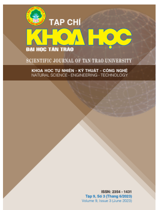
TẠP CHÍ KHOA HỌC ĐẠI HỌC TÂN TRÀO

ISSN: 2354 - 1431

http://tckh.daihoctantrao.edu.vn/



A NOVEL NETWORK BASED ON ATTENTION FOR FALLING DETECTION

Ma Thi Hong Thu[1], *Dinh Thi Lien*[2], *Phung Thi Thu Trang*[2*]

[1]Tan Trao University, Vietnam

[2]Thai Nguyen University, Vietnam

[*]Email address: phungthutrang.sfl@tnu.edu.vn

DOI: 10.51453/2354-1431/2023/960

Article info

Received:06/12/2012

Revised: 25/02/2023

Accepted: 16/5/2023

Keywords:

Falling, Detection, Action, Recognition, Attention.

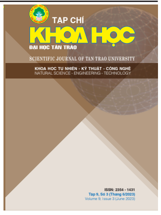
Abstract:

The attention mechanism has been studied extensively in recent years to enhance the model's learning ability. In this paper, we propose a new attention mechanism including the temporal attention module and spatial attention module. These two modules are combined in the 3D ResNet-18 network to provide "attention" to the critical features of the volume. In particular, the temporal attention module exploits the motion relationship between frames, and the spatial attention module is interested in the spatial relationship between features. The experimental results for the proposed model show that our proposed method achieves competitive performance compared with the recently published modern deep and heavy networks.

TẠP CHÍ KHOA HỌC ĐẠI HỌC TÂN TRÀO

ISSN: 2354 - 1431

http://tckh.daihoctantrao.edu.vn/



MỘT MẠNG MỚI DỰA TRÊN SỰ CHÚ Ý ĐỂ PHÁT HIỆN TÉ NGÃ

Ma Thị Hồng Thu[1], Đinh Thị Liên[2], Phùng Thị Thu Trang[2]*

[1]Trường Đại học Tân Trào

[2]Đại học Thái Nguyên

[*]Địa chỉ email: phungthutrang.sfl@tnu.edu.vn

DOI: 10.51453/2354-1431/2023/960

Thông tin bài viết

Ngày nhận bài: 06/12/2012

Ngày sửa bài: 25/02/2023

Ngày duyệt đăng: 16/5/2023

Từ khóa:

Té ngã, Nhận dạng hành động, Sự chú ý.

Tóm tắt

Cơ chế chú ý đã và đang được nghiên cứu rộng rãi trong những năm gần đây nhằm mục đích nâng cao khả năng học tập của mô hình. Trong bài báo này, chúng tôi đề xuất một cơ chế chú ý mới bao gồm mô-dun chú ý thời gian và mô-đun chú ý không gian. Hai mô-đun này được kết hợp với nhau trong mạng 3D ResNet-18 nhằm cung cấp sự "chú ý" vào các đặc trưng quan trọng của khối tích. Cụ thể, mô-dun chú ý thời gian sẽ khai thác mối quan hệ chuyển động giữa các frames và mô-đun chú ý không gian quan tâm đến mối quan hệ không gian giữa các đặc trưng. Kết quả thử nghiệm đối với mô hình được đề xuất cho thấy phương pháp đề xuất đạt được hiệu suất cạnh tranh so với các mạng sâu và nặng hiện đại được công bố gần đây.

1. Introduction

Fall is defined as an event that results in an accidental fall, lying on the ground or other lower position. Often, the causes of falls can be a slip, a seizure, or a stroke.

Stroke, also known as cerebrovascular accident, has become one of the leading causes of death in the world because stroke often occurs suddenly with very few warning signs. It can even appear suddenly in an ordinary person when they are resting or working. It can have very serious consequences such as death, permanent paralysis on one side of the body, or the loss of the ability to speak. Emergency patient care needs to be very urgent, especially if given at the right time of "golden hour", the possibility of saving lives will be very high as well as limiting severe sequelae.

In today's modern society, many elderly people have to stay at home alone. It is a consequence of migration when people of working age gather in big cities in search of job opportunities or even international migration. There have been many cases of elderly people after falling unable to stand up on their own or call for help from others. Therefore, there is a need to develop a solution capable of early detection of falls in the elderly.

