

Video Based Fall Detection Using Human Poses

Ziwei Chen
Southeast University
Nanjing, China

richard_chen@seu.edu.cn

Yiye Wang
Southeast University
Nanjing, China

230208761@seu.edu.cn

Wankou Yang
Southeast University
Nanjing, China

wkyang@seu.edu.cn

Abstract

Video based fall detection accuracy has been largely improved due to the recent progress on deep convolutional neural networks. However, there still exists some challenges, such as lighting variation, complex background, which degrade the accuracy and generalization ability of these approaches. Meanwhile, large computation cost limits the application of existing fall detection approaches. To alleviate these problems, a video based fall detection approach using human poses is proposed in this paper. First, a lightweight pose estimator extracts 2D poses from video sequences and then 2D poses are lifted to 3D poses. Second, we introduce a robust fall detection network to recognize fall events using estimated 3D poses, which increases respective filed and maintains low computation cost by dilated convolutions. The experimental results show that the proposed fall detection approach achieves a high accuracy of 99.83% on large benchmark action recognition dataset NTU RGB+D and real-time performance of 18 FPS on a non-GPU platform and 63 FPS on a GPU platform.

1. Introduction

Nowadays, the ageing of the population has become a global phenomena, there were 727 million persons aged 65 or over in 2020 and the number of the elderly worldwide will be projected to more than double over the next three decades, sharing around 16.0 per cent of the population in 2050 [30]. According to the World Health Organization (WHO) [1], adults older than 65 years suffer the greatest number of fatal falls, which could cause serious injuries and even death. Therefore, intelligent fall detection has drawn increasing attention from both academia and industry and has become an urgent need for vulnerable people, especially the elderly.

The existing fall detection methods can be roughly divided into two categories, which are wearable sensor based methods and vision based methods [9]. Wearable sensors, including accelerometer, gyroscopes, pressure sensor and

microphone, can detect the location change or acceleration change of human body for fall detection. However, inconvenience is still the main problem that many elderly people are unwilling to wear sensors all day. Besides, wearable sensors may be affected by noise and some daily activities like lying or sitting on the sofa quickly may lead to false alarm. With the rapid development of computer vision and deep learning techniques in recent years, the number of proposed vision based methods has increased a lot [46]. Compared with wearable sensor based methods, vision based methods are free from the inconvenience of wearing the device. While the detection accuracy of vision based methods has increased a lot in recent years, false detection still may occur as a result of lighting variation, complex background and so on. Furthermore, vision based methods, especially deep learning based methods, have a large computation cost which makes it hard to achieve real-time performance. To conclude, how to maintain a high detection accuracy while lower the computation cost is a valuable research topic.

Human pose estimation is a fundamental task in computer vision, the goal is to localize human keypoints in a image or 3D space. Human pose estimation has many applications, including human action recognition, human-computer interaction, animation, etc. Nowadays, 2D human pose estimation [28, 44, 42, 5, 47] has achieved convincing performance on some large public datasets [3, 20] and the performance of 3D human pose [25, 15, 48, 41] estimator has been improved a lot. Using human poses can alleviate the problem of lighting variation or complex background in fall detection task, so as to effectively improve the accuracy and generalization ability of fall detection methods.

Although existing vision based methods have brought fall detection accuracy to such a high level, they rely on a large computation [23, 10, 26] and features extracted for classification are not robust for some challenging conditions. Meanwhile, the period of falling differs among people [43], the fall of the elderly lasts longer than other groups and some daily activities such as lying to bed differ from fall. So more video frames should be taken as the input. However, previous works [13, 17, 39] only take a short

