Fall Detection Based on Action Structured Method and Cascaded Dilated Graph Convolution Network

Xin Xiong[1,2], Lei Cao[1], Qiang Liu[1], Zhiwei Tu[1], and Huixia Li[1(✉)]

[1] Information Department, The First Affiliated Hospital of Nanchang University, Nanchang 330000, China
lihuixia0601@163.com
[2] Jiangxi Key Laboratory of Smart City, Nanchang 330000, China

Abstract. The research of fall detection is a hot topic in computer vision. Most existing methods only detect the fall in simple scenes of a single person. Moreover, these methods only extract fall action features from RGB images, and neglect to extract features from human joint coordinates, resulting in a decrease in recognition accuracy. In order to extract discriminative action features, a fall detection method based on action structured method and cascade dilated graph convolution neural network is proposed. The action structured method (ASM) is proposed to model the skeleton of human action through the pose estimation algorithm, which removes the interference of complex background. Besides, the object detection algorithm is utilized to locate multiple people to transfers the fall detection issue of multi-person to single person fall detection. The proposed cascaded dilated graph convolution network (CD-GCN) enlarges the receptive field by the dilated operation, effectively extracts action features from joint node coordinates, and fuses multichannel features with different dilation rates, then finally obtains the classification results. The proposed method achieves the best accuracy on three public datasets and one self-collected dataset, which is out-performing other state-of-art fall detection methods.

Keywords: Fall detection · Action structured method · Pose estimation · Multichannel · Cascaded dilated graph convolution network

1 Introduction

Fall detection has a wide demand of application and research significance in the field of safety monitoring in smart pension, smart city [1] and smart factory [2]. In the field of smart pension, real-time monitoring of the fall action of the elderly can effectively reduce the casualty rate, and enable the elderly to receive treatment in the first time. In the field of public safety monitoring [3] in smart city and smart factory, fall action is also an important detection action for safe production. An effective and rapid detection method for fall can improve people's quality of life and production level. Most existing visual-based fall detection methods can only recognize fall action in simple environments.

S.-H. Wang and Y.-D. Zhang (Eds.): ICMTEL 2022, LNICST 446, pp. 525–535, 2022.
https://doi.org/10.1007/978-3-031-18123-8_41

When there are moving pedestrians or objects in a complex background, the image-based method mixes the features of multiple people to detect falls. These extracted features are inaccurate to classify each person, which leads to failed detection. Some methods use the human skeleton to reduce the interference of background, such as scene or light changes. These methods rely on the pose estimation algorithms [4] to obtain the human skeleton representation. The scene in the background can be eliminated and reserve the skeleton information, but the pose estimation algorithm may incorrectly process some of the back-ground pixels as human skeleton, resulting in inaccurate extraction of action features by deep learning network. As shown in Fig. 1, cartons are mistaken for human skeleton, and the error processing representation is shown on the right part of the image, which makes the feature redundancy unable to discriminate the fall action.

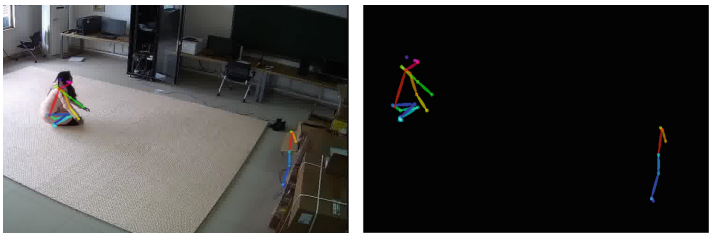

Fig. 1. The pose estimation algorithm error processes the skeleton representation.

In addition, the existing methods only extract action features from RGB images, but neglect to extract features from human joint coordinates, resulting in poor recognition accuracy. These methods use pixel-level features to detect falls, which leads to lack of feature scale invariance and loss of structured information. Human skeleton contains structured information which is hard to extract from image.

In this paper, a fall detection method based on action structured method and cascaded dilated graph convolution neural network is proposed. The contributions of the proposed method are summarized as follows:

(1) An action structured method (ASM) is proposed to model the skeleton of human actions by using pose estimation algorithms, which removes the interference of complex background. Also, the object detection algorithm is utilized to locate multiple people to transfers the fall detection issue of multi-person to single person fall detection.
(2) The cascaded dilated graph convolution network (CD-GCN) is proposed to enlarge the receptive field by the dilated operation, effectively extracts action features from the coordinates of joint nodes, and fuses the multichannel features with different dilation rates. Then finally obtains the classification results.