The problem of detecting people in general and detecting people falling in real-time in particular is one of the problems that have received the attention of many domestic and foreign researchers. Some prominent publications can be mentioned as Trang et al. [1] proposed a deep learning model based on ResNet to solve the problem. In which, the authors propose to use (2+1)D Convolution to replace the traditional 3D Convolution. An et al. [2] proposed LSTM deep learning model to detect falls by accelerometer. Cameiro et al. [3] proposed a multi-stream model to recognize fall events in the input video. Other than [3] focus on many different input objects to build separate recognition models for each input, Chen et al. [4] proposed a fall recognition system on embedded devices based on available CNN models such as LeNet, AlexNet, GoogleNet. Inheriting the idea of distilling knowledge [5, 6], self-identify knowledge [7, 8], Vu et al [9] propose a model of selfdistillation of problem-solving knowledge.

Different from the works mentioned above, in this paper, we propose a fall detection model based on deep learning model combined with attention mechanism. Specifically, we present a new network architecture based on the (2+1)D ResNet network and integrate additional spatial and temporal attention modules to enhance network performance. Compared with recent publications, our model proposes to achieve better accuracy than some networks such as VGGNet, AlexNet, ResNet..

The rest of the paper is organized as follows: Section 2 Recent studies present machine learning-based publications, specifically covering two groups: traditional machine learning methods and deep learning methods. Section 3 The proposed method presents in detail the network architecture proposed in the paper along with the interesting modules. Section 4 Experimental results present the results of the proposed model and compare these results with some recently published models on basic measures such as Accuracy, Sensitivity, Specificity, etc. Finally, the Conclusions and References are given in Section 5.

2. Related Work

2.1. Traditional machine learning methods

One of the traditional methods of machine learning that has been widely used since the 2000s is the SVM algorithm. In 2020, Chen and his associates [4] proposed an IoT model based on two algorithms, HOG and SVM, for early identification of people with falls. Specifically, the author's IoT system is divided into three main components: Local A, Local B and online server. In there: (i) Local A contains a local board connected to the worm sensor. When people are recognized by this worm sensor, the binary images of those recognized people will be stored here. The HOG features of the binary image (from the depth sensor) will be extracted and sent to Local B. To avoid affecting user privacy, Local A is not connected to the Internet. (ii) Local B uses SVM to identify fall events based on features sent from Local A. Note Local B only uses features sent from Local A to identify falls and images (from Local A). are not sent to Local B. In other words, even if Local B is attacked from the Internet, the risk of affecting user privacy is very low. (iii) When the online server receives a fall alarm, family members or the hospital will be notified and provide first aid to the fall person in a timely manner.

Other than [4], Bo-Hua Wang and associates [10] đHe proposed a method to identify people falling through the combination of two algorithms, MLP and Random Forest. Specifically, the authors' approach proposes to divide the fall event into two parts: the falling state and the postfall state. Each algorithm model will be proposed to solve each of these states and finally the combined result of both algorithm models will be used as the final output for the fall/no-fall classification.

2.2. Deep learning-based methods

Different from the traditional models above, the models using Convolutional neural network (CNN) and Long short-term memory (LSTM) have now been and are achieving many impressive achievements in market problems. computer sense and image processing. In which, there have also been many CNN/LSTM models proposed to solve the problem of falling person identification.

Specifically, the group of authors [11] proposed a new model based on the LSTM architecture, and published a dataset of fall actions in complex crowded scenes. First, the Openpos algorithm [12] used to extract 2D human pose from a fall recognition video frame where on top is the Multi-Person Posture Estimator. Body parts of the same person are linked together. Bottom left are the PAFs corresponding to the limb connecting the right elbow and right wrist. Orientation color coding. Bottom right is an enlarged view of the predicted PAFs. At each pixel in the field, a 2D vector encodes the position and orientation of the limbs. Then, the parts that have a significant impact on human falls were selected and the time step of a video sequence of 24 frame poses was formed. Finally, the pose vectors were fed into the two-layer LSTM network to identify the fall event.

Vu et al. [9] proposed a model training method based on self-distillation to improve the accuracy of lightweight networks. In it, the authors propose a method of self-digesting knowledge through the last hidden layer (before entering the classification layer) and using data enhancement to create many different variations of the same video. input. The output of the hidden layer will eventually be normalized so that two variants of the same video will have to give the same output.

