Video Based Fall Detection Using Deep Convolutional Neural Network

Gangireddy Prabhakar Reddy[1], M. Kalaiselvi Geetha[2]

[1] Research Scholar, [2]Professor

[1, 2] Department of Computer Science and Engineering, Annamamai University,
Chidambaram, Tamilnadu, India

[1]prabhakar.sp17@gmail.com, [2]geesiv@gmail.com

Abstract *– Falling often causes deadly conditions such as unconsciousness and related injuries among the elderly population if failing provided with aid and caretakers nearby. In this context, an automatic fall monitoring system gains its popularity by solving the problem with immediate prompting, thereby allowing the caretakers and other persons to get activated with an alarm message. It assists older adults in living without fear of falling and being independent in society. In recent decades, vision-based fall monitoring receiving attention among research communities for its diversified features. It helps identify the human in the intended regions, and by using the collected phenomenon from the area, it trains the fall recognition classifiers. Besides, human detection errors and lack of massive-scale datasets make the vision-based fall monitoring face challenges like robustness and efficiency in performing generalization to invisible regions. Hence a robust learning and classification system is reasonably needed to combat the challenges. In this proposed system, automatic fall detection using deep learning is modeled using RGB images gathered from the single-camera source. More significantly, it determines the sensitive details that prevailed in the original images and ensures privacy, widely considered for safety and protection. Various experiments are carried out using real-time world fall data sets. The results show that the system enhances fall recognition awareness and achieves a high F-Score by performing high accurate fall detection from real-world environments.*

Keywords: *Fall detection, Deep Convolutional Neural Network, Deep learning*

1. INTRODUCTION

Sudden falling on the floor is one of the risks associated with elders' safety, which may lead to physical injury and reduce the daily physical movements resulting from the falling fear [1]. Consecutively, fall detection systems provide automatic collection and monitoring of fall incidents, which paves the way for determining the causes of falls while enhancing quality of elderly people's lives who are alone at home. Wearable computer-based systems in this sense use sensors such as gyroscopes, switches and accelerometers. These devices record high accelerations (which occur during the fall), and provide warnings when the sensor data detects anomalies. While these devices are low-cost, they require regular recharging and thus present problems for elderly people or cognitively disabled persons.

Such devices inflicting non- sensory side that has no impact on people's well-being and have little effect on people's daily routines [2]. For a traditional vision-based fall detection method, attributes are extracted from the visual data and supplied for fall recognition by a machine learning classifier. For instance, methods such as [3]; [4] The human shape details was derived from camera images and various classification models were used to differentiate dropping from other activities. The performance of these automatic systems depends on factors including: i) the quality of scene detection in the human-region, ii) the type of information collected from the detected human-regions, and iii) the classifiers used to learn fall recognition features. In addition , the data used to train the classifiers plays a critical role in learning robust features that can generalize to conditions that are not apparent. Subsequently, the existing systems lack generalization capacity for vision-based fall detection in remote areas that remains unseen and extremely coveted in the healthcare industry for a commercial product.

In this paper, we discuss ways to address the aforementioned obstacles and strengthen the generalization of human fall identification to unknown real-world conditions while protecting people's privacy. To this end, deep learning-based system presented for detecting the fall automatically using human pose and color image segmentation information recorded by a video camera. For summary, this paper's principal contributions are as follows:

Express a human-position dependent depiction of fall which is invariant to alter the identity, surroundings, illuminations, and latitudinal positions of the intended person in human physical appearances. Experimental results demonstrate description of fall by feeding the input to a deep CNN to make it learn through a highly robust features that effectively generalize fall-recognition to unknown real-world environments.

Proposed a FallNet, an ensemble of multiple CNN constructs that learns about human pose and segmentation details based on fall representations. Because of multi-modal input data, FallNet benefits from both modality-specific and complementary information between the two modalities and increases the accuracy of fall predictions as opposed to independent classifiers.