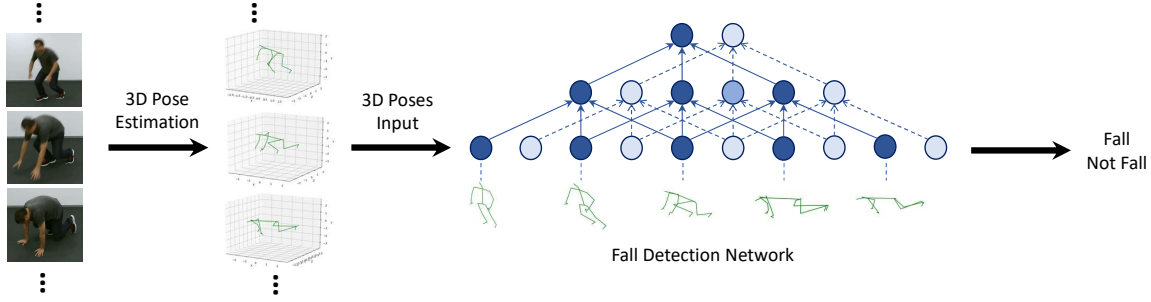


Figure 1. Overall structure of our fall detection approach. Video sequences are first sent into pose estimator to get 3D poses, then fall detection network takes 3D poses to classify action class.

video sequence as the input which may lead to the false detection.

To address the above problems, we propose a video based fall detection approach using human poses in this paper. Our fall detection approach consists of two steps: (1) estimating 3D poses in video sequences (2) recognize fall events from estimated 3D poses. The first step is compatible with any state-of-the-art pose estimator. We formulate 3D pose estimation as 2D pose estimation followed by 2D-to-3D pose lifting. A lightweight 2D pose estimator and a lifting network are adopted to lower the computation cost. At the second step, we present a fall detection network taking 3D poses of each frame as input. In order to achieve a convincing accuracy while maintaining the low computation cost for long video sequences, one-dimensional dilated temporal convolution [12] is adopted.

In summary, this paper has three contributions:

- We propose a fall detection model, which includes a 3D pose estimator and a fall detection network based on human poses.
- We explore the effects of factors which could contribute to the performance of fall detection including input joints, loss function.
- Our approach achieves a high accuracy at 99.83% on NTU RGB+D dataset and real-time performance on non-GPU platform.

2. Related Work

The related work on vision based fall detection are first reviewed and the difference between them and our work are discussed. Then we review some work on human pose estimation.

2.1. Vision Based Fall Detection

Many approaches have been proposed for vision based fall detection [22, 4, 18, 45, 19, 14, 35, 39, 50]. These ap-

proaches differ in terms of the used sensors and classify algorithms.

Sensors for most vision based approaches are RGB cameras, depth cameras, infrared cameras and Kinect. In [22], Lu et al. propose to detect fall on RGB videos using 3D CNN combined with LSTM to address the problem of insufficient fall data. Shojarei et al [35]. obtain 3D coordinates of human keypoints using depth camera and then do fall detection based on these poses. Zhong et al. [50] propose an end to end solution within a multi-occupancy living environment by thermal sensors. Nowadays, the use of Kinect for fall detection [19, 14] has been increased a lot as 3D information can be obtained. However, depth camera in Kinect has a restricted distance which makes it unsuitable for a large space.

Decision trees, SVM and threshold are used to classify fall action categories [27, 33, 31, 36]. Compared with these algorithms, deep neural network can achieve higher classification accuracy and avoiding feature engineering task. Adhikari et al. [2] propose a fall detection system using CNN to recognize Activities of Daily Life (ADL) and fall events. A 3D CNN is developed in [32] to improve fall detection accuracy by exploring spatial and temporal information. In [51], Variational Auto-encoder (VAE) with 3D convolutional residual block and a region extraction technique for enhancing accuracy are used to detect fall actions.

Our approach is similar to Tsai et al. [39], thus we provide more detailed comparison. Tsai et al. [39] propose a traditional algorithm to transform depth information into 3D poses and use 1D CNN to detect fall events. In [39], depth camera is used while our human pose estimator can obtain 3D poses directly from RGB images which makes it free from limited measurement distance. Though 1D CNN is used in both works, only 30 frames are taken as input in their approach while our model can take 300 frames at most as input. Some actions last longer and the elderly fall slower than the young so it is necessary to recognize fall events using long video sequences.

2.2. Human Pose Estimation

Human Pose Estimation is to localize human keypoints, which can be categorized as 2D and 3D human pose estimation according to the output.