Different from the above mentioned methods using CNN and LSTM networks, in [13], The authors propose an AutoEncoder-type hourglass network for fall event recognition called HCAE. To be able to capture more rich information from the input video, research using HRU in the encoder. In addition, the multi-tasking mechanism is designed to improve the ability to learn the representation of network features by assigning the secondary task of frame reconstruction along with the main task of fall recognition.

3. Our Proposed Method

3.1. Problem Setup

Given a training dataset consisting of *N* samples designed in the form D = $\{(x^{(1)},y^{(1)}), (x^{(2)},y^{(2)}),.., (x^{(N)}, y^{(N)})\}$ so that $x^{(i)}$ represents video clips and $y^{(i)}$ are the labels (i.e. fall or not) of the *i* video. Let B = $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ be an input mini-batch from the dataset whether training D. We express the binary cross-entropy loss function (*BCE*) between two probability distributions *p* and *q* as *BCE*(*p*, *q*) as shown below:

$$BCE(p, q) = -n \quad \underset{i=1}{Pi\text{xlog } qi+(1-Pi)\text{xlog}(1-qi))} \quad 1^{n} \qquad (1)$$

Where, *P* is encoded as a binary 0 or 1. With 0 representing the input video containing no fall action and 1 representing the input video containing the fall action.

3.2. Network Architecture

In 2015, He et al have proposed an efficient network, named ResNet [14]. One of the biggest contributors is the skip connection technique. ResNet can avoid the vanishing gradient problem without sacrificing network performance. That keeps the deep layers at least no worse than the shallow ones. Furthermore, the upper layers receive more information directly from the underlying layers so that it adjusts weight more efficiently with the ResNet architecture. Following the success of ResNet, many modeling architectures have been introduced based on the ResNet backbone. Experiments have shown that these architectures can be trained with CNN models have depths of up to thousands of layers. ResNet has quickly become the most popular architecture in deep learning and computer vision. Hara and associates [15] proposed an extended ResNet 3D model from ResNet to perform action recognition on video input objects.

Table 1 describes the overview architecture of the ResNet-18 network in conjunction with the attention modules (att) recommended by us. In which, the network is divided into 4 main convolutional blocks (Conv block) respectively Conv block 2, Conv block 3, Conv block 4 and Conv block 5. Attention mechanism is implemented at the last layer of each block to aim. The goal is to provide “attention” to the important features of the current convolutional block before moving on to the next block.

Table 1: The overall architecture of the proposed network for the fall detection problem is based on ResNet-18 architecture. In which, Conv represents the classes or blocks containing the convolution layer, Max Pool, Ave Pool are respectively Max Pooling and Average Pooling, FC represents the Fully Connected layer.

Layer	Specification		Output size
Input			T X 224 X 224 X 3
Conv1	7 X 7 X 7,64 stride = 1,2,2		T X 112 X 112 X 64
Max Pool	3X3X3 stride = 1,2,2		T X 56 X 56 X 64
Conv block 2	■ 3 X 3 X 3,64 ■ 3 X 3 X 3,64 . *fc*, [16,64] .	X2	TX56X56X64
Conv block 3	3X3X3,128 3X3X3,128 *f c*, [32, 128]	X 2	2 X 28 X 28 X 128
Conv block 4	3X3X3,256 3X3X3,256 *fc*, [64, 256]	X 2	4 X 14 X 14 X 256
Conv block 5	3X3X3,512 3X3X3,512 *f c*, [128, 512]	X 2	T X 7 X 7 X 512 8
Global_Ave_Pool			2 X 128
Drop Out	drop rate = 0.5		1 X 128
FC	unit = 1, sigmoid		1

