial colour wheel (right)

The ringed colour wheels in Figure 5-14 (c) and (d) capture information that allows for poses to be distinguished based on body deformation. The surface area that a pose occupies within the camera frame is reflected in the variety of colours that its key points assume during augmentation. As demonstrated in Figure 5-19 on the next page, a compact pose will tend to have its key point colours limited to the inner rings of the colour wheel, whereas an extended pose is characterized

by colours from the outer rings, as in Figure 5-20. However, unlike the radial colour wheel, the ringed representation susceptible to perspective distortions since it does not account for how the distance between the subject and the camera can impact the size of the silhouette. To minimise such distortions, footage can be recorded in a controlled environment which helps to limit any variance in perspective and distance, as is the case for the Adhikari *et al.* (2017) dataset where similar conditions are maintained for all the recorded footage.

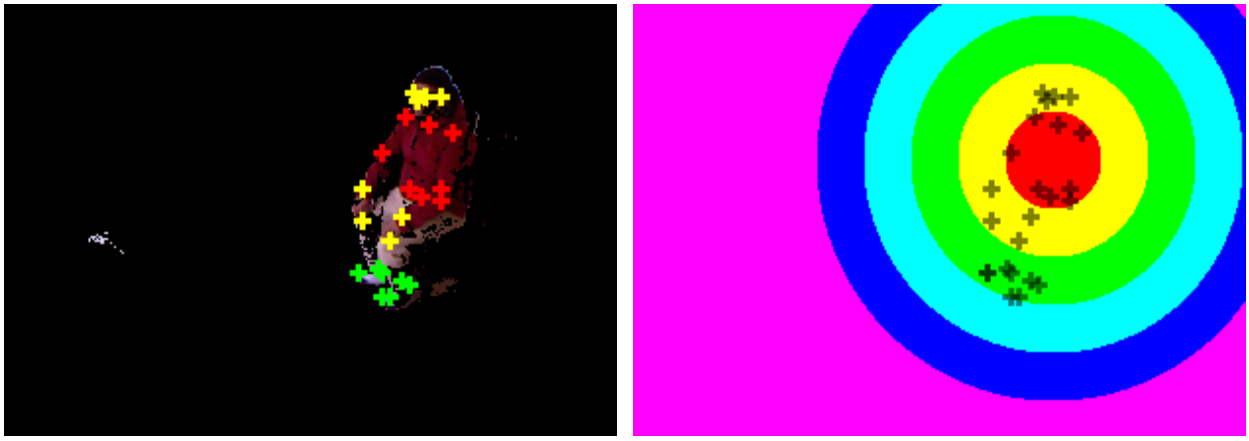


Figure 5-19: Data augmented sitting pose sample (left) populated with key point colours from a ringed colour wheel (right)

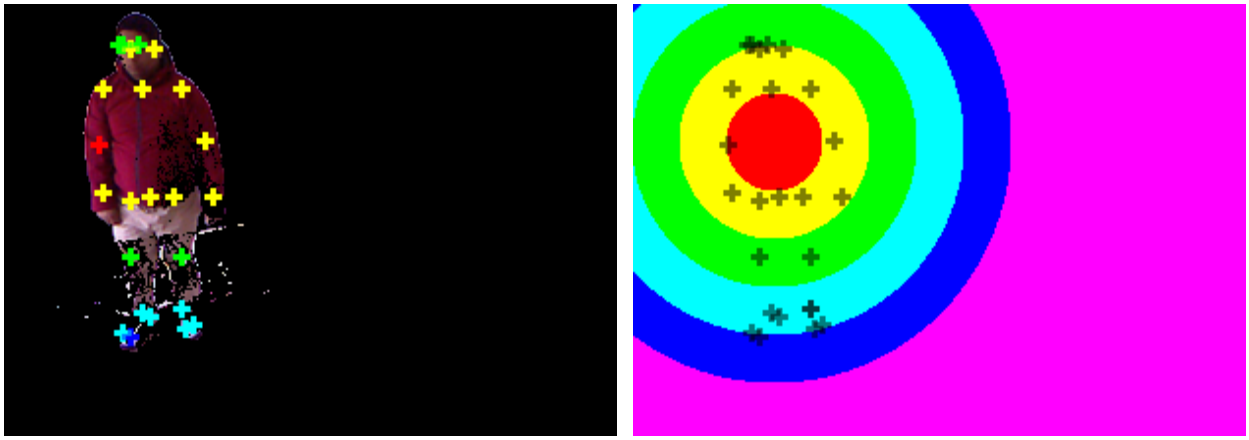


Figure 5-20: Data augmented standing pose sample (left) populated with key point colours from a ringed colour wheel (right)

Depth imaging is included in the dataset as it complements RGB video data and is less affected by information noise. Natural variations in a scene caused by shadows and reflections can influence the RGB foreground during background subtraction, thereby distorting the extracted human silhouette. Depth sensors are resilient to such influences as they solely capture object distances and are therefore unaffected by visual distortions. The depth data mitigates imprecise background subtraction by offering a secondary extracted foreground that shares the same pixel region of the underlying human silhouette.

However, the classifier can only process a single image as input to the CNN. To harness the depth information effectively, augmented pose samples are modified to include a fourth colour channel alongside the existing red, green, and blue channels. Figure 5-21 illustrates how depth images are appended into the alpha channel, which is typically reserved for transparency information in portable network graphic (PNG) images. Similar to the RGB image, the depth data is pre-processed with background subtraction and augmented with corresponding white key points. This format reinforces the depth image's role in delineating the area of interest by expressing same pixel regions across all four image channels.

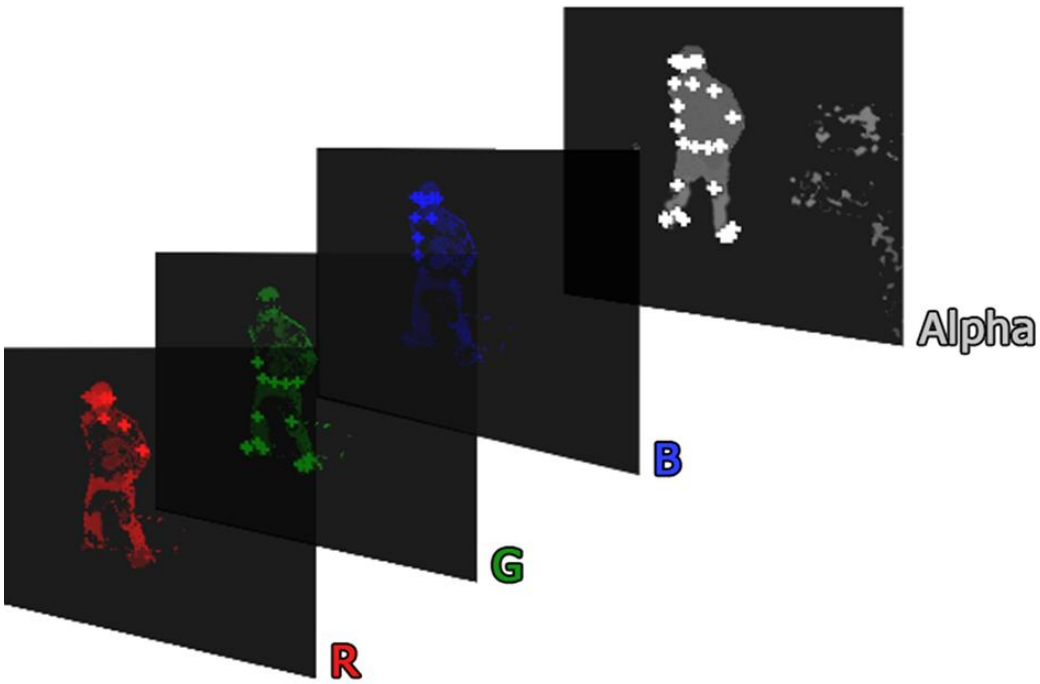


Figure 5-21: The silhouette-extracted and key point-mapped depth image is appended into the alpha channel of the augmented pose samples

In the end, five sets of pose images are generated that make up the experimental dataset, a subset of which can be viewed by referring to Annexure D. The baseline dataset contains non-augmented pose samples that are resized and cropped to the area of interest with a resolution of 156 x 108. The depth images are similarly pre-processed and appended to the non-augmented RGB images without any key point mappings. This format ensures unbiased evaluation by restricting the technical differences between the datasets to only that of the applied augmentation. The other four image datasets are derived from the baseline dataset but are each augmented with projected key points using one of the four proposed colour wheels. The CNN architecture and training process are described in the next subsection.

5.3.4 Convolutional neural network architecture

Classification models are trained using a CNN architecture that is identical to that of Adhikari *et al.* (2017) which was used to perform pose recognition on the same fall dataset. The CNN architecture is implemented using Python, and the code can be obtained by referring to Annexure E. The network was originally inspired by the VGGNet (Simonyan & Zisserman, 2015) which excelled in the 2014 ImageNet Challenge competition, securing second place in the classification task and outperforming other submissions in the localisation task. The distinguishing features of the VGGNet architecture include filters of size 3 x 3, which is the minimum size necessary to capture vertical and horizontal movement, and its deep network of 16 (VGG16) and 19 (VGG19) layers. Figure 5-22 illustrates the VGGNet-inspired architecture, with an adapted input layer that accommodates the dimensions of the augmented datasets, specifically 156 x 108 x 4.

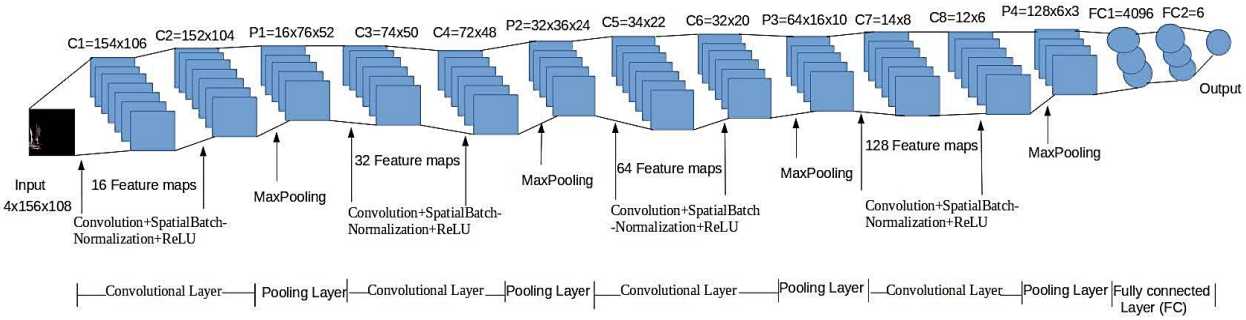


Figure 5-22: Visual representation of the CNN architecture for experimental pose classification (Adhikari *et al.*, 2017)

The initial convolutional layers (C1 and C2) use 16 filters of size 3 x 3 and a stride of 1. These layers employ the *ReLU* activation function to introduce non-linearity and promote sparse activations. Encouraging fewer and selective activations reduces the potential of overfitting and sustains the model's ability to generalise well. Batch normalisation is incorporated to reduce the effects of covariate shift, which manifests when the distribution of variables in the training data differs from the unseen data (Ioffe & Szegedy, 2015). Succeeding every second iteration of convolution and activation is a max-pooling layer (labelled P1 to P4) that aids in sampling the most salient features from the data. This layer order is repeated four times, with the number of convolutional filters doubling from 16 to 32 (C3 and C4), 64 (C5 and C6), and 128 (C7 and C8), respectively. Incrementing the filters in this manner enables the network to learn more detailed and class-specific features in its deeper layers. The final max-pooling operation combines all extracted features before passing through the fully connected layers (FC1 and FC2). FC1 has 4 096 nodes, while FC2 has 6 nodes representing the six possible pose classes. The output layer uses a *Softmax* activation function to yield a probability distribution over the six classes.

5.3.5 Experiment results

This subsection contains the performance results of the colour-based augmentation schemes. The scores are compared to the baseline measure to quantify the improvements in classification accuracy that result from incorporating key points into the fall data.

5.3.5.1 Evaluation methodology

Fall detection, when considered in the context of supervised learning, is mostly observed as a binary classification problem where instances are differentiated as fall or non-fall classes. The efficacy of such classifiers in this area of research is primarily measured in terms of sensitivity (also referred to as *recall* or *true positive rate*) and specificity (also referred to as *true negative rate*). These metrics are impartial to imbalanced class distributions and are appropriate for evaluating performance on a fall dataset that would naturally contain more non-fall class instances than fall instances. Sensitivity is a measure of how accurate a classifier is at predicting a fall instance, whereas specificity measures the performance of predicting non-falls. For the sake of comparison and to account for the impact of class imbalance on evaluation metrics, these measures are reported in addition to accuracy. All three evaluation metrics were defined in Chapter 2 as part of the discussion on relevant evaluation metrics for fall detection. However, instead of fall detection, the primary experiments presented in this section involve a multi-class classification task to assess how the proposed augmentation schemes enhance the classifier's ability to distinguish between different poses. The results of these experiments can be used to ultimately improve other pose-related tasks such as fall detection.

The video dataset is inherently imbalanced due to the infrequency of transitional poses such as bending and crawling, as opposed to the more frequent sitting, standing, and lying poses. When trained on an imbalanced dataset, most classifiers tend to fit to the more prevalent classes, resulting in a biased model. This assumes that the unseen data on which a classifier will be tasked to make predictions are drawn from the same distribution as the training data, which is not always the case and is often preferred (Wasikowski & Chen, 2009). Modern classifiers try to optimise a specific loss function on the given training data to maximise its prediction accuracy, reinforcing a reflection of the training data distribution. Based on this understanding, an imbalanced dataset will almost always produce a weak classifier, especially for tasks that seek to identify a minority class, as in the case for cancer detection and fall detection. Resampling of data is one of the few regularly practised techniques that combat the class imbalance problem. Cross validation, a resampling technique, is therefore applied to all of the experiments reported in this section. The technique helps to maintain a true reflection of the underlying problem without needing to alter the distribution of data classes to compensate for low representation.

The pose classifier models are trained and evaluated using a stratified ten-fold cross-validation strategy. This strategy yields scores with low bias, thereby not overestimating the model performance, and with modest variance, which reflects as relatively consistent scores regardless of the data used to fit the model (Brownlee, 2020). In ten-fold cross-validation, the data is divided into ten subsets, and the model is trained on nine subsets while validating against the remaining subset. This process is repeated ten times to ensure each subset serves as a validation set once and contributes to the model training a total of nine times. Stratification ensures a balanced representation of class samples in each subset, mitigating the effects of a skewed class distribution. A stratified training set also prevents the model from developing a bias towards the majority classes during training.

5.3.5.2 Evaluation results

The initial step in training the pose classifier involves fitting a CNN model to the pose dataset by minimising the categorical cross-entropy validation loss function. Figure 5-23 illustrates the mean validation loss performance on a stratified ten-fold cross-validation training process spanning 20 epochs for both the baseline model and the four augmentation strategies. Each fold (data subset) maintains approximately the same number of data samples and an equally representative distribution of class samples, adhering to an 80% training, 10% validation, and a 10% held-out split for later testing. Consequently, ten models are trained at each epoch and the mean validation score that is reported in Figure 5-23. These results reveal a consistent reduction in validation loss at each epoch as the model performance is iteratively refined against the validation set during training. Notably, convergence is achieved by the 17th epoch, marked by a score of 0.154. The next subsection contains the evaluation results of the models produced at the 17th epoch.

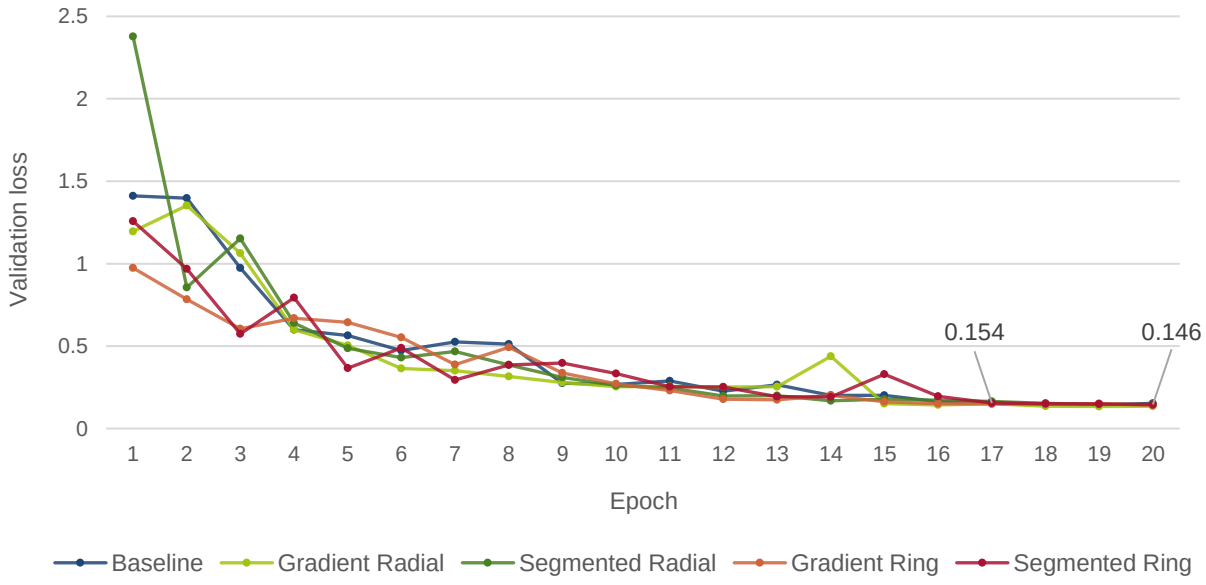


Figure 5-23: Mean ten-fold cross-validation loss of each model trained on respective datasets

The mean classification accuracy, as measured on the held-out test set of 2 344 image samples, is presented in Figure 5-24. Concerning the baseline, the ten-fold models produced accuracy scores ranging from 66% to 76%, which average to the reported score of 70.48%. Each model was individually evaluated on the same held-out test set of non-augmented pose data. In like manner, the augmentation-based models were trained and evaluated on their respective datasets and their averaged scores are also reported in Figure 5-24. The segmented radial augmentation, designed to capture changes in orientation, achieved the most substantial improvement (81.61%), with an increase of 11 percentage points above the baseline (70.48%). Because each augmentation favours the recognition of certain poses over others, the improvement in accuracy is likely concentrated in specific classes. The underlying dataset mostly comprises poses distinguished by orientation (e.g., lying and standing) and less of those differentiated by silhouette deformation (e.g., sitting and standing). Given the fact that standing, sitting, and lying poses make up the majority of the data, it is logical for the radial augmentation model to outperform the others.

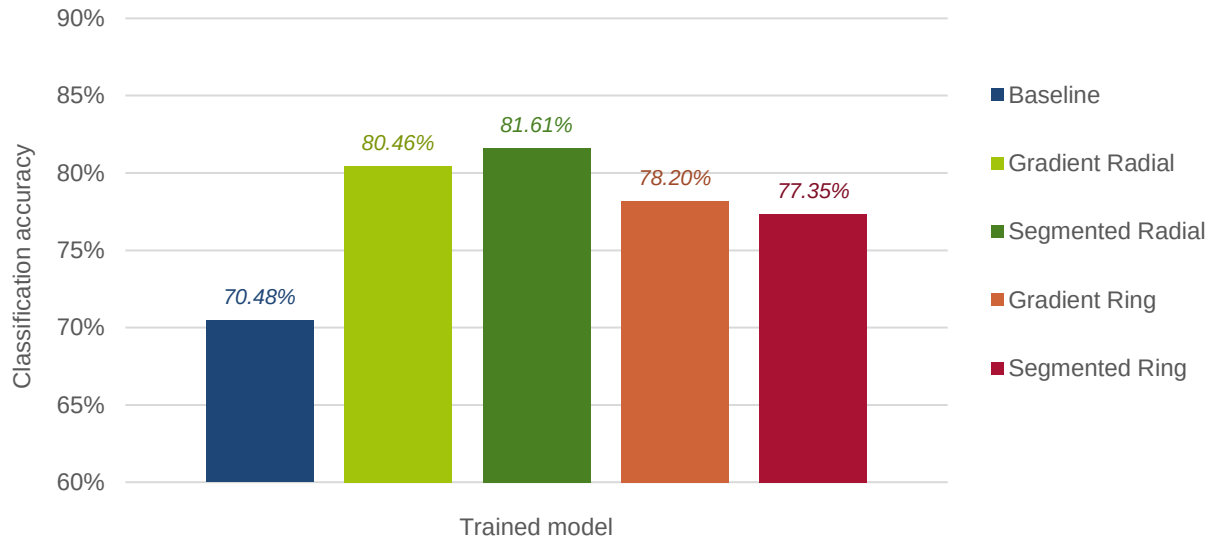


Figure 5-24: Mean accuracy scores for all models trained on respective datasets as measured against the test set

Classification improvements toward specific poses can be observed in the mean recall scores of each data-specific model reported in Figure 5-25 on the next page. Recall (or sensitivity), as defined in Chapter 2, best illustrates the increased sensitivity toward particular poses that the different augmentation schemes afford the CNN model. Figure 5-25 shows that bending and crawling poses benefit the most from the applied augmentation schemes, but they happen to be the least represented classes in the dataset. When compared against the baseline scores, which show only marginal support for the bending (3%) and crawling (7%) poses, the ringed colour wheel provides the highest improvement to recall with respective scores of 54% and 65%. This can be attributed given its capacity for encoding silhouette and size deformations, which is a distinguishing factor between these two poses.

Although the bending and crawling poses share a compact silhouette with the sitting pose, they are not as well represented in the dataset. Therefore, the model learns to recognise sitting poses from the numerous training examples; however, the limited samples of bending and crawling poses likely resulted in the model learning to rely on the unique colour set reflected in their key point augmentation to identify them. This is implied by the improvement in recall over the baseline. Key point colours of compact poses (i.e., crawling, bending, sitting) are uniquely concentrated in the inner rings of the ringed colour wheel, thereby contributing to its improved identification. By contrast, the radial augmentation schemes are less effective in enhancing classification sensitivity for compact poses. This suggests that the classifier recognises deformation in the body's silhouette, rather than its orientation, as a distinguishing attribute for the bending and crawling classes. Improvements in recall of standing, sitting, and lying classes are less pronounced because the classifier is likely already well-attuned to recognising these poses from the abundance of training instances. Consequently, any minor improvement in recall can be attributed to fewer poses being confused or falsely classed as one of the more prevalent classes.

