Image Anal Stereol 2024;43:97-107
Original Research Paper

doi: 10.5566/ias.3087

ADVANCING FALL DETECTION UTILIZING SKELETAL JOINT IMAGE REPRESENTATION AND DEFORMABLE LAYERS

HAMZA ERGÜDER✉,[1], TUNCAY UZUN[1] AND MURAT BADAY[2,3]

[1]Department of Electronics and Communication Engineering, Yildiz Technical University, Istanbul, 34349, Turkiye, [2]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, 94305, CA, USA, [3]Precision Health and Integrated Diagnostics Center, Stanford University School of Medicine, Stanford, 94305, CA, USA
e-mail: hamza.erguder@std.yildiz.edu.tr, uzun@yildiz.edu.tr, muratbaday@stanford.edu

(Received November 22, 2023; revised January 30, 2024; accepted February 3, 2024)

ABSTRACT

Falls are a significant concern among the elderly population, with 25% of individuals over 65 years old experiencing a fall severe enough to require a visit to the emergency department each year. Early detection of falls can prevent serious injuries and complications, making it an important problem to address. There are various methods for detecting falls, utilizing different types of sensor input data. However, when considering factors such as ease of setup, accessibility, and accuracy, utilizing cameras for fall detection is a highly effective approach. In this study, a novel video-based fall detection algorithm that relies on skeleton joints is introduced. The results of pose estimation are preprocessed into an image representation and ShuffleNet V2 model with the addition of a Deformable Layer is employed for classification. Experiments were carried out on four distinct datasets: URFD, UP-Fall Detection, Le2i, and NTU RGB+D 60, which encompass individuals engaged in various activities, including falls. The results showcase exceptional performance across all these datasets, affirming the efficacy of the approach in accurately detecting falls in video footage.

Keywords: Computer Vision, Deep Learning, Fall Detection, Pose Estimation.

INTRODUCTION

According to the Centers for Disease Control and Prevention (CDC), falls are the leading cause of injury-related deaths among individuals aged 65. In 2018, there were an estimated 52 million older adults in the United States, and of those, 36 million experienced falls. In excess of 8 million of these falls led to injuries necessitating medical attention or restricting the individual's activity for a minimum of one day. This number is expected to increase in the future, as the older adult population is projected to reach 73 million in 2030. It's estimated that 52 million falls will occur in that year, resulting in 12 million injuries (Moreland *et al.*, 2020). Falls are also a leading cause of hospitalization and long-term care facility admission among the elderly. This can cause physical and psychological trauma and can potentially be life-threatening, particularly for older individuals (Jager *et al.*, 2000; Sterling *et al.*, 2001). The high number of falls and injuries highlights the importance of fall detection and prevention, as early detection can prevent serious injuries and complications.

Given the significant impact of falls on the health and well-being of elderly individuals, the development of fall detection algorithms has been an active area of research. Throughout the years, various techniques for detecting falls have been proposed. One common approach is the use of sensors, which can be worn by the user or placed within the environment being monitored (Nooruddin *et al.*, 2021). However, sensor-based fall detection methods rely on the user to consistently wear the device or stay within proximity range of the sensors. If the user neglects to wear the device or steps away from a stationary sensor, such as a walker with built-in sensors, the system may misinterpret this as a fall and trigger an unnecessary alarm (Delahoz and Labrador, 2014). An alternative strategy involves the use of cameras, capable of offering continuous data without necessitating the user to wear any devices.

In recent years, researchers have used cameras to detect falls using Convolutional Neural Networks (CNNs) on RGB camera images (Martínez-Villaseñor *et al.*, 2019; Espinosa *et al.*, 2019). While image-based approaches have demonstrated promising results, they are limited by the quality and resolution of the video, as well as the presence of occlusions and clutter in the environment (Singh and Vishwakarma, 2018).

Skeleton-based fall detection methods offer a promising alternative to traditional video-based approaches by relying on information about the body's joint positions, abstracting the representation of human movement and sidestepping issues related to video quality, occlusions, and clutter. By focusing on skeleton data rather than visual information, these methods provide a more efficient and robust solution for fall detection, as they are less susceptible to environmental factors. The abstract nature of skeleton data and the continuous improvement in pose estimation algorithms contribute to the accuracy and reliability of these methods. Moreover, the reduced computational requirements and minimized privacy concerns make skeleton-based approaches a compelling choice for fall detection systems.

Notable efforts, utilizing pose estimation algorithms like AlphaPose (Fang *et al.*, 2016), show promise, with single-frame analyses achieving high accuracy (Ramirez *et al.*, 2021; Serpa *et al.*, 2020). However, these studies focus on single-frame to detect falls, and it is very important to acquire information from multi-frame to understand the change of the body over time, since fall action is a temporal event and it is not enough to look only one frame to classify the action as fall. Additionally, single frame methods may correctly detect falls in some frames, but even one false positive detection in a non-fall sequence can greatly impact the prediction of the entire video, making it an inefficient approach.

Research addressing this gap explores temporal information analysis across multiple frames. Studies employing pose estimation models with Transformer and Long Short-Term Memory (LSTM) models showcase improved accuracy, while others using time-based CNN and Gated Recurrent Unit (GRU) based models yield varying fall detection rates (Juraev *et al.*, 2022; Yadav *et al.*, 2022; Taufeeque *et al.*, 2021). Despite high accuracy, potential trade-offs between precision and recall exist, emphasizing the need for optimization. Moreover, computational intensity and time consumption with LSTM and GRU models may limit real-world applications (Weytjens and De Weerdt, 2020).

