International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 4 Volume No.6 July – 2022
DOI:10.46647/ijetms.2022.v06i04.007 ISSN: 2581-4621

Video Based Human Activity Recognition Surveillance System

Juhi Singh, Shweta Sinha
Department of Computer Science and Engineering
Amity school of Engineering and Technology

ABSTRACT

It is argued that current state-of-the-art methods of home surveillance such as motion detection technology like CCTV intrusion alert are insufficient, in particular to cater the modern need of whole automation with vulnerabilities such as requiring human involvement. We propose an alternative system, a video-based Human Activity Recognition (HAR) approach using the combination of Convolution Neural Network (CNN) and Long Short-Term Memory (LSTM) algorithm, to deal with the identified shortcomings. Our proposal does not require any changes to the existing home security protocols and is easily implemented using only low-cost, commercial-off-the-shelf hardware. We can easily use the traditional surveillance camera for computer vision tasks. We evaluate our approach using real-world activity data collected via video-based sensor. We quantify its effectiveness, by plotting Loss and accuracy curves. Our results show that the video-based HAR approach can provide full automation in home surveillance system compared to conventional CCTV motion detectors by an accuracy of more than 93%. Further system's accuracy can be increased and we can achieve significantly better results by implementing LRCN (Long Term Recurrent Convolution Network) approach.

Keywords: *Human Action Recognition, Home Surveillance, Deep Learning, Computer Vision, Tenserflow.*

I. INTRODUCTION

A home security system's major goals are to protect your home and your family. A home security system provides peace of mind as well as protection for your loved ones and valuables. Protecting what we care about is a natural instinct. Home security systems can now serve as a hub for home automation systems, which add convenience and energy savings. Protection, deterrent to burglars, peace of mind, convenience and energy savings, and lower home insurance premiums are the top five benefits of a home security system. Being able to achieve all this with the help of machine that can both Monitor and execute autonomously without any human involvement will provide an end to end fully automated system.

The flow of project proceeds as described below:

- To analyze the disadvantages of CCTV intrusion alert motion detection technology, the current state of the art solution for Home surveillance.
- Concretely, examine the accuracy in Deep Learning Models, different intrusion

Website: ijetms.in Issue: 4 Volume No.6 July – 2022
DOI:10.46647/ijetms.2022.v06i04.007 ISSN: 2581-4621

activities typically seen around perimeter area, and the efficiency of the video usage.

- To effectiveness of ConvLSTM approach and LRCN approach.
- To evaluate our approach on real-world data and show that it performs better in contrast to already existing system.

The rest of the paper consists of following: Section 2 puts forward the literature review for the articles used and the methods used in them in detail. Section 3 discusses the problem associated and section 4 deals with the methodology consisting of theory for the applications and components implemented. Section 5 discusses the results obtained. Section 6 is a small review of conclusion and future scope for the project and the last section involves the references.

II. RELATED WORK

Zhang et al. [1] review and highlight the advances of state-of-the- art activity recognition approaches, especially for the activity representation and classification methods. For the representation methods, we sort out a chronological research trajectory from global representations to local representations, and recent depth-based representations. For the classification methods, we conform to the categorization of template-based methods, discriminative models, and generative models and review several prevalent methods. Next, representative and available datasets are introduced.

Abbas et al. [2] proposed an advance surveillance system gives us the option to trigger alarm or flashlights in order to scare the intruder. The problem with this system is that it requires the owner to manually trigger the alarm. This system will fail if the owner cannot respond to the notification received on the phone. This system will also fail if the phone has died or is away from the owner. I believe it is a serious problem to solve as it can affect the life of the owner.

The purpose of this paper is to review the recent developments in deep learning and video scene analysis problems is presented. In addition, this paper also briefly describes the most recent used datasets along with their limitations. Moreover, this review provides a detailed overview of the particular challenges existed in real-time video scene analysis that has been contributed towards activity recognition, scene interpretation, and video description and captioning.