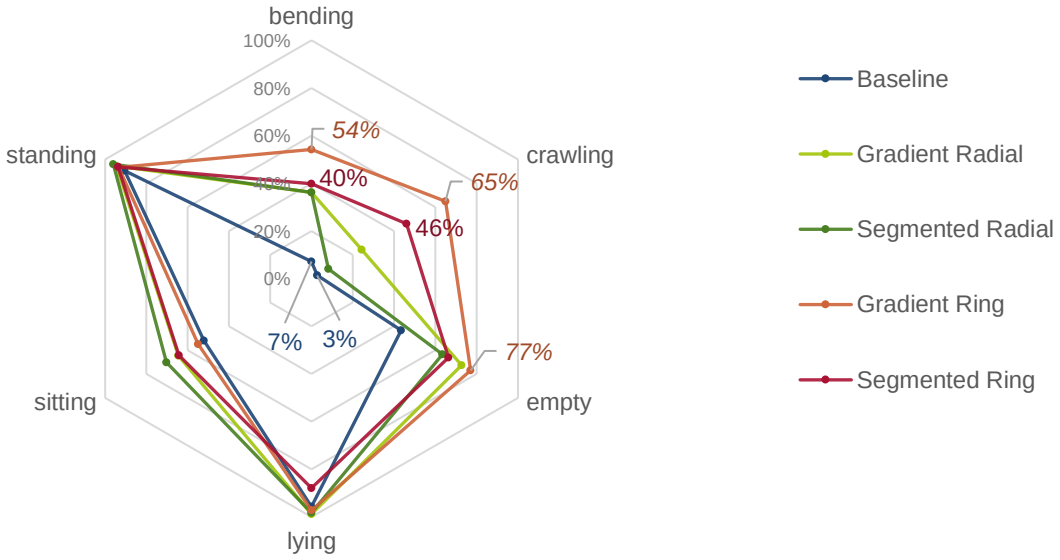


Figure 5-25: Mean recall performance for each pose class across the augmentation-specific models as measured against the test set

Recall is a measure of how well a classifier model can detect positive instances (true positives), which refers to how many of the input poses for a given class are correctly predicted as the expected pose. For instance, when considered in the context of Figure 5-25, the reported recall for bending poses indicates that 54% of the 83 test instances were correctly predicted as bending classes. Specificity complements recall, which indicates how well a classifier model can detect negative instances (true negatives), indicating the precision with which it avoids erroneously predicting non-class-related input poses as belonging to the given class.

When considered in the context of fall detection, recall takes precedence because the model's accuracy in correctly predicting a lying pose is of greater importance than its ability to identify instances that are not lying poses. The respective augmentation schemes are intentionally designed to improve the recall of specific pose classes, thereby enhancing the effectiveness of the classification task. Sensitivity (recall) and specificity are inversely related; it is therefore not possible to optimise both measures simultaneously (Kumar, 2022). In Figure 5-26, the mean specificity scores of the models are reported, indicating that the models consistently and reliably recognise when an input instance does not belong to a particular class.

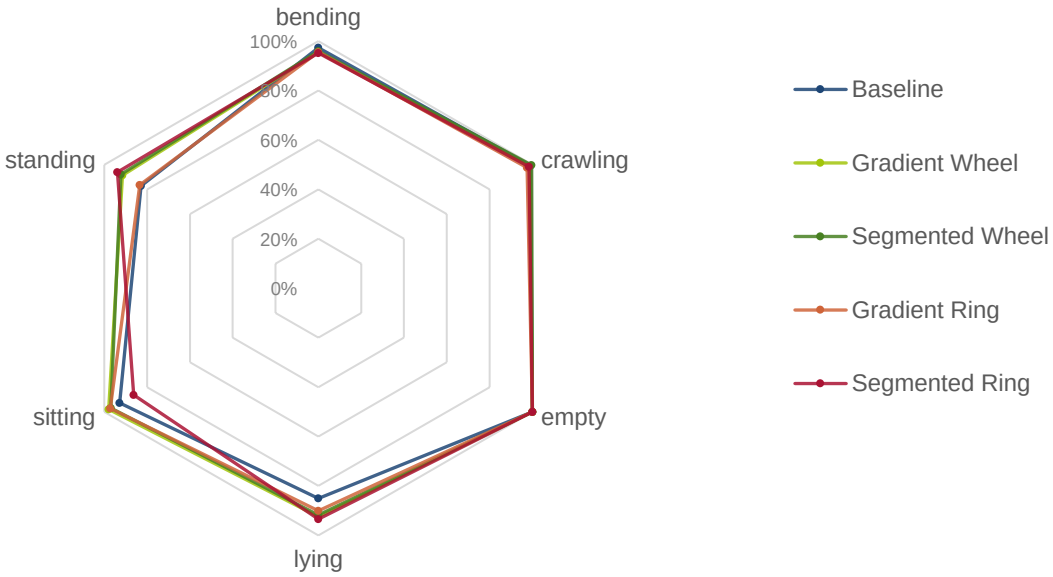


Figure 5-26: Mean specificity performance for each pose class across the augmentation-specific models as measured against the test set

Improved recall is also evident in the reduced confusion among class samples, as depicted in the confusion matrices of Figure 5-27 on the next page. In the case of bending poses, the baseline model correctly predicted only six out of 83 test samples, with most samples being misclassified as lying, sitting, or standing poses. However, with the gradient ring model, the count of correctly predicted bending instances increased to 45, resulting in a 54% recall. Similarly, the baseline model correctly predicted only one out of 37 crawling test samples, but the count improved to 24 with the augmentation. Notably, the gradient ring augmentation introduced increased confusion for sitting samples, misclassifying them as bending poses. In the baseline model, 13 sitting samples were incorrectly predicted as bending poses, but with the gradient ring augmentation, this number increased to 173. The increased confusion suggests that bending and sitting poses share a similar appearance, resulting in similar key point colours in their augmentation. Since the two poses are best differentiated by orientation, the increased confusion is justified as the ringed augmentation primarily encodes variations in silhouette deformation rather than orientation.

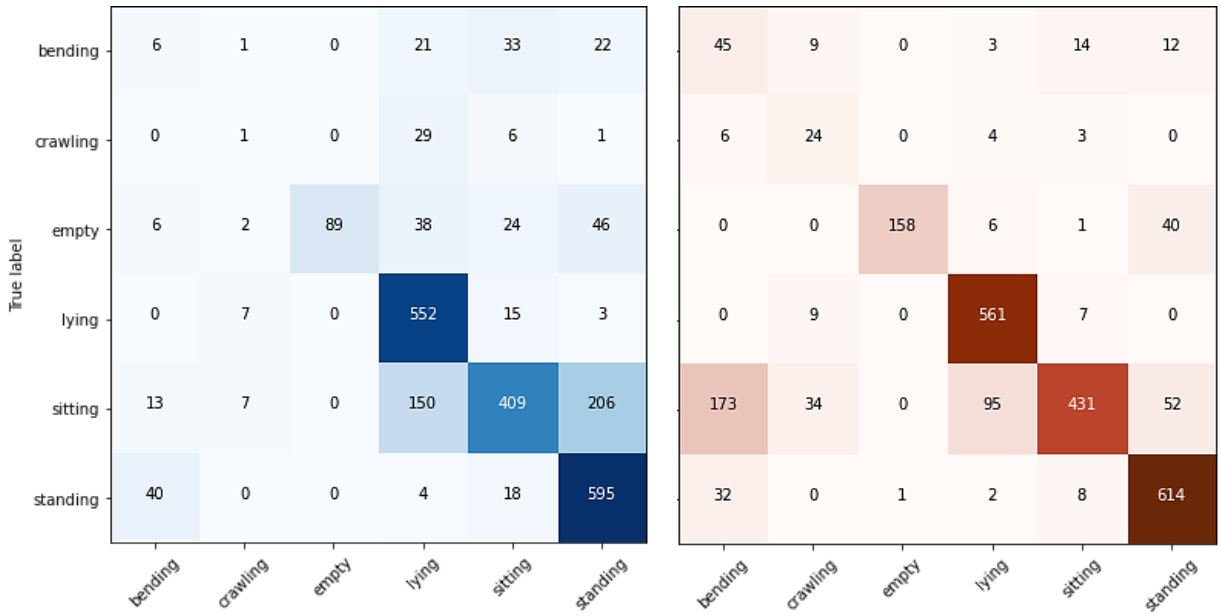


Figure 5-27: Confusion matrix for baseline (left) and gradient ring (right) classification results

Additional insight can be obtained from the true positive count presented in Table 5-4 which further illustrates the class-specific improvements in accuracy. The results confirm the anticipated enhancement in classification by emphasising characteristic pose cues through the various augmentation schemes. Initially, the baseline model only correctly classified one instance of a crawling pose and six of the bending pose from among the 83 and 37 total test samples. However, with the gradient ring representation, the true positive counts for these poses increased to 45 and 24, indicating that these poses are best distinguished based on silhouette size. Similarly, the lying and standing poses benefitted most from the radial augmentation, which encodes differences in orientation. Interestingly, sitting poses were best distinguished also using the segmented radial augmentation, which suggests that the characteristic attribute that distinguishes it from other classes is related to its orientation rather than the expected silhouette deformation.

Table 5-4: Mean true positive counts of each class in the test set

Mean true positive class counts on test set					
Class	Baseline	Gradient Radial	Segmented Radial	Gradient Ring	Segmented Ring
Bending (83)	6	30	30	**45**	33
Crawling (37)	1	9	3	**24**	17
Empty (205)	89	149	130	**158**	136
Lying (577)	552	**570**	567	561	507
Standing (657)	595	621	**631**	614	616
Sitting (785)	409	507	**552**	431	504

The segmented radial augmentation proved to be most effective in helping to correctly identify instances of the sitting class, with a true positive count of 552. This result can be attributed to the reduced confusion between bending and sitting poses. The baseline confusion matrix (Figure 5-27) indicates that bending samples were mostly mistaken for sitting poses. Conversely, sitting samples are mostly confused for bending poses as indicated by the gradient ring confusion matrix. This implies that there is a high visual correlation between the two poses. When viewed laterally, the two poses share a similar silhouette, with the only prominent differences being straighter legs and a slightly forward-leaning torso in the bending pose. The high true positive count of 552 indicates that the segmented radial augmentation effectively captured and helped express this postural difference between the poses and enabled the model to learn and distinguish sitting poses more effectively.

5.3.6 Conclusion of primary experiments

The convergence achieved during model training (depicted in Figure 5-23) indicates that the augmented dataset does not introduce ambiguity that hinders the tuning of model parameters to the training set. Comparing the classification accuracy to the baseline measure (depicted in Figure 5-24) reveals that the CNN classifier heuristically learned to leverage the augmentations and improve the performance of the pose classifier by up to 11 percentage points. The results further demonstrate that the superimposed key points help to improve the correlation and enhance the discriminative information between class labels and extracted features. Evidently, augmenting human pose data to induce supplemental patterns is a viable means of improving pose classification.

Improved recall among underrepresented classes (depicted in Figure 5-25) signifies increased sensitivity in recognising the bending and crawling poses, which comprise only 8.72% of the training and 11.33% of the validation set. The applied augmentations discourage the models from encoding a bias toward the majority classes, thereby safeguarding minority classes from being disadvantaged during classification. When considered in the context of fall detection, recall and specificity are most relevant in measuring the reliability of a classifier since its capacity to positively recognise a fall instance has priority over discerning when an instance is not a fall. Falls are inherently scarce in real-world scenarios and would naturally be underrepresented, yet they were well-represented in the Adhikari *et al.* fall dataset. The disparity likely stems from their specific goal of evaluating fall detection which may have influenced the participants' behaviour and the resulting class distribution of the dataset. This is a common fallacy of such datasets since natural falls and frequencies are difficult to reproduce (see Chapter 2). Regardless, the findings and augmentations remain applicable to increasing a model's sensitivity toward any pose class that would be underrepresented in a dataset or in real life.

Consistent specificity scores across all the classes and models (depicted in Figure 5-26) imply that the augmentations do not introduce ambiguity among the classes and their predictions. In addition, because the two metrics are inversely related, the sustained specificity signifies that recall was correctly targeted for optimisation using data augmentation. High recall indicates that positive samples are predicted as positive, and a high specificity indicates that negative samples (*i.e.*, poses of other classes) are predicted as negative (*i.e.,* not belonging to the given class). Essentially, the augmentations do not impede specificity, ensuring that underrepresented class samples are not inadvertently classified as one of the majority classes. The model therefore maintains its capacity to discern between different poses and consequently maintains the model classification accuracy.

The differences in true positive counts per class for each of the models (depicted in Table 5-4) demonstrate how each augmentation supports encoding the intended class-related cues. The greatest increase in positive predictions for lying and sitting classes were obtained when augmented with a radial augmentation scheme, which denotes its capacity to encode cues based on body orientation. Similarly, the greatest increase in correct predictions for bending and crawling poses is a result of the ringed colour wheel-based augmentation, demonstrating its disposition towards encoding cues based on body and silhouette deformation. Empty class samples also benefit most from the ringed colour wheel, which is most likely a consequence of the reduced class confusion among the underrepresented classes.

5.4 General insights and inferences

The findings of both experiments are summarised in this section and the applicability of the augmentations in pose dependent applications is confirmed. The preliminary experiments were first conducted to determine how colour could be best represented in pose image data to capitalise on its discriminative role in CNN learning, as demonstrated by related literary research. Assigning distinct hues to each key point using a crosshair key point marker offered the most improvement of all the augmentation techniques and indicated that a CNN could heuristically learn to leverage the encoded colour information to better distinguish between different human poses. The success of the crosshair marker was attributed to its larger size, resulting in more opportunity for it to be retained in the abstracted feature space of the CNN after the condensing effects of convolution and pooling. The effects of blending overlapping key point colours could not be effectively observed in the classification results since the occluded key points are not reported by *OpenPose*, leading to sparsity of the augmentation in the data. Additionally, tinting key point colours to reflect the *OpenPose* localisation confidence did not yield any noticeable effect on performance accuracy. This observation aligns with the findings of Krizhevsky *et al.* (2012) who argue that an object's identity remains invariant to fluctuations in its RGB intensities.

An adapted pose descriptor was introduced in the primary experiments that built on the findings of the preliminary experiments. Successful augmentation techniques were retained for the new pose descriptor, including the crosshair body marker and the assignment of colour. However, the range and choice of colours were altered to reflect the structure and arrangement of four distinct styles of colour wheels. The method was inspired by the literary evidence on how CNNs are instilled with translation, scale, and distortion invariance owing to their distinct neural layers and associated operations. Literary evidence also indicated that convolution and pooling contribute to the network's properties of invariance since the position of pixels within the original image is not explicitly regarded in its feature space. To retain the positional information, the location of a key point in the pose descriptor was associated with its colour. This representation allowed for better differentiation of underrepresented classes based on the expressive encoding of postural cues related to orientation and silhouette deformation. The results demonstrated that the augmentation method particularly benefited classes with limited representation in the dataset.

The effect of varying granularity, facilitated by the different number of colours in the gradient and segmented augmentation schemes, could not be demonstrated using the fall dataset in the primary experiments. Its influence would likely only be apparent when combating the distortion of different viewing angles that the camera placement would introduce. For example, in cases where there is an inclined viewing angle from a high vantage point (such as an aerial view), the sought cues that express the pose can be diminished when obfuscated by the perspective. As a result, the difference between standing and lying poses would be most apparent in the silhouette size and less pronounced in orientation. Conversely, differences in standing and sitting poses would be less pronounced in their observed orientation and more apparent in their silhouette size. For poses recorded under these conditions, incorporating more granularity into the projected key point colours may likely provide more distinction between perceptually similar poses that the viewing angle negates.

In conclusion, the results of the study positively address the research question and confirm that pose estimation data can be enhanced through data augmentation by engineering features that a CNN can learn heuristically and leverage in its classification process. Pose-dependent implementations stand to benefit from the proposed augmentation schemes when applied to their incoming data. This is particularly relevant in the context of CNN-based fall detection, where falls are naturally underrepresented or scarce in the training data or in their observations. Delineating the kinetic structure of a pose through coloured key points aids in emphasising the subtle commonalities and differences between poses, aspects that might not be well distinguished in the feature space of a model, especially when dealing with subclasses of the same object. In addition, the simplicity of the augmentation makes it useful for real-time applications that warrant timely action and response.

5.5 Conclusion

The experiments in this chapter demonstrate that purposeful data augmentation is a viable means to improve the correlation and enhance the discriminative information between class labels and extracted features. Pose recognition conducted as an image classification task is naturally encumbered by the loss of spatial information, given the nature of 2D imaging. Evidently, the performed experiments show that it is possible to augment the representation of a pose to partially recover lost spatial cues related to the orientation and shape of the human body. When measured against the baseline model, improved accuracy demonstrates that alternative feature-class associations can be learned from the encoded colour-wheel-based key point augmentations. The improvements were most notable in underrepresented classes, where a unique distribution of key point colours in bending and crawling class samples emphasised their characteristic compact silhouette when encoded with the radial augmentation scheme. In a real-world scenario, such as fall detection, instances of falls are scarce occurrences and would naturally emerge as a minority class. Likewise, the improvements in recall for the bending and crawling poses can be extended to lying poses when augmented with the radial augmentation or, at the very least, would help to reduce the number of confusable class samples. As demonstrated in the experiments, applying the proposed augmentations would be beneficial when dealing with limited and imbalanced datasets to produce a reliable pose classifier, ultimately making it possible to utilise a small dataset to its fullest potential.

CHAPTER 6 CONCLUSION

6.1 Introduction

The goal of this dissertation was to enhance pose estimation data with the aim of improving pose classification using a CNN. To this end, a pose descriptor was developed that can be superimposed onto an image or video frame to delineate points of interest on the human body. Exploratory experimentation of the pose descriptor's configuration informed its design and revealed that colour information was significant in supplementing the feature space of a CNN. The effectiveness of the proposed colour-focused augmentation was assessed by comparatively evaluating models trained on various colour-wheel-based augmentation schemes represented through a pose descriptor. The results demonstrated that colour played a crucial role in capturing and expressing patterns in the human body, enabling the CNN to learn and distinguish visually similar poses more effectively.

This chapter serves to reflect on how the research question posed at the start of this study was answered and how both the primary and secondary objectives, initially formulated in Chapter 1, were achieved. The secondary research objectives and the supporting research findings are discussed in Section 6.2, and the contributions of this dissertation to the body of research and knowledge are discussed in Section 6.3. Additionally, the limitations encountered during the study's execution are described in Section 6.4. Finally, possible future research and applications are suggested in Section 6.5 before the chapter is concluded in Section 6.6.

6.2 Evaluation of research goals

The research question posed in Chapter 1, Section 1.2 was *"Can pose estimation data be enhanced through heuristic data augmentation to obtain an improvement in CNN image classification accuracy when performing pose recognition?"*. Therefore, the primary objective of this study entailed improving pose classification by establishing a set of data augmentation techniques that could be used to enhance the pose-estimated mappings of the human body in an image or video. In support of the research objective, seven secondary objectives were formulated. These objectives are listed below, accompanied by a discussion on how each objective was achieved.

1. **Conduct a literature study of domain-related topics, namely fall detection, pose estimation, CNNs, and data augmentation**

A literature study was undertaken on four topics that pertain to the goal of this dissertation. The four research topics and how they contribute to the study are summarised below.

Chapter 2: Fall detection

A literature study of fall detection was presented in Chapter 2 that contextualised the problem domain and motivated the undertaking of this research study. Population aging was highlighted in Section 2.2 as an impending issue for most developed countries which will result in a high demand for geriatric care by the year 2050. In addition, falling was recognised as a cause for concern that requires specialised care and lengthy hospital stays among the elderly. To combat this, technological innovations regarding fall detection were recommended as reliable solutions that can expedite assistance to the fall victim and minimise the associated medical risks. Fall detection was also credited as a means to alleviate the anticipated pressures on healthcare by limiting the severity of injuries that seniors incur when left unassisted for an extended time. Different approaches to fall detection were also examined in Section 2.4 and notable research from three prominent sensor-based approaches were discussed, namely wearable, ambient, and visual sensor technology. The literature discussed in Chapter 2 helped to substantiate the relevance of automated fall detection and to show the breadth of research performed in the field.

Chapter 3: Visual pose detection

In Chapter 3, a literature study of visual pose detection was presented that explored how visual data from the human body is quantified to detect falls. Regardless of the approach, the research studies discussed in Section 3.2 demonstrated the necessity of extracting features from visual data to obtain pose-related cues that convey fall indicators. To facilitate such feature extraction, HPE was presented in Section 3.3 as a way to identify the probable locations of joints and body parts from images and videos. To this end, *OpenPose* was described in Section 3.4 as a reliable HPE solution that localises the pixel coordinates of key body parts within an image. The chapter helped to define how pose-related information is synthesised from visual data and also highlighted the benefit of deep learning techniques, as used in *OpenPose*, for accurate pose detection.

Chapter 4: Convolutional neural networks and data augmentation

NNs were examined as part of the literature study presented in Chapter 4. This chapter described how NNs are able to automatically learn and abstract features from input data through example-based training. NNs were described in Section 4.2 as being especially useful in modelling complex tasks where relevant features are not easily discernible. In particular, CNNs were highlighted in Section 4.3 and 4.4 for their proficiency in processing visual data using convolution and pooling operations. These operations were credited for inducing CNNs with shift, scale, and rotation equivariance which enable reliable object detection. In addition, the training process of a CNN was discussed, revealing that CNN models learn by identifying the recurrent and common patterns shared by images of the same class. As a result, these kinds of patterns are innately

salient and make for distinguishing features which mostly to relate to texture, shape, and colour information, as demonstrated by the literary evidence and discussion in Section 4.5. The section concluded by proposing that feature saliency be stimulated in pose data by using colour-focused data augmentation based on the literary insights presented in Sections 4.3 to 4.5 on CNN functioning and feature saliency.

Chapter 4 also included a literature study of data augmentation techniques and feature engineering methods that enhance model performance. In Section 4.6, augmentation was described as a technique which introduces modified samples into a dataset to simulate factors of variability. This approach promotes invariance in the model, thereby focusing its learning on unchanging and indicative class features. Feature engineering was examined in Subsection 4.6.3, detailing the creation or transformation of features to better capture and express the underlying patterns of a dataset. In the context of image classification, data augmentation was found to be commonplace in machine learning data preparation, but less so for feature engineering due to the complexity of visual data. The chapter concluded with a proposition to investigate how data augmentation can be applied when engineering a descriptive feature set that contributes to improving pose classification performance. Chapter 4 defined the concept of data augmentation and highlighted its potential for improving classification performance when applied as a form of feature engineering.

2. Collect and assemble a human pose dataset that consists of two classes, namely sitting and standing poses

A dataset of pose images was assembled as part of this dissertation and is described in Chapter 5, Subsection 5.2.2. All images were sourced from the public domain using simple image searches across the internet and from websites that freely offer the use of their images under a creative commons licence. A total of 25 000 pose-related images were collected that depict people in standing and sitting positions which are equally represented in the dataset with 12 500 images each. The key points and localisation confidence scores derived using *OpenPose* were compiled into an Excel document which is available through the NWU research repository[3]. Each entry in the file is also labelled with its corresponding pose class. In Section 5.2, the dataset adequately facilitated a binary classification problem as part of the preliminary experiments which demonstrated the influence of different augmentation techniques on pose classification accuracy.