Deep learning has become a promising technique in 2D human pose estimation in recent years. Firstly, CNN is introduced to solve 2D pose estimation problem by directly regressing the joint coordinates in DeepPose [38]. Then joint heatmaps [37] have been widely adopt as training signals in 2D human pose estimation for great performance. Newell et al. [29] propose an U-shape network by stacking up several hourglass modules to refine prediction. The work by Cao et al. [5] detects all human keypoints first and assembles them to different person by part affinity fields (PAFs), which achieves real-time performance in multi person pose estimation task. HRNet [42] adapts the top-down pipeline and generates convincing performance via maintaining the high resolution of feature maps and multi-scale fusion.

3D human pose estimation is to localize the position of human keypoints in 3D space from images or videos. The early 3D human pose estimation methods directly predict the 3D joint coordinates via deep neural networks [16]. While a set of features suitable for the task can be spontaneously learned, these models usually have large computation cost and high complexity. Due to the development of 2D human pose estimation, methods based on 2D poses [6, 24, 40] have become the main stream. 2D poses are concatenated as the input to predict the depth information of each keypoint, greatly reducing the complexity of the model. In order to overcome the problem of insufficient data in 3D human pose estimation, wandt et al. [40] apply weak supervised learning to 3D human pose estimation by using reprojection method to train 3D mapping model. Cheng et al. [8] explore video information to further modify the human poses to avoid incorrect pose estimation. He et al. [11] use multi view images to overcome the occlusion problem based on epipolar geometry and achieve great performance.

3. Video Based Fall Detection

The goal of this work is to establish a video based fall detection approach using human poses. The general framework of the proposed approach is shown in Figure 1. Firstly, human pose estimator is applied to each video frame to generate a set of human poses. Then fall detection network works to recognize whether there is a fall event based on human poses.

3.1. Human Pose Estimator

Our fall detection approach do not rely on any specific pose estimator. When obtaining 3D poses, we follow the

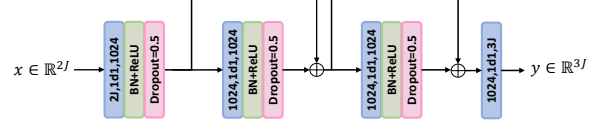


Figure 2. Lifting network. "2J" means 2D coordinates of J joints are concatenated.

widely-used pipeline in 3D human pose estimation [24], which predicts 2D human poses in the first step and lifts 2D poses to 3D poses. For 2D human pose estimation, we adapt off-the-shelf Lightweight Pose Network (LPN) [49] considering its low complexity and adequate accuracy. LPN is pretrained on MS COCO [21] dataset and fine tuned on NTU RGB+D dataset [34] for our task.

Lifting network takes 2D human poses $x \in \mathbb{R}^{2J}$ as input and lifts them to 3D poses $y \in \mathbb{R}^{3J}$. The goal is to find a function $f^* : \mathbb{R}^{2J} \rightarrow \mathbb{R}^{3J}$ that minimizes the prediction error of N poses over a dataset:

$$f^* = \min_f \frac{1}{N} \sum_{i=1}^N \mathcal{L}(f(x_i) - y_i) \quad (1)$$

Figure 2 shows the structure of our lifting network, it consists of two residual blocks. 2D coordinates of joints are concatenated so one-dimensional convolution can be adopt to reduce parameters and complexity. The core idea of our lifting network is to predict depth information of keypoints effectively and efficiently. Note that what we care is not the absolute location of keypoints in 3D space but the relative location between them. So before a 2D pose is lifted to 3D, it is normalized by centering to its root joint and scaled by dividing it by its Frobenius norm. We define the prediction error as the squared difference between the prediction and the ground-truth pose:

$$\mathcal{L}(\hat{y}_i, y_i) = \|\hat{y}_i - y_i\|_2^2 \quad (2)$$

where \hat{y}_i and y_i are estimated and the ground-truth relative position of the i -th pose.

3.2. Fall Detection Network

Our fall detection network is a fully convolutional architecture with residual connections that takes a sequence of 3D poses $X \in \mathbb{R}^{T \times 3J}$ as input where T is the number of frames and predict whether there is a fall behavior. In convolutional networks, the path of gradient between output and input has a fixed length, which mitigates vanishing and exploding gradients. It is important for our task as T was set to 300 to recognize falls in such a long video sequence. Moreover, dilated convolutions are applied in our network to model long-term dependencies while maintaining efficiency.

