Human activity recognition using deep learning

Prof. Hansa Shingrakhia

Electronics & Communication Engineering
Indus University, Ahmedabad, Gujarat, India

Abstract: Activity Recognition is a technology which is based on mathematical interpretation of human body by a computing device. It is the ability to recognize human gestures and movements intelligently and effortlessly with a human-friendly environment. This paper will help in the field of safety of elderly people and people with disabilities, which need special assistance. It will detect the unusual gestures or movements and notify the user about it. Using the machine learning technology, it will determine the gesture which differs from the usual gestures or movements. It will be able to detect movements like person falling off the bed or a chair or a wheelchair, people having heart attack, not being able to stand up etc. Activity Recognition technology enables humans to communicate with the machine and interact naturally without any mechanical devices. The paper will help to increase safety and security in the house.

Index Terms – Activity recognition, Gesture, Machine learning.

I. Introduction

In today's fast-moving world, people are busy with their commitments. It is difficult to take care of both their elders and children when people are away working in their offices. One can solve this problem with the help of Artificial Intelligence. This one will work as a human eye and will keep the user updated with any abnormal activity at home when he/she is away. This project is limited to just two activities. One can increase the scope of the project by adding more classes. In that way, one can be detect any abnormal activity inside home. In this paper main focus is on detection of falling and fighting inside a home. Whenever these occur in specified area, the machine will identify user about that so that the user can take action accordingly. Main aim is reduce the problem of aged people by using this technology which will prevent them from falling down by leveraging java technology along with Machine learning and Image Processing.

II. Scope

The Main purpose is to reduce the errors which are caused by human negligence that results in grave consequences. This work does not only focus on home surveillance but it is also capable of providing service to Hospitals, Schools, Restaurants, Factories and Large-scale Industries to ensure the security as well as the safety of the workers.

III. Block diagram

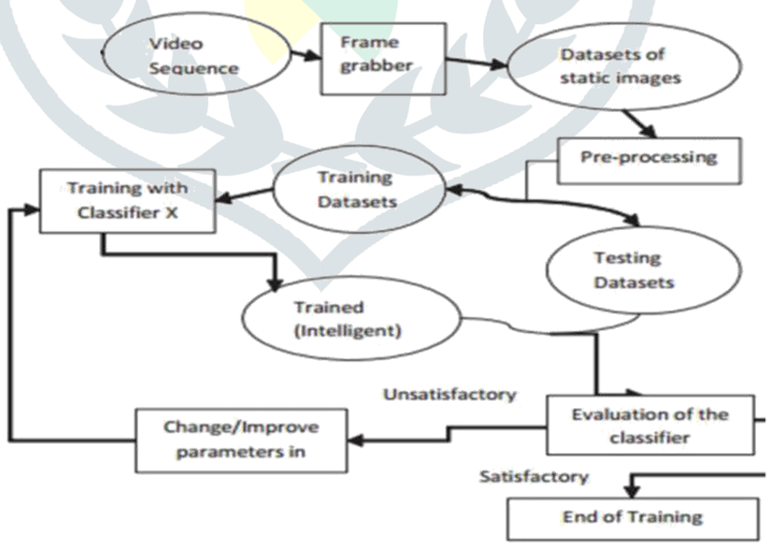


FIGURE 1. TRAINING AND TESTING STAGE

In the training and evaluations stage as shown in figure 1. All the parameters of the system must be incrementally improved to optimal values so that the model would be ready for deployment. The video camera is the data acquisition device which in this case is a digital camera running in video recording mode. The video sequences recorded from the video camera are converted into datasets of static color images (one image corresponds to one frame of a video sequence). The images are then run through the pre-processing step which is a combination of many algorithms and is described in detail in the next section. The outputs of the pre-processing are then randomly categorized into two parts: training and testing data sets. The pre-processed images are trained and a number of classifiers were evaluated. All classifiers have to be trained and tested using the images dataset. Furthermore, the optimal parameters were obtained by re-trained and re-tested each classifier several times. In the literature, there is so far no known or specific method to estimate the optimal parameters that may give satisfied results.

