2025 3rd International Conference on Computer Graphics and Image Processing (CGIP) | 979-8-3315-3790-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/CGIP66596.2025.00028

Fall Detection Using Vision Transformer

Nitu Saha
Department of Computer Science and Engineering (IoT, CS, BT), University of Engineering and Management, Kolkata, India
nitu.saha@uem.edu.in

Tanmoy Kundu
Atria Institute of Technology, India
tanmoykundu1024@gmail.com

Arunima Banerjee
Department of Computer Science and Engineering (IoT, CS, BT), University of Engineering and Management, Kolkata, India
banerji.oru@gmail.com

Ahona Ghosh*
Department of Computer Science and Engineering (IoT, CS, BT), University of Engineering and Management, Kolkata, India
ahonaghosh95@gmail.com
*Corresponding author

Aritrik Ghosh
Department of Computer Science and Engineering (IoT, CS, BT), University of Engineering and Management, Kolkata, India
aritrik.ghosh2005@gmail.com

Abstract*—One of the main reasons older people get hurt and end up in the hospital is falls, which can have long-term effects on their physical and mental health. With the widening ageing population worldwide, the demand for advanced and reliable fall detection systems becomes progressively important in supporting independent life and timely medical intervention. Traditional solutions, like wearable sensors or manual observation, tend to have low accuracy or are even characterized by usability challenges, particularly for elderly people who might be hesitant to wear or forget such devices. The paper describes an automated approach to human fall detection using a Vision Transformer (ViT) model. This deep learning model has recently proven to be a strong contender over convolutional neural networks (CNNs) for vision tasks. It is trained on images depicting fall and non-fall activities. All images are normalized and scaled prior to being fed into the ViT model, which classifies the image as either a fall or non-fall class. The ability of ViT to capture long-range interdependence and global spatial features is especially effective in understanding human movement between frames. Our findings suggest that the method puts high accuracy and generalization capacity to the test data, even when under different lighting conditions and at different locations. This model is a strong choice for real-time falling detection in smart surveillance systems and care homes for elderly.

***Keywords*— Deep learning, human fall, elderly care, Vision Transformer**

I. INTRODUCTION

Falling is one of the major causes of injury in the elderly, commonly putting their safety, health, and independence at risk. The World Health Organisation (WHO) quotes that millions of older people fall every year. When one falls, immediate help greatly minimises the effect of the injury. That is why automatic fall detection systems are becoming more and more valuable—they provide a means to track individuals in environments such as hospitals and nursing homes, ensuring timely help when help is most needed [3]. A typical fall detection system relies primarily on wearable sensors like accelerometers and gyroscopes [4]. The technologies tend to be wearable, which is uncomfortable and needs human intervention to operate successfully. Vision-based systems offer a non-intrusive approach to processing camera data, thus removing the need for human interaction. Further vision-based approaches utilize classic machine learning models, which extract features from silhouettes or motion patterns [5, 6]. With the emergence of deep learning, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have gained popularity due to their superior performance in capturing time-based and space-based features from video [7, 8]. However, these designs' local receptive fields and sequential structure may restrict their ability to describe sophisticated global interactions inside frames.

Transformer-based models, such as the Data-efficient Image Transformer (DeiT), Shifted Window Transformer (Swin), and Pyramid Vision Transformer (PVT), outperform in various vision tasks by successfully capturing global dependencies through self-attention processes. As they are not that old, their application in fall detection is still unexplored. Vision Transformers (ViTs) have attracted s lot of application areas for their impressive performance in image classification problems, and they're also showing promise in transfer learning. In contrast to conventional models, ViTs split an image into patches and process them through self-attention, that allows the model to learn the overall structure of the scene better. This capability of grasping long-range relationships throughout an image is particularly useful for fall detection, as distinguishing between the spatial location of various body parts is critical for effective detection. Furthermore, ViTs provide flexible and scalable structure, making them a suitable choice for enhancing recognition within niche domains with sparse data—something that is typically true in fall detection research [9].

In this work, we investigate the application of a pre-trained ViT model for fall detection. To make the model learn the task at hand, we fine-tuned it over a labelled dataset of both fall and non-fall images. We then benchmarked its performance against common classification metrics and compared it with those for other transformer-based models. This method not only analyzes how good ViTs are in tackling fall detection complexities but also showcases their applications in real-world scenarios [9]. Furthermore, to describe the VIT output, Gradient-weighted Class Activation Mapping (GradCAM) has been used.

979-8-3315-3790-6/25/$31.00 ©2025 IEEE
DOI 10.1109/CGIP66596.2025.00028