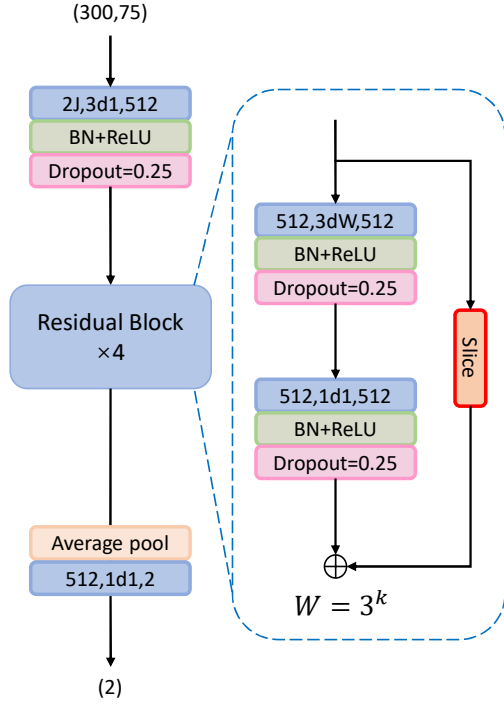


Figure 3. Fall detection network.

When 3D poses are obtained by pose estimator, we also do the centering and scaling. Besides, 3D poses are firstly rotated by paralleling the bone between hip and spine to the z axis, then by paralleling the bone between left shoulder and right shoulder to the x axis, normalized 3D poses can be obtained.

Figure 3 shows our fall detection network. 3D coordinates (x, y, z) of J joints for each frame are concatenated as the network input and a convolution with kernel size 3 and C output channels is applied. This is followed by N residual blocks. Each block includes two one-dimensional convolutions, the first one is dilated and dilation factor is $W = 3^N$, followed by another 1D convolution with kernel size 1. Batch normalization, rectified linear units and dropout are used after every convolution except the last one. With dilation factor, each block increases the receptive field to exploit temporal information without too much computation increase. We use unpadded convolutions so the output size of each block is different, details can be seen in Table 1. Average pool is used to fuse features and change the dimension for the final convolution. The length of video sequence is 300 and we set $N = 4$ to increase the receptive field. For convolutions, we set $C = 512$ output channels to maintain a balance between accuracy and complicity and the dropout rate $p = 0.25$.

layer name	output size	layer
conv_1	(512, 298)	3d1, 512
res_1	(512, 292)	3d3, 512 1d1, 512
res_2	(512, 274)	3d9, 512 1d1, 512
res_3	(512, 220)	3d27, 512 1d1, 512
res_4	(512, 58)	3d81, 512 1d1, 512
pooling	(512)	average pool
conv_2	(2)	1d1, 2

Table 1. This table shows the architecture and output size of each block for our fall detection network. "3d3, 512" means 1D convolution with kernel size 3, dilation factor 3 and 512 channels.

4. Experiments

4.1. Dataset

The proposed fall detection model was trained and evaluated on NTU RGB+D Action Recognition Dataset [34] made available by the ROSE Lab at the Nanyang Technological University, Singapore. This dataset contains 60 action classes and 56,880 video samples including falling. The videos are captured by three synchronous Microsoft Kinect v2 cameras installed at the same height with three different horizontal angles: -45° , 0° , $+45^\circ$. The dataset contains RGB videos, depth map sequences, 3D skeletal data, and infrared (IR) videos for each sample. The resolution of RGB frames are 1920×1080 and the speed are 30 FPS. 3D skeletal data are composed of 3D coordinates of 25 joints. To best of our knowledge, this is the largest action recognition dataset available that contains 3D skeletal data and falling samples. Some of the video samples have missing frames, poses or involve more than one person, we removed these samples. Consequently, the total amount of samples we used was 44372, in which 890 samples were falling samples. Following previous action recognition work [7], we trained on data coming from camera 0° and $+45^\circ$, tested on data from camera -45° .