3.3 Attention Module

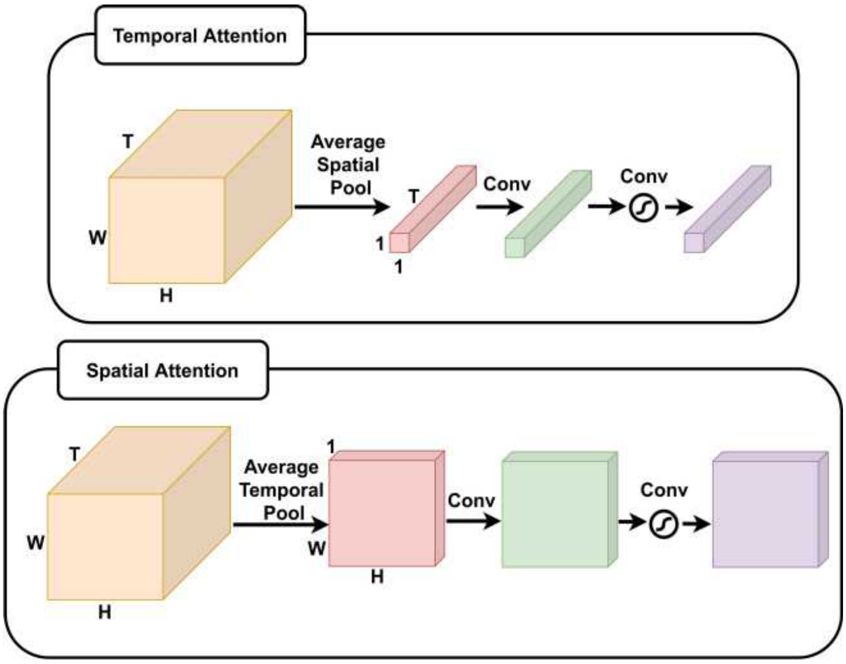
Our Att module is inspired by the attention mechanism in images as in some popular networks such as SENet [16], CBAM [17, 18], etc. As depicted in the picture 1, we propose two attention modules namely temporal attention module and spatial attention module.

We assume that the input to the time attention module is a matrix of size (T, H, W, N). After applying Average Spacial Pool, we have anew matrix with size (T, 1, 1,N). Through two Conv layers we get output with size (T, 1, 1, N). In spatial attention module we use Average Temporal Pool instead of Average Spacial Pool, then through two Conv layers we get output with dimension (1, H, W, N).

Time attention module. We create time attention maps by exploiting relationships between feature movements. Attention over time focuses

on "what" is moving from the sequence of input frames. To calculate attention over time efficiently, we applied the Average Spacial Pool to shrink the spatial size. We then apply convolutional layers to create an attention map over time.

Spatial Attention Module. In addition to temporal attention, spatial attention focuses on "where". This is also an important piece of information. This module is interested in the interspa- tial relationship of the feature. To calculate spatial attention, an average pool is first used along the time axis (Average Temporal Pool). We then apply convolutional layers to create a spatial attention map.



Picture 1: Architectural details of the two modules pay attention. As illustrated, the time module (spatial module) takes the global spatial (temporal) aggregate outputs and forwards them to the convolutional layers. The last layer of each module is the sigmoid function.

4. Experiment

In this section, we first introduce the datasets used and the evaluation metrics. Next we describe in detail the implementation steps of the model implementation, finally we provide the evaluation results of the proposed model and compare it with other modern methods in terms of Accuracy, Recall , and Specificity.

4.1. Datasets and Metrics

In this paper, we used two standard data sets for fall recognition, the FDD set and the URFD set, to evaluate the performance of the proposed model compared with other modern methods.

FDD Dataset: Published in 2013. This dataset contains videos shot in locations such as coffee shops or at home. All videos in the dataset were shot with a single camera at 25 fps and set to a frame resolution of 320 X 240 pixel. The dataset consists of 130 videos with 99 videos containing different fall actions. Actors in each video perform normal actions at home and fall at different times, all actions performed at random.

Datasets URFD: Built by Bogdan Kwolek et al [19] in 2014 to identify people who fell through different types of devices such as cameras, accelerometers, Microsoft Kinect (in this study we only used camera-recorded videos in the dataset and without using information from other devices). The dataset consists of70 videos with 30 videos containing different fall actions and the remaining 40 videos containing normal daily activities, such as: sitting, walking, bending, etc. Details of the parameters of the two datasets are described in the table 2.