This paper introduces an innovative approach dedicated to real-time fall detection in camera feeds. The primary objective of our proposed method is to achieve exceptional accuracy in fall detection while maintaining real-time processing with minimal hardware requirements. Our methodology differs from conventional approaches by eliminating the reliance on sensors or wearable devices, addressing and overcoming limitations present in prior literature methods. A distinctive feature of our method is its capacity to outperform existing approaches in terms of accuracy, without compromising on speed. Our extensive experimental analysis conducted across four public datasets (The University of Rzeszow Fall Detection (URFD) Dataset (Kwolek and Kepski, 2014), UP-Fall Detection (Martínez-Villaseñor *et al.*, 2019), Le2i Fall Detection Dataset (Charfi *et al.*, 2013), and NTU RGB+D 60 (Shahroudy *et al.*, 2016)), attests to the superior performance of our method.

Significantly, our study highlights the effective performance of our model and emphasizes the importance of converting temporal skeletal joint inputs into images for a thorough understanding of fall detection. This transformation improves the strength and adaptability of our approach, establishing a standardized representation of human actions across various datasets. The results confirm the effectiveness of our approach, demonstrating that high-accuracy, low-latency fall detection is attainable solely through camera inputs. By presenting these findings, we contribute to existing research and make a significant advancement in methodology.

MATERIALS AND METHODS

METHODOLOGY OVERVIEW

Our fall detection algorithm, harnessing video-based pose estimation, underwent rigorous evaluation across four diverse datasets: URFD, UP-Fall Detection, Le2i FDD, and NTU RGB+D 60. The methodology encompasses multi-step processes, including object detection, multi-object tracking (MOT), pose estimation utilizing YOLO-V8n-pose (Jocher *et al.*, 2023) with BoT-SORT (Aharon *et al.*, 2022), and subsequent image representation of skeleton joints. These image representations were further processed using the ShuffleNet V2 (Ma *et al.*, 2018) model with Deformable Layers (Dai *et al.*, 2017), showcasing enhanced adaptability to complex transformations.

DATASETS

In our exploration of fall detection using RGB videos, we extensively utilized a variety of datasets, each offering distinctive characteristics and advantages for training and validating fall detection algorithms.

The University of Rzeszow Fall Detection (URFD) Dataset (Kwolek and Kepski, 2014) consists of 30 fall sequences and 40 ADL (Activities of Daily Living) sequences captured by two Kinect sensors. It encompasses falls from both standing and sitting positions, while ADL activities include common movements like walking, sitting, and lying down. With different perspectives and actions from both standing and sitting positions, this dataset aids in training models to distinguish falls from various postures. The videos in this dataset have similar lengths, allowing for the direct selection of the same number of frames from each video.

The UP-Fall Detection Dataset (Martínez-Villaseñor *et al.*, 2019) involves 11 activities performed in three trials per activity, gathered using wearable, ambient sensors, and vision devices. It includes various daily activities and different fall types performed by 17 healthy adults. This dataset offers a wide array of activities and fall types, facilitating the training of models to discern falls amid diverse daily activities and environmental settings. In order to create uniformity in video lengths for training, all videos within this dataset have been standardized to a duration of 10 seconds. While the fall videos were originally 10 seconds in length, the remaining videos were condensed to fit this 10-second duration, capturing individuals performing their respective activities. 10-second segments were manually selected to ensure the presence of ADL (Activities of Daily Living) activities. This careful selection is crucial for the fall detection model to discern the distinction between normal and falling videos especially when there are no other differences in video specifications. Each video is labeled as fallen or unfallen based on the presence of a fall.

The Le2i Fall Detection Dataset (Le2i FDD) (Charfi *et al.*, 2013) comprises 131 fall videos and 59 ADL videos. These videos encompass three distinct types of falls (forward falls, balance loss, falls from sitting) and various ADL activities such as sitting, walking, and standing, among others. These activities were performed in different environmental conditions, considering factors like light, clothing, textures, and camera viewpoints. This dataset provides a diverse range of fall scenarios and daily activities in various settings, enabling effective training of models to recognize falls in diverse environmental conditions. For precise training, the dataset has been meticulously labeled to indicate the specific frame at which a fall begins in the videos where falls are observed. To maintain uniformity in video length for training, 5-second videos were created manually where individuals perform falls and other associated activities. This process has led to the dataset being divided into two segments: videos with falls and those without, both of equal duration. Each video is clearly labeled as ”fallen” or ”unfallen” based on the presence or absence of a fall, respectively.

The NTU RGB+D Dataset (Shahroudy *et al.*, 2016) contains 56,880 samples of 60 action classes from 40 subjects, encompassing daily, health-related, and mutual actions performed in 17 different scene conditions using three cameras with varying horizontal imaging viewpoints. This dataset enables the training of models to recognize falls amid various actions and environmental conditions due to its broad spectrum of actions, including specific health-related actions like falling down. The videos in this dataset have similar lengths, eliminating the need for modifications. While training our model with all labels, our primary focus was on evaluating our approach’s performance in detecting falls.

The diversity in environmental settings, fall types, and ADL activities across these datasets provides a robust foundation for our fall detection algorithm. The inclusion of different viewpoints, varied fall types, and diverse activity contexts demonstrates capacity of the approach for generalization and accurate recognition of falls across varying scenarios, enhancing the robustness and reliability of our fall detection system.

POSE ESTIMATION

The detection of skeleton points in our approach is a multi-step process involving Object Detection, Multi-Object Tracking (MOT), and Pose Estimation. The latest YOLO architecture YOLO-V8 (Jocher *et al.*, 2023) model known for its high accuracy and speed on object detection. This model utilizes advanced techniques such as Cross Stage Partial Network (CSPN) and Feature Pyramid Network (FPN) to enhance performance.