4.2. Training details

We trained our fall detection model step by step. Firstly, for human pose estimation, we adopt off-the-shelf LPN to predict 2D poses from each video frame. LPN was pre-trained on COCO dataset and fine-tuned on NTU RGB+D dataset. When fine-tuning LPN on NTU RGB+D dataset, joint heatmaps were generated according to annotations as the output target which could avoid directly learning the mapping from images to coordinates. Then by calculating the center of mass of heatmaps, 2D joint coordinates could

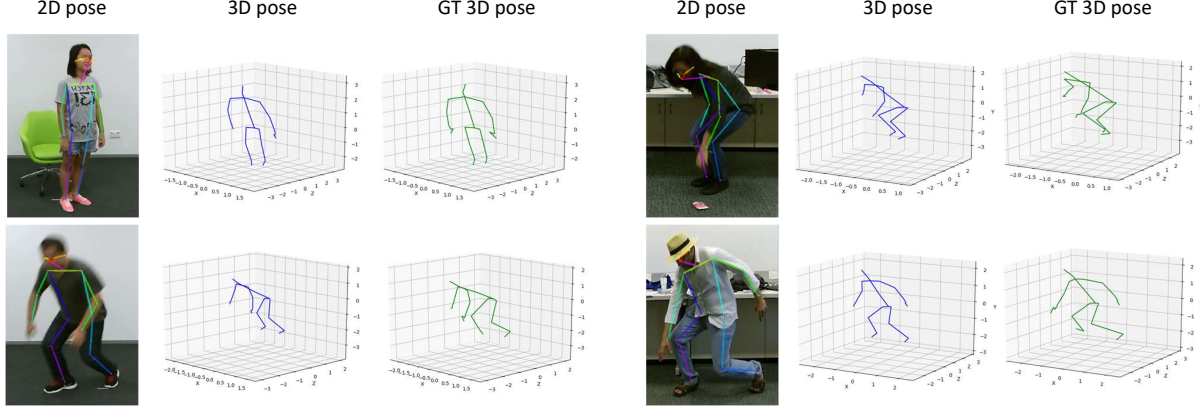


Figure 4. Qualitative results of some example images. Initial image, 2D poses, 3D poses and GT 3D poses are presented. Blue poses are estimated 3D poses and green ones are GT 3D poses.

be obtained.

Before lifting 2D poses to 3D poses, 2D poses were normalized by centering to its root joint and then scaled by dividing its Frobenius norm. Adam optimizer was used to train the lifting network combined with MSE loss for 60 epochs with an initial learning rate of 0.0001 and exponential decay at 20th and 40th epoch.

Similar to the training of lifting network, 3D poses were normalized before being sent to the fall detection network. Video samples from dataset have different frames, so all samples were expanded to 300 frames by padding null frames with previous ones. Adam optimizer and Cross Entropy Loss were used to train fall detection network. We set initial learning rate to 0.0001 with exponential decay. We train the network on one Nvidia GTX 1660 GPU for 20 epochs.

4.3. Results

Human Pose Estimation Most previous skeleton based action recognition works directly use 3D annotations on NTU RGB+D dataset. Here we use predicted 3D poses for fall detection to test feasibility of our approach. Human pose estimation accuracy is measured by Joint Detection Rate (JDR). JDR represents the percentage of successfully detected joints. A joint is regarded as successfully detected if the distance between the estimation and groundtruth is smaller than a threshold. Here we set the threshold to be half of the distance between neck and head.

Table 2 shows pose estimation results of our method on NTU RGB+D dataset. JDR of some joints are presented. For some joints including head, elbow, shoulder, JDRs achieve larger than 90% and the prediction of these joints is accurate. However, JDR of ankle and thumb are not high as a result of frequent occlusion and inaccurate 2D pose. Table 3 shows mean JDR of three aggregations of

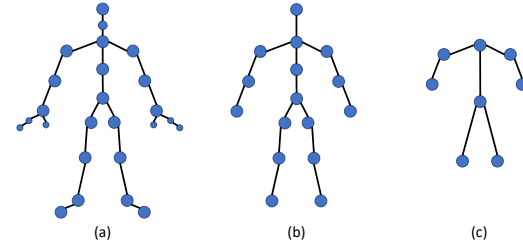


Figure 5. Skeleton information of different inputs. (a) All 25 joints. (b) Selected 16 joints. (c) Selected 8 joints.

joints. Details of three aggregations of joints are showed in Figure 5. The evaluation of our proposed fall detection method using different number of joints as inputs will later be reported.