[3] https://doi.org/10.25388/nwu.23290937

3. Identify a pose estimation implementation that can pre-process pose data to derive points of interest on the human body

OpenPose was identified as a suitable pose-estimation solution for the pre-processing of datasets in this research study. The decision was motivated by the favourable methods for human action recognition and pose estimation discussed in Chapter 3. In Section 3.2, machine-learned analyses were found to be unencumbered by the need for manual feature construction, unlike other methods that rely on establishing thresholds from changes in body shape and head motion. Section 3.3 indicated that both discriminative and machine-learning pose-estimation approaches better represent and approximate human poses when estimated directly from visual data. *OpenPose* aligns with being both a discriminative and machine learning-based pose estimator, resulting in an accurate solution that provides a favourable XY-coordinate output format of body key points. It operates as a multi-stage and two-branching CNN that predicts the location of important human body parts, estimating up to 25 key points that are reported as coordinates within the image frame. At the time of writing, *OpenPose* ranks among the the top 10 best HPE solutions measured against the MPII benchmark for articulated human pose estimation (MPII, 2023), further motivating its use in this study. The *OpenPose* estimated key points were successfully employed as part of the augmentation process in the preliminary and primary experiments presented in Sections 5.2 and 5.3.

4. Propose data augmentation techniques that enhance the feature space of a CNN

The literature study presented in Section 4.5 was used to illustrate how shape, texture, and colour information in images influence the feature space of a CNN and its classification capability. In certain cases, object classes exhibited a strong dependency on one or more of these elements, especially when it coincides with an image attribute that distinguishes it from other classes in the dataset. The dependency was first alluded to when the learned filters were projected back to the image space, revealing colour to be a significant attribute in image classification (Zeiler & Fergus, 2014). Later research finally demonstrated this dependency by diminishing the prevalence of the shape, texture, and colour information through augmentation (Buhrmester *et al.*, 2019). This evidence inspired the idea for augmentation to be used to influence the feature–class dependency of a model to achieve improved classification performance. Rather than encouraging invariance through augmentation, it was hypothesised that feature saliency could be generated for pose data by introducing a less prevalent but inherently salient feature such as colour. As a result, the augmentation would theoretically introduce an added dimension of patterned body behaviour that could be retained in the feature space of a CNN and help it more effectively learn the intended pose classes from a pose dataset.

To this end, a pose descriptor was engineered through exploratory experimentation in Section 5.2 which indicated that augmentations related to colour were the most advantageous for improving classification performance. The designed pose descriptor represented the skeletal outline of the human body using key points that stimulate visual patterns for identical poses. Superimposing the key points onto pose images helped to express the kinematic arrangement of the body and expose relevant positional cues to the classifier during training and classification. On examining the blending of overlapping key point colours and the adjustment of their saturation, the outright use of colour was still found to be the most prevalent in aiding pose classification when trained using a CNN. These findings helped to determine that colour-focused augmentations would enhance the feature space of a CNN-based pose classifier.

5. Determine how data augmentation can be applied to pose data for improved pose classification

An initial investigation was conducted in Chapter 5, Section 5.2 regarding the use of data augmentation for improved pose classification. The study evaluated the viability of augmenting pose images with a pose descriptor using different hues, body marker sizes, and saturation intensities. Results demonstrated that the pose descriptor aids pose classification, and certain augmentation methods provided more improvement than others. Assigning a colour to each key point was the most advantageous augmentation for classification performance. In addition, the crosshair-style key point marker was found to be the more favourable representation format of the two since its larger size allowed for the pose descriptor to be retained in the condensed feature space of the CNN. However, blending the colours of overlapping key points to minimise the effects of occlusion was not observable from the classification results, since only visible and partially occluded key points are identified by *OpenPose*. Similarly, tinting key point colours to reflect its localisation confidence yielded no significant influence on classification performance since a CNN disregards fluctuations in the RGB intensity of a hue. These findings were a result of the first attempt to apply data augmentation to pose data to improve pose classification performance.

The primary experiments outlined in Chapter 5, Section 5.3 expanded on the findings of the preliminary experiments. All the successful augmentation techniques from the preliminary experiments, namely the crosshair body marker and the assignment of colour, were retained in the design of a new pose descriptor. Literature concerning CNNs and how visual information is processed informed how colour could be best adopted in the pose descriptor. While the preliminary experiments indicated that coloured key points were a favourable augmentation, the literature suggested its effects were likely diminished due to the property of translation invariance that results in positional information not being retained in the CNN feature space. To address this issue, key point colours were assigned based on ringed and radial colour wheels to directly

correlate the position of a key point with its colour. This approach helped to correlate a highly salient feature, namely colour, with the kinematic arrangement of the body, which is highly relevant in deducing human pose. The augmentation emphasised and exposed underlying and significant pose information that could then be heuristically learned by the CNN during training to aid classification performance. These findings demonstrated how data augmentation could be applied to pose data to improve the accuracy of pose classification.

6. Establish a suitable CNN architecture and hyperparameter configuration that supports pose classification

A CNN architecture was established from a set of exploratory experiments conducted in Chapter 5, Subsection 5.2.4. A non-exhaustive search was performed using various network configurations to determine a suitable and unbiased CNN architecture. By varying the number of convolutional layers, fully connected layers, dense nodes, and trainable filters per layer, 32 potential network architectures were generated. Each of the potential architectures was trained on two sets of divergent data, resulting in 64 models. These models were evaluated based on their mean classification accuracy between each of the two dataset-specific models that share the same architecture. Because the two datasets expressed different amounts of encoded information through their unique pose descriptor, the approach helped to ensure that the network was optimised for the task of pose classification rather than for a specific augmentation. The highest mean classification accuracy was achieved using a CNN architecture with two convolutional and two fully connected layers, both with 64 trainable filters and dense nodes. These findings helped to establish a suitable CNN architecture and hyperparameter configuration that supports pose classification. The architecture also ensured a fair comparison of different augmentation techniques investigated in Section 5.2.

7. Evaluate the effectiveness of the data augmentation techniques when applied to a real-world dataset

The primary experiments presented in Chapter 5, Section 5.3 were performed on a real-world fall detection dataset developed by Adhikari *et al.* (2017) that consists of five pose classes (sitting, standing, lying, bending, and crawling) and an empty class that represents out-of-frame instances. Each video frame in the dataset was pre-processed using background subtraction and minimal cropping to isolate the human silhouette. Using the proposed pose descriptor presented in Subsection 5.3.3, four datasets were generated from the pre-processed data by augmenting the human silhouette with key points and assigning hues derived from different colour wheel structures. Models were then trained on both the pre-processed dataset and the augmented

derivatives, and their classification performance was comparatively evaluated using three performance measures to assess their effectiveness regarding pose classification performance.

The models trained on the augmented data consistently outperformed the baseline model, achieving different levels of accuracy depending on the specific augmentation method employed. These findings have been detailed in the experimental results presented in Subsection 5.3.5. The baseline model achieved a classification accuracy score of 70.48%. In comparison, the model trained on the segmented ring augmentation achieved the least improvement with an accuracy score of 77.35%, and the segmented radial augmentation obtained the greatest improvement with 81.61%. The results indicate that a significant improvement in pose classification accuracy is obtainable when applying the novel augmentations. Moreover, the results indicated that the different augmentations favoured different poses, thereby making the improvements in performance proportional to the underlying class distribution of the dataset. This is ascribed to the to the distinct structural organisation of colour hues within the different colour wheels, each of which is suited for encoding positional cues associated with either body orientation or silhouette deformation. Recall scores for each class revealed that the radial colour wheel augmentations improved the recognition of poses that share similar silhouettes but differ in orientation, while the ringed colour wheel augmentations improved the classification of poses with the same orientation but different silhouette sizes. The most significant improvement was observed in underrepresented pose samples (*i.e.*, bending and crawling), where the augmentation helped to mitigate the misclassification of these poses as one of the majority classes. These results favourably demonstrate the effectiveness of augmenting pose data on a real-world dataset and indicate that reducing misclassification for confusable pose classes is especially useful when dealing with limited and imbalanced data.

6.3 Research contributions

The main contribution of this study is apparent in the research question and primary objective, which sought to investigate the use of data augmentation for improving CNN-based pose classification. This research established a set of novel augmentation schemes that effectively enhances pose-estimated mappings of the human body in an image or video to enable accurate pose recognition. The results demonstrate that a CNN classifier is able to better differentiate between pose classes when learning class-related cues encoded by the augmentation schemes. In addition to this main contribution, this study has also yielded several related findings that further contribute to the body of knowledge. These additional insights are detailed below:

- The literature studies performed as part of this dissertation provided an overview of existing research in fall detection, pose estimation, CNNs, and data augmentation.

Important findings, existing challenges, and best practices were highlighted for each of the literature topics. Additionally, state-of-the-art solutions and advances were identified where applicable, and the strengths and weaknesses of previous studies and proposed solutions were also captured. This body of knowledge functions as a literary source for future reference and potential further research efforts.

- A dataset of 25 000 pose images that consists of people pictured in either sitting or standing poses was manually curated and processed using *OpenPose* to derive body key points. The dataset is available through the NWU research repository[4].
- The viability of data augmentation using pose-estimated body key points to enhance the accuracy of human pose classification was effectively demonstrated in two sets of experiments using both a trial dataset and a real-world dataset.
- A novel data augmentation technique that derives key point colours from different colour wheel structures was proposed and successfully implemented. The technique was shown to support the encoding of cues related to either body orientation or silhouette deformation and help to express the underlying human pose.
- The results of the study demonstrate how the novel augmentation schemes help to enhance the CNN's capacity to differentiate between confusable poses that share a similar appearance, thereby improving the classification accuracy of human poses. This was most beneficial in underrepresented poses where too few training samples were available for the classifier to generalise well for those classes during training.
- The improved pose classification performance demonstrated that a CNN is able to learn the unique colour distribution encoded in augmented samples of the same pose class. This indicated that the colour-focused augmentation schemes enhance the feature space of a CNN with discriminative information related to the classification task since colour acts as a salient feature in CNN-based classification.

The data augmentation techniques proposed in this study have promising applications that extend beyond pose classification. These techniques can be advantageous for other pose-dependent tasks beyond fall detection, where enhanced performance is highly valuable. Moreover, the simplicity of the augmentation method makes it suitable for use in real-time pose evaluation. This means it is practical and feasible for real-world scenarios and applications. Despite the success of this study, there are notable limitations which are discussed in the next section.

[4] https://doi.org/10.25388/nwu.23290937

6.4 Research limitations

The proposed augmentation approach contributed significantly to improving pose classification and demonstrated the potential of enhancing fall detection. Despite the imbalanced class distribution of the fall dataset, the models accurately identified fall poses. It is worth noting that fall events are rare occurrences in everyday life and when realistically represented in a pose dataset would naturally result in data sparsity for that class. The fall dataset used in this study exhibited a natural sparsity of transitional poses (*i.e.*, crawling and bending) but less so for fall instances, likely due to the influence of research bias and participant expectations that influenced its creation. In keeping with realistic time and scope constraints, the proposed models were not designed to address this sparsity. Therefore, this study did not investigate the impact of data sparsity or provide recommendations on how to compensate for any class imbalance when performing pose classification.

Various forms of augmentation for improving pose classification were explored in this study and those that exhibited promise were used to design the eventual augmentation schemes. Colour augmentation emerged as a particularly promising technique, ultimately inspiring the use of structured colour wheels for key point colour assignment. Although the literature suggested that texture and shape information were also salient in CNN classification, the study did not focus on investigating the opportunity to produce augmentations related to these features for improving pose classification. Time and scope limitations prevented further exploration of other potential opportunities for meaningful augmentation. This study has introduced novel and highly effective augmentation schemes capable of enhancing pose estimation in CNN-based pose classification tasks. Moreover, it has illuminated potential opportunities for future research in terms of exploring data augmentation techniques that are applicable to visual data.

6.5 Potential future research

The research conducted in this dissertation focused on visual approaches to pose classification using a CNN. While the proposed approach was successful using a CNN, other NNs and derivatives of the CNN that operate on visual data, such as the long short-term memory NN and the recurrent CNN, may also benefit from similar augmentations when dealing with pose-related classification tasks. For example, deep learning models such as the long short-term memory NN can enable the contextual analysis of human pose by taking multiple video frames into account preceding a pose. Augmentation in this context may be valuable when relevant indicators underlying a specific action are emphasised by the pose descriptor. Different and alternative learning algorithms may provide opportunities for future research in pose classification, especially when applied to tasks beyond fall detection, such as gesture recognition or action recognition.

The pre-processing of the assembled dataset and fall dataset was performed using the *OpenPose* (Cao *et al.*, 2021) implementation to extract the location of body parts and joints. Other pose estimation systems, such as DeepPose (Toshev & Szegedy, 2014) and AlphaPose (Fang *et al.*, 2022), may also offer strong accuracy on different or a more expansive sets of key points. These systems could be used to obtain an alternative skeletal mapping of the human body. The proposed augmentation schemes can then be applied to the output of one such system and is likely an avenue for potential future research.

Although initially applied to the problem of fall detection, the proposed augmentations may be relevant to other pose-related tasks. Fitness tracking, for instance, relies heavily on pose information to monitor body position and proper form during exercise routines. With the aid of augmentation techniques, joint angles and body positions could be analysed more accurately to provide better feedback to the user. Similarly, for the purpose of safety and security, action recognition could benefit from enhanced pose information where suspicious behaviour can be readily detected. Therefore, the proposed augmentations have the potential to be highly beneficial in various pose-related applications beyond fall detection. Exploring these alternative applications is also a promising avenue for future research.

6.6 Conclusion

In this dissertation, the research question of *"Can pose estimation data be enhanced through heuristic data augmentation to obtain an improvement in CNN image classification accuracy when performing pose recognition?"* was investigated. It was shown that colour-based augmentations can be applied to a skeletal mapping of the human body and superimposed onto image and video data to improve the accuracy of a CNN pose classification model. The augmentation technique is readily implementable and can support real-time pose classification in time-sensitive applications. Improved classification results obtained on the fall dataset is evidence of the approach's potential and its suitability for real-world use cases. In summary, the proposed augmentation-based approach emerges as a reliable technique for enhancing overall pose classification accuracy.

In conclusion, this study successfully addressed the research question and achieved the primary objective by demonstrating how data augmentation can be used to improve CNN-based pose classification. Additionally, each of the secondary research objectives formulated in Chapter 1 were successfully achieved and the research methods and findings that support these outcomes were discussed. The limitations and opportunities for future research were also presented, including the potential of texture and shape information in improving pose classification. The study contributes to the body of knowledge regarding CNN-based pose classification and provides a foundation for future research regarding pose classification and visual data augmentation.

BIBLIOGRAPHY

Abbate, S., Avvenuti, M., Corsini, P., Light, J. & Vecchio, A. 2010. Monitoring of human movements for fall detection and activities recognition in elderly care using wwireless sensor network: a survey. In: Tan, Y.K. & Merrett, G., eds. *Wireless Sensor Networks: Application - Centric Design*. London, UK: IntechOpen. pp. 147-166.

Abdel-Hamid, O., Mohamed, A.R., Jiang, H., Deng, L., Penn, G. & Yu, D. 2014. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1533-1545. https://doi.org/10.1109/TASLP.2014.2339736

Addlesee, M.D., Jones, A., Livesey, F. & Samaria, F. 1997. The ORL active floor [sensor system]. *IEEE Personal Communications*, 4(5):35-41. https://doi.org/10.1109/98.626980

Adhikari, K., Bouchachia, H. & Nait-Charif, H. 2017. Activity recognition for indoor fall detection using convolutional neural network. In: Ishikawa, H. & Okada, R., eds. *Conference Proceedings*. 15th IAPR International Conference on Machine Vision Applications (MVA 2017), Nagoya, Japan. Piscataway, NJ: IEEE. pp. 81-84.

Alwan, M., Rajendran, P.J., Kell, S., Mack, D., Dalal, S., Wolfe, M. & Felder, R. 2006. A smart and passive floor-vibration based fall detector for elderly. In: Solaiman, B., ed. *Conference Proceedings*. 2nd International Conference on Information & Communication Technologies: From Theory to Applications (ICTTA 2006), Damascus, Syria. Piscataway, NJ: IEEE. pp. 1003-1007.

Andriluka, M., Roth, S. & Schiele, B. 2009. Pictorial structures revisited: people detection and articulated pose estimation. In: Essa, I., Kang, S.B. & Pollefeys, M., eds. *Conference Proceedings*. 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL. Piscataway, NJ: IEEE. pp. 1014-1021.

Andriluka, M., Pishchulin, L., Gehler, P. & Schiele, B. 2014. 2D human pose estimation: new benchmark and state of the art analysis. In: Basri, R., Fermuller, C., Martinez, A. & Vidal, R., eds. *Conference Proceedings*. 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014), Columbus, OH. Piscataway, NJ: IEEE. pp. 3686-3693.

Arcelus, A., Goubran, R., Sveistrup, H., Bilodeau, M. & Knoefel, F. 2010. Context-aware smart home monitoring through pressure measurement sequences. In: Dajani, H. & Rapuano, S., eds. *Conference Proceedings*. 5th IEEE International Workshop on Medical Measurements and Applications (MeMeA 2010), Ottawa, Canada. Piscataway, NJ: IEEE. pp. 32-37.

Arcelus, A., Holtzman, M., Goubran, R., Sveistrup, H., Guitard, P. & Knoefel, F. 2009. Analysis of commode grab bar usage for the monitoring of older adults in the smart home environment. In: Pan, X., Worrell, G.A., Saranummi, N., McCullough, J. & Tranquillo, R., eds. *Conference Proceedings*. 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS 2009), Minneapolis, MN. Piscataway, NJ: IEEE. pp. 6155-6158.

Arcelus, A., Herry, C.L., Goubran, R.A., Knoefel, F., Sveistrup, H. & Bilodeau, M. 2009. Determination of sit-to-stand transfer duration using bed and floor pressure sequences. *IEEE Transactions on Biomedical Engineering*, 56(10):2485-2492. https://doi.org/10.1109/TBME.2009.2026733

Auvinet, E., Multon, F., Saint-Arnaud, A., Rousseau, J. & Meunier, J. 2010. Fall detection with multiple cameras: an occlusion-resistant method based on 3-D silhouette vertical distribution. *IEEE Transactions on Information Technology in Biomedicine*, 15(2):290-300. https://doi.org/10.1109/TITB.2010.2087385

Baraka, A., Shokry, A., Omar, I., Kamel, S., Fouad, T., El-Nasr, M.A. & Shaban, H. 2012. A WBAN for human movement kinematics and ECG measurements. *E-Health Telecommunication Systems and Networks*, 1(2):19-25. https://doi.org/10.4236/etsn.2012.12004

Bian, Z.P., Hou, J., Chau, L.P. & Magnenat-Thalmann, N. 2015. Fall detection based on body part tracking using a depth camera. *IEEE Journal of Biomedical and Health Informatics*, 19(2):430-439. https://doi.org/10.1109/JBHI.2014.2319372

Birku, Y. & Agrawal, H. 2018. Survey on fall detection systems. *International Journal of Pure and Applied Mathematics*, 118(18):2537-2543. http://dx.doi.org/10.5220/0011674800003417

Bloom, D.E. & LeeLuca, D.L. 2016. The global demography of aging: facts, explanations, future. In: Piggott, J. & Woodland, A., eds. *Handbook of the Economics of Population Aging*. Amsterdam, The Netherlands: Elsevier. pp. 3-56.

Braun, A., Heggen, H. & Wichert, R. 2012. CapFloor – a flexible capacitive indoor localization system. In: Chessa, S. & Knauth, S., eds. *Conference Proceedings*. 2011 International Competition on Evaluating AAL Systems through Competitive Benchmarking (EvAAL 2011), Valencia, Spain. Berlin, Heidelberg: Springer. pp. 26-35.

Brownlee, J. 2017. *Introduction to time series forecasting with Python*. Melbourne: Machine Learning Mastery.

Brownlee, J. 2020. *Imbalanced classification with Python*. Melbourne: Machine Learning Mastery.

Buhrmester, V., Münch, D., Bulatov, D. & Arens, M. 2019. Evaluating the impact of color information in deep neural networks. In: Morales, A., Fierrez, J., Sánchez, J.S. & Ribeiro, B., eds. *Conference Proceedings*. 9th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2019), Madrid, Spain. Cham, Switzerland: Springer. pp. 302-316.

Campbell, J., Dussault, G., Buchan, J., Pozo-Martin, F., Guerra Arias, M., Leone, C., Siyam, A. & Cometto, G. 2013. A universal truth: no health without a workforce. *Third Global Forum on Human Resources for Health*, https://cdn.who.int/media/docs/default-source/health-workforce/ghwn/ghwa/ghwa_auniversaltruthreport.pdf Date of access: 16 Feb. 2020.

Cao, Z., Simon, T., Wei, S.E. & Sheikh, Y. 2017. Realtime multi-person 2D pose estimation using part affinity fields. In: Liu, Y., Rehg, J.M., Taylor, C.J. & Wu, Y., eds. *Conference Proceedings*. 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, Hawaii. Piscataway, NJ: IEEE. pp. 7291-7299.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.E. & Sheikh, Y. 2021. OpenPose: realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172-186. https://doi.org/10.1109/TPAMI.2019.2929257

Carreira, J., Agrawal, P., Fragkiadaki, K. & Malik, J. 2016. Human pose estimation with iterative error feedback. In: Agapito, L., Berg, T., Kosecka, J. & Zelnik-Manor, L., eds. *Conference Proceedings*. 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV. Piscataway, NJ: IEEE. pp. 4733-4742.

Casilari, E., Luque, R. & Morón, M.J. 2015. Analysis of android device-based solutions for fall detection. *MDPI Sensors Journal*, 15(8):17827-17894. https://doi.org/10.3390/s150817827

Chua, J.L., Chang, Y.C. & Lim, W.K. 2015. A simple vision-based fall detection technique for indoor video surveillance. *Signal, Image and Video Processing*, 9(3):623-633. https://doi.org/10.1007/s11760-013-0493-7

Daher, M., Diab, A., El Najjar El Badaoui, M., Khalil, M.A. & Charpillet, F. 2017. Elder tracking and fall detection system using smart tiles. *IEEE Sensors Journal*, 17(2):469-479. https://doi.org/10.1109/JSEN.2016.2625099

Dake, K. 2018. Chinese gait recognition tech IDs people by how they walk. *The Associated Press*, 6 Nov. https://apnews.com/article/china-technology-beijing-business-international-news-bf75dd1c26c947b7826d270a16e2658a Date of access: 5 Aug. 2020.

Dang, Q., Yin, J., Wang, B. & Zheng, W. 2019. Deep learning based 2D human pose estimation: a survey. *Tsinghua Science and Technology*, 24(6):663-676. https://doi.org/10.26599/TST.2018.9010100

De Miguel, K., Brunete, A., Hernando, M. & Gambao, E. 2017. Home camera-based fall detection system for the elderly. *MDPI Sensors Journal*, 17(12):1-21. art. #2864. https://doi.org/10.3390/s17122864

Dhiman, C. & Vishwakarma, D.K. 2019. A review of state-of-the-art techniques for abnormal human activity recognition. *Engineering Applications of Artificial Intelligence*, 77:21-45. https://doi.org/10.1016/j.engappai.2018.08.014

Du Toit, J.S., Du Toit, J.V. & Kruger, H.A. 2019a. Improving human pose classification using a convolutional neural network through data augmentation. In: Venter, L., ed. *Conference Proceedings*. 48th Annual Conference of the Operations Research Society of South Africa (ORSSA 2019), Cape Town, South Africa. Stellenbosch, South Africa: ORSSA.