Multi-Object Tracking (MOT) is essential for maintaining continuity across frames in scenarios with multiple individuals. MOT involves detecting and predicting the spatial-temporal trajectories of multiple objects within a video stream. BoT-SORT (Aharon *et al.*, 2022), the state of the art in MOT, known for its robustness in challenging situations with crowded scenes and occlusions, combining motion and appearance information for precise tracking.

Pose estimation serves as a pivotal methodology for fall detection, leveraging its ability to encapsulate comprehensive human body structure, posture, and motion information. This technique adeptly captures the spatial and temporal characteristics of human actions, encompassing essential elements like body part locations, orientations, and movements. The significance of pose estimation lies in its capacity to facilitate a multifold analysis of human actions. This analytical versatility allows for a reduction in input data complexity and noise, enabling a focused study on the fundamental features inherent in human actions.

The chosen pose estimation architecture is Ultralytics’ YOLO-V8n-pose model with BoT-SORT, known for its accuracy and speed. This model does not require an additional object detection model and achieves an 80.1 mAP 50 on the MS COCO val dataset (Lin *et al.*, 2014). It runs at 10 ms per frame on an NVIDIA T4 GPU. The YOLO-V8n-pose model provides x and y coordinates for each keypoint, indicating spatial information, along with a confidence score reflecting the accuracy of keypoint detection. This pre-trained model forms the basis for our pose estimation process, extracting crucial information on human body poses necessary for subsequent stages in our fall detection algorithm.

IMAGE REPRESENTATION OF SKELETON JOINTS

The datasets contain videos of varying lengths, capturing falling actions with additional frames of preceding and subsequent activities. To ensure uniformity in training data length for effective comparison, videos underwent specific editing procedures outlined in the datasets section.

Following data editing, we performed pose estimation using multi-object tracking to track the skeleton points of individuals in the video, represented in the widely accepted COCO pose annotation format. This format employs (x, y) coordinates to define 17 keypoints covering crucial body parts such as the head, neck, shoulders, elbows, wrists, hips, knees, and ankles. The pose estimation results, locations of keypoints, were normalized based on the maximum and minimum x and y values of each individual, ensuring a uniform range between 0 and 1. The normalized results were then scaled by 255 to convert them into RGB format.

In situations where multiple individuals are present in a video, certain person-tracking algorithms have the capability to identify and track more than one person concurrently. However, when exclusively tracking the keypoints of a single individual in such instances, there exists a potential oversight in accurately capturing fall events. Moreover, within the NTU RGB+D 60 dataset, activities portraying interactions among multiple individuals further underscore the need for a comprehensive approach.

To address this challenge, our methodology incorporates the inclusion of keypoints associated with other identified individuals into the array. Recognizing that the videos in the datasets under consideration may involve up to 5 individuals, we standardized the arrays to accommodate a maximum of 5 people, ensuring the inclusion of pertinent information in scenarios involving multiple individuals. In instances where the number of detected individuals was less than 5, keypoint values were designated as 0, 0, 0, representing the x, y coordinates, and confidence score. It is important to note that, for datasets with greater crowd density, adjustments to the array length may be made to suit the specific characteristics of the dataset. This adaptive approach ensures a robust and comprehensive representation, particularly when dealing with diverse and dynamic scenarios.

The aggregated results were used to generate images with dimensions corresponding to the number of frames, 17 keypoints of 5 individuals, location and accuracy values. The standardization of image size necessitated a careful selection of frames, and various methodologies for achieving this were considered. These methods encompassed options such as choosing a random interval, employing repetitions appended to the end of the video, and adopting uniform sampling between frames. In the context of these alternatives, the study conducted by Duan et al (Duan *et al.*, 2022). prominently highlighted that uniform sampling yields optimal results. Consequently, our approach aligns with this recommendation, utilizing the uniform sampling method for standardizing image sizes.

This comprehensive approach transformed each video into an image, establishing a standardized representation of human actions. The resulting matrix, with dimensions reflecting frames, joint groups, and position with confidence scores, served as the foundation for our supervised learning-based fall detection algorithm.

The labeled representations, categorized as ”fallen” or ”unfallen,” constituted the training dataset. Converting pose estimation data into an image-like format facilitated the application of established image processing techniques and convolutional neural networks, capitalizing on their efficacy in identifying patterns crucial for accurate fall detection.

SHUFFLENET V2 MODEL WITH DEFORMABLE LAYERS

The moderate depth of ShuffleNet V2, paired with its adeptness in capturing crucial features while avoiding challenges such as vanishing gradients, establishes a balance between computational efficiency and representation capacity. This characteristic makes ShuffleNet V2 an ideal foundation for our fall detection algorithm, ensuring an effective framework.

In our pursuit to enhance the efficacy of the fall detection algorithm, we incorporated Deformable Layers from the Deformable Convolutional Neural Network architecture (Dai *et al.*, 2017). These layers address challenges in handling large, unknown transformations in visual recognition tasks. Deformable Layers improve upon traditional CNN limitations by introducing deformable convolution and deformable Region-of-Interest (RoI) pooling.