B spi	Head	L elb	L wri	R elb	R wri	R ank
99.69	98.02	98.45	97.91	94.22	90.82	71.06

Table 2. This table shows pose estimation accuracy on NTU RGB+D dataset. JDR (%) of seven joints are showed due to limited space. "B spi" means base spine and "L wri" means left wrist.

Joints	mJDR
25 joints	86.02
16 joints	94.07
8 joints	94.52

Table 3. This table shows mean JDR (%) of three aggregations of joints.

fall detection Table 4 shows the result of our proposed method and other fall detection methods. It can be seen that our proposed method achieve a quite high performance of

Methods	Input	Feature	Network	Accuracy
Xu et al. [45]	RGB	Pose	2D conv	91.70%
Anahita et al. [35]	Depth	Pose	LSTM	96.12%
Han et al. [39]	Depth	Pose	1D conv	99.20%
Ours	RGB	Pose	1D conv	99.83%

Table 4. Fall detection accuracy of different methods on NTU RGB+D dataset.

Method	Accuracy	Precision	Recall
8 joints-CEL	99.72%	97.15%	89.74%
8 joints-WCEL	99.29%	98.70%	74.32%
16 joints-CEL	99.83%	97.47%	94.25%
16 joints-WCEL	99.50%	98.73%	80.67%
25 joints-CEL	99.77%	97.79%	91.35%
25 joints-WCEL	99.66%	97.47%	87.11%

Table 5. Fall detection results of different methods on NTU RGB+D dataset. The naming convention of the methods follows the rule of “A-B” where “A” indicates how many joints are used in fall detection. “B” denotes the loss. “CEL” means Cross Entropy Loss and “WCEL” means weighted Cross Entropy Loss.

99.83% accuracy and outperforms other methods.

We also evaluate the influence of the number of input joints, as shown in Figure 5, we select 8 and 16 joints from all the 25 joints as input and calculate the classification accuracy. Table 5 shows the results of different joint input. It can be seen that when using 16 joints as input, the model achieves the highest accuracy of 99.83%. Using all the 25 joints or 8 joints achieve a lower but still high accuracy. We consider that some joints like eyes, hands could disturb the model to learn action features, which could lead to the degradation of accuracy. Besides, few joints may not be able to model the variance of different actions.

Considering that the number of falling samples only takes a part of 1.67 percentage of the whole dataset, it is necessary to test whether our model can truly classify fall from other actions rather than just classifying all the samples to not fall class. So Precision is also calculated to test whether our model can really recognize fall behaviour. Moreover, we use Weighted Cross Entropy Loss to train this fall detection network follow previous method and evaluate it on this dataset. For falling class, we set $\alpha = 59/60$ and $\beta = 1/60$ for other samples. Table 5 shows the evaluation results, we can see that our network truly learns how to classify a fall behavior and the precision achieves 1.26% improvement using Weighted Cross Entropy Loss while accuracy decreased by 0.83%. The variation of Recall is very large as the number of falling samples is much smaller than it of other actions.

Actual inference speed is what we care about which defines whether our fall detection method can achieve real

Part	Params	FLOPs	non-GPU	GPU
LPN	2.7M	1.0G	20 FPS	74 FPS
LN	2.2M	0.28G	560 FPS	1450 FPS
FDN	4.2M	0.9G	260 FPS	590 FPS
Whole	9.1M	2.18G	18 FPS	63 FPS

Table 6. Measurement of Params, FLOPs and Speed of each part of our proposed fall detection method on different platforms. “LN” means lifting network and “FDN” means fall detection network.

time. We test number of parameters, FLOPs and inference speed of every part of our fall detection approach. The speed test is based on two platforms. The one is a non-GPU platform with Intel Core i5-9400F CPU (2.9GHZ \times 6) and the other is one Nvidia GTX 1660 GPU. Table 6 shows the measurements and we can find that our method has a speed of 18 FPS on a non-GPU platform and 63 FPS on one Nvidia GTX 1660 GPU. Moreover, the inference speed of lifting network and fall detection network is very fast that only takes a few milliseconds. The 2D pose estimator, LPN, is the one mainly limits the inference speed. It is worth mentioning that our fall detection method does not rely on any specific 2D human pose estimator, LPN can be changed to other pose estimator for more efficiency and robustness.