Du Toit, J.S., Du Toit, J.V. & Kruger, H.A. 2019b. Heuristic data augmentation for improved human activity recognition. In: Lewis, J. & Balmahoon, T., eds. *Conference Proceedings*. 2019 Southern Africa Telecommunication Networks and Applications Conference (SATNAC 2019), Ballito, South Africa. Pretoria, South Africa: SATNAC. pp. 264-269.

Du Toit, J.S., Du Toit, J.V. & Kruger, H.A. 2021a. Improved human pose differentiation in convolutional neural network classification using colour-based data augmentation. In: Smuts, M., ed. *Conference Proceedings*. 50th Annual Conference of the Operations Research Society of South Africa (ORSSA 2021), Virtual. Stellenbosch, South Africa: ORSSA.

Du Toit, J.S., Du Toit, J.V. & Kruger, H.A. 2021b. Colour-based encoding schemes for improved human pose recognition using a convolutional neural network. In: Smuts, M. & Moorcroft, R., eds. *Conference Proceedings*. 2021 Southern Africa Telecommunication Networks and Applications Conference (SATNAC 2021), Central Drakensberg, South Africa. Pretoria, South Africa: SATNAC. pp. 178-183.

Epshtein, B. & Ullman, S. 2007. Semantic hierarchies for recognizing objects and parts. In: Baker, S., Matas, J. & Zabih, R., eds. *Conference Proceedings*. 25th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007), Minneapolis, MN. Piscataway, NJ: IEEE. pp. 1-8.

Fang, H.-S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.-L. & Lu, C. 2022. AlphaPose: whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7157-7173. https://doi.org/10.48550/arXiv.2211.03375

Felzenszwalb, P.F. & Huttenlocher, D.P. 2005. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61:55-79. https://doi.org/10.1023/B:VISI.0000042934.15159.49

Fisher, C.J. 2010. Using an accelerometer for inclination sensing. *AN-1057, Application Note, Analog Devices*, https://www.analog.com/media/en/technical-documentation/app-notes/an-1057.pdf Date of access: 20 Jul. 2020.

Fu, Z., Culurciello, E., Lichtsteiner, P. & Delbruck, T. 2008. Fall detection using an address-event temporal contrast vision sensor. In: Moon, U.-K. & Setti, G., eds. *Conference Proceedings*. 2008 IEEE International Symposium on Circuits and Systems (ISCAS 2008), Seattle, WA. Piscataway, NJ: IEEE. pp. 424-427.

Galdran, A., Alvarez-Gila, A., Meyer, M.I., Saratxaga, C.L., Araújo, T., Garrote, E., Aresta, G., Costa, P., Mendonça, A.M. & Campilho, A. 2017. Data-driven color augmentation techniques for deep skin image analysis. *arXiv preprint arXiv:1703.03702*, https://doi.org/10.48550/arXiv.1703.03702

Gong, W., Zhang, X., Gonzàlez, J., Sobral, A., Bouwmans, T., Tu, C. & Zahzah, E.H. 2016. Human pose estimation from monocular images: a comprehensive survey. *MDPI Sensors Journal*, 16(12):1-39. art. #1966. https://doi.org/10.3390/s16121966

González-Cañete, F.J. & Casilari, E. 2020. Consumption analysis of smartphone based fall detection systems with multiple external wireless sensors. *MDPI Sensors Journal*, 20(3):1-27. art. #622. https://doi.org/10.3390/s20030622

Goodfellow, I., Bengio, Y. & Courville, A. 2016. *Deep learning*. 3rd ed. Cambridge, Massachusetts: MIT Press. (Adaptive computation and machine learning series).

Gutiérrez, J., Rodríguez, V. & Martin, S. 2021. Comprehensive review of vision-based fall detection systems. *MDPI Sensors Journal*, 21(3):1-50. art. #947. https://doi.org/10.3390/s21030947

Guyon, I. & Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(7-8):1157-1182. http://dx.doi.org/10.1162/153244303322753616

Habib, M.A., Mohktar, M.S., Kamaruzzaman, S.B., Lim, K.S., Pin, T.M. & Ibrahim, F. 2014. Smartphone-based solutions for fall detection and prevention: challenges and open issues. *MDPI Sensors Journal*, 14(4):7181-7208. https://doi.org/10.3390/s140407181

Hazelhoff, L. & Han, J. 2008. Video-based fall detection in the home using principal component analysis. In: Blanc-Talon, J., Bourennane, S., Philips, W., Popescu, D. & Scheunders, P., eds. *Conference Proceedings*. 10th International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS 2008), Juan-les-Pins, France. Berlin, Heidelberg: Springer. pp. 298-309.

He, Y., Li, Y. & Bao, S.-D. 2012. Fall detection by built-in tri-accelerometer of smartphone. In: Zhang, Y.T. & Naghavi, M., eds. *Conference Proceedings*. 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI 2012), Hong Kong, China. Piscataway, NJ: IEEE. pp. 184-187.

Heinrich, S., Rapp, K., Rissmann, U., Becker, C. & König, H.H. 2010. Cost of falls in old age: a systematic review. *Osteoporosis International*, 21(6):891-902. https://doi.org/10.1007/s00198-009-1100-1

Henry, R., Matti, L. & Raimo, S. 2008. Human tracking using near field imaging. In: Saranummi, N. & Weerasinghe, D., eds. *Conference Proceedings*. 2nd International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth 2008), Tampere, Finland. Piscataway, NJ: IEEE. pp. 148-151.

Hernández-García, A., König, P. & Kietzmann, T.C. 2019. Learning robust visual representations using data augmentation invariance. *arXiv preprint arXiv:1906.04547*, https://doi.org/10.48550/arXiv.1906.04547

Huang, X., Wang, F., Zhang, J., Hu, Z. & Jin, J. 2019. A posture recognition method based on indoor positioning technology. *MDPI Sensors Journal*, 19(6):1-14. art. #1464. https://doi.org/10.3390/s19061464

Hyndman, D., Ashburn, A. & Stack, E. 2002. Fall events among people with stroke living in the community: circumstances of falls and characteristics of fallers. *Archives of Physical Medicine and Rehabilitation*, 83(2):165-170. https://doi.org/10.1053/apmr.2002.28030

Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M. & Schiele, B. 2016. Deepercut: a deeper, stronger, and faster multi-person pose estimation model. In: Leibe, B., Matas, J., Sebe, N. & Welling, M., eds. *Conference Proceedings*. 14th European Conference on Computer Vision (ECCV 2016), Amsterdam, The Netherlands. Cham, Switzerland: Springer. pp. 34-50.

Ioffe, S. & Szegedy, C. 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Bach, F. & Blei, D., eds. *Conference Proceedings*. 32nd International Conference on Machine Learning (ICML 2015), Lille, France. Maastricht, The Netherlands: ML Research Press. pp. 448-456.

Jarrett, K., Kavukcuoglu, K., Ranzato, M.A. & LeCun, Y. 2009. What is the best multi-stage architecture for object recognition? In: Matsuyama, T., Cipolla, R., Hebert, M., Tang, X. & Yokoya, N., eds. *Conference Proceedings*. 12th IEEE International Conference on Computer Vision (ICCV 2009), Kyoto, Japan. Piscataway, NJ: IEEE. pp. 2146-2153.

Ji, S., Xu, W., Yang, M. & Yu, K. 2013. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221-231. https://doi.org/10.1109/TPAMI.2012.59

Johnson, S. & Everingham, M. 2011. Learning effective human pose estimation from inaccurate annotation. In: Rosenberg, C. & Chellappa, R., eds. *Conference Proceedings*. 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011), Colorado Springs, CO. Piscataway, NJ: IEEE. pp. 1465-1472.

Jurafsky, D. & Martin, J.H. 2000. *Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition*. New Delhi, India: Pearson Education India.

Kalula, S.Z., Ferreira, M., Swingler, G.H. & Badri, M. 2016. Risk factors for falls in older adults in a South African urban community. *BMC Geriatrics*, 16:1-11. art. #51. https://doi.org/10.1186%2Fs12877-016-0212-7

Kangas, M., Vikman, I., Wiklander, J., Lindgren, P., Nyberg, L. & Jämsä, T. 2009. Sensitivity and specificity of fall detection in people aged 40 years and over. *Gait & Posture*, 29(4):571-574. https://doi.org/10.1016/j.gaitpost.2008.12.008

Kannus, P., Niemi, S., Sievänen, H. & Parkkari, J. 2018. Declining incidence in fall-induced deaths of older adults: Finnish statistics during 1971–2015. *Aging Clinical and Experimental Research*, 30(9):1111-1115. https://doi.org/10.1007/s40520-018-0898-9

Kheradpisheh, S.R., Ghodrati, M., Ganjtabesh, M. & Masquelier, T. 2016. Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific Reports*, 6:1-24. art. #32672. https://doi.org/10.1038/srep32672

Kotsiantis, S. 2011. Feature selection for machine learning classification problems: a recent overview. *Artificial Intelligence Review*, 42:157-176. https://doi.org/10.1007/s10462-011-9230-1

Kreiss, S., Bertoni, L. & Alahi, A. 2019. PifPaf: composite fields for human pose estimation. In: Gupta, A., Hoiem, D., Hua, G. & Tu, Z., eds. *Conference Proceedings*. 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019), Long Beach, CA. Piscataway, NJ: IEEE. pp. 11977-11986.

Kreković, M., Čerić, P., Dominko, T., Ilijaš, M., Ivančić, K., Skolan, V. & Šarlija, J. 2012. A method for real-time detection of human fall from video. In: Biljanovic, P., ed. *Conference Proceedings*. 35th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2012), Opatija, Croatia. Piscataway, NJ: IEEE. pp. 1709-1712.

Krizhevsky, A., Sutskever, I. & Hinton, G.E. 2012. Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J., Bottou, L. & Weinberger, K.Q., eds. *Conference Proceedings*. 26th Conference on Neural Information Processing Systems (NIPS 2012), Lake Tahoe, Nevada. Red Hook, NY: Curran Associates Inc. pp. 1097-1105.

Kumar, A. 2022. Machine learning – sensitivity vs specificity difference. *Data Analytics*, https://vitalflux.com/ml-metrics-sensitivity-vs-specificity-difference/ Date of access: 12 Nov. 2022.

Lan, X., Li, H., Wang, Z. & Chen, Y. 2020. Frailty as a predictor of future falls in hospitalized patients: a systematic review and meta-analysis. *Geriatric Nursing*, 41(2):69-74. https://doi.org/10.1016/j.gerinurse.2019.01.004

Lee, S., Jha, D., Agrawal, A., Choudhary, A. & Liao, W.K. 2017. Parallel deep convolutional neural network training by exploiting the overlapping of computation and communication. In: Sur, C.Ç., Ümit V. & Xia, Y., eds. *Conference Proceedings*. 24th IEEE International Conference on High Performance Computing, Data, and Analytics (HiPC 2017), Jaipur, India. Piscataway, NJ: IEEE. pp. 183-192.

Lee, S.W., Kim, Y.J., Lee, G.S., Cho, B.O. & Lee, N.H. 2007. A remote behavioral monitoring system for elders living alone. In: Joo, D.Y. & Park, Y., eds. *Conference Proceedings*. 7th International Conference on Control, Automation and Systems (ICCAS 2007), Seoul, South Korea. Piscataway, NJ: IEEE. pp. 2725-2730.

Li, Y., Ho, K. & Popescu, M. 2012. A microphone array system for automatic fall detection. *IEEE Transactions on Biomedical Engineering*, 59(5):1291-1301. https://doi.org/10.1109/TBME.2012.2186449

Liu, L., Popescu, M., Skubic, M., Rantz, M., Yardibi, T. & Cuddihy, P. 2011. Automatic fall detection based on Doppler radar motion signature. In: Augusto, J.C., Maitland, J. & Caulfield, B., eds. *Conference Proceedings*. 5th International Conference on Pervasive Computing Technologies for Healthcare and Workshops (PervasiveHealth 2011), Dublin, Ireland. Piscataway, NJ: IEEE. pp. 222-225.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y. & Alsaadi, F.E. 2017. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11-26. https://doi.org/10.1016/j.neucom.2016.12.038

Lu, K.L. & Chu, E.T.H. 2018. An image-based fall detection system for the elderly. *MDPI Applied Sciences Journal*, 8(10):1-31. art. #1995. https://doi.org/10.3390/app8101995

Martin, T., Majeed, B., Lee, B.S. & Clarke, N. 2006. Fuzzy ambient intelligence for next generation telecare. In: Yen, G.G. & Bonissone, P., eds. *Conference Proceedings*. 2006 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2006), Vancouver, Canada. Piscataway, NJ: IEEE. pp. 894-901.

Martinez, J., Hossain, R., Romero, J. & Little, J.J. 2017. A simple yet effective baseline for 3D human pose estimation. In: Cucchiara, R., Matsushita, Y., Sebe, N. & Soatto, S., eds. *Conference Proceedings*. 16th IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy. Piscataway, NJ: IEEE. pp. 2640-2649.

MathWorks (The MathWorks Inc.). 2017. *What is a convolutional neural network?* https://www.mathworks.com/discovery/convolutional-neural-network-matlab.html Date of access: 29 Aug. 2022.

Medrano, C., Plaza, I., Igual, R., Sánchez, Á. & Castro, M. 2016. The effect of personalization on smartphone-based fall detectors. *MDPI Sensors Journal*, 16(1):1-14. art. #117. https://doi.org/10.3390%2Fs16010117

Moeslund, T.B. & Granum, E. 2001. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81(3):231-268. https://doi.org/10.1006/cviu.2000.0897

Mohamed, O., Choi, H.J. & Iraqi, Y. 2014. Fall detection systems for elderly care: a survey. In: Badra, M. & Alfandi, O., eds. *Conference Proceedings*. 6th International Conference on New Technologies, Mobility and Security (NTMS 2014), Dubai, United Arab Emirates. Piscataway, NJ: IEEE. pp. 1-4.

MPII (Max Planck Institute for Informatics). 2023. *Evaluation Results*. (human pose dataset). http://human-pose.mpi-inf.mpg.de/#results Date of access: 4 Jun. 2023.

Mubashir, M., Shao, L. & Seed, L. 2013. A survey on fall detection: principles and approaches. *Neurocomputing*, 100:144-152. https://doi.org/10.1016/j.neucom.2011.09.037

Munea, T.L., Jembre, Y.Z., Weldegebriel, H.T., Chen, L., Huang, C. & Yang, C. 2020. The progress of human pose estimation: a survey and taxonomy of models applied in 2D human pose estimation. *IEEE Access*, 8:133330-133348. https://doi.org/10.1109/ACCESS.2020.3010248

Nait-Charif, H. & McKenna, S.J. 2004. Activity summarisation and fall detection in a supportive home environment. In: Kittler, J., Petrou, M. & Nixon, M., eds. *Conference Proceedings*. 17th International Conference on Pattern Recognition (ICPR 2004), Cambridge, UK. Piscataway, NJ: IEEE. pp. 323-326.

Newell, A., Huang, Z. & Deng, J. 2017. Associative embedding: end-to-end learning for joint detection and grouping. In: Bengio, S., Wallach, H., Fergus, R. & Vishwanathan, S.V.N., eds. *Conference Proceedings*. 31st International Conference on Neural Information Processing (NIPS 2017), Long Beach, CA. Red Hook, NY: Curran Associates Inc. pp. 2277-2287.

Ntinou, I. 2018. *Human pose estimation using convolutional neural networks.* Patras, Greece: University of Patras. (Thesis – MSc). http://hdl.handle.net/10889/11344

Nunez-Marcos, A., Azkune, G. & Arganda-Carreras, I. 2017. Vision-based fall detection with convolutional neural networks. *Wireless Communications and Mobile Computing*, 2017:1-16. art. #9474806. https://doi.org/10.1155/2017/9474806

Oates, B.J., Griffiths, M. & McLean, R. 2022. *Researching information systems and computing*. 2nd ed. London, UK: Sage Publications.

OECD (Organisation for Economic Co-operation and Development). 2017. *Pensions at a glance 2017: OECD and G20 indicators*. Paris, France: OECD Publishing.

Okafor, E., Schomaker, L. & Wiering, M.A. 2018. An analysis of rotation matrix and colour constancy data augmentation in classifying images of animals. *Journal of Information and Telecommunication*, 2(4):465-491. https://doi.org/10.1080/24751839.2018.1479932

Oksanen, R., Paldanius, S., Nykänen, J., Linnavuo, M., Raivio, K., Segerstam, K.E., Sepponen, R., Pohjola, L. & Finne-Soveri, H. 2009. Testing and adopting floor-sensor solutions in daily practice for patient safety in Kustaankartano nursing home. Poster presented at the *19th IAGG World Congress of Gerontology and Geriatrics*, Paris, France.

Orr, R.J. & Abowd, G.D. 2000. The smart floor: a mechanism for natural user identification and tracking. In: Tremaine, M., ed. *Conference Proceedings*. 2000 Conference on Human Factors in Computing Systems (CHI 2000), The Hague, The Netherlands. New York, NY: Association for Computing Machinery. pp. 275-276.

Panahi, L. & Ghods, V. 2018. Human fall detection using machine vision techniques on RGB–D images. *Biomedical Signal Processing and Control*, 44:146-153. https://doi.org/10.1016/j.bspc.2018.04.014

Paradiso, J., Abler, C., Hsiao, K.Y. & Reynolds, M. 1997. The magic carpet: physical sensing for immersive environments. In: Edwards, A. & Pemberton, S., eds. *Conference Proceedings*. 1997 Conference on Human Factors in Computing Systems (CHI 1997), Atlanta, Georgia. New York, NY: Association for Computing Machinery. pp. 277-278.

Pierleoni, P., Belli, A., Palma, L., Pellegrini, M., Pernini, L. & Valenti, S. 2015. A high reliability wearable device for elderly fall detection. *IEEE Sensors Journal*, 15(8):4544-4553. https://doi.org/10.1109/JSEN.2015.2423562

Pierleoni, P., Belli, A., Maurizi, L., Palma, L., Pernini, L., Paniccia, M. & Valenti, S. 2016. A wearable fall detector for elderly people based on AHRS and barometric sensor. *IEEE Sensors Journal*, 16(17):6733-6744. https://doi.org/10.1109/JSEN.2016.2585667

Pinto, N., Doukhan, D., DiCarlo, J.J. & Cox, D.D. 2009. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Computational Biology*, 5(11):1-12. art. #1000579. https://doi.org/10.1371/journal.pcbi.1000579

Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P.V. & Schiele, B. 2016. Deepcut: joint subset partition and labeling for multi person pose estimation. In: Agapito, L., Berg, T., Kosecka, J. & Zelnik-Manor, L., eds. *Conference Proceedings*. 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV. Piscataway, NJ: IEEE. pp. 4929-4937.

Planinc, R. & Kampel, M. 2013. Introducing the use of depth data for fall detection. *Personal and Ubiquitous Computing*, 17(6):1063-1072. https://doi.org/10.1007/s00779-012-0552-z

Popescu, M. & Mahnot, A. 2009. Acoustic fall detection using one-class classifiers. In: Pan, X., Worrell, G.A., Saranummi, N., McCullough, J. & Tranquillo, R., eds. *Conference Proceedings*. 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS 2009), Minneapolis, MN. Piscataway, NJ: IEEE. pp. 3505-3508.

Popescu, M., Li, Y., Skubic, M. & Rantz, M. 2008. An acoustic fall detector system that uses sound height information to reduce the false alarm rate. In: Vicini, P. & Principe, J., eds. *Conference Proceedings*. 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2008), Vancouver, Canada. Piscataway, NJ: IEEE. pp. 4628-4631.

Poppe, R. 2007. Vision-based human motion analysis: an overview. *Computer Vision and Image Understanding*, 108(1-2):4-18. https://doi.org/10.1016/j.cviu.2006.10.016

Ramachandran, A. & Karuppiah, A. 2020. A survey on recent advances in wearable fall detection systems. *BioMed Research International*, 2020:1-17. art. #2167160. https://doi.org/10.1155/2020/2167160

Ren, L. & Peng, Y. 2019. Research of fall detection and fall prevention technologies: a systematic review. *IEEE Access*, 7:77702-77722. https://doi.org/10.1109/ACCESS.2019.2922708

Rimminen, H., Lindström, J., Linnavuo, M. & Sepponen, R. 2010. Detection of falls among the elderly by a floor sensor using the electric near field. *IEEE Transactions on Information Technology in Biomedicine*, 14(6):1475-1476. https://doi.org/10.1109/TITB.2010.2051956

Ropponen, A., Rimminen, H. & Sepponen, R. 2011. Robust system for indoor localisation and identification for the health care environment. *Wireless Personal Communications*, 59:57-71. https://doi.org/10.1007/s11277-010-0189-z

Rougier, C. & Meunier, J. 2006. Fall detection using 3D head trajectory extracted from a single camera video sequence. In: Yoshiura, H., Sakurai, K., Rannenberg, K., Murayama, Y. & Kawamura, S., eds. *Conference Proceedings*. 1st International Workshop on Video Processing for Security (IWSEC 2006), Kyoto, Japan. Berlin, Heidelberg: Springer. pp. 7-9.

Rougier, C., Auvinet, E., Rousseau, J., Mignotte, M. & Meunier, J. 2011. Fall detection from depth map video sequences. In: Abdulrazak, B., Giroux, S., Bouchard, B., Pigot, H. & Mokhtari, M., eds. *Conference Proceedings*. 9th International Conference on Smart Homes and Health Telematics (ICOST 2011), Montreal, Canada. Berlin, Heidelberg: Springer. pp. 121-128.

Roy, D., Murty, K.S.R. & Mohan, C.K. 2015. Feature selection using deep neural networks. In: Choe, Y., He, H. & Roy, A., eds. *Conference Proceedings*. Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN 2015), Killarney, Ireland. Piscataway, NJ: IEEE. pp. 1-6.

Salzmann, M. & Urtasun, R. 2010. Combining discriminative and generative methods for 3D deformable surface and articulated pose reconstruction. In: Darrell, T., Hogg, D. & Jacobs, D., eds. *Conference Proceedings*. 23rd IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2010), San Francisco, CA. Piscataway, NJ: IEEE. pp. 647-654.

Santoyo-Ramón, J.A., Casilari, E. & Cano-García, J.M. 2018. Analysis of a smartphone-based architecture with multiple mobility sensors for fall detection with supervised learning. *MDPI Sensors Journal*, 18(4):1-29. art. #1155. https://doi.org/10.3390/s18041155

Saxe, A.M., Koh, P.W., Chen, Z., Bhand, M., Suresh, B. & Ng, A.Y. 2011. On random weights and unsupervised feature learning. In: Getoor, L. & Scheffer, T., eds. *Conference Proceedings*. 28th International Conference on Machine Learning (ICML 2011), Bellevue, Washington. Madison, WI: Omnipress. pp. 1089–1096.

Shen, R.K., Yang, C.Y., Shen, V.R. & Chen, W.C. 2016. A novel fall prediction system on smartphones. *IEEE Sensors Journal*, 17(6):1865-1871. https://doi.org/10.1109/JSEN.2016.2598524

Shorten, C. & Khoshgoftaar, T.M. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:1-48. art. #60. https://doi.org/10.1186/s40537-019-0197-0

Simonyan, K. & Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D. & Weinberger, K.Q., eds. *Conference Proceedings*. 27th International Conference on Neural Information Processing Systems (NIPS 2014), Montreal, Canada. Cambridge, MA: MIT Press. pp. 568-576.