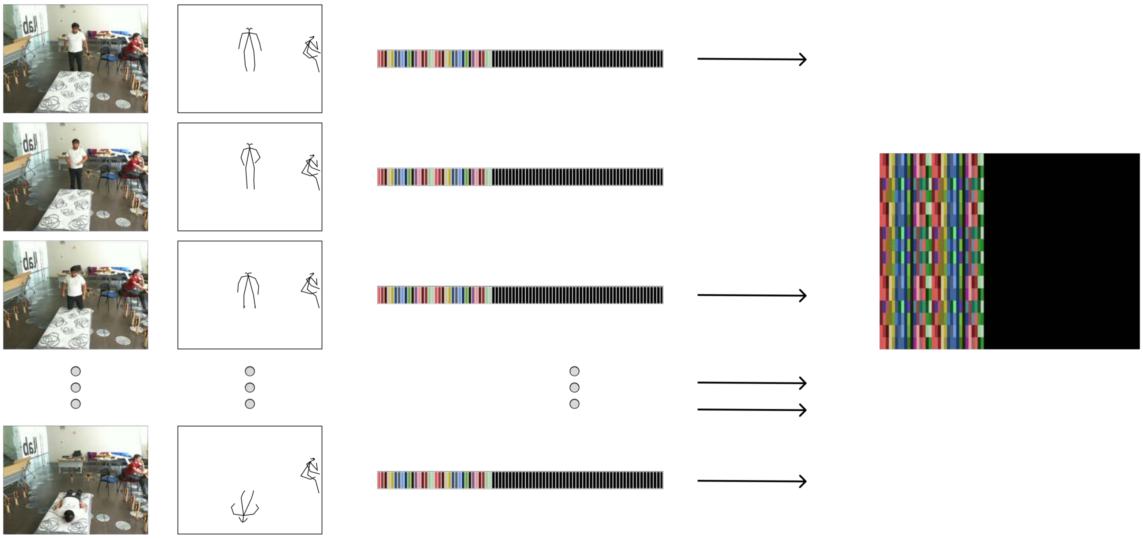


Fig. 1. *Image Representation Processes - A visual depiction of the key steps involved in converting video inputs into standardized image representations for fall detection.*

The deformable convolution allows for flexible sampling, adapting to complex transformations in human actions, facilitating local, dense, and adaptive deformations based on input features. Similarly, the deformable RoI pooling enables a more adaptable localization of body parts, crucial for accurately identifying variations in body poses associated with falls among individuals with different shapes and sizes.

EXPERIMENTS

EVALUATION METRICS

In order to evaluate the performance of our fall detection algorithm, we used several standard evaluation metrics: accuracy (1), precision (2), recall (3), and F1-score (4). These metrics are calculated as follows:

$$A = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

$$P = TP/(TP + FP) \quad (2)$$

$$R = TP/(TP + FN) \quad (3)$$

$$F1 = 2x(PxR)/(P + R) \quad (4)$$

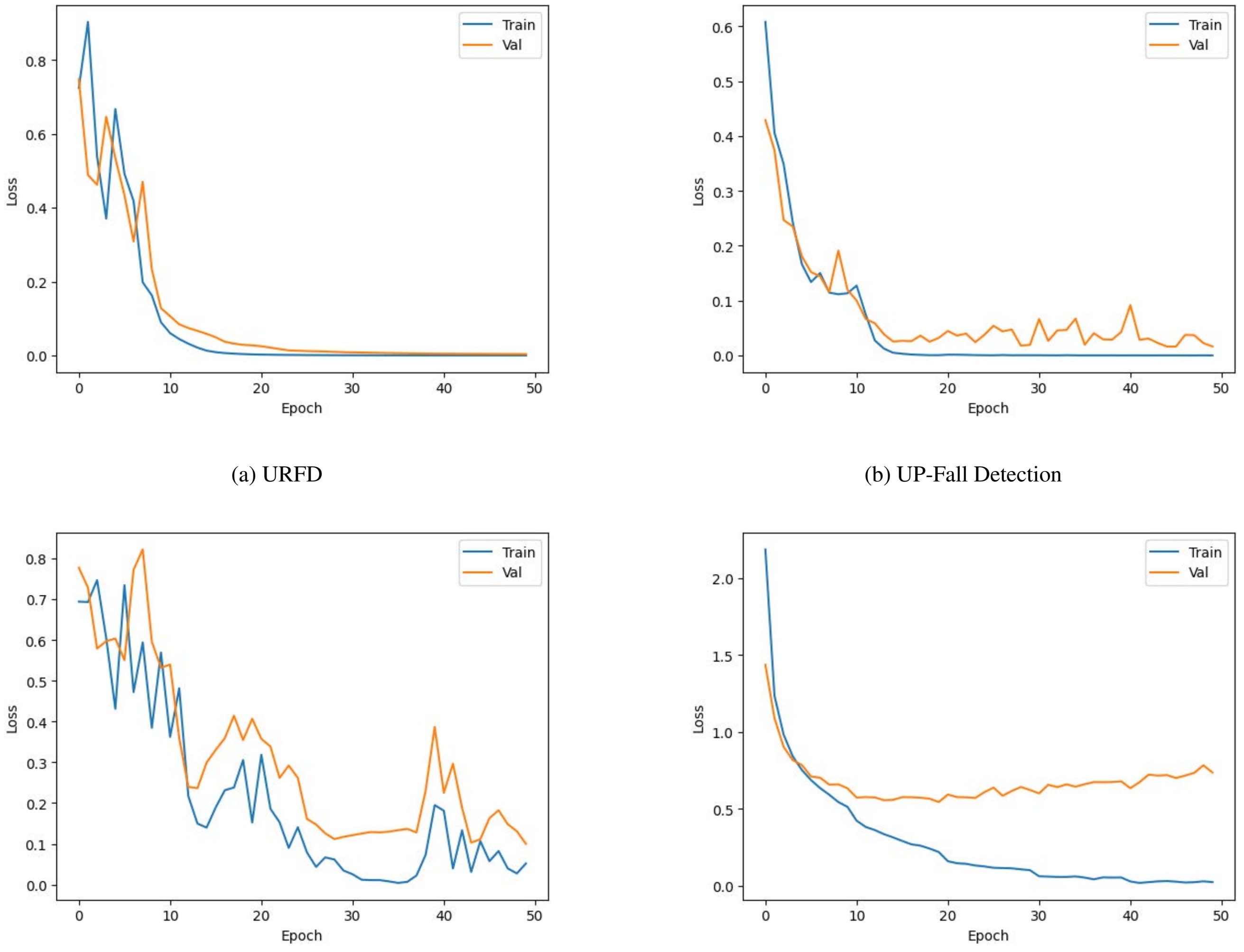
where TP, TN, FP, and FN represent the number of true positive, true negative, false positive, and false negative predictions, respectively.