5. Conclusion

In this paper, we propose an approach to recognize fall events from video sequences. More specifically, our approach includes a 3D pose estimator based on lifting 2D poses to 3D poses and a fall detection network using dilated convolution. Our approach achieves a high accuracy of 99.83 on NTU RGB+D dataset and realtime performance of 18 FPS on a non-GPU platform and 63 FPS on a GPU platform.

References

- [1] WHO (2021). Fall. <https://www.who.int/news-room/factsheets/detail/falls>. 1
- [2] Kripesh Adhikari, Hamid Bouchachia, and Hammadi Nait-Charif. Activity recognition for indoor fall detection using convolutional neural network. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pages 81–84, 2017. 2
- [3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt 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 (CVPR)*, June 2014. 1
- [4] Sarah Almeida Cameiro, Gabriel Pellegrino da Silva, Guilherme Vieira Leite, Ricardo Moreno, Silvio Jamil F. Guimarães, and Helio Pedrini. Multi-stream deep convolutional network using high-level features applied to fall detection in video sequences. In *2019 International Conference*

- on Systems, Signals and Image Processing (IWSSIP), pages 293–298, 2019. 2
- [5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2021. 1, 3
- [6] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3
- [7] Ke Cheng, Yifan Zhang, Congqi Cao, Lei Shi, Jian Cheng, and Hanqing Lu. Decoupling gc3n with dropgraph module for skeleton-based action recognition. *European conference on computer vision*, pages 536–553, 2020. 4
- [8] Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10631–10638, 2020. 3
- [9] Jesús Gutiérrez, Víctor Rodríguez, and Sergio Martín. Comprehensive review of vision-based fall detection systems. *Sensors*, 21(3), 2021. 1
- [10] Fouzi Harrou, Nabil Zerrouki, Ying Sun, and Amrane Houacine. An integrated vision-based approach for efficient human fall detection in a home environment. *IEEE Access*, 7:114966–114974, 2019. 1
- [11] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoubo-I Yu. Epipolar transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3
- [12] Matthias Holschneider, Richard Kronland-Martinet, Jean Morlet, and Ph Tchamitchian. A real-time algorithm for signal analysis with the help of the wavelet transform. In *Wavelets*, pages 286–297. Springer, 1990. 2
- [13] Seokhyun Hwang, DaeHan Ahn, Homin Park, and Taejoon Park. Poster abstract: Maximizing accuracy of fall detection and alert systems based on 3d convolutional neural network. In *2017 IEEE/ACM Second International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pages 343–344, 2017. 1
- [14] Sowmya Kasturi, Alexander Filonenko, and Kang-Hyun Jo. Human fall recognition using the spatiotemporal 3d cnn. In *Proc. IW-FCV*, pages 1–3, 2019. 2
- [15] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1
- [16] Sijin Li and Antoni B. Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In Daniel Cremers, Ian Reid, Hideo Saito, and Ming-Hsuan Yang, editors, *Computer Vision – ACCV 2014*, pages 332–347, Cham, 2015. Springer International Publishing. 3
- [17] Shengchao Li, Hao Xiong, and Xiumin Diao. Pre-impact fall detection using 3d convolutional neural network. In *2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR)*, pages 1173–1178, 2019. 1
- [18] Shengchao Li, Hao Xiong, and Xiumin Diao. Pre-impact fall detection using 3d convolutional neural network. In *2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR)*, pages 1173–1178, 2019. 2
- [19] Xiaogang Li, Tiantian Pang, Weixiang Liu, and Tianfu Wang. Fall detection for elderly person care using convolutional neural networks. In *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–6, 2017. 2
- [20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 1
- [21] T. Y. Lin, M. Maire, S. Belongie, J. Hays, and C. L. Zitnick. Microsoft coco: Common objects in context. *European Conference on Computer Vision*, 2014. 3
- [22] Na Lu, Yidan Wu, Li Feng, and Jinbo Song. Deep learning for fall detection: Three-dimensional cnn combined with lstm on video kinematic data. *IEEE Journal of Biomedical and Health Informatics*, 23(1):314–323, 2019. 2
- [23] Chao Ma, Atsushi Shimada, Hideaki Uchiyama, Hajime Nagahara, and Rin ichiro Taniguchi. Fall detection using optical level anonymous image sensing system. *Optics & Laser Technology*, 110:44–61, 2019. Special Issue: Optical Imaging for Extreme Environment. 1
- [24] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 3
- [25] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 International Conference on 3D Vision (3DV)*, pages 506–516, 2017. 1
- [26] Carlos Menacho and Jhon Ordoñez. Fall detection based on cnn models implemented on a mobile robot. In *2020 17th International Conference on Ubiquitous Robots (UR)*, pages 284–289, 2020. 1
- [27] Weidong Min, Leiyue Yao, Zhenrong Lin, and Li Liu. Support vector machine approach to fall recognition based on simplified expression of human skeleton action and fast detection of start key frame using torso angle. *IET Computer Vision*, 12(8):1133–1140, 2018. 2
- [28] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hour-glass networks for human pose estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 483–499, Cham, 2016. Springer International Publishing. 1
- [29] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hour-glass networks for human pose estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 483–499, Cham, 2016. Springer International Publishing. 3
- [30] United Nations Department of Economic and Population Division (2020) Social Affairs. World population ageing 2020 highlights: Living arrangements of older persons. 1