Table 2: Detailed description of the two data sets FDD and URFD used

Specification	FDD	URFD
Year publication	2013	2014
Number of videos	130	70
Number of fall videos	99	30
Average number of frames per video	239.4	139.4
Average fall action frames	31	30

Rating Index: The fall action recognition task can be thought of as a supervised learning problem with binary classification. In particular, the model needs to decide whether an input video clip should be labeled as a fall. To evaluate classifier performance, Accuracy, Recall and Specificity are common metrics among the methods used in this framework.

4.2. Imprementation Detail

The proposed network is trained from scratch with the Stochastic Gradient Descent optimizer. The training videos are divided into clips of 16 frames in length and each frame size is 224 X 224 X 3. The batch size is set to 32 clips. The learning rate is initialized to 0.001 and decreases 10 times if the model does not improve Accuracy for 10 consecutive epochs. All models are trained with 100 epochs and the indices are calculated on the test set. In this work, we use the set of data augmentation strategies proposed by Vu et al [7]. Each increment operator has a probability of being chosen of 0.5. To evaluate the network performance, we compare the results of the model with the independent training methods (base method) and the methods proposed recently in [20, 3, 21] About Accuracy, Recall, Specificity. The entire model is trained on a computer configured with 2 Xeon E5- 2673 v3 2.4GHz CPUs with 12 cores for each CPU (total 24 cores and 48 threads), 64GB of DDR4 RAM and 2 GPUs 2070S.

4.3. Performance Comparision

We present the model test proposed by us and compare it with other modern methods on the FDD dataset in the table. 3. As shown in the table 3, Our model obtains modern performance compared to existing methods such as VGG, Multi-Stream, MobileNet3D. Specifically, our method achieves 97.7% Accuracy, 99.2% Recall và 100.0% Specificity. Furthermore, our model reaches 100.0% Specificity, This proves that the specificity of the model is completely accurate for cases other than falls. For other measurements, our proposed method is only poor 1.8% Accuracy (compared to MobileNet3D [21]) và 0.8% Recall (compared to RGB [3]).

We continue to compare our proposed model with existing methods on the URFD dataset in the table 4. Our method suggests the best performance in terms of Recall and Specificity (ie 100.0% Recall and 100.0% Specificity). These results show that the proposed network and our approach are very effective for fall recognition with any dataset, especially the URFD dataset. Notably, our methods all achieve 100.0% Specificity on both FDD and URFD datasets. This proves that the specificity of the model is completely accurate for cases other than falls.

5. Conclusion

In this paper, we have proposed a simple but effective model for fall action recognition. Specifically, our model is based on the ResNet-18 network architecture incorporating two more attention mod- ules namely temporal attention module and spatial attention module. These two modules help the network to enhance the ability to learn features and improve the generalization performance of the model. Experimental results have shown that the proposed method achieves competitive performance compared to the recently published modern deep and heavy networks. In the future, we continue to improve the networks and combine them with more modern attention models such as Transformer.

Table 3: Comparison of Accuracy, Recall, Specificity and model size between the proposed Self-KD method and the SOTA method on the FDD dataset.

Method	Accuracy	Recall	Specificity
VGG [20]	97.00%	99.00%	97.00%
RGB [3]	80.52%	100%	79.20%
OF [3]	96.43%	99.90%	96.17%
PE [3]	63.01%	100%	60.15%
OF&PE&RGB [3]	98.43%	99.90%	98.32%
MobileNet3D [21]	99.50%	99.00%	99.80%
Our method	97.70%	99.20%	100%

Table 4: Comparison of Accuracy, Recall and Specificity and model size between the proposed Self-KD method and the SOTA method on the URFD dataset.

Method	Accuracy	Recall	Specificity
VGG [20]	95.00%	100%	92.00%
RGB [3]	96.99%	100%	96.61%
OF [3]	96.75%	100%	96.34%
PE [3]	93.24%	94.41%	93.09%
OF&PE&RGB [3]	98.77%	100%	98.62%
MobileNet3D [21]	99.90%	100%	99.90%
Our method	98.10%	100%	100%

Acknowledgment This study is the result of the grassroots project code 2022.2.02, supported by Tan Trao University, Tuyen Quang, Vietnam.

Refences

[1] P. T. T. Trang, M. T. H. Thu, "*MỘT MÔ HÌNH HỌC SÂU CHO BÀI TOÁN PHÁT HIỆN NGƯỜI BỊ NGÃ,*" TNU Journal of Science and Technology, 225(14), 48-53, 2020.