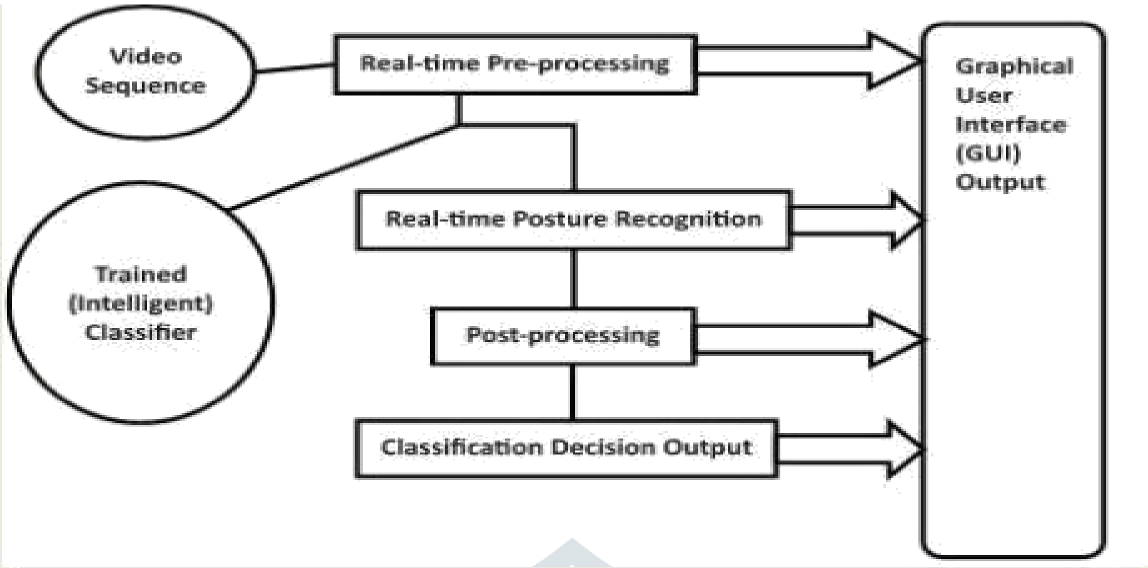


FIGURE 2 DEPLOYMENT STAGE

After the system has been trained and evaluated many times and optimal parameters for a model have been obtained, the trained model is then ready to be deployed. Whilst the inputs to the system are static images (samples) in the dataset during the training and evaluation stage, the inputs to the system are video sequences in the deployment stage. The deployment stage is depicted in Figure 2. In the deployment stage of the proposed system, each frame of a recorded video sequence which is running at 30 frames per second (fps) is grabbed at a lower speed such as 6 fps. The reason why a lower sampling rate can be used is because, to continuously recognize human postures in a video sequence, most of the frames do not need to be processed since they are redundant or each frame contain very similar posture to the next. Thus, for any input video sequence which contains a human moving or changing from pose to pose at a reasonable speed, it is sufficient to process only a few frames in a second. For each frame that has been grabbed, the posture recognition (which includes pre-processing, feature extraction and classification) is performed. And then the post processing step is done.

IV. ALGORITHMS

A Convolutional Neural Network (CNN) are a special kind of multi-layer neural networks, designed to recognize visual patterns directly from pixel images with minimal preprocessing [1]. It's a deep, feed-forward artificial neural network. Feed-forward neural networks are also called multi-layer perceptron(MLPs), which are the quintessential deep learning models. The models are called "feed-forward" because information flows right through the model. There are no feedback connections in which outputs of the model are fed back into itself.

These neural networks have proven to be successful in many different real-life case studies and applications, like:

- Image classification, object detection, segmentation, face recognition
- Self-driving cars that leverage CNN based vision systems;
- Classification of crystal structure using a convolutional neural network;

1**] LeNet**

LeNet-5, a pioneering 7-level convolutional network by LeCun et al in 1998, that classifies digits, was applied by several banks to recognize hand-written numbers on checks (cheques) digitized in 32x32 pixel images. The ability to process higher resolution images requires larger and more convolutional layers, so this technique is constrained by the availability of computing resources.