2. RELATED WORKS

Deep learning-based automatic fall detection using RGB images proposed by Umar Asif, et.al(2019) presents the pictures captured from a single camera. In this study, the system learns from synthetic data in nature, often prevailed in the virtual world, body posture, and movement segmentation during falls. This framework analyzes and identifies the person's original details found in the picture with privacy-preserving measures, which is highly concerned with computer safety. Later, a five-point inverted pendulum model using enhanced 2-branch Multi-stage CNN(M-CNN) proposed by Jin Zhang et al. (2020) presents an enhanced posture representation model fall-behavior. It extracts and builds the inverted pendulum structure of body posture from real-world scenarios. Typically, the vision-based models for fall detection focusing on human region detection in the intended areas relatively typical scene using motion segmentation or subtracting the contents. Mainly it exploits the information gathered from the detection areas for training the fall recognition classifiers. Consecutively, the methods proposed by Miaou et.al (2018); Laura, et.al.(2018) enhances the exploration of fall detection models by dividing the methods into four axes. First, the emergency response builds on Computer Vision(CV). Second, algorithms modeled for the target; third, the hardware types and the algorithms used. Also, the survey offers a diversified view of CV with all types of emergencies.

T¨oreyin et al. (2005) Used background subtraction to detect human bounding boxes and compared boxes to different thresholds in consecutive frames of the Auvinet et al. (2010) MultiCam fall dataset to detect fall events. Suad Albawendi, et al. (2018) An enhanced vision-based fall detection schemes have been introduced for promoting independent living for elderly population. This method uses three specific features; motion details, variation in body posture shape, and histogram prediction to detect a fall. Mirmahboub et al. (2013) methods incorporated type and context knowledge from human bounding boxes, and used fall recognition support vector machine (SVM). Fouzi Harrou, et al. (2017) Focuses on detecting and classifying falls which solely relies on the changes in the form of a human figure, a crucial computer vision task. The detection is accomplished in this study with a multivariate exponentially weighted moving average (MEWMA) monitoring scheme remains successful in accurate fall detection, since it pose some minor changes. Huang et al. (2004)'s approach used extreme learning machines with shape features and got fast computing. Most of the above methods depend heavily on the assumption that the shift in visual information between subsequent image frames is important in order to achieve sufficient segmentation of the motion. This limits their use in cases where there is inadequate knowledge shift between subsequent frames.3D vision related approaches used information from several cameras or depth sensors (such as Microsoft Kinect) to solve this constraint, and studied 3D features for fall detection. For example, the Hung et al. (2013) method used visual data from multiple cameras and made decisions by voting from different points of view.

Kun Wang, et al. have developed an automated video surveillance-based human fall detection system that can enhance the health of elders in indoor environments. Gasparrini et al. (2014) methods; Mastorakis and Makris (2014) used Kinect depth maps to remove 3D silhouettes and 3D bounding box based features to detect falls. Although these multi-camera-based systems produce more reliable fall detection results compared to single-camera-based methods, hardware limitations affect the performance of these methods in large measure. Multiple camera-based techniques , for example, require precise synchronization between the individual cameras. On the other hand, approaches focused on the depth camera are influenced by inherent noise from the sensor, small fields of view and minimal constraints on depth sensing. In addition , due to safety-related issues, many public places such as elderly care centers and child care facilities are limiting the use of depth-based camera systems. Accurate identification of the fall from monocular images is therefore considered a highly important technology area in the healthcare industry. Angela Sucerquia. et al. (2017) models a dataset (ADL) consists of daily life fall events and related activities that are acquired by a self-developed model encompassing dual forms of accelerometer and gyroscope. More significantly, it consists of 19 ADL and 15 fall events made by 23 adults. Also, 15 ADL forms made by 14 stable and live participants who are exceeding the age of 62 and data from a single 60-year-old participant having both ADL and fall. Consecutively, Eduardo casilari. et.al. (2017) explores 12 data which are existing in the available data repository. It is used for evaluating the fall detection algorithms in wearable encompassing ADL measurements and emulated falls. The detailed analysis of captured datasets is placed in a well determined manner by taking multiple factors into account used in defining the test data which is meant for generating the mobility sample data.

3. PROPOSED METHODOLOGY

In this section we give a brief introduction to the proposed fall detection system and the description of the detection method.

3.1 Detect Human

In this paper we propose a deep learning-based architecture that uses RGB images to detect a fall in the scene. Compared with current approaches our work varies in many respects. Next, our system integrates numerous human-identification using Hare Cascade classifiers. It uses a refinement method to correct pose and segmentation errors and generates high-quality human proposals, particularly for scenes with multiple people or partially occlusive scenes, compared to methods Miaou et al. (2006); T¨oreyin et al. (2005) that use background-foreground subtraction techniques for the identification of human regions and yield low true positive results in these challenging situations. Second, we use visual representations based on human-skeletons and segmentation for profound learning of features. Our visual representations preserve human privacy and are invariant to variations in appearance and spatial translations of people in the scene. This allows our framework to generalize unseen real-world environments with success compared to the methods (e.g., Mirmahboub et al. (2013); Miaou et al. (2006) Which learn to recognize dropping with the aid of appearance data and suffer from poor generalization in the presence of significant changes in appearance properties.