- [31] Koray Ozcan and Senem Velipasalar. Wearable camera- and accelerometer-based fall detection on portable devices. *IEEE Embedded Systems Letters*, 8(1):6–9, 2016. 2
- [32] Maryam Rahnemoonfar and Hend Alkittawi. Spatio-temporal convolutional neural network for elderly fall detection in depth video cameras. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2868–2873, 2018. 2
- [33] Benaoumeur Senouci, Imen Charfi, Barthelemy Heyrman, Julien Dubois, and Johel Miteran. Fast prototyping of a soc-based smart-camera: a real-time fall detection case study. *Journal of Real-Time Image Processing*, 12(4):649–662, 2016. 2
- [34] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. *CVPR*, pages 1010–1019, 2016. 3, 4
- [35] Anahita Shojaei-Hashemi, Panos Nasiopoulos, James J. Little, and Mahsa T. Pourazad. Video-based human fall detection in smart homes using deep learning. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, 2018. 2, 6
- [36] Erik E. Stone and Marjorie Skubic. Fall detection in homes of older adults using the microsoft kinect. *IEEE Journal of Biomedical and Health Informatics*, 19(1):290–301, 2015. 2
- [37] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 3
- [38] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 3
- [39] Tsung-Han Tsai and Chin-Wei Hsu. Implementation of fall detection system based on 3d skeleton for deep learning technique. *IEEE Access*, 7:153049–153059, 2019. 1, 2, 6
- [40] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3
- [41] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13294–13304, June 2021. 1
- [42] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 1, 3
- [43] Xueyi Wang, Joshua Ellul, and George Azzopardi. Elderly fall detection systems: A literature survey. *Frontiers in Robotics and AI*, 7:71, 2020. 1
- [44] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1
- [45] Qingzhen Xu, Guangyi Huang, Mengjing Yu, and Yanliang Guo. Fall prediction based on key points of human bones. *Physica A: Statistical Mechanics and its Applications*, 540:123205, 2020. 2, 6
- [46] Tao Xu, Yun Zhou, and Jing Zhu. New advances and challenges of fall detection systems: A survey. *Applied Sciences*, 8(3), 2018. 1
- [47] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1
- [48] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1
- [49] Zhe Zhang, Jie Tang, and Gangshan Wu. Simple and lightweight human pose estimation. *arXiv preprint arXiv:1911.10346*, 2019. 3
- [50] Cankun Zhong, Wing W. Y. Ng, Shuai Zhang, Chris D. Nugent, Colin Shewell, and Javier Medina-Quero. Multi-occupancy fall detection using non-invasive thermal vision sensor. *IEEE Sensors Journal*, 21(4):5377–5388, 2021. 2
- [51] Jiaxin Zhou and Takashi Komuro. Recognizing fall actions from videos using reconstruction error of variational autoencoder. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3372–3376, 2019. 2