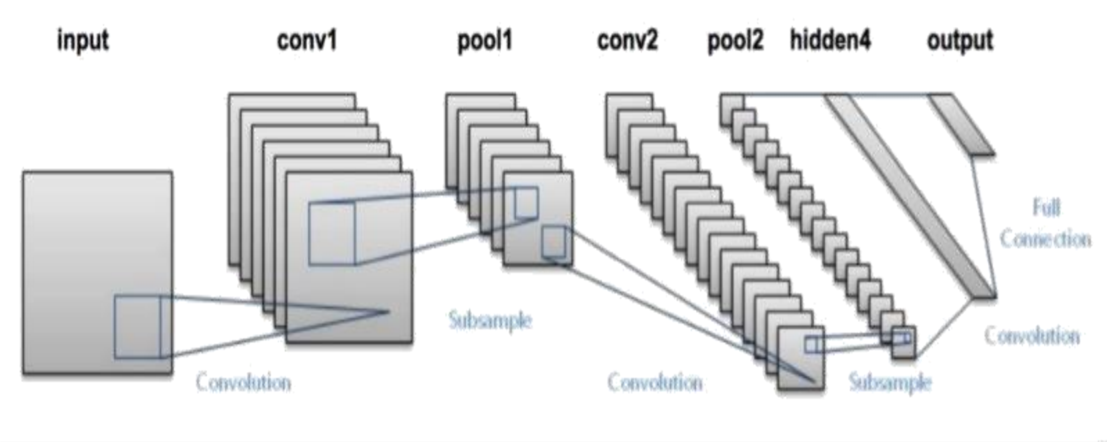


FIGURE 3 NEURAL NETWORKS

2] Inception v3

Inception is Google's pre-trained classifier used to distinguish between thousands of classes. In the actual case, a solution such as a powerful image classifier can help the company track shelf inventory, categorize products, record product volume, etc. from raw products images captured in real time by dedicated devices (drones, robots).

FIGURE 4 CONSUMER GOODS FOR CLASSIFICATION

First of all, yes, it is obvious that every **adult** should be able to correctly identify a product from its image. The first observation is that this assumption holds true as far as we consider only **adults**. A 5-years-old child, for example, can't beat an experienced person in this task. And an "experienced person" in this case, when we zoom in, is simply someone who has seen enough product images in his life that he can easily recognize any product he has seen before, from a given image.

This concept of experience is what we try to transfer to robots when we train them with existing labelled data so that they can learn on their own how to accurately distinguish each image from a training dataset. In this sense, we use **Artificial Neural Networks** which are nothing but an imitation of how human brains actually work. The knowledge that robots build using these algorithms are later tested on unlabeled observations. They label given images based on what they have learned and that is why this kind of problem is usually referred to as **supervised learning:** robots try to map a new image instance to one of the class labels they encountered during the training phase [7]. And talking about performance, it has been noticed that well trained robots tend to give better accuracy than humans in most cases of supervised learning.

After knowing about how to train and test a model, we started with the first activity detection. First, we captured six video sequences of different persons falling on the floor. We converted those videos into frames and we had about 2500 images. We trained the network with neural network and tested five random images of person fallen on the floor. One of the tested image had 98.13 % of accuracy of falling.