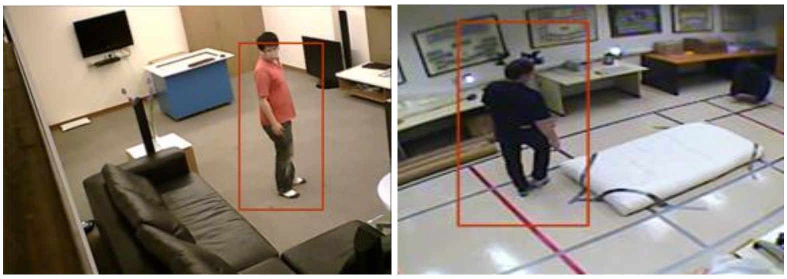


Figure 1: Human Detection

3.2 Deep Convolutional Neural Network (DCNN)

Figure 2 shows our Framework's overall architecture which has three main components like Convolution layer, sub-sampling layer and fully connected layer. FallNet, a CNN model that uses visual representations based on the RGB image and segmentation and learns high-level embedding features for fall recognition. We explain the individual components of the conceptual structure in detail below. Figure 3 shows the architecture of CNN.

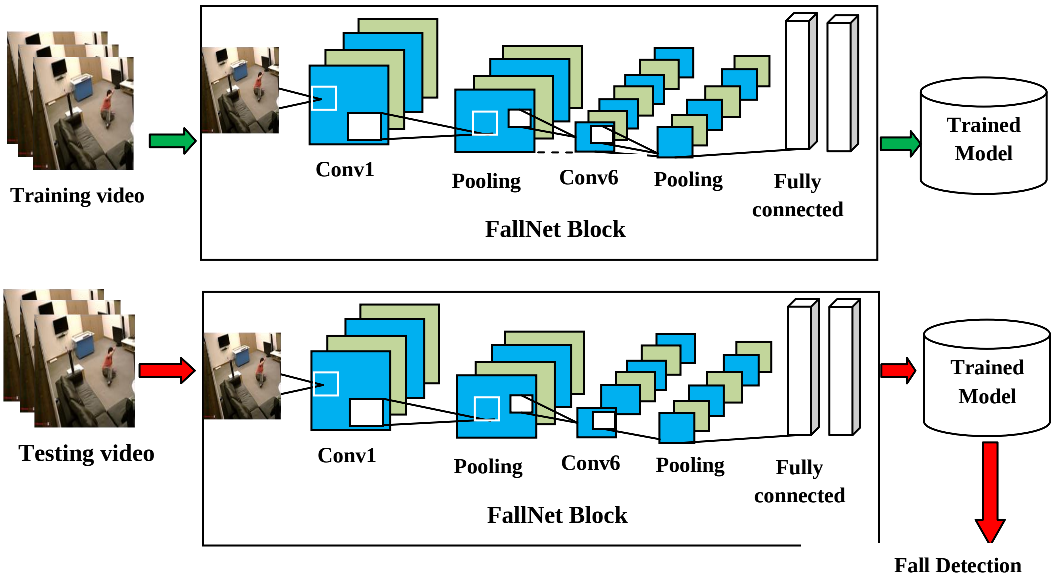


Figure 2: Overall Architecture for FallNet model

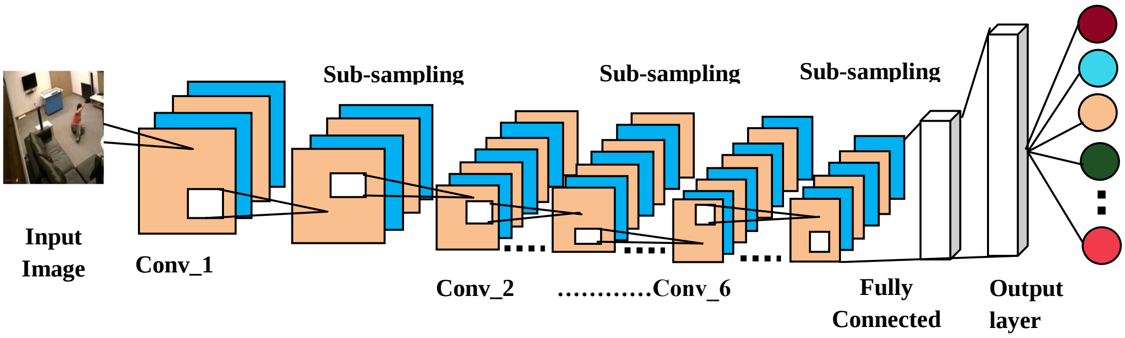


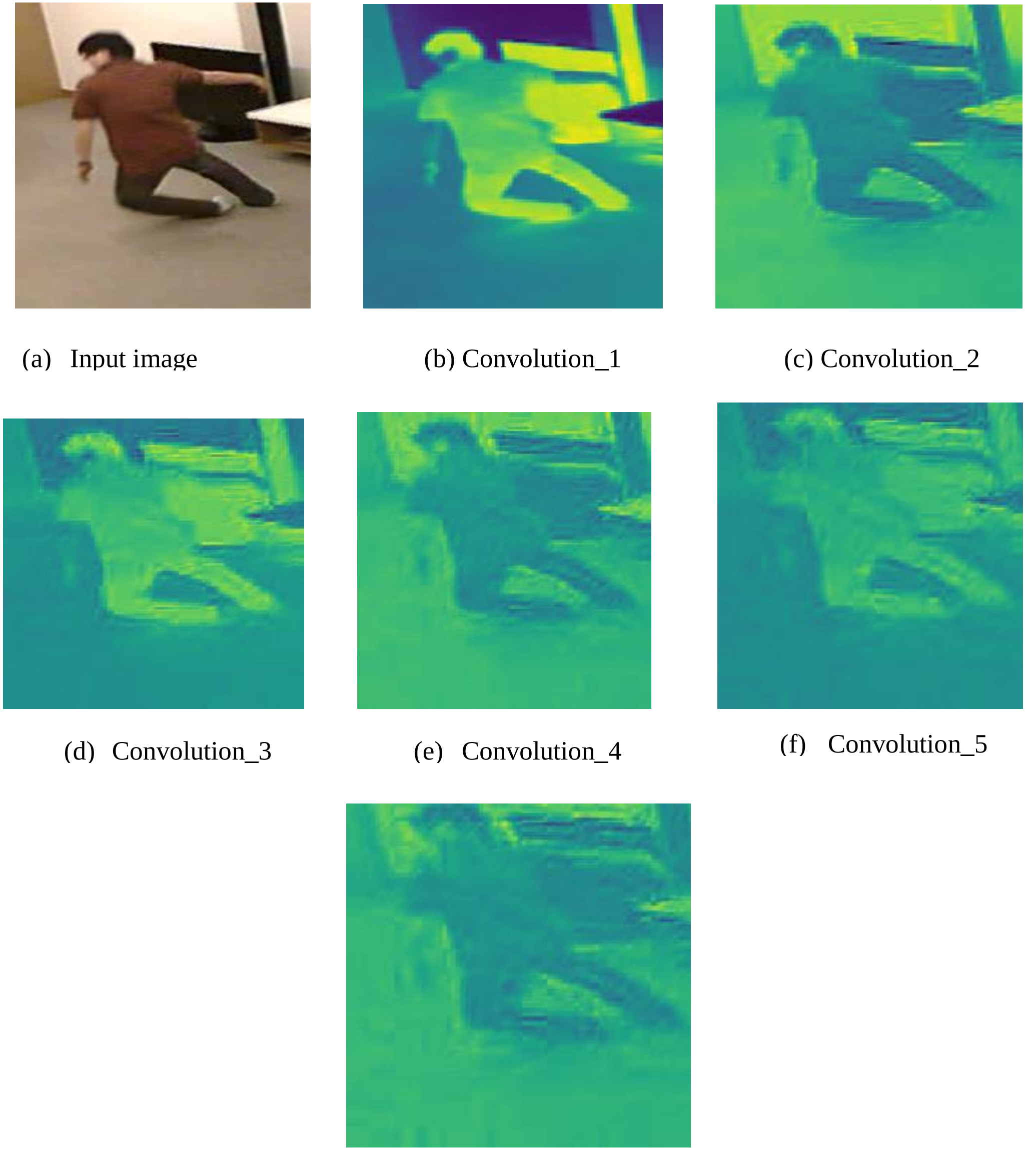
Figure 3: Architecture of CNN

3.3 Training and Implementation

At the outset, initialize the weights of the nodes in the FallNet network model. In this model, the weights are initialized for the convolutional layer using the weights in the next embedding layer that possessing zero-mean gaussian distribution (SD= 0.01 and bias=0). For instance, 20 epochs, the trained convolutional layer and embedding layers are from end to another end. Here, the learning rate at the initial state is set it as 0.01 and is divided by 10 at the 50th percentile and 75th percentile of totality of epochs. Then the decay parameter is set as 0.0003 of the total weighs and changes. The proposed system is built on the basis of Torch library Paszke et al.(2017) where training performed by ADAM optimizer. Figure 4 illustrates the convolution layers ranging from 1-6 convolutions respectively.