2 Related Works

Automatic fall detection technology has become a hot research topic in recent years [5–7]. With the aging of Chinese population and the development of public security

technology in smart city, the automatic detection of human falls is of great importance for protecting vulnerable groups such as the elderly and children and ensuring public safety. Falling is the second cause of accidental injury for people of all ages, and it is the primary cause of accidental injury for people over 79 years old [8]. Fall detection methods are mainly based on three categories of wearable sensors [9], environmental sensors [10] and machine vision [11]. In addition, many published methods focus on general motion recognition and video understanding, rather than specific fall detection. There is a problem of inaccurate extraction due to the small gap between feature classes. The wearable sensor method is to detect the acceleration and position of the human movement by wearing one or more sensors on the pedestrian to analyze whether it conforms to the motion characteristics of human falling [12–14]. These methods are uneconomical and can only detect the wearer of the sensor, which cannot cover all people. Environmental sensor method uses multiple sensors arranged in the environment to detect floor vibration, the current generated near the fall, falling sound and other information to judge human fall [15]. This method has poor anti-interference and high misjudgement rate, and cannot be widely used. The machine vision method judges human fall action through video images. Literature [16] designed local spatio-temporal points of interest were designed to represent features, and then the support vector machine (SVM) was used to classify and identify fall. In literature [17], human joint information is obtained by object detection algorithm, and feature vectors are formed, and then the integrated classifier is used to detect fall. Literature [18] proposed a multi-feature fusion detection method. In literature [19], a deep neural network-based method for identifying fall behaviour is proposed. Abobakr et.al [20] proposed a method based on skeleton information and random decision forest to extract fall features, and then a method using human skeleton information is proposed to remove background information through support vector machine classification. Xin et al [21] proposed a method of removing background information by using human skeleton information, and then extracting temporal and spatial features of actions through three-dimensional convolution neural network. Mastorakis et al. [22] proposed a human modeling method with other occlusion and used Hausdorff distance measure for fall detection. Panahi et al. [23] used support vector machine to classify the action features of depth images. When there are moving pedestrians in complex backgrounds, the image-based methods mixed the feature of multiple people to detect the fall. These extracted features are inaccurate, leading to detection failure. Moreover, the existing methods only extract action features from RGB images, neglecting to extract features from human joint coordinates, resulting in poor recognition accuracy. These methods use pixel-level features to detect falls, resulting in the invariance missing of feature scale and loss of structured information.

In order to solve the abovementioned problems, a fall detection method based on ASM and CD-GCN is proposed. The proposed ASM models the skeleton of human action through the pose estimation algorithm, which removes the interference of complex background. In addition, the object detection algorithm is utilized to locate multiple people to transfers the fall detection issue of multi-person to single person fall detection. The proposed CD-GCN enlarges the receptive field by the dilated operation, effectively extracts the action features from the coordinates of joint nodes, and fuses the multichannel feature with different dilation rate, then finally obtains the classification results.

3 Overview of Proposed Method

In this paper, a fall detection method based on ASM and CD-GCN is proposed. As shown in Fig. 2. The ASM is proposed to conduct skeleton model of human action through pose estimation algorithm [24] to remove the interference of background, and classify pedestrians by using the object detection algorithm [25] to transfer the fall detection issue of multi-person to single person fall detection. The proposed deep learning network based on the CD-GCN enlarges the receptive field and effectively extracts the fall features from joints coordinates. Finally, the classification results are obtained.

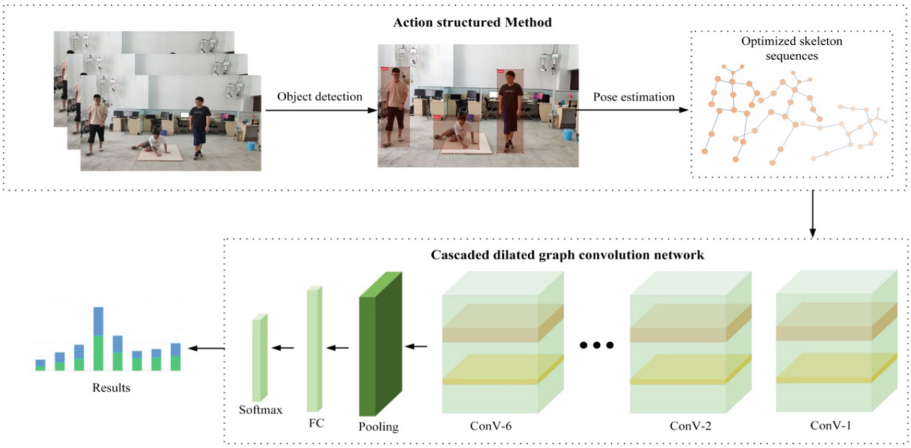

Fig. 2. The structure of the proposed method.