EVALUATION

Experiments to assess the efficacy of our proposed fall detection approach were conducted on Google Colab (Bisong and Bisong, 2019) using the NVIDIA T4 GPU. The dataset sizes varied, with 70 videos from URFD, 1118 from UP-Fall, 190 from Le2i, and 56880 from NTU RGB+D. A 70/30 train/validation split was applied to all four datasets, ensuring a judicious partitioning strategy for powerful model training and effective performance evaluation. This division ensures a substantial portion for training the algorithm while maintaining a separate set for validation, contributing to the overall reliability and generalizability of the model. Our ShuffleNet V2 model with Deformable Layers underwent training on relevant sets and evaluation on validation sets, with performance metrics including accuracy, precision, recall, and F1 score.

During training, a fixed batch size of 64 was employed. It was found to be effective in achieving a balance between computational efficiency and model convergence. After thorough experimentation, we found that the ADAM (Kingma and Ba, 2014) optimizer provided the best results for our ShuffleNet V2 model with Deformable Layers. With a systematic search, we identified that a learning rate of 0.005 yielded optimal performance. This value was chosen based on its ability to converge effectively during training. To further enhance the learning process, we employed a learning rate step planner with a gamma value of 0.75. This strategy involved decreasing the learning rate by a factor of 0.75 every 10 epochs. This dynamic adjustment contributed to the model's adaptability over the course of training, potentially improving overall performance.

The loss plots in Fig. 2 indicate training over



(a) URFD

(b) UP-Fall Detection

(c) Le2i FDD

(d) NTU RGB+D 60

Fig. 2. *Loss Plots - The training and validation losses for our proposed fall detection method are illustrated across four datasets.*

50 epochs with tracking of loss values. Stabilization around the 30th epoch suggests that the model achieved a certain level of convergence. To strike a balance between model convergence and preventing overfitting, we determined that training the model for 50 epochs produced the best results.

Additionally, for real-time application considerations, our model's computational speed was analyzed. In real-time, the object detection and pose estimation models for each frame take approximately 10 milliseconds on the T4 GPU. The fall detection model, deployed on this subset of frames, operates in less than 1 milliseconds.

MODEL SELECTION

During the model selection process, we assessed the performance of three distinct models. Initially, a basic CNN model was implemented, featuring three convolutional layers and two fully connected layers. However, the results obtained fell short of expectations, as evidenced by the CNN model's accuracy of 82.3% on the Le2i FDD dataset, thereby underscoring its limitations.

Subsequently, our evaluation focused on ShuffleNet V2 and ResNet18 (He *et al.*, 2015), both recognized for their efficacy in feature extraction and recognition tasks. Unexpectedly, despite ResNet18's deeper architecture and increased parameters, ShuffleNet V2 demonstrated marginally superior performance. On the same Le2i FDD dataset, ShuffleNet V2 achieved an accuracy of 95.2%, while ResNet18 reached 94.4%. Given ShuffleNet V2's efficiency, it was selected over ResNet18 and the CNN model.

On the Le2i FDD dataset, integrating Deformable Layers into the ShuffleNet V2 model significantly improved accuracy from 95.2% to 98.95%. This underscores the effectiveness of Deformable Layers in handling intricate transformations in fall detection tasks, empowering the system to accurately identify falls with adaptability to pose variations.

RESULTS

To further substantiate the effectiveness of our proposed model, we conducted a comparative analysis against several other fall detection algorithms, as revealed in Table 1. This comparison showcases the performance of our approach concerning existing methodologies in the literature. The experimental results indicate the exceptional performance of our proposed fall detection algorithm across four diverse datasets.

In the URFD dataset, our method achieved a flawless 100% accuracy, surpassing or matching established approaches (Zhao *et al.*, 2022; Wang *et al.*, 2020; Dentamaro *et al.*, 2021; Li *et al.*, 2022; Zahan *et al.*, 2023; Galvao *et al.*, 2021). Precision, recall, and F1-Score also hit the 100% mark, showcasing the model's robustness in accurately identifying falls with minimal false positives.

Similar success was observed in the UP-Fall Detection dataset, where our algorithm outperformed other methods (Zahan *et al.*, 2023; Yadav *et al.*, 2022; Taufeeque *et al.*, 2021; Galvao *et al.*, 2021; Zhao *et al.*, 2022; Ramirez *et al.*, 2023; Li *et al.*, 2022). The accuracy reached an impressive 99.7%, demonstrating exceptional performance in conjunction with high precision, recall, and F1-Score metrics. This demonstrates the consistency and effectiveness of our algorithm across different datasets with varying activities and fall types.

In the Le2i Fall Detection Dataset, our approach attained an accuracy of 98.95%, surpassing competitors (Wang *et al.*, 2020; Dentamaro *et al.*, 2021; Yuan *et al.*, 2022). Precision, recall, and F1-Score metrics matched the high accuracy at 98.95%, reinforcing our algorithm's ability to detect falls in varied scenarios.

In the NTU RGB+D 60 Dataset presented a more challenging environment with diverse actions and conditions. Yet, our approach demonstrated outstanding performance in the falling subset with a 99.98% accuracy.Precision, recall, and F1-Score metrics were also impressive,standing at 99.65%, 98.94%, and 99.29%, respectively, highlighting the adaptability and reliability of our algorithm in complex situations.