FIGURE 5 PERSON FALLING ON THE FLOOR

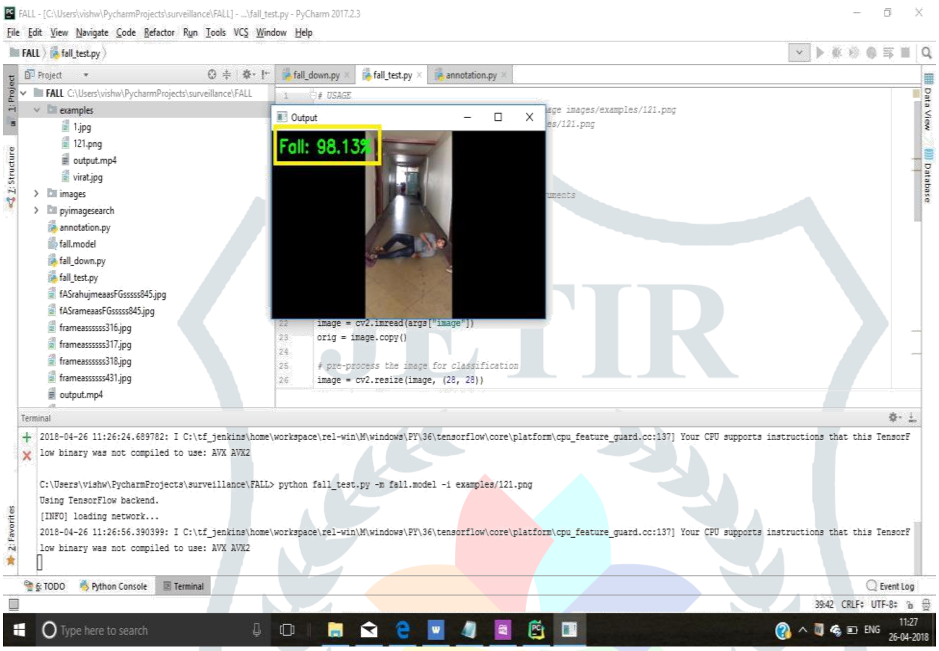


FIGURE 6 FALLING DETECTED

DRAWBACKS OF THIS METHOD

One of the problems with LeNet v5 is that it is not possible to define region of interest for the model to detect the fall. For example, if someone is sleeping on the bed or on the floor, the trained model will detect it as a fall because it would look like someone fallen. The solution to this problem can be defining the region of interest in the room and only detect the fall in that specified area. This would ensure that the model would not detect someone sleeping on the bed as fallen. Another problem with LeNet is that it does not provide localization. i.e. It does not tell where in the room the person has fallen. [7] Both are major drawbacks and it can raise concerns over the credibility of the project. Also, it would be difficult to implement it in real world because of these drawbacks.

SOLUTION TO THE PROBLEM

INCEPTION V3

It is a pre-trained image classifier and widely used to distinguish among various objects. In our ProJet, it would provide us with the solutions to both our problems. i.e. Defining region of interest and localization

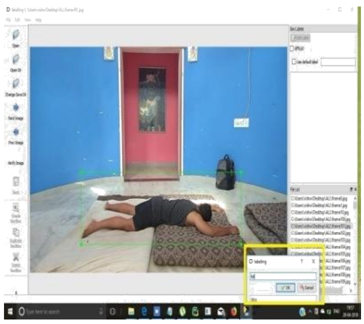


FIGURE 7 ANNOTATIONS

One can again train the network with the images and created a model. Here tested 5 videos with that model and detected fall. And also tested a random video of someone walking on the road to check whether it detects fall or not. From all this experiment successfully completed detection in all the videos and completed.

Here tested multiple fall videos of different persons and obtained the desired results with all of them. Here are some of the screenshots of results.