[2] T. C. Án, L. M. Phúc, Đ. T. Đức, N. B. Hùng, L. Đ. Chiến, P. T. X. Diễm, S. B. Pha, "*Phát hiện té ngã cho người cao tuổi bằng gia tốc kế và mô hình học sâu Long Short-Term Memory*." Tạp chí Khoa học Trường Đại học Cần Thơ, (CĐ Công nghệ TT 2017), 65-71, 2017.

[3] S. A. Carneiro, G. P. da Silva, G. V. Leite, R. Moreno, S. J. F. Guimarães, H. 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), 293-298, IEEE, 2019.

[4] Y. Chen, X. Kong, L. Meng, H. Tomiyama, "*An edge computing based fall detection system for elderly persons,"* Procedia Computer Science, 174, 9-14, 2020.

[5] D.-Q. Vu, N. T. Le, J.-C. Wang, "*(2+ 1) D Distilled ShuffleNet: A Lightweight Unsupervised Distillation Network for Human Action Recognition*," in 2022 26th International Conference on Pattern Recognition (ICPR), 3197-3203, IEEE, 2022.

[6] J. Gou, B. Yu, S. J. Maybank, D. Tao, "*Knowledge distillation: A survey*," International Journal of Computer Vision, 129, 1789-1819, 2021.

[7] D.-Q. Vu, N. Le, J.-C. Wang, "*Teaching yourself: A self-knowledge distillation approach to action recognition,*" IEEE Access, 9, 105711-105723, 2021.

[8] D.-Q. Vu, J.-C. Wang, et al., "*A novel self-knowledge distillation approach with siamese representation learning for action recognition*," in 2021 International Conference on Visual Communications and Image Processing (VCIP), 1-5, IEEE, 2021.

[9] Q. V. Duc, T. Phung, M. Nguyen, B. Y. Nguyen, T. H. Nguyen, "*Self-knowledge Distillation: An Efficient Approach for Falling Detection*," in Artificial Intelligence in Data and Big Data Processing: Proceedings of ICABDE 2021, 369-380, Springer, 2022.

[10] B.-H. Wang, J. Yu, K. Wang, X.-Y. Bao, K.-M. Mao, "*Fall detection based on dual-channel feature integration,*" IEEE Access, 8, 103443-103453, 2020.

[11] M. M. Hasan, M. S. Islam, S. Abdullah, "Robust posebased human fall detection using recurrent neural network," RAAICON, 48-51, 2019.

[12] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, "*Realtime multi-person 2d pose estimation using part affinity fields,*" CVPR, 7291-7299, 2017.

[13] X. Cai, S. Li, X. Liu, G. Han, "*Vision-based fall detection with multi-task hourglass convolutional autoencoder,*" IEEE Access, 8, 44493-44502, 2020.

[14] K. He, X. Zhang, S. Ren, J. Sun, "*Deep residual learning for image recognition*," in CVPR, 770-778, 2016.

[15] K. Hara, H. Kataoka, Y. Satoh, "*Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?"* in CVPR, 6546-6555, 2018.

[16] J. Hu, L. Shen, G. Sun, "*Squeeze-and-excitation networks*," in Proceedings of the IEEE conference on computer vision and pattern recognition, 7132-7141, 2018.

[17] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, "*Cbam: Convolutional block attention module*," in Proceedings of the European conference on computer vision (ECCV), 3-19, 2018.

[18] T. Phung, V. T. Nguyen, T. H. T. Ma, Q. V. Duc, "*A (2+ 1) D attention convolutional neural network for video prediction*," in Artificial Intelligence in Data and Big Data Processing: Proceedings of ICABDE 2021, 395406, Springer, 2022.

[19] B. Kwolek, M. Kepski, "*Human fall detection on embedded platform using depth maps and wireless accelerometer,*" Computer methods and programs in biomedicine, 117(3), 489-501, 2014.

[20] A. Nunez-Marcos, G. Azkune, I. Arganda-Carreras, "*Vision-based fall detection with convolutional neural networks*," Wireless communications and mobile computing, 2017, 2017.

[21] T. X. Hoa, V. R. M. Madrid, E. A. Albacea, "*A Lightweight Model for Falling Detection*," RIVF, 15, 2021.