Our study demonstrates a harmonious balance between accuracy, speed, and adaptability across diverse datasets, outperforming or matching established methodologies such as OpenPose-based skeleton extraction (Zhao *et al.*, 2022), handcrafted feature approaches (Wang *et al.*, 2020), and advanced techniques like Kinematic Theory (Dentamaro *et al.*, 2021) and adaptive keypoint attention modules (Li *et al.*, 2022). The proposed methodology, with its emphasis on efficient pose estimation and subsequent image representation, stands as a promising solution for real-time fall detection applications.

In terms of preprocessing, our model integrates object detection, multi-object tracking, and pose estimation processes seamlessly, showcasing a comprehensive and efficient pipeline. Compared to models relying on handcrafted features (Wang *et al.*, 2020) or complex feature extractions involving the Kinematic Theory (Dentamaro *et al.*, 2021), our approach streamlines the preprocessing stage, contributing to the model's overall speed and efficiency.

In assessing model complexity, our study strikes a balance, achieving high accuracy with a streamlined architecture. While advanced models, such as those incorporating adaptive keypoint attention modules (Li *et al.*, 2022) or complex GCN architectures (Zahan *et al.*, 2023), may have higher parameter counts, our model's efficient design, featuring YOLO-V8n-pose and BoT-SORT, demonstrates competitive accuracy with lower computational complexity. This makes our approach not only effective but also resource-efficient.

The real-time computational efficiency of our model, operating in less than 1 milliseconds, further emphasizes its practical applicability for real-world scenarios. This quick processing time, combined with high accuracy, positions our algorithm as a promising solution for real-time fall detection applications.

DISCUSSION

To reinforce the robustness of our fall detection algorithm, we tested it across diverse datasets, including URFD, UP-Fall Detection, Le2i FDD, and NTU RGB+D 60. The effectiveness of our proposed fall detection algorithm is evident in a comparative analysis against multiple fall detection algorithms, as presented in Table 1. This analysis not only provides technical insights into our approach but also demonstrates its superior performance in accuracy, precision, recall, and F1-Score. Furthermore, the computational efficiency and real-time performance of our model enhance its practical utility, making it a

Table 1. *Comparison of fall detection algorithms across 4 Datasets.*

Method	Accuracy	Precision	Recall	F1-Score
URFD				
(Zhao *et al.*, 2022)	97	-	98.5	-
(Wang *et al.*, 2020)	97.33	97.78	97.78	97.78
(Dentamaro *et al.*, 2021)	99.6	98.28	98	99.6
(Li *et al.*, 2022)	99.73	-	99.74	-
(Zahan *et al.*, 2023)	100	100	100	100
(Galvao *et al.*, 2021)	100	100	100	100
Ours	100	100	100	100
UP-Fall Detection				
(Zahan *et al.*, 2023)	88.71	90.55	92.94	88.27
(Yadav *et al.*, 2022)	96.7	96.9	96.7	96.6
(Taufeeque *et al.*, 2021)	98.22	89.76	95.62	92.56
(Galvao *et al.*, 2021)	98.62	92.5	92	93
(Zhao *et al.*, 2022)	98.85	-	95.43	-
(Ramirez *et al.*, 2023)	99.5	86.49	85.79	87.2
(Li *et al.*, 2022)	99.62	-	99.26	-
Ours	99.7	100	99.37	99.68
Le2i FDD				
(Wang *et al.*, 2020)	96.91	96.79	96.51	97.08
(Dentamaro *et al.*, 2021)	98	97.6	97.2	98
(Yuan *et al.*, 2022)	98.43	-	-	-
Ours	98.95	98.95	98.95	98.95
NTU RGB+D 60				
Ours	85.63	85.78	85.63	85.64
Falling Subset of NTU RGB+D 60				
(Zhao *et al.*, 2022)	94.5	-	97.5	-
(Tsai and Hsu, 2019)	99.2	99.1	98.9	99
(Gutiérrez *et al.*, 2023)	99.24	-	99.16	-
Ours	99.98	99.65	98.94	99.29

promising candidate for real-world applications where timely and accurate fall detection is crucial.