4 Action Structured Method

In order to remove the background interference in the image, as shown in Fig. 3, an action structured optimization method (ASM) is proposed in this paper. Firstly, the proposed method utilizes the YOLOv5 [25] object detection algorithm to classify the people in image, and uses the detection coordinates to reduce the image size. The ASM reduces the computational complexity of subsequent processing and improves the processing efficiency. Secondly, the coordinate information of each human skeleton in the image is obtained through pose estimation algorithm, OpenPose [24], and finally the skeleton sequence with spatio-temporal action features is obtained. The pose estimation algorithm has the problem of misrecognizing the background object as human body, which has been illustrated in Sect. 1, which results in redundant and erroneous human action information in the data. The object detection algorithm recognizes each person's region and then performs the pose estimation process, so that the skeleton information of each person can be obtained without the interference of background misjudgment, which transfers the fall detection issue of multi-person to single person fall detection and improves the feature extraction. Finally, the optimized skeleton sequences of each person with spatio-temporal features are obtained.

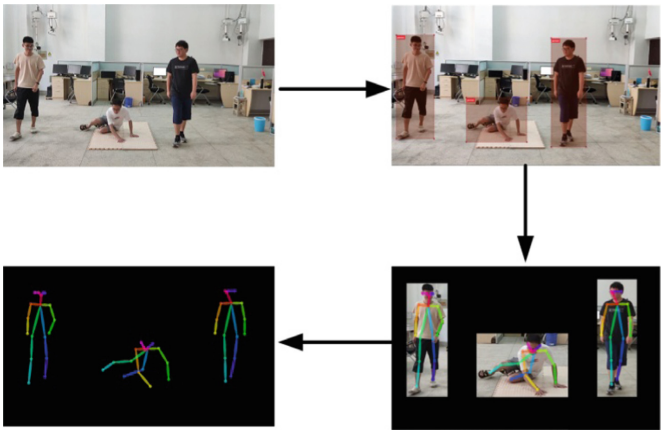

Fig. 3. The process of ASM.

5 Cascaded Dilated Graph Convolution Neural Network

In recent years, graph convolution network (GCN) theory has attracted more and more attention from researchers [26–29]. The advantage of graph convolution is that it can process data with non-Euclidean structure. The skeleton structure of human body is a natural non-Euclidean data, and the coordinates of each joint are native and original feature vectors. However, the existing methods extract action features from RGB images, which are not only susceptible to background interference, but also have a large gap in the extracted fall features, resulting in poor accuracy of detection methods. A cascaded dilated graph convolution neural network (CD-GCN) for feature extraction and fall detection is proposed. The proposed dilated graph convolution is shown in Fig. 4. The dilated theory is transformed from CNN to GCN.

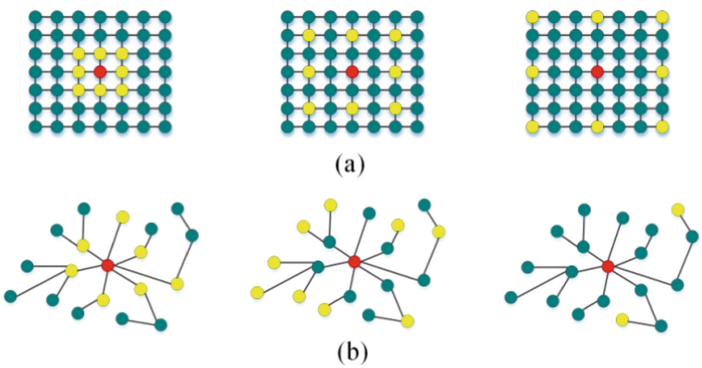

Fig. 4. Figure (a) is the traditional CNN with regular structure; (b) is the non-Euclidean structure. The red node is the central node, the yellow node is the aggregated node with different dilation rate. (Color figure online)