Problem statement: To be able to develop an appropriate activity recognition method, in particular for home security, it is crucial to first identify the challenges of the image processing. The following characteristics distinguish the video-based HAR problem from other motion detection problems such as cctv intrusion alert system

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 4 Volume No.6 July – 2022
DOI:10.46647/ijetms.2022.v06i04.007 ISSN: 2581-4621

III. PROPOSED METHODOLOGY

- As in previous system there was no automation and the person had to respond to the indications received by mere sensors. In this system alarm is triggered automatically on the detection of the intruder
- **Changes in scale, viewpoint andlighting**:
  It is important to maintain uniformity in size of images for the system to give accurate results. Other factors like viewpoint of the camera and lighting conditions may also affect the result.

Basic Architecture

The project is developed primarily using Anaconda Jupiter Notebook, Python 3.7 runtime, Tensor Flow keras and OpenCV tools. The system architecture is based on a CNN and LSTM based Deep learning model which acts as thecore.

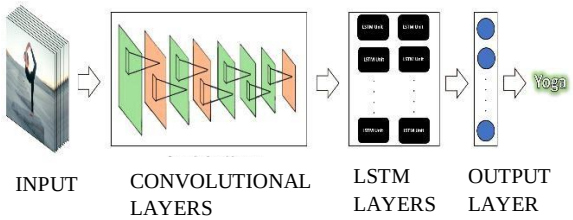


Fig 4.1: LRCN Architecture System

Pre-Processing the DataAccumulated

- Desired Data-Set has been collected through various means. Data is in the form of video that mimics particular action of an intruder. For example, a person trying to cross the fence of the home or jump the wall and try to enter the perimeter of the house in an unconventional way.
- Since we are going to use classification architecture to train on a video classification dataset, we had applied few techniques to pre-process the dataset first.
- To resized all frames of the video to width of the image being 64 units and height being 64 units. This toavoid unnecessary computation.
- We took a video file as input. Then read the video file frame by frame, resized each frame, normalized the resized frame. by dividing it with 255 so that each pixelvalue then lies between 0 and 1

Using One Hot Encoding

- It converted class labels that contain different types of actions to one hot encoded vectors using Kera's Method.

Website: ijetms.in Issue: 4 Volume No.6 July – 2022
DOI:10.46647/ijetms.2022.v06i04.007 ISSN: 2581-4621

- Machine learning algorithms cannot operate on label data directly as the data is categorical.
- In general, this is mostly a constraint of the efficient implementation of machine learning algorithms rather than hard limitations on the algorithms themselves.
- This means that categorical data must be converted to a numerical form. In order to do so we have used One Hot Encoding.

Constructing the Model

In this model Convolution Neural Network (CNN) with Long Short Term Memory (LSTM) network to perform Action Recognition while utilizing the Spatial-temporal aspect of the videos.

- A Convolutional Neural Network (CNN or ConvNet) is a type of deep neural network that is specifically designed to work with image data and excels when it comes to analyzing the images and making predictions on them.
- It works with kernels (called filters) that go over the image and generates feature maps (that represent whether a certain feature is present at a location in the image or not) and initially it generates few feature maps and as we go deeper in the network the number of feature maps is increased and the size of maps is decreased using pooling operations without losing critical information.

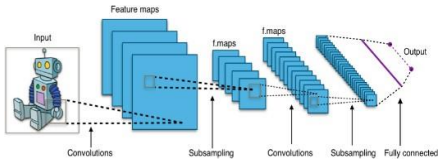


Fig 4.4 CNN Architecture

- An LSTM network is specifically designed to work with a data sequence as it takes into consideration all of the previous inputs while generating an output. So an LSTM cell can remember context for long input sequences.

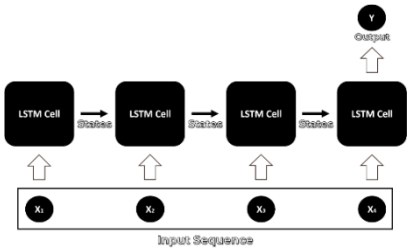


Fig 4.4 LSTM Architecture

- We have used a CNN to extract spatial features at a given time step in the input sequence (video) and then an LSTM to identify temporal relations between frames.