Simonyan, K. & Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In: Kingsbury, B., Bengio, S., De Freitas, N. & Larochelle, H., eds. *Conference Proceedings*. 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA.

Solbach, M.D. & Tsotsos, J.K. 2017. Vision-based fallen person detection for the elderly. In: Cucchiara, R., Matsushita, Y., Sebe, N. & Soatto, S., eds. *Conference Proceedings*. 16th IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy. Piscataway, NJ: IEEE. pp. 1433-1442.

Sposaro, F. & Tyson, G. 2009. iFall: an android application for fall monitoring and response. In: Pan, X., Worrell, G.A., Saranummi, N., McCullough, J. & Tranquillo, R., eds. *Conference Proceedings*. 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS 2009), Minneapolis, MN. Piscataway, NJ: IEEE. pp. 6119-6122.

SU (Stanford University). 2023. *Convolutional neural networks for visual recognition*. School of Engineering Information Technology. San Jose: Stanford University. (Study guide, CS231n).

Su, B.Y., Ho, K., Rantz, M.J. & Skubic, M. 2014. Doppler radar fall activity detection using the wavelet transform. *IEEE Transactions on Biomedical Engineering*, 62(3):865-875. https://doi.org/10.1109/TBME.2014.2367038

Sung, G.Y., Kwak, D.M. & Lyou, J. 2010. Neural network based terrain classification using wavelet features. *Journal of Intelligent & Robotic Systems*, 59(3):269-281. https://doi.org/10.1007/s10846-010-9402-2

Sze, V., Chen, Y.H., Yang, T.J. & Emer, J.S. 2017. Efficient processing of deep neural networks: a tutorial and survey. *Proceedings of the IEEE*, 105(12):2295-2329. https://doi.org/10.1109/JPROC.2017.2761740

Taylor, L. & Nitschke, G. 2018. Improving deep learning with generic data augmentation. In: Sundaram, S., ed. *Conference Proceedings*. 8th IEEE Symposium Series on Computational Intelligence (SSCI 2018), Bangalore, India. Piscataway, NJ: IEEE. pp. 1542-1547.

Terroso, M., Freitas, R., Gabriel, J., Marques, A.T. & Simões, R. 2013. Active assistance for senior healthcare: a wearable system for fall detection. In: Rocha, Á., Calvo-Manzano, J.A., Reis, L.P. & Pérez Cota, M., eds. *Conference Proceedings*. 8th Iberian Conference on Information Systems and Technologies (CISTI 2013), Lisboa, Portugal. Piscataway, NJ: IEEE. pp. 1-6.

Thammasat, E. & Chaicharn, J. 2012. A simply fall-detection algorithm using accelerometers on a smartphone. In: Gale, T., Stack, C. & Dargaville, P., eds. *Conference Proceedings*. 5th 2012 Biomedical Engineering International Conference (BMEiCON 2012), Muang, Thailand. Piscataway, NJ: IEEE. pp. 1-4.

Thomas, P.Y. 2010. *Towards developing a web-based blended learning environment at the University of Botswana.* Pretoria: University of South Africa. (Thesis – Phd). http://hdl.handle.net/10500/4245

Tian, Y., Zitnick, C.L. & Narasimhan, S.G. 2012. Exploring the spatial hierarchy of mixture models for human pose estimation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y. & Schmid, C., eds. *Conference Proceedings*. 12th European Conference on Computer Vision (ECCV 2012), Florence, Italy. Berlin, Heidelberg: Springer. pp. 256-269.

Tomii, S. & Ohtsuki, T. 2012. Falling detection using multiple Doppler sensors. In: Yang, J., Sezaki, K. & Agoulmine, N., eds. *Conference Proceedings*. 14th IEEE International Conference on e-Health Networking, Applications and Services (Healthcom 2012), Beijing, China. Piscataway, NJ: IEEE. pp. 196-201.

Tompson, J.J., Jain, A., LeCun, Y. & Bregler, C. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D. & Weinberger, K.Q., eds. *Conference Proceedings*. 27th International Conference on Neural Information Processing Systems (NIPS 2014), Montreal, Canada. Cambridge, MA: MIT Press. pp. 1799-1807.

Tong, L., Song, Q., Ge, Y. & Liu, M. 2013. HMM-based human fall detection and prediction method using tri-axial accelerometer. *IEEE Sensors Journal*, 13(5):1849-1856. https://doi.org/10.1109/JSEN.2013.2245231

Toshev, A. & Szegedy, C. 2014. DeepPose: human pose estimation via deep neural networks. In: Basri, R., Fermuller, C., Martinez, A. & Vidal, R., eds. *Conference Proceedings*. 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014), Columbus, OH. Piscataway, NJ: IEEE. pp. 1653-1660.

Trochim, W. & Donnelly, J. 2006. *The research methods knowledge base*. 3rd ed. Mason, Ohio: Atomic Dog Publishing.

Uddin, M., Khaksar, W. & Torresen, J. 2018. Ambient sensors for elderly care and independent living: a survey. *MDPI Sensors Journal*, 18(7):1-31. art. #2027. https://doi.org/10.3390/s18072027

Vaidehi, V., Ganapathy, K., Mohan, K., Aldrin, A. & Nirmal, K. 2011. Video based automatic fall detection in indoor environment. In: Mannar, J.P., Shanmugavel, S., Joseph, S.A. & Vaidehi, V., eds. *Conference Proceedings*. 2011 International Conference on Recent Trends in Information Technology (ICRTIT 2011), Chennai, India. Piscataway, NJ: IEEE. pp. 1016-1020.

Vallabh, P. & Malekian, R. 2018. Fall detection monitoring systems: a comprehensive review. *Journal of Ambient Intelligence and Humanized Computing*, 9(6):1809-1833. https://doi.org/10.1007/s12652-017-0592-3

Van den Bergh, M., Koller-Meier, E., Kehl, R. & Van Gool, L. 2009. Real-time 3D body pose estimation. In: Aghajan, H. & Cavallaro, A., eds. *Multi-Camera Networks: Principles and Applications*. Oxford, UK: Academic Press. pp. 335-361.

Walczak, S. & Cerpa, N. 2003. Artificial neural networks. In: Meyers, R.A., ed. *Encyclopedia of Physical Science and Technology*. 3rd ed. New York, NY: Academic Press. pp. 631-645.

Wang, K., Lin, L., Jiang, C., Qian, C. & Wei, P. 2019. 3D human pose machines with self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(5):1069-1082. https://doi.org/10.1109/TPAMI.2019.2892452

Wang, Y., Wu, K. & Ni, L.M. 2016. WiFall: device-free fall detection by wireless networks. *IEEE Transactions on Mobile Computing*, 16(2):581-594. https://doi.org/10.1109/TMC.2016.2557792

Wasikowski, M. & Chen, X.W. 2009. Combating the small sample class imbalance problem using feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1388-1400. https://doi.org/10.1109/TKDE.2009.187

Wei, S.E., Ramakrishna, V., Kanade, T. & Sheikh, Y. 2016. Convolutional pose machines. In: Agapito, L., Berg, T., Kosecka, J. & Zelnik-Manor, L., eds. *Conference Proceedings*. 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV. Piscataway, NJ: IEEE. pp. 4724-4732.

WHO (World Health Organization). 2008. *WHO global report on falls prevention in older age*. Geneva, Switzerland: WHO Press.

Wingfield, N., Mozur, P. & Corkery, M. 2018. Retailers race against Amazon to automate stores. *The New York Times*, 1 Apr. https://www.nytimes.com/2018/04/01/technology/retailer-stores-automation-amazon.html Date of access: 4 Sep. 2021.

Xu, T., Zhou, Y. & Zhu, J. 2018. New advances and challenges of fall detection systems: a survey. *MDPI Applied Sciences Journal*, 8(3):1-11. art. #418. https://doi.org/10.3390/app8030418

Yacchirema, D., de Puga, J.S., Palau, C. & Esteve, M. 2019. Fall detection system for elderly people using IoT and ensemble machine learning algorithm. *Personal and Ubiquitous Computing*, 23(9):801-817. https://doi.org/10.1007/s00779-018-01196-8

Yang, W., Wang, K. & Zuo, W. 2012. Neighborhood component feature selection for high-dimensional data. *Journal of Computers*, 7(1):161-168. https://doi.org/10.4304/jcp.7.1.161-168

Yang, Y. & Ramanan, D. 2011. Articulated pose estimation with flexible mixtures-of-parts. In: Rosenberg, C. & Chellappa, R., eds. *Conference Proceedings*. 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011), Colorado Springs, CO. Piscataway, NJ: IEEE. pp. 1385-1392.

Yazar, A., Keskin, F., Töreyin, B.U. & Çetin, A.E. 2013. Fall detection using single-tree complex wavelet transform. *Pattern Recognition Letters*, 34(15):1945-1952. https://doi.org/10.1016/j.patrec.2012.12.010

Yu, X. 2008. Approaches and principles of fall detection for elderly and patient. In: Merrell, R.C., ed. *Conference Proceedings*. 10th International Conference on e-health Networking, Applications and Services (HealthCom 2008), Singapore. Piscataway, NJ: IEEE. pp. 42-47.

Zeiler, M.D. & Fergus, R. 2014. Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T., eds. *Conference Proceedings*. 13th European Conference on Computer Vision (ECCV 2014), Zurich, Switzerland. Cham, Switzerland: Springer. pp. 818-833.

Zhang, C., Tian, Y. & Capezuti, E. 2012. Privacy preserving automatic fall detection for elderly using RGBD cameras. In: Miesenberger, K., Karshmer, A., Penaz, P. & Zagler, W., eds. *Conference Proceedings*. 13th International Conference on Computers for Handicapped Persons (ICCHP 2012), Linz, Austria. Berlin, Heidelberg: Springer. pp. 625-633.

Zhang, L. & Suganthan, P.N. 2016. A survey of randomized algorithms for training neural networks. *Information Sciences*, 364:146-155. https://doi.org/10.1016/j.ins.2016.01.039

Zhang, Z., Conly, C. & Athitsos, V. 2015. A survey on vision-based fall detection. In: Makedon, F., Mariottini, G.-L., Korn, O., Maglogiannis, I. & Metsis, V., eds. *Conference Proceedings*. 8th ACM International Conference on Pervasive Technologies Related to Assistive Environments (PETRA 2015), Corfu, Greece. New York, NY: Association for Computing Machinery. pp. 1-7.

Zheng, A. & Casari, A. 2018. *Feature engineering for machine learning: principles and techniques for data scientists*. 1st ed. Sebastopol, CA: O'Reilly Media Inc.

Zheng, C., Wu, W., Yang, T., Zhu, S., Chen, C., Liu, R., Shen, J., Kehtarnavaz, N. & Shah, M. 2020. Deep learning-based human pose estimation: a survey. *arXiv preprint arXiv:2012.13392*, https://doi.org/10.48550/arXiv.2012.13392

Zivkovic, Z. & Van Der Heijden, F. 2006. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters*, 27(7):773-780. https://doi.org/10.1016/j.patrec.2005.11.005

ANNEXURE A: CONFERENCE PAPER PRESENTED AT SATNAC 2021

Colour-based encoding schemes for improved human pose recognition using a Convolutional neural network

J.S. du Toit, J.V. du Toit, H.A. Kruger

This paper was presented at the SATNAC 2021 Conference, hosted at the Champagne Sports Resort, Central Drakensberg, KwaZulu-Natal, South Africa. 21-23 November 2021

https://doi.org/10.25388/nwu.23284388

Colour-Based Encoding Schemes for Improved Human Pose Recognition Using a Convolutional Neural Network

J.S. du Toit[1], J.V. du Toit[2], H.A. Kruger[3]

School of Computer Science and Information Systems
North-West University, Potchefstroom Campus
Private Bag X6001, Potchefstroom, 2520, South Africa
Tel: +27 18 2992548, Fax: +27 18 2992570

[1]Jaco.DuToit@nwu.ac.za
[2]Tiny.DuToit@nwu.ac.za
[3]Hennie.Kruger@nwu.ac.za

Abstract—**The goal of computer vision is to provide computers with the perceptual capability to process and understand visual data and is achieved by synthesizing information from raw videos and images. Humans are among the most frequently analysed subjects of computer vision because activity recognition and pose identification are applicable in various critical industries. This study examines a novel data augmentation technique that can aid in the machine-learned interpretation of the human form for improved pose recognition by superimposing joint markers on humans detected in video footage. The joint locations are derived from a pose estimator and are supplemented with additional information through the use of colour. The keypoint colour augmentations are applied based on either a radial or ringed colour wheel, which is notionally intended to encode spatial information based on the position of the joint. An improvement in classification accuracy of up to 11 percentage points over a baseline model was achieved when applied to a pose dataset and classified using a convolutional neural network. Furthermore, different augmentation schemes were found to favour the recognition of certain poses over others. These augmentation schemes can be diversely applied in human activity recognition and tailored to foster improved accuracy for pose-dependent classification tasks.**

Keywords—**Convolutional neural network, Data augmentation, Image classification, Pose estimation, Pose recognition.**

I. Introduction

Automated human activity and human pose recognition are among the technologies used to assist and cooperate with humans in meaningful ways. Human-orientated machinery and assistive technology systems often apply person recognition to enable systems that involve patient monitoring and health evaluation [1], human-computer interfacing [2], smart home technology [3], and limb and recovery rehabilitation therapy [4]. These systems operate by acquiring situational awareness through scene understanding based on the presence and actions of humans in the immediate environment [5]. Depending on the nature of the task, these systems often need to operate in real-time. Video-based human activity recognition (HAR) is one such application that is frequently relied upon for time-critical tasks, especially in healthcare monitoring. A HAR solution operates by learning the routine activities of daily life to recognise any abnormal behaviour, like a person falling down. An event of this kind should ideally be met with immediate assistance to the victim.

HAR achieves scene understanding by synthesising information in video footage [5]. To date, attempts to identify visual cues have relied on body shape analysis, where deformations in the silhouette of a person are measured to infer a specific action or movement. This involves monitoring changes in the proportions of a projected bounding box drawn around the individual. Alternatively, contemporary model-based approaches obtain cues by encoding a human pose as a compact feature vector which maintains a measure of similarity with a database of known poses, thus allowing for comparative distance measurements when estimating new pose instances [5]. Preserving pose similarity makes it possible for known poses to be easily recognised and unknown poses to be readily estimated. Similarly, modern machine learning methods maintain pose similarity by mapping an abstracted feature space to a common class label as is the case for a Convolutional Neural Network (CNN).

This paper investigates the use of data augmentation in pose recognition to compose a favourable feature encoding that a machine-learned classifier can leverage for improved classification accuracy. The augmentation techniques embed spatial information into each input instance and consequently encourage class similarity among similar poses. This approach entails identifying individuals in an incoming video feed and plotting a skeletal mapping of keypoints across their approximated silhouette. The keypoints are assigned colours based on their position, thereby supplementing the visual information available to the classifier. These descriptors can be leveraged during feature extraction and thereby enhance the discriminative cues for pose recognition within the CNN feature space.

Two colour wheel configurations are proposed for the positional assignment of each keypoint colour, namely a radial and a ringed structure, which are presented in Section IV. The distinct colour arrangement of each wheel helps to encode cues for silhouette deformation (ringed) and orientation changes (radial). The fixed position of keypoints on the colour wheel results in identical poses expressing a similar range of keypoint colours. Ultimately, this approach allows a pose classifier the freedom to selectively learn which visual cues are salient features that best describe a pose based on either the isolated human silhouette or the colour-encoded keypoints.

The remainder of this paper is organised as follows. Section II provides an overview of advances in pose estimation and recognition and how colour has been discovered to be a salient feature in image classification. Insight into how a CNN functions and performs image recognition is presented in Section III. Section IV contains details on the experimental design, the dataset, the colour-based augmentation techniques, and the structure of the implemented CNN. The evaluations performed on each keypoint colour encoding are summarised in Section V, whereafter the study is concluded in Section VI.

II. Related Work

Different approaches exist for human pose estimation which is decided by the choice of sensors and the intended application. The kinematic configuration, or pose, of the human body is often only approximated given the complexity of its articulation within 3D space [5]. Simplifying the human form to an abstract shape or a set of keypoints in 2D space makes for a practical approach to estimating a pose. This is because a 2D representation is subject to a reduced search space and is the result of forgoing contextual and spatial information that would otherwise be captured in 3D imaging [5].

Pons-Moll and Rosenhahn [6] identified a selection of keypoints to model a 3D representation of the human body. Their approach relied on depth sensors and the complex coordination of multiple cameras aligned to observe the same scene. This 3D abstraction allowed for an accurate representation of a pose as it relates to the real-world configuration of the human body. However, accurately modelling the human form in 3D affords a greater degree of freedom in permissible poses and consequently more opportunity for possible errors. Additionally, optimising such a representation to adapt to incoming instances for pose estimation is computationally expensive. For these reasons, a simplistic 2D representation is often preferred given its realistic potential for real-time applications, albeit at the expense of 3D information that would otherwise support greater accuracy. Therein lies the motivation of this study: to experimentally enhance a 2D pose representation with a pose-sensitive descriptor that can mimic forfeited 3D information. Colour is a prime candidate for this type of enhancement given recent insights into the functionality of CNNs and how colour has been discovered to be a valuable feature in image classification.

A recent study showcases how adept CNNs are in processing colour. Buhrmester *et al.* [7] examined the effects of colour omission when classifying images from multiple datasets. The significance of colour-dependency among samples was most prominently expressed in an image dataset of different terrains that depict desert, forest, snow, and urban scenery. Samples from the desert and snow classes were among the most dependent on colour information, likely due to the similarity in their appearance. Their identical landscapes prompted the classifier to learn characteristic hues of warm or cool colours as distinguishing features between them. Interestingly, the urban category was least affected by the omission of colour, suggesting that perhaps colours varied too often among its samples and thus learnt characteristic objects and shapes within the cityscapes as cues for that class. These findings suggest that CNNs learn adaptively based on the most discriminate features of each class, including colour. This study capitalises on this insight by introducing colour-based keypoint mappings that act as cues in the classification of human poses. These mappings are based on a fixed geometrical shape that simulates 3D spatial information regarding variations in the orientation and shape deformation of the human body.

III. Convolutional Neural Network

A CNN is a multi-layer neural network that effectively learns visual patterns directly from images. The initial layers in the CNN act as a hierarchical feature extractor where a set of learnt filters sequentially convolve across the entirety of an image to yield a feature vector that is iteratively refined as it is passed through each layer. This connection structure essentially performs feature filtering on the output of each preceding layer to help isolate salient features that suggest the input class. The filters are adapted during training to become sensitive to discriminative patterns and activate when a valuable feature is detected.

Convolution is traditionally followed by a sub-sampling layer, such as pooling, which reduces the dimensionality of the feature vector and instils the classifier with translation invariance. This makes the classifier robust against minor distortions and variations for the sought features in an image [8]. The effect is a consequence of condensing the feature vector to contain only the most significant features that contribute to the input's eventual classification. Translation invariance is advantageous for pose recognition because a person can be positioned in any location and in various poses within the recorded video frames.

A CNN concludes in either a single or a set of fully connected layers that acts as a general-purpose classifier over the extracted features. These layers learn an association between a collection of features and its predefined class. This association is self-elected and can be based on a concentration of certain abstracted textures, shapes, or even colours in the compiled feature vector [7]. For this study, adopting data augmentation as a pre-processing step is intended to benefit the classifier by providing it with more freedom in electing a favourable class association. This would improve the classification accuracy for easily confused classes such as the sitting (on a chair) and crawling pose which are identical except in their orientation. Given the translation invariance inherent in a CNN, it is not sensitive to differentiating classes based on the position or orientation of a pose. However, encoding such positional and orientational cues using colour-based augmentations may equip the classifier with this capacity.

IV. Experimental Design

This section contains an overview on each aspect of the experiments conducted in this study, including the dataset of human poses, the method used for human detection and silhouette extraction, how keypoints are identified and augmented to encode additional information through colour, and finally the CNN structure that facilitates pose recognition.

A. Dataset

The dataset was obtained from Adhikari *et. al* [9], which was compiled for their own fall detection experiments. It consists of five different poses (sitting, standing, laying, bending, crawling), which allows for the influence of colour-based augmentations to be observed on each of these poses. The five poses were manually labelled for each captured video frame that was recorded using a Microsoft Kinect sensor. Variability within the dataset was also minimised by limiting real-world complications such as extreme occlusions, changes in the background of the scene, and the possibility of multiple people within a scene. However, out of frame observations were retained, which naturally occur as an individual walks into and out of the camera's field of view, resulting in a sixth class free of any pose and labelled as an empty instance.

The resolution of the Kinect sensor yields frames of 640 x 480 pixels and captures both Red Green Blue (RGB) and depth images. These were recorded from a vantage point of approximately 2.4m above ground which is the height of the ceiling for each room in which the recordings were made. A total of five different participants were monitored from eight distinct viewing angles within selected rooms. The training set of 14,938 frames features two participants whose movements were recorded in separate rooms. The validation set of 3,063 frames again features one of these participants but recorded from a different viewing angle not contained within the training set. The remaining three participants make up the test set of 2,344 frames recorded in a room not included in either the training or validation sets. Combining these different angles, rooms, and participants into the dataset helps ensure that it expresses various perspectives for all of the five poses.

B. Silhouette Extraction

The first step toward pose estimation entails detecting any humans in the video footage using static background subtraction. It is a simple technique that discounts the changeless elements in a scene to identify motion and extract a silhouetted person from the video frame. The method is typically only applied when monitoring a controlled environment to avoid any confusion that pets, or other moving objects, may introduce. Adaptive background subtraction is a similar but more advanced technique that helps counteract irrelevant scene changes such as shifting shadows by weighting incoming observations in a video feed more heavily than preceding observations [10]. Using this technique, the dataset is pre-processed to extract the silhouette of any person present in each video frame. The original video frames in the RGB and depth footage are depicted alongside their background subtracted counterparts in Fig. 1 and Fig. 2, respectively.

Figure 1: Example of recorded RGB frame (left) alongside its background-subtracted counterpart (right).

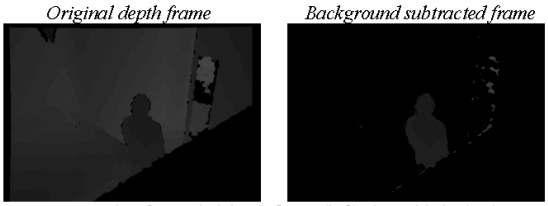


Figure 2: Example of recorded depth frame (left) alongside its background-subtracted counterpart (right).

The information loss suffered by RGB images due to scene variations and changes in lighting during background subtraction is supplemented by depth images, which are robust against these influences. Depth imaging is, however, subject to a limited viewing range given the sensor's observable distance (typically less than 5m for the Kinect). Both sets of images are complementary in the information they provide and including both in the dataset allows the pose classifier the opportunity to utilise RGB information when depth imaging is affected by its distance limitation. Similarly, it can rely on depth information when RGB imaging is affected by a malformed silhouette caused by variations in lighting and shadows.