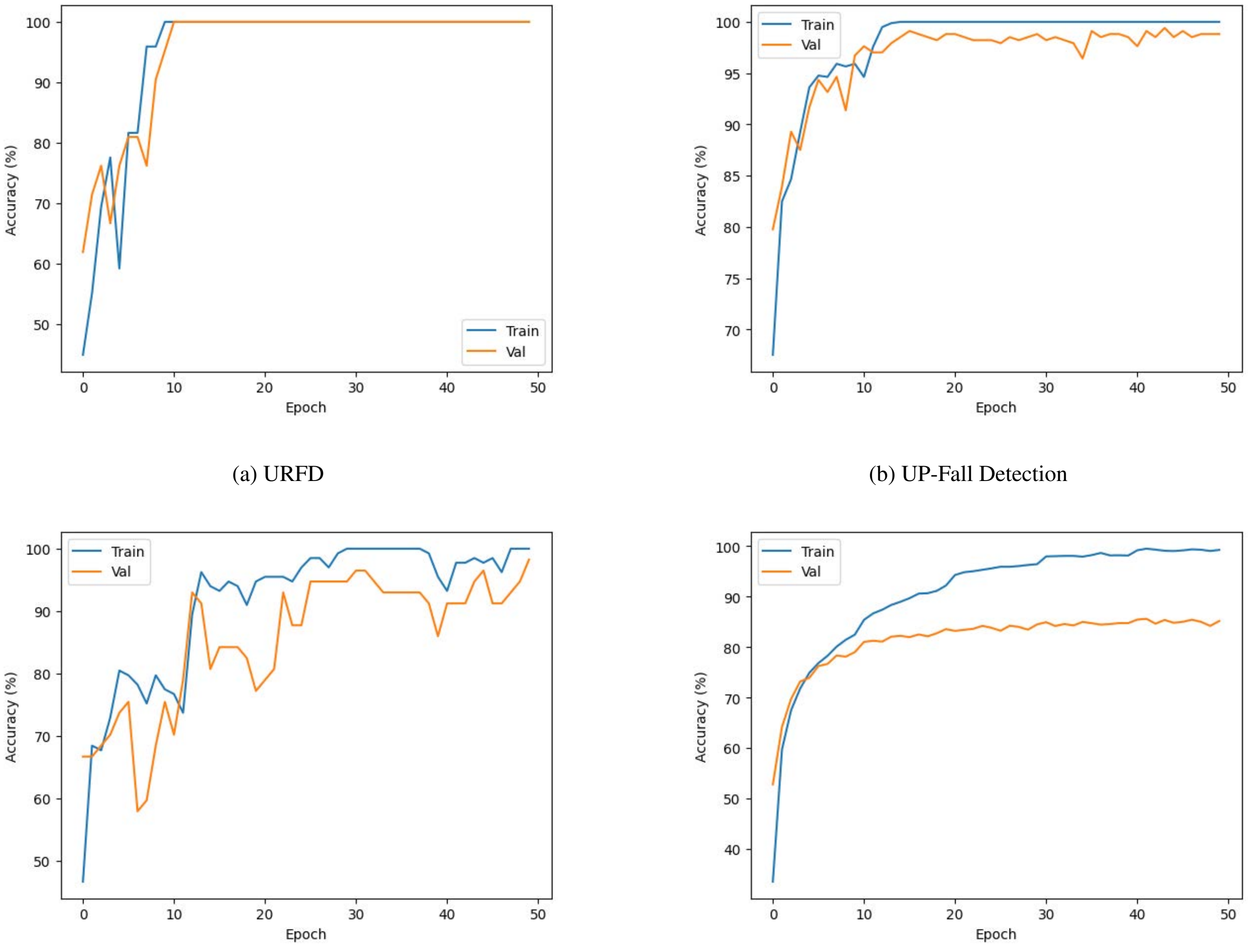
The success of our fall detection algorithm can be attributed to several key factors. The utilization of skeleton joints for fall detection provides a robust and efficient representation of human actions. Pose estimation, facilitated by the YOLO-V8n-pose model with BoT-SORT, accurately captures spatial and temporal characteristics, allowing for a focused analysis of essential features in fall detection.

The decision to transform skeleton joint inputs into image representations enhances the algorithm's strength and adaptability. This transformation, achieved through scaling, normalization and sampling, enables the use of established CNN based image processing techniques.

Challenges exist in classifying falls using images derived from skeleton points due to the uneven distribution of joint positions and orientations in the feature matrices. However, our incorporation of deformable convolution and deformable RoI pooling modules in the ShuffleNet V2 with Deformable Layers model effectively addresses this challenge, enabling the model to adaptively sample input data and achieve high accuracy in fall detection.

While our study yields promising results, there are avenues for further improvement. An identified issue is the difficulty the pose estimation model faces in capturing individuals at the camera's field of view limits, leading to classification errors. To mitigate this, further research and refinement of the pose estimation algorithms could enhance accuracy in these scenarios. Additionally, optimizing camera angles to comprehensively cover the action area proves effective in minimizing errors attributed to incomplete scene coverage.

While the choice of neural network architecture is effective, it can be further explored to optimize for specific datasets. Experimenting with different architectures and more comprehensive hyperparameter tuning could potentially yield even better results.



(a) URFD

(b) UP-Fall Detection

(c) Le2i FDD

(d) NTU RGB+D 60

Fig. 3. *Accuracy Plots - The training and validation accuracies for our proposed fall detection method are illustrated across four datasets.*

In terms of computational efficiency, the model runs in real time, but there is also room for additional optimization to shorten processing times. This optimization can increase the practicality of deploying the algorithm when there are constrained resources.

Future research could also evaluate the robustness of our approach in different scenarios, such as outdoor settings or cluttered backgrounds. Additionally, extending our approach to diverse populations, like children or athletes, would provide valuable insights into its generalizability.

CONCLUSION

In this paper, we presented a novel approach for detecting falls in video footage of individuals using machine learning techniques. Our approach involves the use of image representation of skeleton joints, and a ShuffleNet V2 with Deformable Layers model for fall detection. Our results demonstrate the effectiveness of our approach in detecting falls with high speed, high accuracy and low false positive rate, and we have shown that it outperforms existing fall detection approaches in the literature. Our approach has the potential to significantly improve the safety and well-being of elderly individuals, and we plan to further develop and apply in real-world scenarios.

ACKNOWLEDGEMENTS

This research was conducted as part of Hamza Ergüder's PhD Study at Yildiz Technical University.

REFERENCES

Aharon N, Orfaig R, Bobrovsky BZ (2022). Bot-sort: Robust associations multi-pedestrian tracking. arXiv:2206.14651.

Bisong E, Bisong E (2019). Google colaboratory. Building machine learning and deep learning models on google cloud platform a comprehensive guide for beginners :59–64.

Charfi I, Miteran J, Dubois J, Atri M, Tourki R (2013). Optimized spatio-temporal descriptors for real-time fall detection: comparison of support vector machine and adaboost-based classification. Journal of Electronic Imaging 22:041106.

Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y (2017). Deformable convolutional networks. arXiv:1703.06211.

Delahoz Y, Labrador M (2014). Survey on fall detection and fall prevention using wearable and external sensors. Sensors 14:19806–42.

Dentamaro V, Impedovo D, Pirlo G (2021). Fall detection by human pose estimation and kinematic theory. In: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE.