Table 1: Summary of the FallNet Model

Layer type	**Filter size & Stride**	**Details**	**Output Shape**
Conv1	3x3 & s =1	Conv1 (16)	255,225, 16
Activation	ReLU		255,225, 16
MaxPooling		Pooling Size (2,2)	127,127, 16
Conv2	3x3 & s =1	Conv2 (16)	127,127, 16
Activation	ReLU		127,127, 16
MaxPooling		Pooling Size (2,2)	63,63,16
Conv3	3x3 & s =1	Conv3 (32)	63,63,32
Activation	ReLU		63,63,32
MaxPooling		Pooling Size (2,2)	31,31, 32
Conv4	3x3 & s =1	Conv4 (32)	31,31, 32
Activation	ReLU		31,31, 32
MaxPooling		Pooling Size (2,2)	15,15,32
Conv5	3x3 & s =1	Conv5 (64)	15,15,64
Activation	ReLU		15,15,64
MaxPooling		Pooling Size (2,2)	7,7,64
Conv6	3x3 & s =1	Conv6 (64)	7,7,64
Activation	ReLU		7,7,64
MaxPooling		Pooling Size (2,2)	3,3,64
Flatten	Flatten to a vector		96,756
Dense	Dense Input =256		256
Dense	Input Classes = 6		6
Activation	Softmax		6



(a) Input image

(b) Convolution_1

(c) Convolution_2

(d) Convolution_3

(e) Convolution_4

(f) Convolution_5

(g) Convolution_6

Figure 4: One to six convolution visualization for input image

4. EXPERIMENTAL RESULTS

4.1 Datasets:

For analysis purpose, two datasets are taken from the survey, which found to be well suited for making benchmarking strategy. The data set consists of fall video can be accessed in the literature [27]. The link http:// foe.mmu.edu.my/digitalhome/FallVideo.zip provides the suggested fall video data samples. It comprises of nearly 30 activities like walking, sitting, squatting and nearly 21 fall