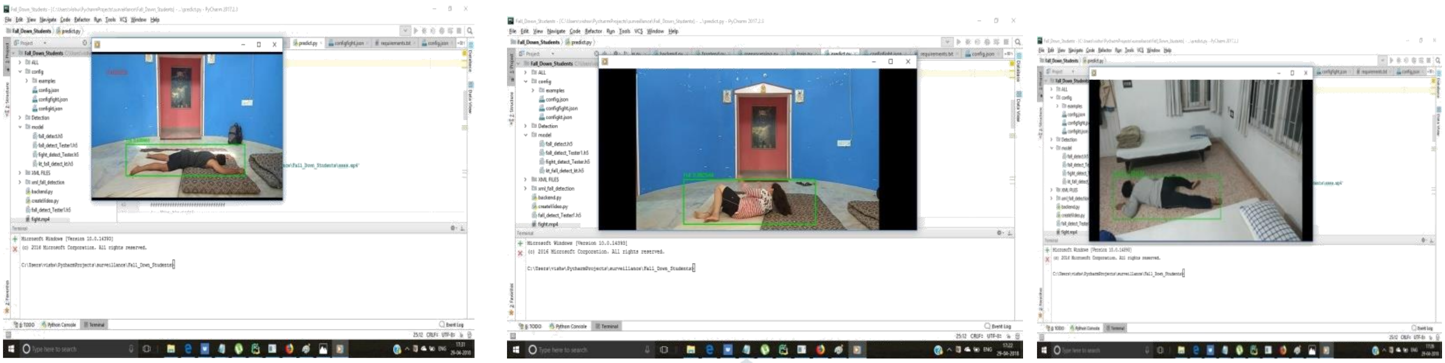
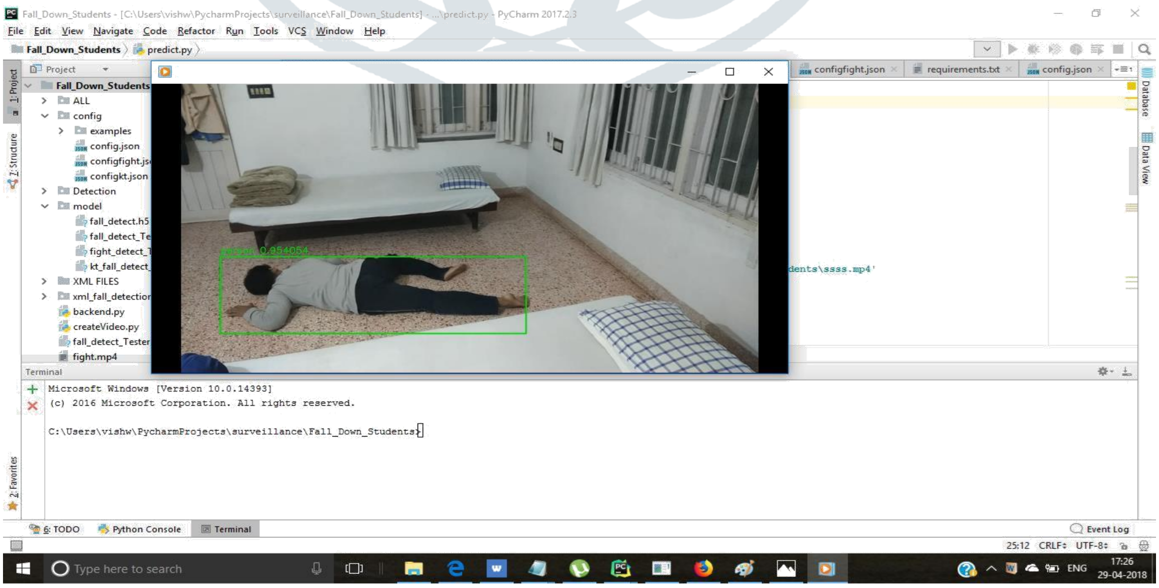
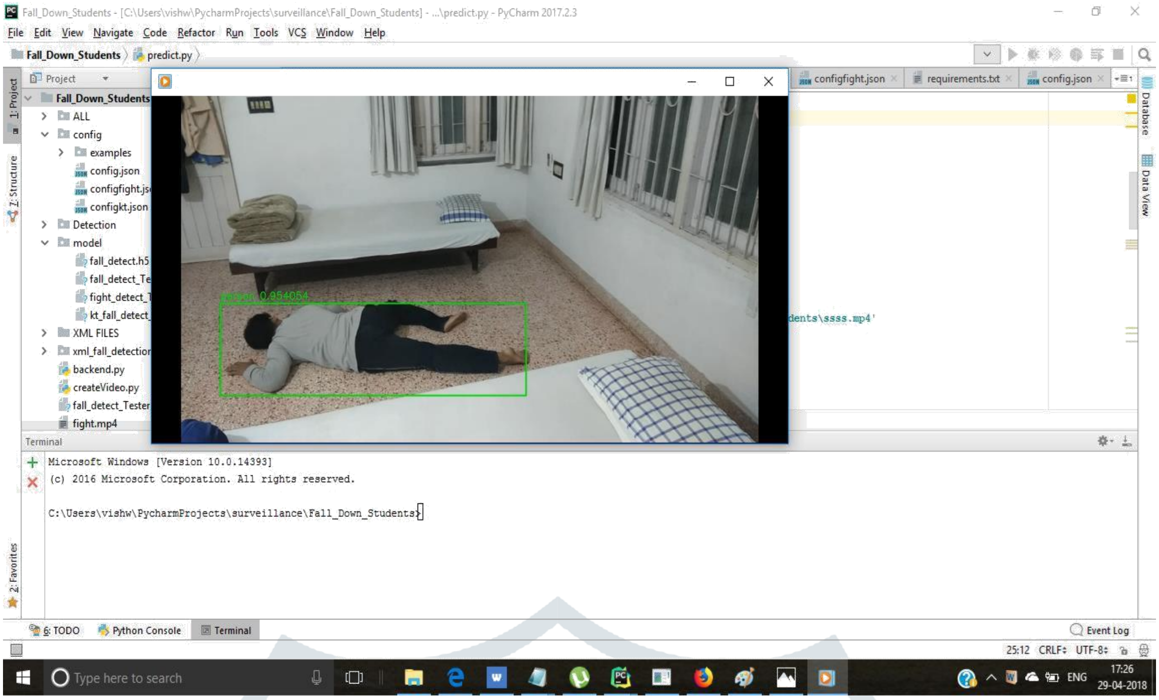