C. Keypoint Colour-based Augmentation

OpenPose [11] is a reliable vision-based pose estimator used to derive a set of 25 skeletal keypoints from each identified person in the dataset. It performs this estimation using an iteratively refined heat map of the most probable locations for every joint of the human body. The positional likelihood of each joint is refined by considering its relation to surrounding joints. In the end, the pixels associated with the highest degree of probability for their associated joint are then output as a set of XY coordinates. These represent the keypoints that are mapped onto the human silhouettes and assigned a colour based on their position within a projected colour wheel. Fig. 3 illustrates the four colour wheels used to this end. In addition, the extracted human silhouettes are used to establish a comparable baseline dataset without keypoint mappings to further highlight the effectiveness of this approach. Finally, the baseline images and applied augmentations yield five separate datasets that are comparatively evaluated in Section V.

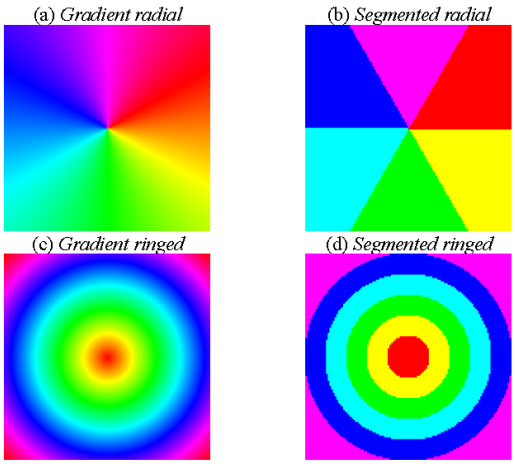


Figure 3: The proposed colour wheels that are used to assign keypoint colour based on their position within the colour wheel.

Using the obtained OpenPose XY coordinates, the given colour wheel is fixed to the centre of a person's torso when assigning colours to the keypoints. Examples of this mapping are illustrated in Figs. 4-7. The type of information encoded through each of the colour wheels is determined by the available spectrum of colours and the structural arrangement of those colours. The gradient-based colour wheels in Fig. 3(a) and Fig. 3(c) supports greater granularity that emphasizes slight changes in a pose through the use of 360 distinct colours. The segmented colour wheels in Fig. 3(b) and Fig. 3(d) instead account for only six colours, helping to disregard any insignificant movement between poses. Using less granularity through a limited spectrum of colours helps to encourage CNN generalisation among poses of the same class. This is best illustrated in Fig. 4 where the projected keypoint mapping of various standing poses will mostly express the same six colours for the same joints across most instances. In contrast, varying gradients of similar colours will instead be expressed in the same set of joints when encoded using a gradient colour wheel as demonstrated in Fig. 5.

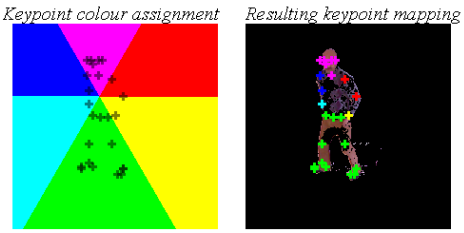


Figure 4: Positional keypoint colour assignment for a standing pose is demonstrated on a colour wheel of low colour granularity (left) alongside its projected mapping onto a human silhouette (right).

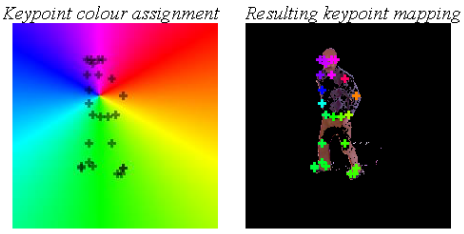


Figure 5: Positional keypoint colour assignment for a standing pose is demonstrated on a colour wheel of high colour granularity (left) alongside its projected mapping onto a human silhouette (right).

The structural arrangement of colours in the colour wheel helps to mimic spatial cues that denote postural changes that would otherwise go unrecorded in the abstracted feature space of a CNN. The radial colour wheels in Fig. 3(a) and Fig. 3(b) are to help encode changes in the orientation of a pose, whereas the ringed colour wheels in Fig. 3(c) and Fig. 3(d) are to help capture silhouette deformations. Encoding any orientation changes should assist the classifier in distinguishing poses that share the same silhouette, such as sitting (on a chair) and crawling. Fig. 6 illustrates this in that a crawling pose will align horizontally with the colour wheel, thus saturating the keypoints of similarly oriented poses with the same colours: cyan and yellow. Again, a standing pose will align vertically and adopt mostly green and magenta keypoints.

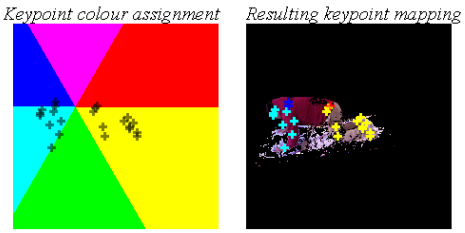


Figure 6: Positional keypoint colour assignment for a crawling pose is demonstrated on a radial colour wheel that encodes orientation (left) alongside its projected mapping onto a human silhouette (right).

On the other hand, the ringed colour wheel is expected to help capture size deformations related to the surface area that a pose assumes within a frame (e.g. standing compared to sitting). As demonstrated in Fig. 7, the range or number of colours expressed in the keypoint mapping is dependent on the size of the pose. The keypoint colours of a shrunken pose such as sitting will tend to be limited to colours in the inner rings of the colour wheel, whereas a larger pose will capture colours in the outer rings. However, unlike the radial colour wheel, the ringed representation is easily affected by size distortions caused by the distance between the subject and the camera. This influence can be limited by recording poses in a controlled environment, like in the given pose dataset [9], where consistent distances are maintained throughout all instances.

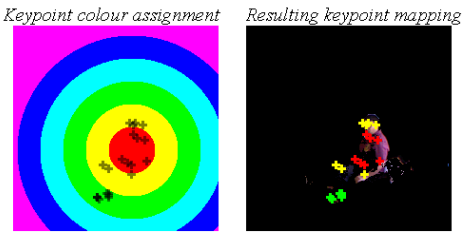


Figure 7: Positional keypoint colour assignment for a sitting pose is demonstrated on a ringed colour wheel that encodes size deformation (left) alongside its projected mapping onto a human silhouette (right).

In the end, the final output image is resized to 156 x 108 pixels to help reduce the computational load when training and executing the classifier. The images were also adapted to include four channels. The first three accommodate the RGB colour information for the background-subtracted image and encoded keypoint markers, like the resulting mappings illustrated in Figs. 4-7. The fourth channel accommodates the background-subtracted depth image, shown earlier in Fig. 2. Keypoint markers are also superimposed onto the silhouette depth images to highlight the points of interest that the CNN can use to classify the instances.

D. Convolutional Neural Network Architecture

The structure of the CNN is based on that of Adhikari *et. al* [11], which performed pose recognition on the same dataset. Their CNN was inspired by the VGGNet [12] which outperformed all other submissions in the ImageNet Challenge 2014 for image classification. As illustrated in Fig. 8, the input layer of the network accepts an input image of size 156 x 108 x 4. The remainder of the network replicates the VGGNet with only its hyperparameters adjusted to favour the pose dataset.

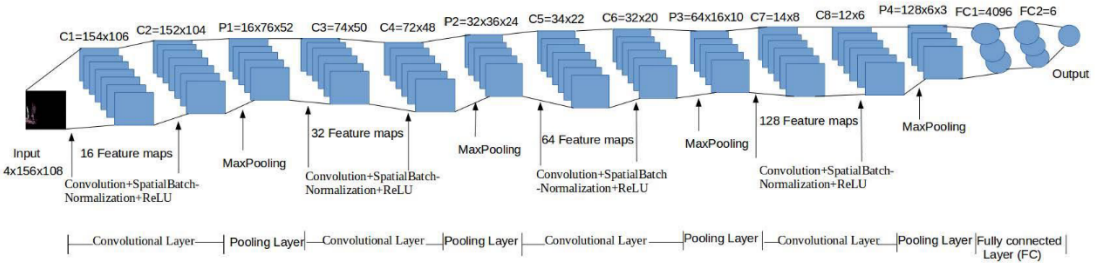


Figure 8: The composition of the CNN used to perform pose recognition as an image classification task in this study [9].

V. Results

The performance results for the colour-based augmentation schemes in pose recognition are documented in this section. The scores are evaluated against a baseline measure of non-augmented pose representations to denote the improvements in classification accuracy.

A. Validation Loss During Training

The averaged validation loss performance for 20 training epochs for the five types of CNN classifiers are shown in Fig. 9. Each classifier was trained on separate datasets augmented with a chosen augmentation scheme represented by each line in the graph. These and succeeding metrics presented in this section were computed using a stratified ten-fold cross-validation strategy where the dataset was divided into ten subsets. Each set consisted of approximately the same number of data samples and an equally representative distribution of classes according to an approximate 75% train, 15% validation split, and a held-out 10% testing split. Fig. 9 reveals a steady performance improvement as the models are optimised with every training epoch. All models achieve convergence by the 17th epoch.

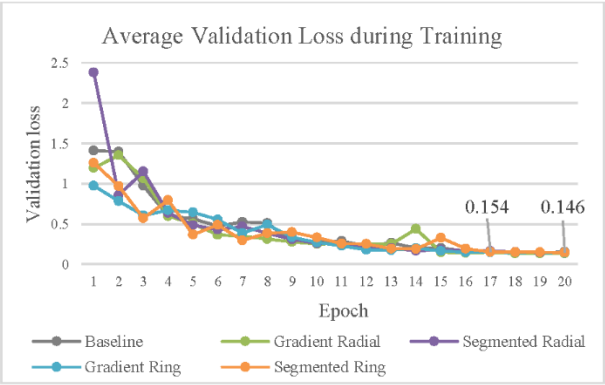


Figure 9: Averaged ten-fold cross-validation loss for each model type trained on their respective datasets.

B. Pose Classification Accuracy

The averaged classification accuracy for the ten-fold models measured on a single set of 2,344 test images is shown in Fig. 10. The baseline models achieved performance scores in the range of 66% to 76%, with an average score of 70.48%. Each colour-based augmentation improves on the baseline scores and in the case of the segmented radial wheel improves performance by up to 11 percentage points (81.61%). However, these improvements are concentrated in specific classes since each augmentation tends to favour the recognition of certain poses over others.

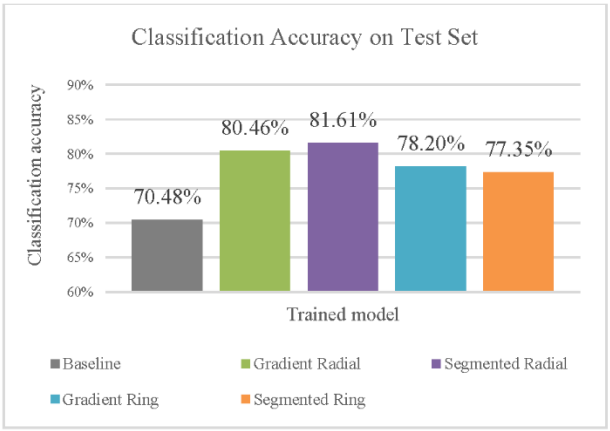


Figure 10: Averaged test accuracy performance scores across all ten-fold cross-validation models trained on each augmented dataset.

C. Pose Class Recall

The improvements that colour-coded keypoint markers afford classification accuracy is most obvious when observed using the averaged recall values depicted in Fig. 11. This illustrates how each augmentation scheme facilitates a model's improved sensitivity toward certain poses. The baseline scores show marginal support for the bending (3%) and crawling (7%) poses. They are also the least represented poses in the dataset but evidently benefit the most from the applied augmentation schemes. Keypoint colours assigned based on the ringed colour wheel provide the greatest increase in recall likely due to its support for encoding size deformations. Since these poses demonstrate a more compact silhouette than any of the other poses, their keypoints acquire a greater concentration of colours within the inner rings of the colour wheel. This results in a unique colour set for the crawling and bending class instances.

The radial colour-based augmentation scheme does not afford much improvement in classification sensitivity. This reveals that the most distinguishing attribute of bending and crawling poses is their compact shape, rather than their orientation. Improvements in recall for other poses namely, standing, sitting, and laying are not as significant because the classifier is likely well attuned to recognising these poses given the abundance of training instances. Thus, any minor improvement in sensitivity may be attributed to fewer poses being confused or falsely classed as one of these prevalent classes.

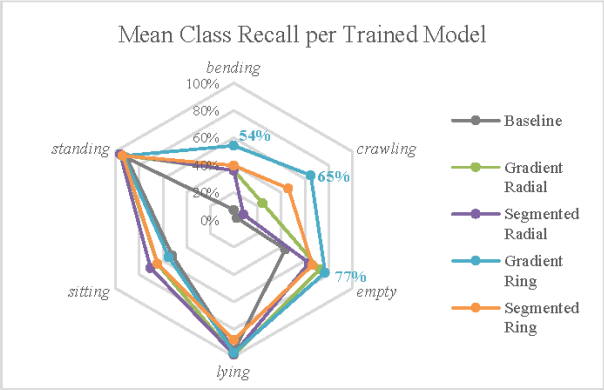


Figure 11: Mean recall performance on the test set for each pose across all ten-fold cross-validation models per augmented dataset.

D. True Positive Class Count

Examining the true positive count can further demonstrate how assigning a colour to keypoints based on a colour wheel encodes discriminative features into the pose representation. The counts are compiled in Table 1 and reveal how the structural arrangement of colours supports the classifier in learning to recognise poses that are mainly differentiated based on orientation or body shape deformation. The bending and crawling instances have the highest true positive counts when represented using the ringed augmentation scheme, given its aptitude for encoding variations in silhouette size. Laying and standing poses have the highest counts when keypoints are assigned colours from the radial colour wheel which encodes the characteristic vertical or horizontal orientation of these poses. Interestingly, sitting poses seem to benefit from the radial colour wheel augmentation scheme instead of the expected ringed colour wheel. It would appear that size deformation is not considered a significant feature in its classification and thus demonstrates an underlying and unknown feature set and class association learnt by the CNN.

TABLE I
TRUE POSITIVE COUNTS ON TEST DATASET

Class	Baseline	Gradient Radial	Segmented Radial	Gradient Ring	Segmented Ring
Crawling (37)	1	9	3	**24**	17
Bending (83)	6	30	30	**45**	33
Empty (205)	89	149	130	**158**	136
Lying (577)	552	**570**	567	561	507
Standing (657)	595	621	**631**	614	616
Sitting (785)	409	507	**552**	431	504

VI. CONCLUSION

Pose recognition conducted as an image classification task is encumbered by the loss of spatial information given the nature of 2D imaging. The experiments conducted as part of this study show that it is possible to augment the representation of a pose to partially recover lost spatial cues related to the orientation and changes in the shape of the human body. The projection of keypoints onto a human silhouette and positional colour association within a colour wheel demonstrates the ability to mimic spatial cues. These cues emphasise postural changes that would otherwise go unnoticed in the abstracted feature space of a CNN image classifier. The structural arrangement of these colours in either a radial or ringed pattern encodes a similarity among instances of respective pose classes based on changes in orientation or shape. These colour-based signals can be leveraged by a CNN to better recognise poses that are characterised by these features. In addition, this form of augmentation can be readily applied to an incoming data stream, thereby helping to enhance a pose-dependent implementation that may require accurate and real-time classification when performing a time-critical task.

The influence that the level of granularity (facilitated by the number of colours) has on classification performance could not be demonstrated using this dataset. Its influence would likely be most apparent when combating the distortion of different viewing angles that the placement of a camera may introduce. An inclined viewing angle from a high vantage point diminishes the cues for silhouette deformation while emphasising changes in orientation. Incorporating more granularity into the projected keypoint colours may offer more differentiation among poses that the viewing angle negates.

REFERENCES

[1] T. Banerjee, M. Enayati, J. M. Keller, M. Skubic, M. Popescu, M. Rantz, "Monitoring patients in hospital beds using unobtrusive depth sensors", in 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2014, pp. 5904-5907.

[2] K. Lai, J. Konrad, P. Ishwar, "A gesture-driven computer interface using Kinect", in 2012 IEEE Southwest Symposium on Image Analysis and Interpretation, 2012, pp. 185-188.

[3] T. Sato, T. Harada, T. Mori, "Environment-type robot system 'Robotic Room' featured by behavior media, behavior contents, and behavior adaptation, *IEEE/ASME Transactions on Mechatronics*, vol 9, no. 3, pp. 529-534, Sep. 2004.

[4] R. Komatireddy, A. Chokshi, J. Basnett, M. Casale, D. Goble, T. Shubert, "Quality and quantity of rehabilitation exercises delivered by a 3-D motion controlled camera: A pilot study", *International Journal of Physical Medicine & Rehabilitation*, vol. 2, no. 4, pp. 1-14 , 2014.

[5] M. Stommel, M. Beetz, W. Xu, "Model-Free detection, encoding, retrieval, and visualization of human poses from kinect data", *IEEE/ASME Transactions on Mechatronics*, vol. 20, no. 2, pp. 865-875, Apr. 2015.

[6] G. Pons-Moll, B. Rosenhahn, Model-based pose estimation. In *Visual analysis of humans*, London, England, Springer-Verlag, 2011, pp. 139-170.

[7] V. Buhrmester, D. Münch, D. Bulatov, M. Arens, "Evaluating the Impact of Color Information in Deep Neural Networks", in Iberian Conference on Pattern Recognition and Image Analysis, 2019, pp. 302-316.

[8] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, London, England: MIT press, 2016.

[9] K. Adhikari, H. Bouchachia, H. Nait-Charif, "Activity recognition for indoor fall detection using convolutional neural network", in 2017 Fifteenth IAPR International Conference on Machine Vision Applications, 2017, pp. 81-84.

[10] Z. Zivkovic, F. Van Der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction", *Pattern Recognition Letters*, vol. 27, no. 7, pp. 773-780, May 2006.

[11] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172-186, July 2018.

[12] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition", in International Conference on Learning Representations, 2014.

Jaco du Toit received his B.A. in Language Technology in 2017 and his B.Sc. Hons. In Computer Science in 2018 from the North-West University, Potchefstroom Campus. He is presently studying towards his Master of Science degree at the same institution in the field of Artificial Intelligence.

ANNEXURE B: CONFERENCE PAPER PRESENTED AT SATNAC 2019

Heuristic Data Augmentation for Improved Human Activity Recognition

J.S. du Toit, J.V. du Toit, H.A. Kruger

This paper was presented at the
SATNAC 2019 Conference, hosted at the
Fairmont Zimbali Resort, Ballito,
KwaZulu-Natal, South Africa
1-4 September 2019

https://doi.org/10.25388/nwu.23284544

Recognised as first runner-up for the outstanding student paper award.

Heuristic Data Augmentation for Improved Human Activity Recognition

J.S. du Toit[1], J.V. du Toit[2], H.A. Kruger[3]

School of Computer Sciences and Information Systems
North-West University, Potchefstroom Campus
Private Bag X6001, Potchefstroom, 2520, South Africa
Tel: +27 18 2992548, Fax: +27 18 2992570

[1]DuJToit@gmail.com

[2]Tiny.DuToit@nwu.ac.za

[3]Hennie.Kruger@nwu.ac.za

Abstract— **Human pose estimation has been an important area of research in recent years due to its applicability in various everyday scenarios. Video surveillance is one such application where people detection and pose analysis can support safety and risk monitoring. Modern pose estimation implementations have attained notable success in terms of accuracy and efficiency thus affording the opportunity for viable solutions that enable such systems. In this study, data augmentation is employed to help improve real-time human activity recognition such as fall detection and other smart surveillance applications. Different augmentation techniques are applied to joint locations recovered from a pose estimator. The representation of the mapped joints is altered to include colour, the blending of colours where joints overlap, and the tinting of colours based on the degree of confidence for their approximated position. These augmentations and their combined effect are comparatively evaluated to yield the pose representation that is most beneficial for image classification when using a convolutional neural network. The greatest improvement in classification accuracy of 5% was attained using a combination of the examined augmentation techniques. These techniques can be diversely applied in different monitoring applications to enhance the detection of machine-learned behavioural patterns defined by pose estimations.**

Keywords— **data augmentation, image classification, pose estimation, smart surveillance**

I. Introduction

The demand for autonomous systems and the convenience that they afford everyday life is what drives the advances being made in technology. Breakthroughs in this field provide solutions to current problems especially in areas where systems can be freed of human-related vulnerabilities such as fatigue and human error. Safety and security as well as healthcare belong to such areas that lead in their demand for improved technologies and smart self-operating systems [1].

One such valuable area of innovation is assistive technology. Modern population growth and the increased life expectancy chiefly contribute to its increased demand, most notably among the growing elderly population [2]. Fall detection has been an active area of research since major risks are associated with seniors experiencing falls late in life [1]. Past solutions that ensure timely aid is provided in the event of a fall have had varying degrees of success since they are generally dependent on the technology's perceived ease of use, intrusion on privacy, and overall implementation costs [1]. However, fall detection realised through smart surveillance can operate independent of human interaction and has recently proven capable of achieving feasible accuracy [3].

Human pose estimation is utilised in most such systems where its purpose is to estimate the localisation of joints on a detected human body within an image. Its incorporation extends to implementations that aim to accomplish human-computer interaction, human behaviour analysis, and other forms of automated video surveillance [4]. Many of these implementations rely upon the skeletal mapping of the human frame to accurately recognise and analyse the posture or actions being performed by a given individual.

The work presented in this study examines data augmentation techniques on pose estimation data with the aim to improve classification accuracy. These techniques are evaluated on pose classification using a convolutional neural network (CNN) and can be extended to fall detection and other pose-related classification problems. The augmentation techniques examined in this study supplement the key point mapping of the human skeleton produced by OpenPose [5] with additional information which is predicted to enhance the discriminative features of the pose representation when performing classification. Varying amounts of heuristic information is encoded by each technique and are comparatively evaluated in order to determine which modifications can aid in human activity and pose detection.

The remainder of this paper is organised as follows. In Section II advances made in the field of data augmentation with respect to image classification is presented. In Section III the structural properties of CNNs and the advantages they provide for image classification are considered. The OpenPose implementation is discussed in Section IV where its efficient operation and accuracy provide motivation for its implementation for pose estimation. The dataset used to train the pose detection classifier is discussed in Section V. In sections VI and VII, the experimental design and evaluation of each form of data augmentation that offers the best separation among image samples of people in sitting and standing poses are explored. The subsequent results provide evidence for the inclusion of data augmentation and other possible future adaptions for pose estimation reliant systems and are discussed in Section IX. Finally, the paper concludes in Section VIII.

II. Research focus and contribution

Research on vision-based fall detection has grown popular in recent years with the advances made in computer vision technology and the advent of deep learning [3]. The general aim of modern implementations involves the improvement of detection accuracy and has more recently extended to include the diminishing of their computational complexity for more efficient detection. These goals are relevant for most research that pertains to image classification. Various methods and techniques that help to achieve such goals have been explored and documented within the field of machine learning and data augmentation is one such method.

Common practices when employing a CNN-based classifier involve applying augmentation to the dataset in order to artificially expand or enhance the training data in a bid to improve its performance. CNN-based pose estimation implementations have shown to benefit from common augmentation practices such as shifts, rotations, scaling, and transforming of its training data, all of which aid the classifier in generalising well and inhibiting overfitting [6]. Such methods yield a satisfactory classification accuracy and are thus deemed sufficient.