images of forward, backward and sideway. We adopted the method shown in section 3.1 for the implementation of training and testing the fall detection. Consecutively two actions are performed. First, the human detection and extraction is performed. Second, the extracted images are reframed and scaled to 64 x 64-pixel resolution. http://www.iro.umontreal.ca/~labimage/ Dataset/ is the Multicam fall dataset used here. It normally contains 24 variety of performances in which 22 with one fall and 2 consists of confounding events. Here, each performance is acquired from 8 variety of perspectives where same stage is used and some instances uses the furniture reallocation.

Table 2: Number of frames of each dataset, distribution of frames per class (fall and “no fall”), and number of fall/“no fall” samples (sequences of frames corresponding to a fall or a “no fall” event).

Dataset	Total Frames	Fall Frames	No Fall Frames	Falls	No Falls
Fall Video Dataset	11312	3462	7850	577	1308
Multiple Cameras Fall Dataset	26137	7880	253257	184	376

4.2 Evaluation Metrics

The confusion matrix of the given problem is illustrated in figure 5.The matrix comprises of true positive, true negative, false positive and false negative values of fall detection scenario. The success state is considered as the state when the classifier correctly identifies the issue and vice versa denotes the failure state. The overall performance of the classifier is obtained by the error rate. It is usually the proportion of the errors over the instances’ set.

.