Website: ijetms.in Issue: 4 Volume No.6 July – 2022
DOI:10.46647/ijetms.2022.v06i04.007 ISSN: 2581-4621

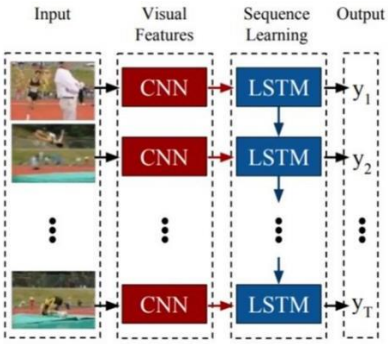


Fig 4.4 ConvLSTM Architecture

IV. IMPLAEMENTATION AND RESULT

The two architectures that we will be using to useCNN along with LSTM are:

- **ConvLSTM**
- **LRCN**

Both of these approaches can be used usingTensorFlow

a) **ConvLSTM Approach:** This approach combines Convolution and LSTM layers in a single model. Another similar approach can be to use a CNN model and LSTM model trained separately. The CNN model can be used toextract spatial features from the frames in the video, and for this purpose, a pre- trained model can be used, that can be fine-tuned for the problem. And the LSTM model can then use the features extracted by CNN, to predict the action being performed in the video.

- A ConvLSTM cell is a variant of an LSTM network that contains convolutions operations in the network. it is an LSTM with convolution embedded in the architecture, which makes it capable of identifying spatial features of the data while keeping into account the temporal relation.
- For video classification, this approach effectively captures the spatial relation in the individual frames and the temporal relation across the different frames. As a result of this convolution structure, the ConvLSTM is capable of taking in 3- dimensional input (width, height, num_of_channels) whereas a simple LSTM only takes in 1-dimensional input hence an LSTM is incompatible for modeling Spatio-temporal data on its own.
- To construct the model, Keras is used with ConvLSTM2D recurrent layers. The ConvLSTM2D layer also takes in the number of filters and kernel size required for applying the convolution operations. The output of the layers is flattened in the end and is fed to the dense layer with soft-max activation which outputs the probability of each action category.
- The MaxPooling3D layers to reduce the dimensions of the frames and avoid unnecessary

computations and dropout layers to prevent over fitting the model on the data.

b) LRCN Approach

- But here, we have implemented another approach known as the Long-term Recurrent Convolution Network (LRCN), which combines CNN and LSTM layers in a single model. The Convolution layers are used for spatial feature extraction from the frames, and the extracted spatial features are fed to LSTM layer(s) at each time-steps for temporal sequence modeling. This way the network learns spatiotemporal features directly in an end-to-end training, resulting in a robustmodel.
- To implement our LRCN architecture, wewill use time-distributed Conv2D layers which will be followedby MaxPooling2D and Dropout layers.
- The feature extracted from the Conv2D layers will be then flattened using the Flatten layer and will be fed toa LSTM layer. The Dense layer with soft max activation will then use the outputfrom the LSTM layer to predict the actionbeing performed.

V. EXPERIMENTAL RESULTS

a) Loss and Accuracy curves of ConvLSTM Model

By implementing the ConvLstm Model and plotting its Loss and Accuracy Curves we found that the difference between Loss and Validation Loss is of 0.3. Here Loss value implies how poorlyor well our model behaves after each iteration of optimization.

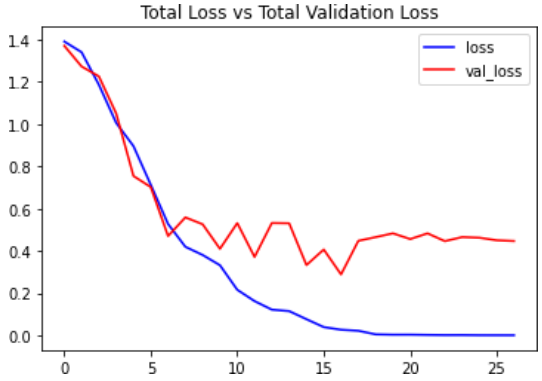


Fig 5.1.1 Loss Curves for ConvLSTM

Here in the below the Accuracy vs. Total validation Accuracy graph shows that there is very less difference between the two, of about 0.1. The accuracy is used to measure the performance of the model in an interpretable way. Here, the accuracy of ConvLSTM is high.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 4 Volume No.6 July – 2022
DOI:10.46647/ijetms.2022.v06i04.007 ISSN: 2581-4621