The idea of dilated graph convolution is enlightened from traditional convolution neural network. As shown in Fig. 4, (a) shows the dilated convolution under the grid structure data based on RGB pixel rules, (b) shows the dilated graph convolution of non-European structured data. It can be seen that the aggregated pixels or nodes (yellow) of the central node or pixel point (red) are gradually diffused around. The degree of diffusion is defined as dilation rate, and the distance from the central node to aggregated node is taken as the quantitative index. Dilation rate $dr = 1$ means that the distance from the central node to the aggregated node is 1. In this paper, the feature of a cascaded dilated graph convolution is an aggregated eigenmatrix of different dilation rate. As shown in Fig. 5, the structure of the cascaded graph convolution operation is proposed. Features of different dilated rate are cascaded and fused.

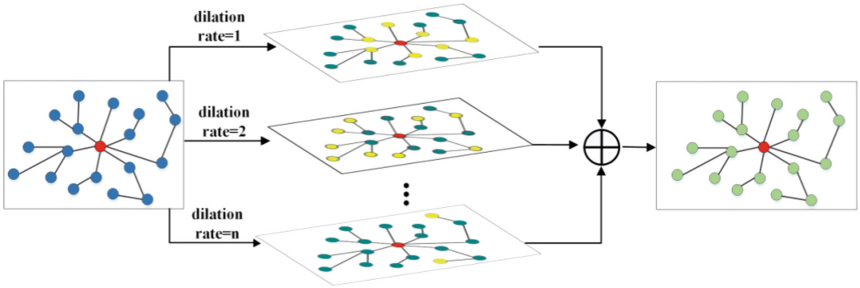

Fig. 5. The structure of cascaded dilated graph convolution operation.

As shown in Fig. 6, the proposed cascaded dilated graph convolution neural network in this paper consists of six convolution layers, one pooling layer, one full connection layer and one Softmax layer. The joint coordinates of the optimized skeleton sequence into the network as the input, and a $C \times T \times V$ tensor is designed, where T means the number of video frames, V means the number of human joints, and C means the joint coordinate data dimension. In this paper, C = 3, T = 40, V = 18 are set. The mathematical expression of graph convolution operation is shown in formula (1):

$$f = \Lambda^{-\frac{1}{2}}(A_m + 2I)\Lambda^{-\frac{1}{2}} f_{in} W \tag{1}$$

where f means the feature matrix, A_m means the adjacency matrix of human body structure, here is the cascaded graph data, and I means the identity matrix with the same dimension as A_m. In this paper, the mathematical expression of A_m is shown in formula (2):

$$A_m = A_1 + A_2 + \cdots + A_{\mathrm{n}} \tag{2}$$

where A_1 means the adjacency matrix with dilation rate $dr = 1$, A_2 means the adjacency matrix with the $dr = 2$, and A_{n} means the adjacency matrix with $dr = \mathrm{n}$. Then the formula (1) transfers to formula (3):

$$f = \sum_{m=1}^{n} \Lambda^{-\frac{1}{2}}(A_m + 2I)\Lambda^{-\frac{1}{2}} f_{in} W \tag{3}$$

In this paper, the accuracy performance with different dilation rate is studied. When the convolution strategy of $dr = 1$ and $dr = 2$ are fused to extract features, the optimal recognition accuracy is obtained.

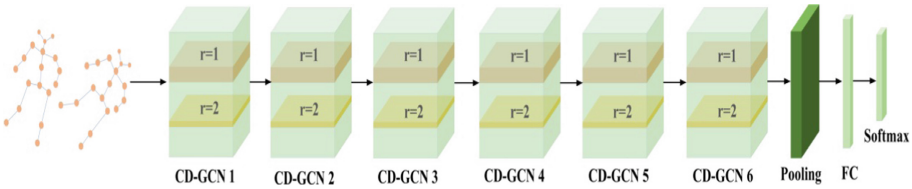

Fig. 6. The structure of proposed cascaded dilated graph convolution network.

6 Experiments

6.1 Datasets and Implementation Setup

Experiments and analyses are implemented on three public data sets and one self-collected dataset. These three public datasets are single-person fall datasets in a simple environment. UR dataset [30] contains 30 falling videos and 40 daily action videos. FDD dataset [31] contains 191 videos. MCFD dataset [32] contains 192 videos, including 7 types of actions such as fall, sit, walk and run. In this paper, the UR, MCFD and FDD datasets are divided into two categories: fall and daily behaviour. The self-collected dataset, for 5 classes, uses HIKVISION DS camera to collect video data. As shown in Fig. 7, we collected more than 150 videos with complex scenarios, with each video length of 150–250 frames and each video lasts 5–9 s. We also collected 300 daily non-fall videos with 4 classes, including normal walking, squatting, lying down and sitting.