		Actual Value	
		Positive	Negative
Predicted Value	Positive	TP	FP
	Negative	FN	TN

Figure 5: Confusion Matrix for fall detection

Precision (P) or detection rate defines as the ratio of correctly labeled instances to the total labeled instances. Typically, P can measure the prediction’s model which denotes the true positive value which is illustrated below:

$$\text{Precision(P)} = \frac{TP}{TP + FP} \qquad (1)$$

Recall (R) or Sensitivity defines as the ratio of labelled instances and the total instances. R measure usually denotes the predictions’ model and defined as the true positive figure which is defined by:

$$\text{Recall (R)} = \frac{TP}{TP + FN} \qquad (2)$$

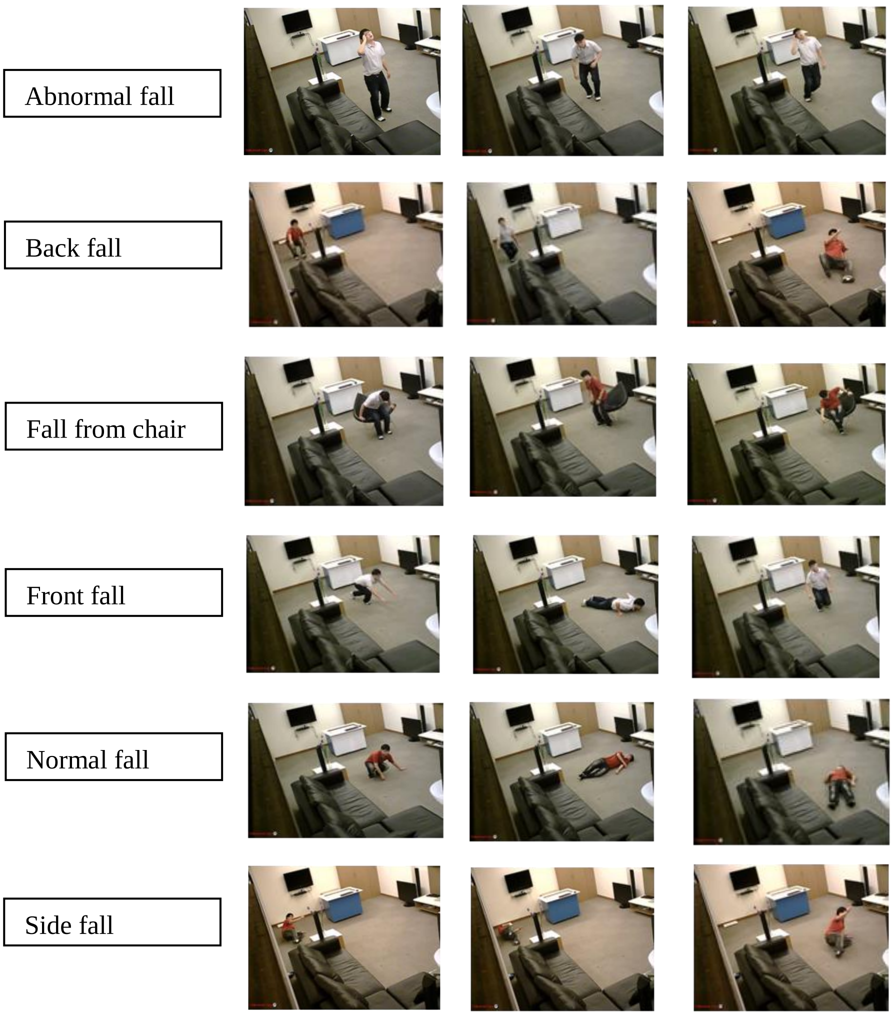
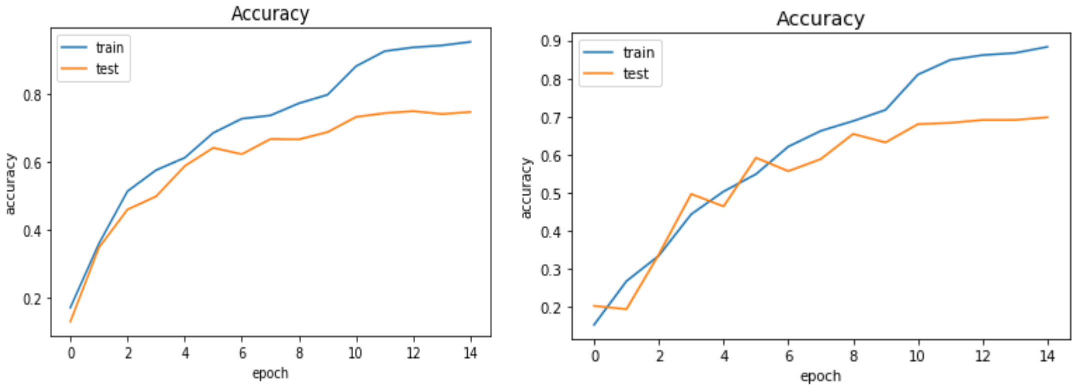