Other techniques are not as commonly practiced but have been proven to be beneficial when dealing with complex classification tasks as demonstrated in the work of Okafor *et al.* [7]. Their study evaluates the improvement in classification on different datasets which are respectively made up of images of a variety of Croatian fish species, images different bird species, and pasture aerial shots where some images include grazing cows. These image sets were intended for the development of respective classifiers to perform object identification. Augmentation was applied in in these tasks to achieve colour constancy within the images in order to preserve colour information. This technique is most applicable to objects with similar appearance across each of its instances or where objects are presented under varying illumination conditions [7]. Another example of colour constancy applied as a method of data augmentation is that of Galdran *et al.* [8] where skin lesion analysis utilising a neural network is performed and the colour information is normalised for its training data.

Colour information has been proven to be a contributing factor when performing image classification. Zeiler and Fergus [9] proposed an architecture that performs deconvolution and un-pooling that was used to visualise the learned feature maps of a trained CNN's intermediate convolutional layers. When demonstrated on a renowned deep CNN implementation, AlexNet [10], visualisations revealed that colour information which was generally considered a low-level feature, relatable to edge and shape detection learned in shallow levels of a CNN, showed up in deeper succeeding layers as mid-level learned features. This alludes to colour being a discriminating feature for image classification. Krizhevsky *et al.* [10] also employed data augmentation for their AlexNet classifier which achieved state-of-the-art classification accuracy on the benchmark ImageNet dataset in 2012. Augmentation in the form of conventional label-preserving transformations were applied to the image set along with alterations to the intensity of each image's RGB channels which both helped to artificially enlarge the training set and diminish overfitting. The augmentation of colour exemplifies an important property of object recognition in that an object's identity and/or classification is independent of changes in the intensity and colour of its illumination [10] making it a viable data augmentation practice for image classification.

The contribution of this paper is thus the adoption of data augmentation in pose representation in order to improve pose classification with the eventual goal of being utilised in a fall detection system. In addition, a novel image dataset of people is also assembled which will provide the eventual fall detection classifier with varied samples of individuals in either standing or sitting poses and will later be expanded to include lying poses. It is this compiled dataset that is used to evaluate each form of data augmentation observed in this study.

III. Convolutional Neural Network Architecture

Convolutional neural networks (CNNs) have proven to be highly successful in many applications by achieving state-of-the-art performance in areas such as computer vision, speech recognition, and natural language processing [11]. Their superior performance lies in their ability to extract high-level abstractions from raw unprocessed input data which effectively represents the input space. This is achieved through iterative learning over large amounts of data and is realised by utilising a distinct hierarchical neural network architecture.

A. Leveraging advantages of CNN

CNNs are a type of artificial neural network that consists of three characteristic neural layers, namely convolutional layers, pooling layers, and fully connected layers. It is the combination of these layers through which characteristic CNN concepts such as shared weights, sub-sampling (pooling), and local receptive fields are realised. These concepts instil CNN's with some degrees of shift (translation), scale, and distortion invariance [12] all of which contribute greatly to its robust performance.

The work done in [13] showcases CNN architectures as being inherently translation equivariant which can be contributed to weight sharing where the same weight vector is reapplied for every convolution of the input image or subsequent input vector. Thus, the weights learnt for a given feature map (filter) will be invariant to position since the same detection is evaluated at various locations in the input image or feature vector [14]. Pooling itself (particularly max pooling) approximates translation invariance since the exact position of each of the pooled pixels are not explicitly regarded by succeeding layers in the CNN in that the output vector condenses the image according to the most important and distinguishing detected features for the given receptive field [14]. That is to say if the value being pooled, whether shifted in any direction, remains within the receptive field then the pooling layer will still output the same result. These qualities allow for object detection to remain possible and accurate regardless of its position within the image frame. This quality can be capitalised on in automated video surveillance where the subject may be positioned anywhere within the camera's field of view.

In addition, CNN architectures can be constructed and adapted to take advantage of the discriminating features inherent in a given dataset for the purpose of classification. The composition of a CNN architecture in terms of its depth and the organisation of intermediate layers aids in a trained classifier's discriminative classification capability more so than the selective assignment of filter hyperparameters such as the size and number of filters to be learnt during training. This is supported by the findings of Jarrett *et al.* [15], and is again verified in the empirical approach of Pinto *et al.* [16] where the comparative evaluation of numerous CNN architectures applied to object detection tasks is performed. In like manner, this study makes use of experimentation to determine which CNN architecture best adapts to performing classification of human poses without exploring a comprehensive selection of adaptions to its per-layer hyperparameters.

Although convolution and pooling condense the input image by way of feature extraction, convolution does however preserve spatial information among local pixel arrangements. If the relative position of a captured lower-level feature (e.g. an eye) attributes to the image's classification in some way (e.g. classification of a human face), then the CNN will tend to capture this information in a deeper layer (e.g. the positioning of the eyes above the nose, or to the side of the ears). This characteristic can be benefited from in human activity recognition and pose classification since the kinematic limitations of the human body (e.g. the fixed points at which limbs attach to the torso) and the positional organisation of detected joints (i.e. the wrist being located after the elbow on the arm) obtained through pose estimation can jointly help indicate a person's current posture or stance.

B. CNN internal operation

A CNN model can be made up of three different types of layers, namely a convolutional layer, pooling layer, and a fully connected layer as illustrated in Figure 1. The first set of layers prior to the fully connected layer act as a hierarchical feature extractor which can be composed of numerous stacked convolutional and pooling layers.

Further referring to Figure 1, the input image (represented as a matrix of pixel values and colour channels) is passed to the CNN in its raw state. The convolutional layer learns a predefined number of filters/feature maps which function with a receptive field smaller than the image's dimensions. The filters sequentially convolve across the entire height and width of the input image in strides by computing the dot product between the filter weights and the pixel values which yields the first iteration of a feature vector. The convolutional layer thus learns filters that activate when it detects a certain feature at some position in the input image. These weights are adjusted during backpropagation to be sensitive to certain patterns.

Typically, the convolved feature vector is then passed to a subsequent pooling layer. Pooling is a form of non-linear down sampling which reduces the input vector's dimensionality and condenses it to hold key elements that aid in the image's eventual classification. When convolution and pooling layers are stacked, as represented by the ellipse between the pooling layer and the convolutional layer, then the operations will be repeated over the entire input volume, further condensing its dimensionality. The final layer can be a single fully connected layer or a set thereof, each with a defined number of neurons implementing a general-purpose classifier over the features extracted from the previous layers. This final fully-connected layer outputs the classification of the initial input image as a confidence value for each possible class. Both the convolutional layers and fully-connected layers employ non-linear activation functions such as the popularised *Softmax* and rectified linear unit (*ReLu*), which essentially provides the classifier with the capability of learning high-level abstractions provided enough layers are present within the architecture [12].

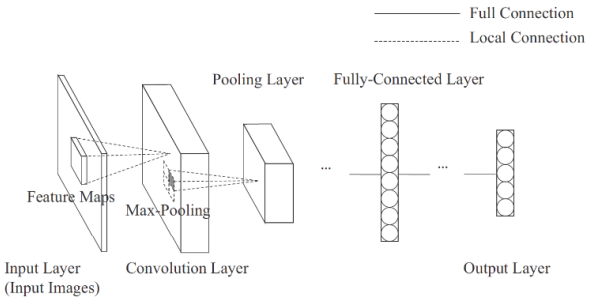


Figure 1: Example convolutional neural network architecture containing a convolutional layer, pooling layer, and fully-connected layers.

Recent advances in deep learning allow for further performance benefits through adaptions to the convolutional architecture in the form of batch normalisation layers, dropout, exponential learning rate decay, and Adam optimisation in place of stochastic gradient descent. Such enhancements will be considered in future implementations as described in Section IX.

IV. OpenPose

Pose estimation or human pose recovery refers to approximating the kinematic structure of a person from a sensor input such as a camera. Wei *et al.* [17] adopted a detection-based strategy with a technique they refer to as Part Affinity Fields which bases the location of each joint on a generated heatmap of its possible locations. Their deployed system, entitled OpenPose, achieves this by means of a deep CNN architecture which yields the XY-coordinates of joints on any detected human body present in an input image or video frame. It is also listed as one of the top ten performing multi-person pose estimation implementations as measured on the MPII benchmark dataset [18].

The position of each joint is realised by refining a heatmap of its most probable locations within the given image. The heatmap of each joint takes into consideration previously generated heatmaps of other joints in order to refine this positional likelihood. The greater the probability surrounding a given joint's location, the greater its heat signature becomes across the corresponding set of pixels. In the end, the pixels with the highest associated degree of probability for each joint's location is then output to a set of normalised XY-coordinates. It is these coordinates and their associated confidence which are used to generate the training and test sets as described in the next section.

In addition, OpenPose achieves accurate pose estimation thus making it an attractive choice for its incorporation as a pose estimator in this study. OpenPose's accuracy can be attributed to its ability to overcome some of the prevalent challenges involved in pose estimation such as the high variability present in a complex articulated object like the human body. This complexity is further expressed in its varied appearance due to differences in clothing, body shapes, inconsistent lighting, image noise, changing backgrounds, variable viewing angles, and the occurrence of occlusions [19].

V. Dataset

The dataset used in this study is made up of images collected from the public domain which were obtained through ordinary internet searches and batch downloading. It contains 37,000 unique images of people standing and 12,500 unique images of people sitting. Each image contains individuals performing either of the two before-mentioned poses in diverse environments as depicted in Figure 2. The dataset is first pre-processed by OpenPose where pose estimation is performed and a skeletal outline based on 18 identified joints is mapped onto the image. These landmark mappings are illustrated in the corresponding images depicted in Figure 3.

Figure 2: Collected dataset samples of sitting and standing poses.

Figure 3: Dataset sample after OpenPose skeletal mapping.

The OpenPose skeletal mappings as depicted in Figure 3 are all extracted to a list of 18 normalised XY-plane coordinates. Each extracted joint mapping also has a degree of associated confidence which expresses the probability of the given joint's position in the image. Some images include occlusions where objects, such as furniture, can obstruct the view of the human body. These joints can thus not be localised by OpenPose and are omitted from the generated mapping.

The XY-coordinates are used to generate a collection of seven different datasets where each incorporates a different form of data augmentation. All of the sets depict the same collection of OpenPose joint mappings where the XY-coordinates are standardised to fit within a 32x32 image frame. Keeping in mind the equivariance property of CNNs, the mappings need not be centred and are rendered in relation to the person's position within the original image. Each generated dataset differs in the amount of additional information encoded in the pose mapping based on the employed data augmentation. The first dataset offers a baseline pose representation in that it is encoded with the least amount of information. It depicts each joint position as a single white pixel on a black background as illustrated in Figure 4.

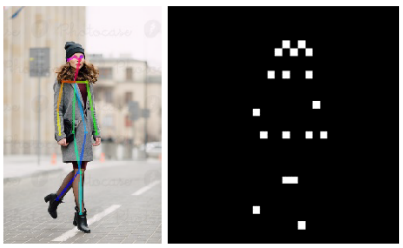
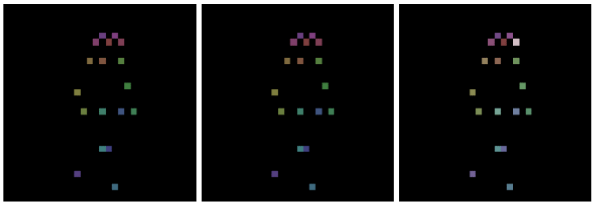


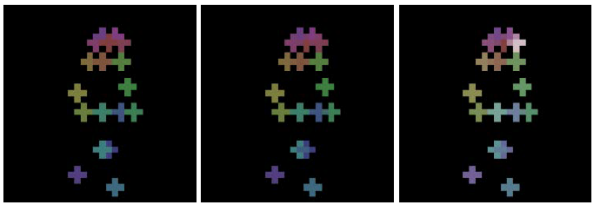
Figure 4: Source data sample (*left*) alongside its generated joint mapping as part of the baseline dataset (*right*).

The next group of generated mappings encode additional information through the use of colour. Each joint is marked as a single pixel with a distinct and joint-specific colour as illustrated in Figure 5(a). The chosen colours are all selected from a uniform distribution of RGB colour gradients with equal distribution. This essentially employs colour constancy since the same colours are reused. Figure 5(b) depicts the same pose but allows for the colours of overlapping joints to be blended, thus preventing obscured joints from going unobserved. Figure 5(c) again depicts the same pose and the same amount of information as in (b) but encodes OpenPose's degree of confidence for the given joint's position through tinting. Lower degrees of confidence result in a lighter shade of the joint's associated colour. Figure 6 represents another three datasets that correspond with the level of encoding expressed in Figure 5, but instead makes use of a larger, crosshair-style marker for the joint positions.



(a) (b) (c)

Figure 5: Corresponding samples of the Dot dataset (a), Dot with Blending dataset (b), and Dot with Blending and Confidence (c).



(a) (b) (c)

Figure 6: Corresponding samples of the Crosshair dataset (a), Crosshair with Blending dataset (b), and Crosshair with Blending and Confidence (c).

VI. Experimental Design

A non-exhaustive approach is taken to determine a CNN architecture that promotes classification accuracy for the

generated datasets. This architecture allows a basis for evaluative comparison between the various pose representations presented in Section V with respect to their influence on classification accuracy. In the leading experiment, a selection of CNN architectures are trained and evaluated on two distinct datasets which comprise of corresponding poses but differ in the amount of encoded information. The first dataset's pose representation is the least informative with only a single white pixel indicating joint positions (Figure 4: Baseline), while the other dataset is considered the most informative (Figure 6(b): Cross with blending) in that it includes colour-coded markers, blending of overlapping markers, and a larger crosshair-style indicator for each of the joint markers. Deducing a beneficial architecture based on the two most dissimilar generated data representations offers a classifier optimised for the classification problem at hand and not a specific pose representation.

For the purpose of this study's set of experiments, each dataset is constrained to 12,500 samples of sitting and standing classes respectively. This allows for a balanced dataset of 25,000 samples and prevents the learning of any classification bias. The two selected datasets are divided into training sets of 20,000 samples and validation sets of 5,000 samples. The same poses are expressed in both corresponding training and validation sets thus facilitating fair comparison.

A total of 32 different CNN architectures are both trained and evaluated on their respective datasets. The model architectures are made up of all combinations of the following set of hyperparameter specifications as listed in Table 1. The number of convolutional layers in each model ranges from a minimum of one to a maximum of four. Filters learned by each convolutional layer are collectively adjusted to powers of 2 for each model. A fully connected layer acting as a classifier is appended to the end of each network where the outputs of this layer yields the class prediction confidence. An additional dense layer is experimentally prepended to the final dense layer with its number of nodes equal the number of learned filters. The goal is to identify the best performing CNN architecture from the given set under consideration.

Convolutional layers	Number of filters	Dense layers
[1, 2, 3, 4]	[32, 64, 128, 256]	[1, 2]

Table 1: Experimental CNN architecture hyperparameter specifications.

VII. Results

The first set of experiments help propose a CNN architecture that can serve as a medium for comparison between the different data augmentations. Keeping in mind that the training sets are relatively small with only 20,000 samples and overfitting is likely to occur early on, each model is evaluated on their respective validation sets after each training epoch. Figure 7 displays the highest average classification accuracy scores across both selected datasets for only the top five performing CNN architectures at 10 and 15 epochs. After 15 epochs, each of the models show clear signs of overfitting.

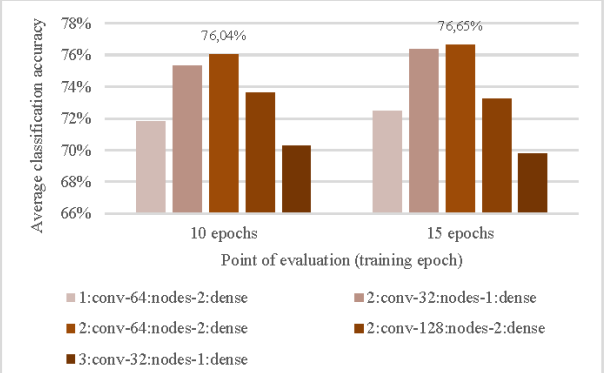


Figure 7: Average classification accuracy across both selected datasets of the top five performing classifiers.

Figure 7 exhibits the *2:conv-64:nodes-2:dense* CNN architecture as being the best adapted to the classification problem in that it obtains the highest average classification accuracy of 76,65% after the 15th training epoch. This architecture is illustrated in Figure 8. It employs two convolutional layers with a filter size of 3x3, a stride of 1, and learns 64 filters in each of the two convolutional layers during training. The structure includes max pooling after each convolution with a pool size of 2x2. After the second and final pooling layer, the feature vector is passed to the set of 3 fully connected layers. *ReLu* is employed as the non-linear activation function in each of the convolutional and fully-connected layers. The last fully connected layer outputs the classification as a confidence relating to each of the two possible classes, namely sitting or standing.

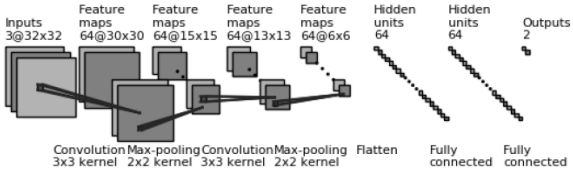


Figure 8: The CNN architecture determined to be the most adapted to the classification problem.

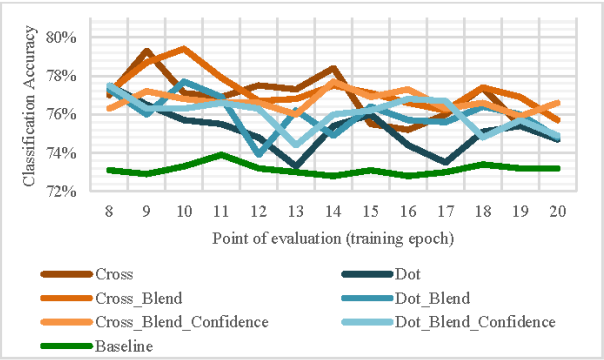


Figure 9: The classification accuracy of each generated dataset evaluated using the ideal CNN architecture.

A model is trained for each dataset using the ideal CNN architecture where the different lines in Figure 9 correspond with the different datasets. The results demonstrate how encoding additional information in the pose representation for

human activity detection is beneficial for improving classification accuracy. This is especially the case when measured against the baseline representation where an increase in classification accuracy of approximately 5% can be achieved. This supports the findings of Zeiler and Fergus [9] in that colour was found to be a discriminating factor for classification in images. In general, the crosshair-style marker proves to be more beneficial than the single dot pixel representation as portrayed in the set of overarching orange lines in Figure 9. This can be contributed to the larger surface area that each joint marker covers which allows a greater probability for joint markers to overlap and blend. These results serve as a basis for further improvements that are discussed in Section IX.

It is worth noting from Figure 9 that encoding the degree of confidence in the joint mapping as derived from OpenPose does not significantly impact nor improve the classification accuracy of either sets of pose representation models. This also supports Krizhevsky *et al.*'s [10] claim in that an object's identity is invariant to adjustments of its RGB intensities such as would be expressed in varying illumination conditions. This property can be taken advantage of when mapping missing joints according to their confidence for the purpose of visual interpretability. It would be possible to interpolate missing joints based on the already detected joints and their own confidences. The interpolated joints can too be encoded with a degree of certainty in the pose representation image before being passed to the CNN for classification. The classifier would inherently learn to also adapt to the interpolator's proposed joint markers even when their positional certainty is low thus making it possible to add more discriminating features without them imposing on classification accuracy.

VIII. Conclusion

The methods for data augmentation explored in this paper provide evidence for adapting pose estimation to both perform and improve human action recognition. Fall detection is a viable application and an area of interest when considering enhancing pose estimation given the increased demand in assistive technology and its reliability in that it operates autonomously. These methods can however also serve other areas of activity recognition where posture or stance play a role. Implementations of this kind can help support safety and risk monitoring where event detection requires an immediate response or intervention.

IX. Future Work

Convolutional neural network-based pose estimation implementations have shown to benefit from common augmentation practices such as rotations, scaling, and transforming of its training data [6]. Such methods can support the learning of invariance to individuals' orientation and aids in good generalisation. One method for achieving this is by rendering images from three-dimensional objects like in the work of [20]. Methods currently exist that allow three-dimensional poses to be recovered from 2D images [21] and can be used to apply rotations and scaling in the three-dimensional domain before being mapped back to a two-dimensional image.

References

[1] X. Yu, "Approaches and principles of fall detection for elderly and patient," in HealthCom 2008-10th International Conference on e-health Networking, Applications and Services, 2008, p. 42-47.

[2] World Health Organization and World Health Organization. Ageing, Life Course Unit, *WHO global report on falls prevention in older age*, Switzerland: WHO Press, 2007.

[3] Y. Birku, H. Agrawal, "Survey on fall detection systems," *International Journal of Pure and Applied Mathematics*, vol. 118, pp. 2537-2543, 2018.

[4] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M.S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing.* vol. 187, pp. 27-48, 2016.

[5] M. Pismenskova, O. Balabaeva, V. Voronin, V. Fedosov, "Classification of a two-dimensional pose using a human skeleton, in *MATEC Web of Conferences*, 2017, pp. 5016.

[6] X. Cui, V. Goel, B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, pp. 1469-1477, 2015.

[7] E. Okafor, L. Schomaker, M.A. Wiering, "An analysis of rotation matrix and colour constancy data augmentation in classifying images of animals," *Journal of Information and Telecommunication,* vol. 2, pp. 456-491, 2018.

[8] A. Galdran, A. Alvarez-Gila, M.I. Meyer, C.L. Saratxaga, T. Araújo, E. Garrote, G. Aresta, P. Costa, A.M. Mendonça, A. Campilho, "Data-driven color augmentation techniques for deep skin image analysis," *CoRR*, vol. 1703, pp. 3702, 2017.

[9] M.D. Zeiler, R. Fergus, "Visualizing and understanding convolutional networks," *European conference on computer vision*, vol. 8689, pp. 818-833, 2014.

[10] A. Krizhevsky, I. Sutskever, G.E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097-1105, 2012.

[11] V. Sze, Y.-H. Chen, T.-J. Yang, J.S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," in *Proceedings of the IEEE*, 2017, paper 105.12, pp. 2295-2329.

[12] L. Zhang, P.N. Suganthan, "A survey of randomized algorithms for training neural networks," *Information Sciences*, vol. 264, pp. 146-155, 2016.

[13] A.M. Saxe, P.W. Koh, Z. Chen, M. Bhand, B. Suresh, A.Y. Ng, "On Random Weights and Unsupervised Feature Learning," *ICML*, vol. 2, pp. 6, 2011.

[14] I. Goodfellow, Y. Bengio, A. Courville, *Deep learning*, London, England: MIT press, 2016.

[15] K. Jarrett, K. Kavukcuoglu, Y. LeCun, "What is the best multi-stage architecture for object recognition?," *IEEE 12th international conference on computer vision*, vol. 4, pp. 2146-2153, 2009.