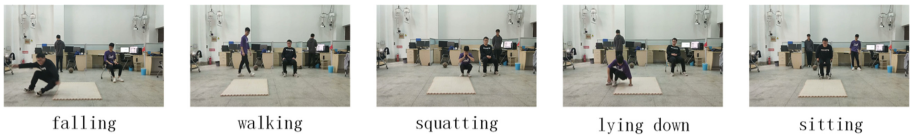

Fig. 7. Example frames of self-collected dataset.

The experiment is implemented on Ubuntu18.04 system, Intel Xeon E2 CPU and NVIDIA 16 GB RTX5000 GPU. The program is implemented in Pytorch 0.4.1 and Pycharm. The initial learning rate of the cascaded dilated graph convolution network is 0.1, the weight decay is 0.0001, and the batch size is 64. Atotal 30 epochs are trained, and the learning rate decayed every 10 epochs.

6.2 Comparison with State-of-Art Methods

A fall detection method based on ASM and CD-GCN is proposed. The ASM models the skeleton of human action through the pose estimation algorithm, which removes the interference of complex background. Also, the object detection algorithm is utilized to locate multiple people to transfers the fall detection issue of multi-person to single person fall detection. The proposed CD-GCN enlarges the receptive field by the dilated operation, effectively extracts the action features from the coordinates of joint nodes, and fuses the multichannel feature with different dilation rates, then finally obtains the classification results.

Evaluation of ASM. The ASM is proposed to conduct skeleton modeling of human action through pose estimation algorithm, and to remove the interference of background. The pose estimation algorithm has the problem of misrecognizing background objects as human body, which leads to redundant and erroneous human action information in the data. The object detection algorithm recognizes each person's region and then performs the pose estimation processing, so that the skeleton information of each person can be obtained without the interference of background misjudgment, which transfers the fall detection issue of multi-person to single person fall detection and improves the efficiency of feature extraction. Experiments results on four datasets are shown in Table 1. As shown in the table, experiments were performed with and without ASM, respectively, and performance evaluations were performed. The experiments show that the proposed method achieves the highest accuracy on four datasets. When only the human skeleton is input and ASM is not used, there is the possibility of multi-person skeleton interference and pose estimation misrecognizing the limb skeleton. The accuracy of the proposed method is slightly improved on the single-person fall dataset in a simple environment, while it is greatly improved on the self-collected dataset in a complex multi-person scenario.

Table 1. The accuracy evaluation of ASM (%).

Method	UR	FDD	MCFD	Self-collected
Without ASM	98.9	98.0	99.3	88.8
With ASM	**99.3**	**98.9**	**99.5**	**95.6**

Evaluation of CD-GCN. The proposed cascaded dilated graph convolution network (CD-GCN) is proposed to enlarge the receptive field by the dilated operation, effectively extracts the action features from the coordinates of joint nodes, and fuses the multichannel feature with different dilation rate. Then finally obtains the classification results. The evaluation of CD-GCN is shown in Table 2, the dilation combination of the cascade dilated graph convolution is evaluated. Dilation rate $dr = 1$ means the convolution without any dilation process. The experimental results show that when the features of the cascade dilation rate are 1 and 2, the proposed method achieves the highest accuracy.

Table 2. The accuracy evaluation of CD-GCN with different dilation rate. (%)

r	UR	FDD	MCFD	Self-collected
1	99.0	97.3	97.9	94.9
2	90.9	85.0	90.9	86.3
3	70.1	67.6	73.7	62.5
1+2	**99.3**	**98.9**	**99.5**	**95.6**
1+3	98.2	96.5	94.4	91.6
1+2+3	96.9	96.7	90.3	90.4

6.3 Comparison with State-Of-Art Methods

In this paper, the proposed method is compared with the state-of-the-art methods in recent years, as shown in Table 3. Literature [7–14] uses a method based on machine vision. [16–23]. These methods ignore the structured optimization of human body information, resulting in inaccurate detection, especially in a multi-person scene with a complex background, which is susceptible to interference from other action features. In addition, these methods only extract action features from RGB pixels, and neglect to extract features from human joint coordinates. The proposed method achieves accuracy improvement about 0.7%, 0.6% and 0.5% on UR, FDD and MCFD datasets, and 1.9% on self-collected datasets. The experiment proves the superiority of the proposed method.