Figure 6: Sample frames for fall detection dataset

Recall normally exhibits the single measure of performance and F-score is stated as the harmonic mean of precision

$$F_{\beta} = \frac{(1+\beta)^2.TP}{(1+\beta)^2.TP+\beta^2.FN+FP} orF1 = 2.\frac{P.R}{P+R} \qquad (3)$$



(a) Accuracy for 'ReLU' **(b) Accuracy for 'ELU'**

Figure 7: Accuracy of proposed model for different activation function

Hence in this juncture, 70% of fall video datasets are used for training the detection model and consecutively, 30% is meant for testing. Parameters such as precision, recall and f-score is used for analyzing the efficiency. The F-score normally used for accurate classification of predicting the fall as an exact fall. Likewise, accuracy used for classifying the fall as a dropping [28].

Table 3: Confusion Matrix for fall detection in (%)

	Abnormal_fall	**Back_fall**	**Fall_from_chair**	**Front_fall**	**Normal_fall**	**Side_fall**
Abnormal_fall	91	0	4	5	0	0
Back_fall	2	93	1	0	0	4
Fall_from_chair	0	0	100	0	0	0
Front_fall	5	0	0	95	0	0
Normal_fall	0	0	0	0	100	0
Side_fall	0	4	0	6	0	90

Table 4: Performance Metrices of Fall Detection using FallNet

	Precision	**Recall**	**F1-score**
Abnormal_fall	0.80	0.91	0.87
Back_fall	0.81	0.93	0.88
Fall_from_chair	1.00	1.00	1.00
Front_fall	0.77	0.95	0.89
Normal_fall	1.00	1.00	1.00
Side_fall	0.79	0.90	0.89

Thus, from the table 4, it is represented that the average sensitivity of precision is recorded as 95.7 % and the F1-score is 97.6%. Experiment 1 shows the highest performance, since fall video data set consists of a single human entity. And the dataset encompasses an elderly person, hence it is speculated that the fall detection seems to be hard for this experiment.

5. CONCLUSION

This paper proposes deep learning architecture for automated detection of human fall from frames taken by a single camera. With body joint locations and information on segmentation, our framework produces human propositions. Such ideas are refined and converted into multimodal visual representations for input into FallNet, a model CNN that uses modality-specific and multimodal layers and utilizes highly discriminative embedding features for fall recognition. And also present a human fall dataset consisting of synthetically generated human pose and segmentation data under various camera viewpoints. Experiments on complex public fall data sets show that our qualified system uses only synthetically generated pose data to generalize effectively to unknown environments and achieve high accuracy and recall scores for fall recognition. Our picture, trained on pure synthetic data, is highly resilient due to variations in appearance characteristics, shifts in size and different camera viewpoints. This opens up new possibilities for advancing privacy to maintain highly beneficial human fall detection in health information technology. Expand our strategy to the structure for identification of other events in the future in order to improve its scope for identification of general human activity. We expect to reduce our fall detector's computational burden by parameter-pruning we memory-efficient CNN structures, too.

REFERENCES

[1] Umar Asif, Benjamin Mashford, Stefan von Cavallar, Shivanthan Yohanandan, Subhrajit Roy, Jianbin Tang, Stefan Harrer, “Privacy Preserving Human Fall Detection using Video Data” Machine Learning for Health (ML4H) at NeurIPS 2019, Proceedings of Machine Learning Research , pp. 1–12, 2019

[2] Jin Zhang, Cheng Wu, Yiming Wang, “Human Fall Detection Based on Body Posture Spatio-Temporal Evolution” Sensors, MDPI Publisher, 20, 946, 2020, pp. 1-21

[3] Laura Lopez-Fuentes, Joost van de Weijer, Manuel Gonz´alez-Hidalgo, Harald Skinnemoen, Andrew D. Bagdanov, “Review On Computer Vision Techniques In Emergency Situations” Multimedia Tools and Applications, 2018

[4] Cmu graphics lab: Carnegie-mellon motion capture (mocap) database. http://mocap.cs.cmu.edu,2003.