[16] N. Pinto, D. Doukhan, J.J. DiCarlo, D.D. Cox, "A high-throughput screening approach to discovering good forms of biologically inspired visual representation," *PLoS computational biology*, vol. 5, pp. e1000579, 2009.

[17] S.E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724-4732.

[18] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3686-3693.

[19] T.B. Moeslund, E. Granum, "A survey of computer vision-based human motion capture," *Computer vision and image understanding*, vol. 81, np. 3, pp. 231-268, 2001.

[20] S. Gupta, P. Arbeláez, R. Girshick, J. Malik, "Inferring 3d object pose in RGB-D images," *CoRR*, vol. 1502, pp. 4652, 2015.

[21] A. Agarwal, B. Triggs, "Recovering 3D human pose from monocular images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 1, 2005.

Jaco du Toit received his B.A. in Language Technology in 2017 and his B.Sc. Hons. In Computer Science in 2018 from the North-West University, Potchefstroom Campus. He is presently studying towards his Master of Science degree at the same institution in the field of Artificial Intelligence.

The authors gratefully acknowledge the financial support of this study by the Telkom CoE at the NWU and the National Research Foundation under grant nr TP14081892668.

ANNEXURE C: DATA SAMPLES FROM THE POSE DATASET

Sample images from the **pose dataset** used in the preliminary experiments are illustrated in this annexure. The original images are depicted alongside their key point-augmented versions.

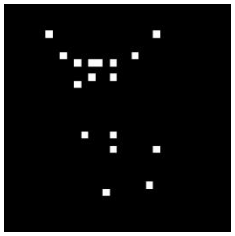
The pose dataset can be downloaded from the NWU research repository:

https://doi.org/10.25388/nwu.23290937

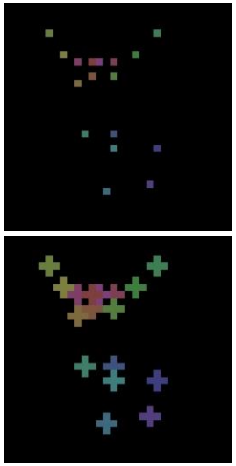
Openpose mapping

Sitting pose

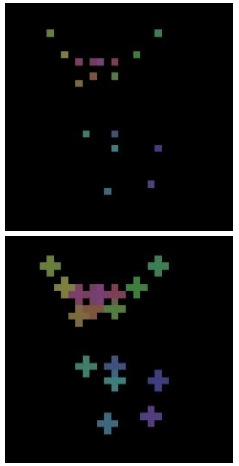
White key points (Baseline)



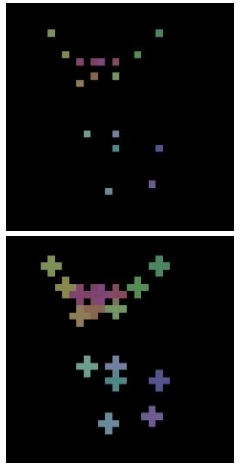
Key point



Key point + Blend

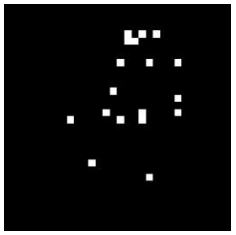
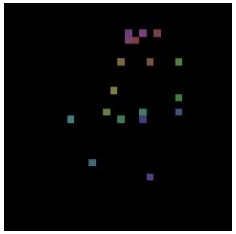
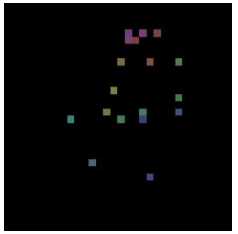
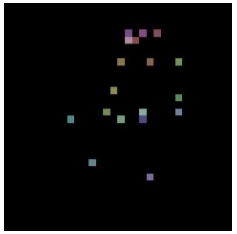
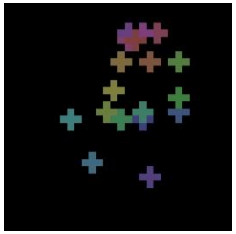
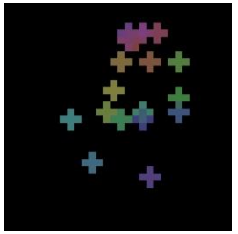
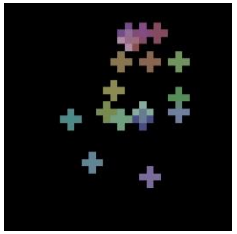
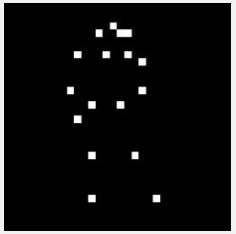
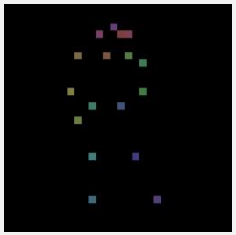
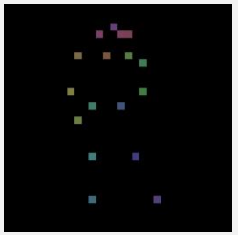
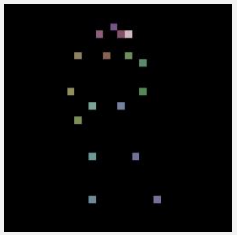
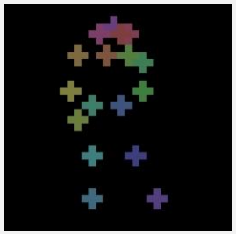
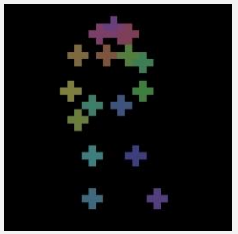
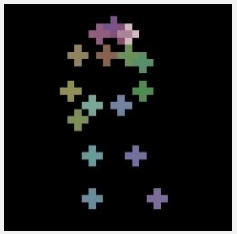


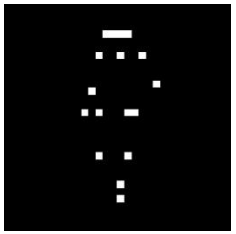
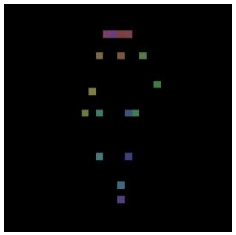
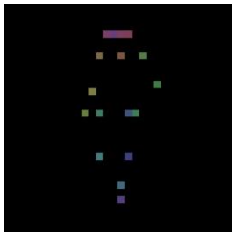
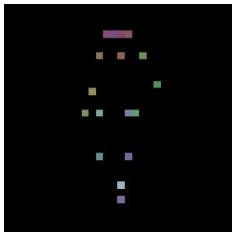
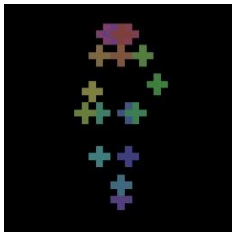
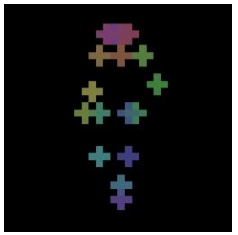
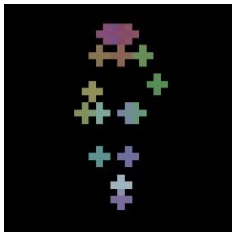
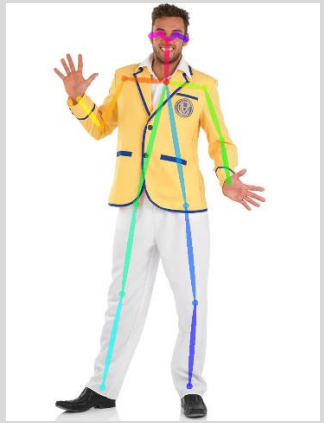
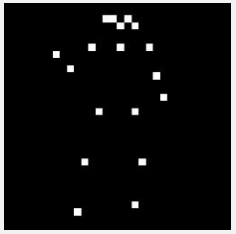
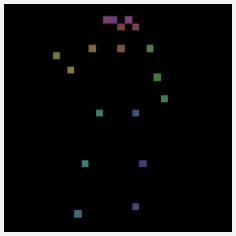
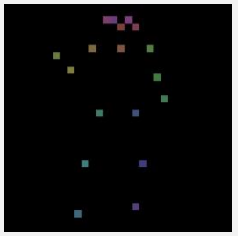
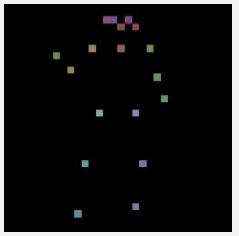
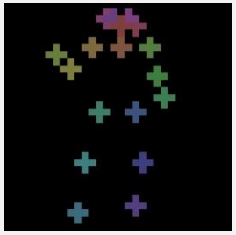
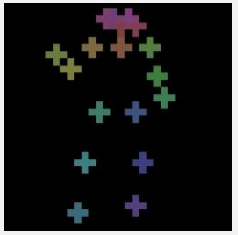
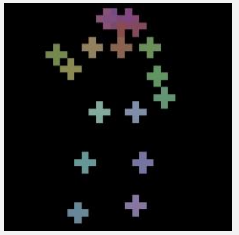
Key point + Blend + Confidence



Dot

Crosshair

Openpose mapping	White key points (Baseline)	Key point	Key point + Blend	Key point + Blend + Confidence	
 *Sitting pose*					Dot
					Crosshair
Openpose mapping	White key points (Baseline)	Key point	Key point + Blend	Key point + Blend + Confidence	
 *Standing pose*					Dot
					Crosshair

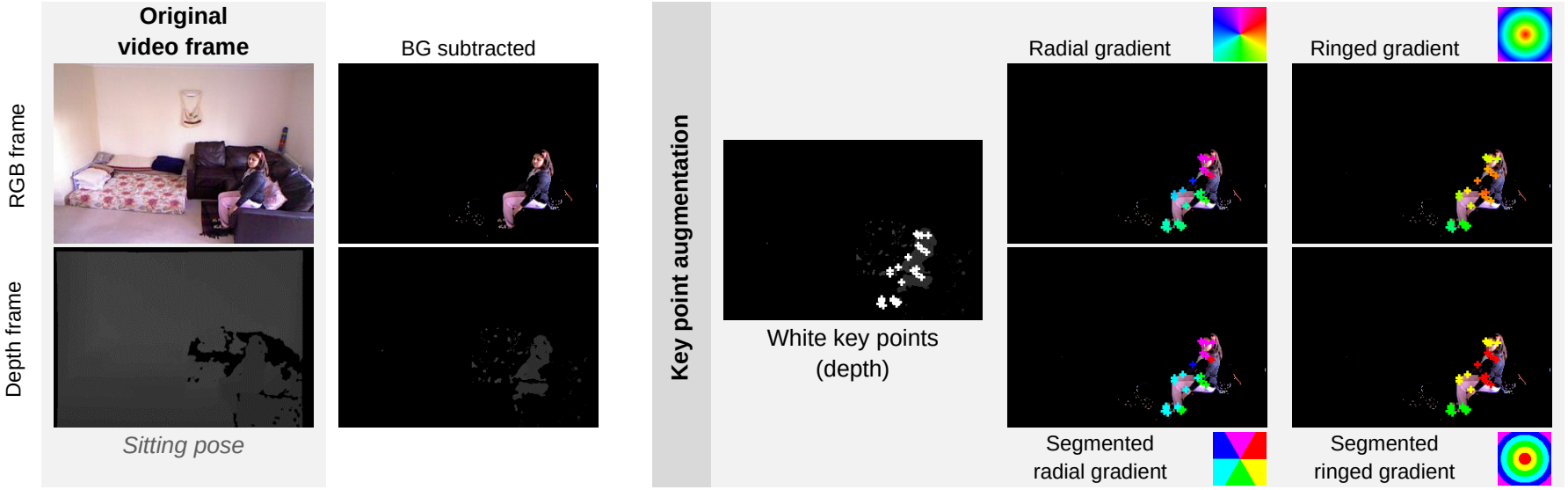
Openpose mapping	White key points (Baseline)	Key point	Key point + Blend	Key point + Blend + Confidence	
 *Standing pose*					Dot
					Crosshair
Openpose mapping	White key points (Baseline)	Key point	Key point + Blend	Key point + Blend + Confidence	
 *Standing pose*					Dot
					Crosshair

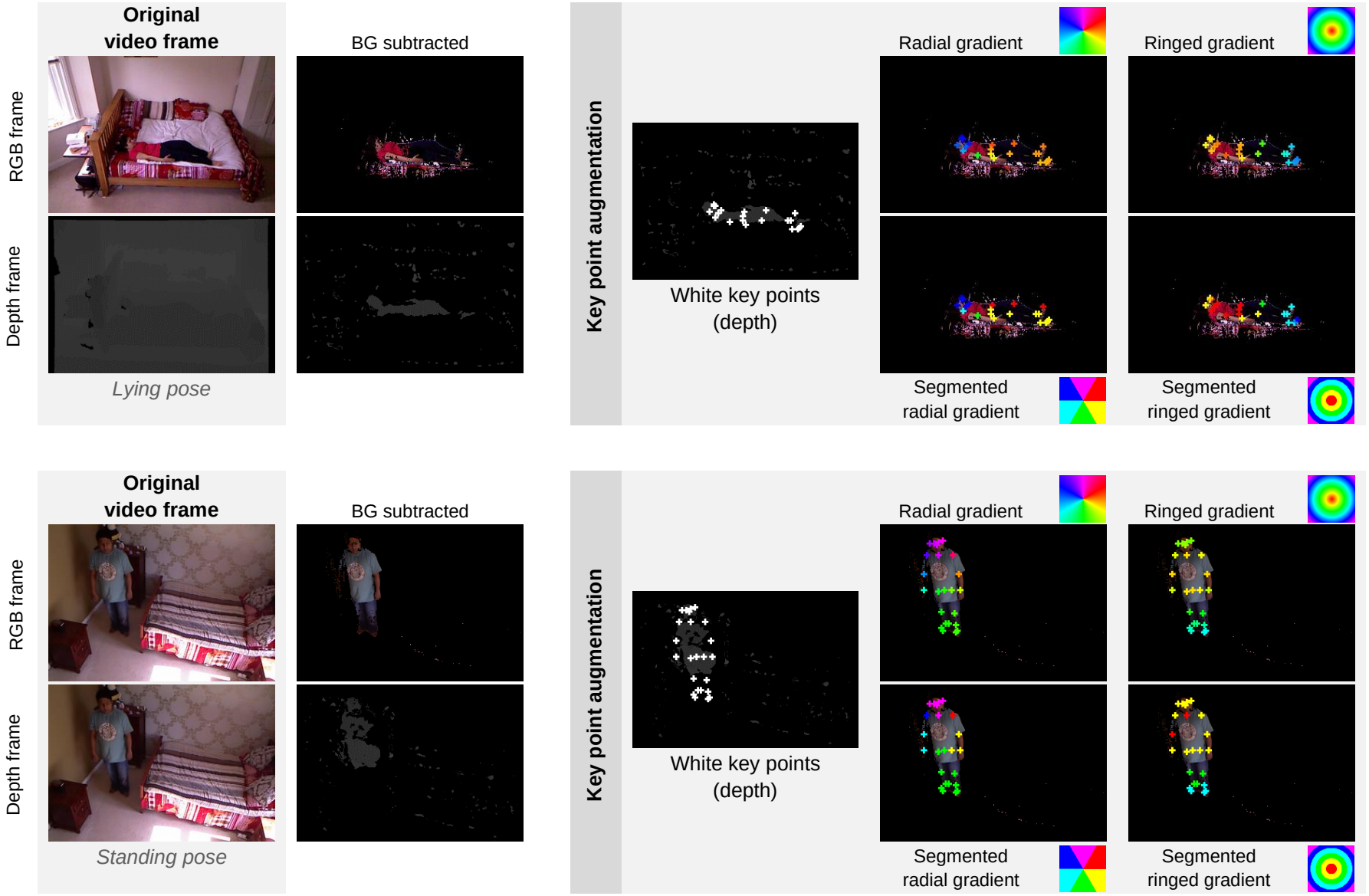
ANNEXURE D: SAMPLES FROM THE FALL DATASET

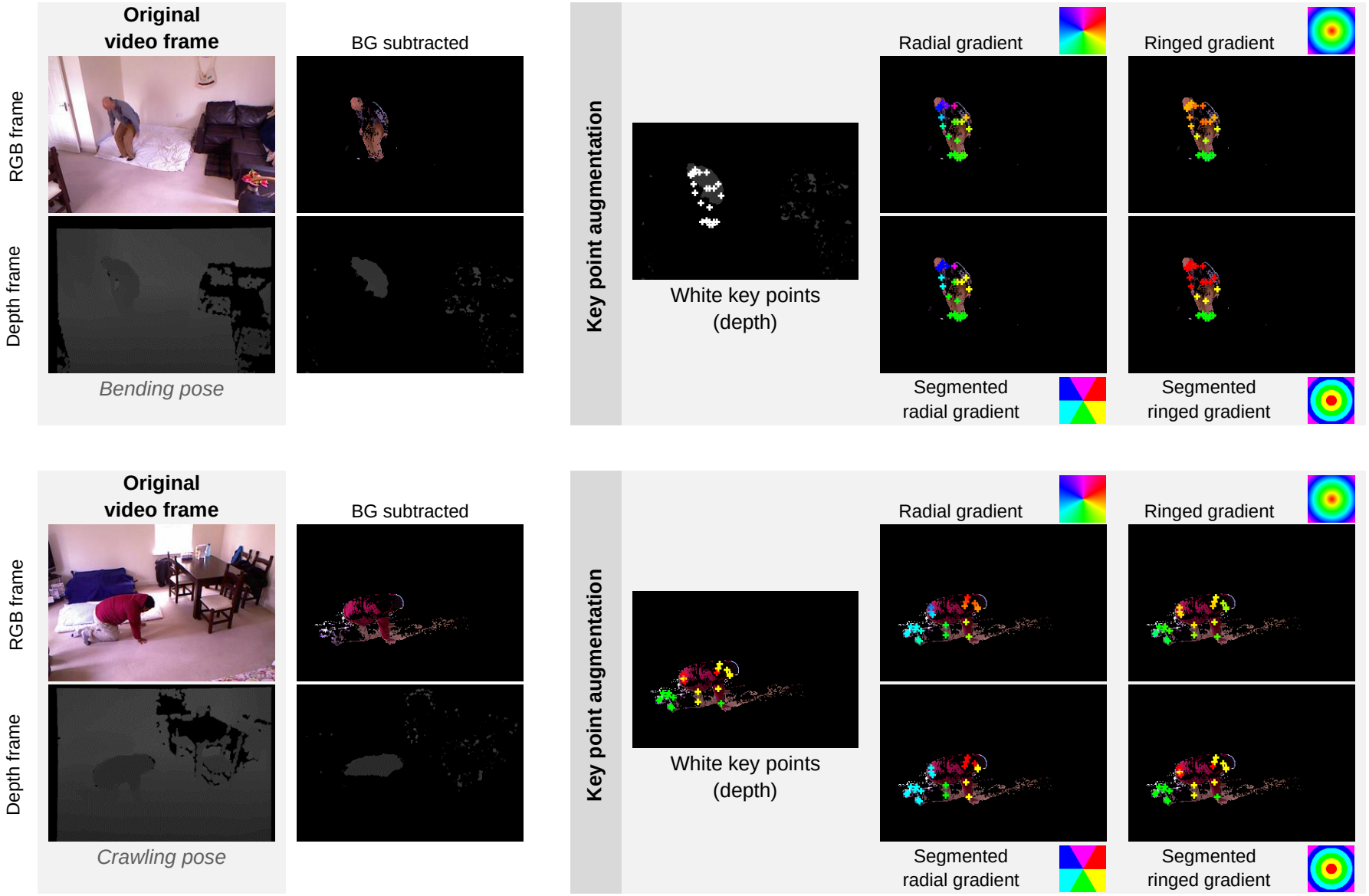
Sample images from the fall dataset used in the primary experiments are illustrated in this annexure. The original RGB and depth video frames are depicted alongside their background-subtracted counterparts and the key point-augmented versions.

The fall dataset can be downloaded from its original host website:

https://falldataset.com/




Original video frame
BG subtracted
RGB frame
Depth frame
Lying pose
Key point augmentation
White key points (depth)
Radial gradient
Ringed gradient
Segmented radial gradient
Segmented ringed gradient
Original video frame
BG subtracted
RGB frame
Depth frame
Standing pose
Key point augmentation
White key points (depth)
Radial gradient
Ringed gradient
Segmented radial gradient
Segmented ringed gradient


Original video frame
BG subtracted
RGB frame
Depth frame
Bending pose
Key point augmentation
White key points (depth)
Radial gradient
Ringed gradient
Segmented radial gradient
Segmented ringed gradient
Original video frame
BG subtracted
RGB frame
Depth frame
Crawling pose
Key point augmentation
White key points (depth)
Radial gradient
Ringed gradient
Segmented radial gradient
Segmented ringed gradient

ANNEXURE E: PYTHON CODE FOR NEURAL POSE ARCHITECTURE

The project files for the preliminary and primary experiments are hosted on **GitHub**. The project is also accessible through the **NWU research repository**.

The project files are available from the NWU research repository:

https://doi.org/10.25388/nwu.23614977

Repository table of contents

Preliminary experiments

- **script.A1.DataGeneration.py**
  Converts *OpenPose* coordinates into data samples. Pose descriptors are projected onto a black background.
- **script.A2.CNNExploration.py**
  Generate, train, and evaluate 32 CNN architectures with various hyperparameter and layer configurations on two datasets (*baseline & crosshair + blend + confidence*).
- **script.A3.CNNPoseClassifier.py**
  Generate, train, and evaluate the preferred CNN architectures (*2:conv-64:nodes-2:dense*) and on all generated datasets.

Primary experiments

- **script.B1.DataGeneration.py**
  Converts *OpenPose* coordinates into data samples. The fall dataset video frames are individually augmented with pose descriptors.
- **script.B2.VGGNetPoseClassifier.py**
  Replicate the VGGNet CNN architecture. Train and evaluate models on each of the augmented fall datasets.

ANNEXURE F: CONFIRMATION OF LANGUAGE EDITING

RUTH COETZEE
Accredited Text Editor (English)
Full member: Professional Editors' Guild NPC
Academic specialist

34 Heritage Village
Tzaneen 0850
Cell: 072 9339417
Home: 015 0650145
Email: ruthc111@gmail.com

3 June 2023

I am an experienced English language editor, accredited by the Professional Editors' Guild, South Africa.

I hereby confirm that I have completed a language edit of the master's thesis written by **Jaco du Toit** titled:

Novel data augmentation schemes for pose classification using a convolutional neural network.

The work was edited to achieve

- clarity of expression and style;
- accuracy of grammar, spelling and punctuation;
- consistency in all aspects of language and presentation.

The author was requested to attend to suggestions for improvement of the text, and is responsible for the quality and accuracy of the final document. References were not included in the language edit.

RCoetzee

Ruth Coetzee (Mrs)

Professional EDITORS' Guild 30 1993–2023
Promoting excellence in editing

Ruth Coetzee
Accredited Text Editor (English)

National treasurer
Membership number: COE004
Membership year: March 2023 to February 2024

ruthc111@gmail.com
treasurer@editors.org.za

www.editors.org.za