Table 3. The accuracy comparison of proposed method with state-of-art methods. (%)

Method	UR	FDD	MCFD	Self-collected
Su [16]	96.3	97.5	98.1	93.7
Zhao [17]	97.2	98.3	99.0	89.9
Li [18]	83.5	87.9	90.0	80.9
Fan [19]	94.3	92.9	97.5	91.1
Abobakr [20]	96.1	95.2	96.1	92.3
Xin [21]	83.5	87.9	90.0	89.4
Mastorakis [22]	98.6	94.3	97.3	93.2
Panahi [23]	97.1	90.9	93.6	90.5
The proposed method	**99.3**	**98.9**	**99.5**	**95.6**

7 Conclusions

In this paper, a cascade dilated graph convolution neural network for fall detection in complex scenes is proposed. In order solve the problem that existing methods are susceptible to the interference of complex backgrounds, the proposed ASM uses pose estimation

algorithms to model human action skeleton, and combines object detection algorithm to remove the interference of objects and pedestrians in the background. Aiming at the problem of inaccurate extraction of action structure features in existing methods, a cascaded dilated graph convolution network is proposed to expand the receptive field and effectively extract the action features from joint point coordinates. In the future, human body occlusion and multi-limb overlap action recognition methods will be studied.

Acknowledgement. This research is supported by the National Natural Science Foundation of China (Grant No. 52107081); Young Talents Scientific Research Cultivation Project of the First Affiliated Hospital of Nanchang University (No. 701566001).

References

1. Buzachis, A., Celesti, A., Galletta, A., Fazio, M., Fortino, G., Villari, M.: A multi-agent autonomous intersection management (MA-AIM) system for smart cities leveraging edge-of-things and blockchain. Inf. Sci. **522**, 148–163 (2020)
2. Zhang, R.: Improved control for industrial systems over model uncertainty: a receding horizon expanded state space control approach. IEEE Trans. Syst. Man Cybern. Syst. **50**(4), 1343–1349 (2020)
3. Ahire, S.K., Wankhade, N.R.: Context-aware local binary feature learning for face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(5), 1139–1153 (2019)
4. Fang. H.S., Xie. S., Tai. Y.W., Lu. C.: RMPE: regional multi-person pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2353–2362 (2017)
5. Sadreazami, H., Bolic, M., Rajan, S.: Contactless fall detection using time-frequency analysis and convolutional neural networks. IEEE Trans. Industr. Inf. **17**(10), 6842–6851 (2021)
6. Tahir, A., Morison, G., Skelton, D.A., Gibson, R.M.: A novel functional link network stacking ensemble with fractal features for multichannel fall detection. Cogn. Comput. **12**(5), 1024–1042 (2020). https://doi.org/10.1007/s12559-020-09749-x
7. Mrozek, D., Koczur, A., Maysiak-Mrozek, B.: Fall detection in older adults with mobile IoT devices and machine learning in the cloud and on the edge. Inf. Sci. **537**(5), 132–147 (2020)
8. Mubashir, M., Shao, L., Seed, L.: A survey on fall detection: principles and approaches. Neurocomputing **100**, 144–152 (2013)
9. Qian, X., Chen, H., Jiang, H., Green, J., Cheng, H., Huang, M.: Wearable computing with distributed deep learning hierarchy: a study of fall detection. IEEE Sens. J. **20**(16), 9408–9416 (2020)
10. Liu, J., Tan, R., Han, G., Sun, N., Kwong, S.: Privacy-preserving in-home fall detection using visual shielding sensing and private information-embedding. IEEE Trans. Multimedia **23**, 3684–3699 (2020)
11. Khan, S.S., Hoey, J.: Review of fall detection techniques: a data availability perspective. Med. Eng. Phys. **39**, 12–22 (2017)
12. Medrano, C., Plaza, I., Igual, R., Sanchez, A., Castro, M.: The effect of personalization on smartphone-based fall detectors. Sensors **16**(1), 117 (2016)
13. Cola, G., Avvenuti, M., Vecchio, A., Yang, G.Z., Lo, B.: An on-node processing approach for anomaly detection in gait. IEEE Sens. J. **15**(11), 6640–6649 (2015)
14. Wei, W., Song, H., Li, W., Shen, P., Vasilakos, A.: Gradient-driven parking navigation using a continuous information potential field based on wireless sensor network. Inf. Sci. **408**, 100–114 (2017)