[5] Edouard Auvinet, Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. Multiple cameras fall dataset. DIRO-Universit´e de Montr´eal, Tech. Rep, 1350, 2010.

[6] Suad Albawendi, Ahmad Lot, Heather Powell, Kofi Appiah "Video Based Fall Detection using Features of Motion, Shape and Histogram" Proceeding of PETRA'18, June 26–29, 2018, Corfu, Greece, pp. 529-536

[7] Imen Charfi, Johel Miteran, Julien Dubois, Mohamed Atri, and Rached Tourki. Optimized spatiotemporal descriptors for real-time fall detection: comparison of support vector machine and adaboost-based classification. Journal of Electronic Imaging, 22(4):041106, 2013.

[8] Fouzi Harrou, Nabil Zerrouki, Ying Sun, and Amrane Houacine, "Vision-Based Fall Detection System for Improving Safety of Elderly People" IEEE Instrumentation & Measurement Magazine, December 2017, pp. 49-55

[9] Jane Fleming and Carol Brayne. Inability to get up after falling, subsequent time on floor, and summoning help: prospective cohort study in people over 90. Bmj, 337:a2227, 2008.

[10] Kun Wang, Guitao Cao, Dan Meng, Weiting Chen, Wenming Cao, Automatic fall detection of human in video using combination of features.

[11] Samuele Gasparrini, Enea Cippitelli, Susanna Spinsante, and Ennio Gambi. A depth-based fall detection system using a kinect sensor. Sensors, 14(2):2756–2775, 2014.

[12] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Doll´ar, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016.

[14] Kaiming He, Georgia Gkioxari, Piotr Doll´ar, and Ross Girshick. Mask r-cnn. In ICCV, pages 2961–2969, 2017.

[15] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: a new learning scheme of feedforward neural networks. In Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on, volume 2, pages 985–990. IEEE, 2004.

[16] Angela Sucerquia, José David López, Jesús Francisco Vargas-Bonilla, "SisFall: A Fall and Movement Dataset" 17, 198, Sensors 2017, pp. 1-15

[17] Dao Huu Hung, Hideo Saito, and Gee-Sern Hsu. Detecting fall incidents of the elderly based on human-ground contact areas. In 2013 2nd IAPR Asian Conference on Pattern Recognition, pages 516–521. IEEE, 2013.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll´ar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, pages 740–755. Springer, 2014.

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In ECCV, pages 21–37. Springer, 2016.

[20] Georgios Mastorakis and Dimitrios Makris. Fall detection system using kinect's infrared sensor. Journal of Real-Time Image Processing, 9(4):635–646, 2014.

[21] S-G Miaou, Pei-Hsu Sung, and Chia-Yuan Huang. A customized human fall detection system using omni-camera images and personal information. In Distributed Diagnosis and Home Healthcare, 2006. 1st Trans disciplinary Conference on, pages 39–42. IEEE, 2006.

[22] Behzad Mirmahboub, Shadrokh Samavi, Nader Karimi, and Shahram Shirani. Automatic monocular system for human fall detection based on variations in silhouette area. IEEE Transactions on Biomedical Engineering, 60(2):427–436, 2013.

[23] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In ECCV, pages 483–499. Springer, 2016.

[24] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[25] B U˘gur T¨oreyin, Yi˘githan Dedeo˘glu, and A Enis C¸ etin. Hmm based falling person detection using both audio and video. In International Workshop on Human-Computer Interaction, pages 211–220. Springer, 2005.

[26] Shengke Wang, Long Chen, Zixi Zhou, Xin Sun, and Junyu Dong. Human fall detection in surveillance video based on pcanet. Multimedia tools and applications, 75(19):11603–11613, 2016.

[27] Guoru Zhao, Zhanyong Mei, Ding Liang, Kamen Ivanov, Yanwei Guo, Yongfeng Wang, and Lei Wang. Exploration and implementation of a pre-impact fall recognition method based on an inertial body sensor network. Sensors, 12(11):15338–15355, 2012.

[28] Eduardo Casilari, José-Antonio Santoyo-Ramón and José-Manuel Cano-García, "Analysis of Public Datasets for Wearable Fall Detection Systems" sensors, MDPI, 17, 1513, 2017, pp. 1-28