Figure 8

To know whether this model was accurate, model is tested it with some random videos. One such video was of a person walking on the road and successfully tested the video without detecting anything.

Figure 9





CONCLUSION

After successfully completed falling detection. The models are trained with a thousand to fifteen hundred images. In real life scenario, to get a sufficient level of efficiency, require a very large dataset. This system can be more efficient by adding more images. This model can be very helpful for home surveillance system. One can increase the classes by adding more abnormal activities and secure the premises with the help of Artificial intelligence. The scope of this project in not only limited to home surveillance as it can be used in hospitals, too. Due to fast image processing capabilities of this system, companies can move to automated employee recognition profile systems. Accurate helmet detection models can save a lot of government money and increase the revenue. It will also help increasing safety and security of citizens.

FUTURE SCOPE

The existing system for monitoring videos takes a lot of time to detect an activity. Real time monitoring system can be adopted instead of the existing system. It will make the system more supportive and reliable. A larger dataset can make the system more accurate. Also, if we can add images of all the detected videos in the training dataset daily, the system can become more accurate with every passing day. Number of classes should be increased to make the system more efficient for both home and road surveillance.

REFERENCES

[1] Ling Bao, and Stephen S. Inutile, " Activity recognition from user annotated acceleration data & quot; Pervasive computing, pp. 1-17, 2004.

[2] Oscar D. Lara, and Miguel A. Labrador, " A survey on human activity recognition using wearable sensors, & quot; Communications Surveys & Tutorials, vol.15, no.3, pp. 1192-1209, 2013.

[3] Akram Bayat, Marc Pomplun, and Duc A. Tran, " A study on human activity recognition using accelerometer data from smartphones, " Procedia Computer Science, vol.34, pp. 450-457, 2014.

[4] Tam Huynh, and Brent Schiele, " Analyzing features for activity recognition, " Proceedings of the joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies, pp. 159-164, 2005

[5] Z Yong-Quang, " Human physical activity recognition algorithm based on Hu moment invariants " Computer Science, vol. 41, no. 3, pp. 306-309, 2014.

[6] X. Xiao, D. Xu and W. Wan, " Overview: Video recognition from handcrafted method to deep learning method & quot; 2016 International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, 2016, pp. 646-651.

[7] M. Asadi-Aghbolaghi et al., " A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences & quot; 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, 2017, pp. 476-483.

[8] A. F. Bobick and J. W. Davis, " The recognition of human movement using temporal templates & quot; in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 3, pp. 257-267, Mar 2001.

[9] D. Ciregan, U. Meier and J. Schmidhuber, " Multi-creolumn deep neural networks for image Classification " 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, 2012, pp. 3642-3649.

[10] S. Ji, W. Xu, M. Yang and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition & quot; in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221-231, Jan. 2013.

[11] Joe Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga and G. Toderici, " Beyond short snippets: Deep networks for video classification & quot; 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 4694-4702.