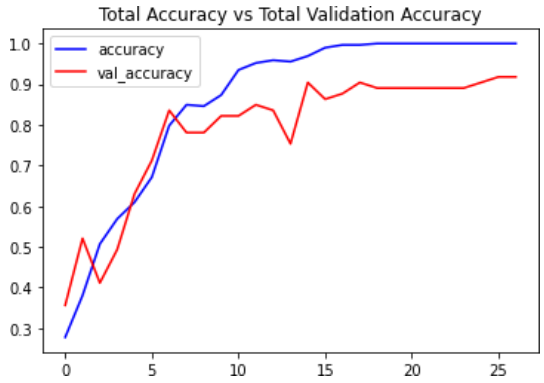


Fig 5.1.2 Accuracy Curves for ConvLSTM

Shown significantly less loss. Therefore the LRCN model is more robust.

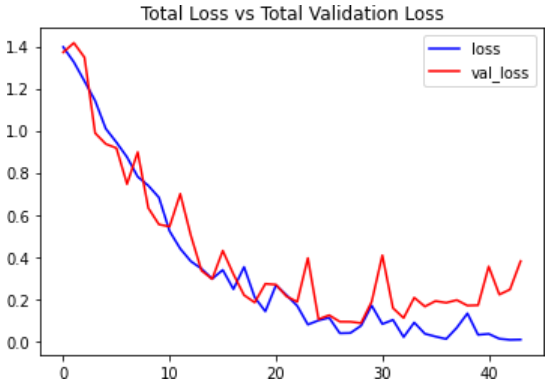
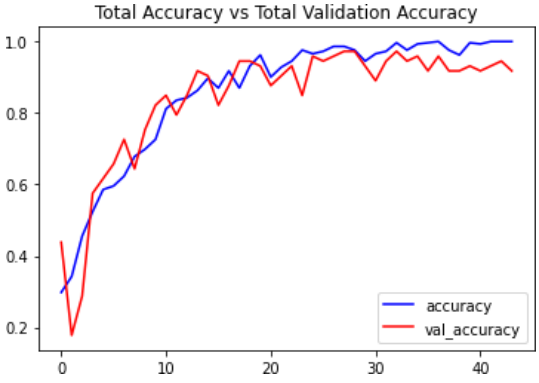


Fig 5.1.3 Loss Curves for LRCN

Here in the below Accuracy vs Total validation Accuracy graph shows that there is very less difference between the two, of about less than 0.1. Therefore accuracy of LRCN is higher than that of ConvLSTM model.



Website: ijetms.in Issue: 4 Volume No.6 July – 2022
DOI:10.46647/ijetms.2022.v06i04.007 ISSN: 2581-4621

Fig 5.1.1 Accuracy Curves for LRCN

b) Loss and Accuracy Curves of LRCN Model

By implementing the ConvLstm Model and plotting its Loss and Accuracy Curves we found that the difference between Loss and Validation Loss is of 0.1 to 0.2.
This shows that when compared to ConvLSTM model's Loss curves, the LRCN model has

VI. CONCLUSION

The proposed an application of LRCN algorithm in Human activity recognition for Home surveillance to evaluated the scheme with real-world video data from our large-scale network. It find that outperforms the popular CCTV Intrusion alert approach in terms of Automation. It is possible to increase the overall Performance by adding more data. In terms of accuracy, our results show that the ConvLSTM Model's accuracy is not less than 93% and the accuracy of LRCN's Model is not less than 96%. Therefore we can conclude that the LRCN Model is more Robust and accurate. We successfully demonstrated the possibility of Automation in Home surveillance system By Using Human Activity Recognition.

References

[1] Shugang Zhang , Zhiqiang Wei , Jie Nie, Lei Huang, Shuang Wang, and **Zhen Li**. "A Reviw on Human Activity Recognition using Vision based method." 2017 "Journal of Health Care Engineering" Article ID 3090343

[2] Abbas, Qaisar et al. "Video scene analysis: an overview and challenges on deep learning algorithms." *Multimedia Tools and Applications* 77 (2017): 20415-20453.

[3] "AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions", (CVPR 2018) by Chunhui Gu

[4] *Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A. Ross, Jia Deng, Rahul Sukthankar*; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1130-1139

[5] *Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, Li Fei-Fei*; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1725-1732

[6] https://learnopencv.com

[7] https://openaccess.thecvf.com