Duan H, Wang J, Chen K, Lin D (2022). Dg-stgcn: Dynamic spatial-temporal modeling for skeleton-based action recognition. arXiv:2210.05895.

Espinosa R, Ponce H, Gutiérrez S, Martínez-Villaseñor L, Brieva J, Moya-Albor E (2019). A vision-based approach for fall detection using multiple cameras and convolutional neural networks: A case study using the UP-fall detection dataset. Computers in Biology and Medicine 115:103520.

Fang HS, Xie S, Tai YW, Lu C (2016). Rmpe: Regional multi-person pose estimation. arXiv:1612.00137.

Galvao YM, Portela L, Ferreira J, Barros P, De Araujo Fagundes OA, Fernandes BJT (2021). A framework for anomaly identification applied on fall detection. IEEE Access 9:77264–74.

Gutiérrez J, Martin S, Rodriguez V (2023). Human stability assessment and fall detection based on dynamic descriptors. IET Image Processing 17:3177–95.

He K, Zhang X, Ren S, Sun J (2015). Deep residual learning for image recognition. arXiv:1512.03385.

Jager TE, Weiss HB, Coben JH, Pepe PE (2000). Traumatic brain injuries evaluated in u.s. emergency departments, 1992-1994. Academic Emergency Medicine 7:134–40.

Jocher G, Chaurasia A, Qiu J (2023). YOLO by Ultralytics. https://github.com/ultralytics/ultralytics.

Juraev S, Ghimire A, Alikhanov J, Kakani V, Kim H (2022). Exploring human pose estimation and the usage of synthetic data for elderly fall detection in real-world surveillance. IEEE Access 10:94249–61.

Kingma DP, Ba J (2014). Adam: A method for stochastic optimization. arXiv:1412.6980.

Kwolek B, Kepski M (2014). Human fall detection on embedded platform using depth maps and wireless accelerometer. Computer Methods and Programs in Biomedicine 117:489–501.

Li J, Gao M, Li B, Zhou D, Zhi Y, Zhang Y (2022). Kamtfenet: a fall detection algorithm based on keypoint attention module and temporal feature extraction. International Journal of Machine Learning and Cybernetics 14:1831–44.

Lin TY, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, Perona P, Ramanan D, Zitnick CL, Dollár P (2014). Microsoft coco: Common objects in context. arXiv:1405.0312.

Ma N, Zhang X, Zheng HT, Sun J (2018). Shufflenet v2: Practical guidelines for efficient cnn architecture design. arXiv:1807.11164.

Martínez-Villaseñor L, Ponce H, Brieva J, Moya-Albor E, Núñez-Martínez J, Peñafort-Asturiano C (2019). UP-fall detection dataset: A multimodal approach. Sensors 19:1988.

Moreland B, Kakara R, Henry A (2020). Trends in nonfatal falls and fall-related injuries among adults aged ≥65 years — united states, 2012–2018. MMWR Morbidity and Mortality Weekly Report 69:875–81.

Nooruddin S, Islam MM, Sharna FA, Alhetari H, Kabir MN (2021). Sensor-based fall detection systems: a review. Journal of Ambient Intelligence and Humanized Computing 13:2735–51.

Ramirez H, Velastin SA, Cuellar S, Fabregas E, Farias G (2023). Bert for activity recognition using sequences of skeleton features and data augmentation with gan. Sensors 23:1400.

Ramirez H, Velastin SA, Meza I, Fabregas E, Makris D, Farias G (2021). Fall detection and activity recognition using human skeleton features. IEEE Access 9:33532–42.

Serpa YR, Nogueira MB, Neto PPM, Rodrigues MAF (2020). Evaluating pose estimation as a solution to the fall detection problem. In: 2020 IEEE 8th International Conference on Serious Games and Applications for Health (SeGAH).

Shahroudy A, Liu J, Ng TT, Wang G (2016). NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE.

Singh T, Vishwakarma DK (2018). Human activity recognition in video benchmarks: A survey. In: Lecture Notes in Electrical Engineering. Springer Singapore, 247–59.

Sterling DA, O'Connor JA, Bonadies J (2001). Geriatric falls: Injury severity is high and disproportionate to mechanism. The Journal of Trauma Injury Infection and Critical Care 50:116–19.

Taufeeque M, Koita S, Spicher N, Deserno TM (2021). Multi-camera, multi-person, and real-time fall detection using long short term memory. In: Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications.

Tsai TH, Hsu CW (2019). Implementation of fall detection system based on 3d skeleton for deep learning technique. IEEE Access 7:153049–59.

Wang BH, Yu J, Wang K, Bao XY, Mao KM (2020). Fall detection based on dual-channel feature integration. IEEE Access 8:103443–53.

Weytjens H, De Weerdt J (2020). Process Outcome Prediction: CNN vs. LSTM (with Attention), vol. 397. Cham: Springer International Publishing, 321–33.

Yadav SK, Luthra A, Tiwari K, Pandey HM, Akbar SA (2022). ARFDNet: An efficient activity recognition & fall detection system using latent feature pooling. Knowledge Based Systems 239:107948.

Yuan J, Liu C, Liu C, Wang L, Chen Q (2022). Real-time human falling recognition via spatial and temporal self-attention augmented graph convolutional network. In: 2022 IEEE International Conference on Real-time Computing and Robotics (RCAR).

Zahan S, Hassan GM, Mian A (2023). Sdfa: Structure-aware discriminative feature aggregation for efficient human fall detection in video. IEEE Transactions on Industrial Informatics 19:8713–21.

Zhao Z, Zhang L, Shang H (2022). A lightweight subgraph-based deep learning approach for fall recognition. Sensors 22:5482.