15. Rimminen, H., Lindstrom, J., Linnavuo, M., Sepponen, R.: Detection of falls among the elderly by a floor sensor using the electric near field. IEEE Trans. Inf Technol. Biomed. **14**(6), 1475–1476 (2010)
16. Su, S., Wu, S.-S., Chen, S.-Y., Duh, D.-J., Li, S.: Multi-view fall detection based on spatio-temporal interest points. Multimedia Tools Appl. **75**(14), 8469–8492 (2015). https://doi.org/10.1007/s11042-015-2766-3
17. Zhao, X., Hu, A., He, W.: Fall detection based on convolutional neural network and XGBoost. Laser Optoelectron. Progress **57**(16), 248–256 (2020)
18. Li, Y., Yang, B.: Fall detection method based on ViBe algorithm and multi-feature fusion. Chin. J. Electron Devices **42**(6), 1583–1589 (2019)
19. Fan, Y., Levine, M.D., Wen, G., Qiu, S.: A deep neural network for real-time detection of falling humans in naturally occurring scenes. Neurocomputing **260**, 43–58 (2017)
20. Abobakr, A., Hossny, M., Nahavandi, S.: A skeleton-free fall detection system from depth images using random decision forest. IEEE Syst. J. **12**(3), 2994–3005 (2018)
21. Xiong, X., Min, W., Zheng, W.-S., Liao, P., Yang, H., Wang, S.: S3D-CNN: skeleton-based 3D consecutive-low-pooling neural network for fall detection. Appl. Intell. **50**(10), 3521–3534 (2020). https://doi.org/10.1007/s10489-020-01751-y
22. Mastorakis, G., Ellis, T., Makris, D.: Fall detection without people: a simulation approach tackling video data scarcity. Expert Syst. Appl. **112**, 125–137 (2018)
23. Panahi, L., Ghods, V.: Human fall detection using machine vision techniques on RGB-D images. Biomed. Signal Process **44**, 146–153 (2018)
24. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D Pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 172–186 (2019)
25. Tan. S., Lu. G., Jiang. Z., Huang. L.: Improved YOLOv5 network model and application in safety helmet detection. In: Proceedings of the IEEE International Conference on Intelligence and Safety for Robotics, pp. 330–333 (2021)
26. Kipf. T.N., Welling M.: Semi-supervised classification with graph convolutional networks. In: Proceedings of the International Conference on Learning Representations (2016)
27. Albert, M.M., Javier, R.H.: 2D–3D geometric fusion network using multi-neighbourhood graph convolution for RGB-D indoor scene classification. Inf. Fus. **76**, 46–54 (2021)
28. Qin. L., Che, W., Ni. M., Li. Y., Liu. T.: Knowing where to leverage: context-aware graph convolution network with an adaptive fusion layer for contextual spoken language understanding. In: Proceedings of the IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 1280–1289 (2021)
29. Zi, W., Xiong, W., Chen, H., Chen, L.: TAGCN: station-level demand prediction for bike-sharing system via a temporal attention graph convolution network. Inf. Sci. **561**, 274–285 (2021)
30. Kwolek, B., Kepski, M.: Human fall detection on embedded platform using depth maps and wireless accelerometer. Comput. Methods Programs Biomed. **117**(3), 489–501 (2014)
31. Charfi. I., Miteran. J., Dubois. J., Atri. M., Tourki. R.: Definition and performance evaluation of a robust SVM based fall detection solution. In: Proceedings of the IEEE Eighth International Conference on Signal Image Technology and Internet Based Systems, pp. 218–224 (2012)
32. Auvinet, E., Multon, F., Alain, S.A., Rousseau, J., Meunier, J.: Fall detection with multiple cameras: an occlusion-resistant method based on 3D silhouette vertical distribution. Proc. IEEE Trans. Inf. Technol. Biomed. **15**(2), 290–300 (2011)
33. Lu, N., Wu, Y.D., Feng, L., Song, J.B.: Deep learning for fall detection: three-dimensional CNN combined with LSTM on video kinematic data. IEEE J. Biomed. Health Inform. **23**(1), 314–323 (2019)