Human Activity Recognition for Office Surveillance

*

P J Subrahmanya Hande
Computer Science and Engineering
PES University
Bangalore, Karnataka
hjairama@gmail.com

Rakeshgowda D S
Computer Science and Engineering
PES University
Bangalore, Karnataka
abhirakeshgowdads@gmail.com

Naveen Kumar
Computer Science and Engineering
PES University
Bangalore, Karnataka
naveenrampure2001@gmail.com

Nandana K A
Computer Science and Engineering
PES University
Bangalore, Karnataka
nandanaka10@gmail.com

Prof. Preet Kanwal
Computer Science and Engineering
PES University
Bangalore, Karnataka
preetkanwal@pes.edu

***Abstract*—Human activity surveillance video systems are gaining popularity in the field of computer vision due to user demands for security as well as their growing importance in many applications such as elder care, home nursing, and unusual event alarming. Automatic activity recognition is the key to video surveillance.This paper presents a method for human activity recognition in office surveillance videos using machine learning models including convLSTM,GRCNN and LRCN with three main steps: pre-processing, feature extraction and activity classification.The main targeted activities are walking, sleeping on desk, handshaking, typing, opening or closing door. Experimental results demonstrate the effectiveness of the proposed LRCN approach in accurately recognizing human activities in office surveillance videos with acceptable training and testing accuracy.**

I. Introduction

The last 10–20 years have witnessed significant progress in the automation of human activity recognition in order to realize intelligent environments that are capable of detecting users' actions and gestures so that the required services can be provided automatically and instantly for improving the user's comfort, increasing efficiency, and providing a sense of security. It is also being increasingly used in the detection of unusual activities, where activities that differ from day-to-day activities are categorized as unusual and hence detected. A system capable of inferring the behavior of humans and recognizing or even predicting their activities can have a wide range of applications, from surveillance to more complex functions like an automatic commentary on sporting events like cricket, football, etc. even better when these activities are recognized instantly and automatically. One such environment where human activity recognition plays a vital role is the office environment. People across the world work in different office-related jobs and perform a wide range of office activities, but there are some common activities that are similar in most workplaces. In the paper, there has been an attempt to classify some of the common office activities being performed by office employees on a day-to-day basis.

Recognizing these common office activities can aid in employee monitoring and detecting unusual behavior. Unusual activities can vary from employees slacking off, an employee falling to the ground due to fatigue, heart-related disease, etc., or even detecting intruders behaving suspiciously, or rapid movement in the form of running by employees indicating some kind of trouble and thereby taking immediate, appropriate actions to tackle the problems.

There are several major gaps in the human activity recognition from video for office surveillance using machine learning (ML) models. One of the significant gaps is the lack of large-scale annotated datasets that cover a diverse range of human activities in office environments. Another gap is the challenge of dealing with occlusions, where parts of the body are not visible, leading to incomplete information for activity recognition. The generalizability of ML models is also a major concern as models trained on one dataset may not perform well on different data sets, leading to poor real-world performance. Additionally, the deployment of ML models in real-world scenarios involves challenges related to power consumption, processing speed, and hardware requirements. Addressing these gaps requires further research and development of more robust and effective human activity recognition systems for office surveillance using ML models.

To overcome the dataset collection gap we have collected more diverse and representative datasets from online by considering copyright and privacy issues. It's essential to collect more diverse and representative dataset that cover a wide range of office activities, environmental conditions, and camera angles. This can be achieved by collecting data from multiple offices, locations, and times of day. The entire frame of the video is processed to reduce the error due to occlusion and lighting effects. To achieve generalizability of model it has trained over large and diverse office video data. This paper deals with the ML models that are comparatively efficient in terms of speed, power and other resources by reducing training time by processing an entire video in single pass.

II. Related Work

Literature survey is a crucial step in any research project.A thorough literature survey has been conducted to gain a comprehensive understanding of the existing work in the field, identify the gaps in the research, and develop a novel and effective approach to solve the problem.A literature survey can help to identify the strengths and weaknesses of different approaches and select the most appropriate method for their specific application. Moreover, a literature survey helps to establish the state-of-the-art performance on benchmark dataset, set a baseline for their proposed approach, and evaluate the effectiveness of their method compared to the existing work.Some of the major approached method key points of our literature survey:

[1] explains the framework with pretrained models to classify activity efficiently,it uses YOLO v4,ResNet,VGG pretrained models for target localization and feature extraction, and LSTM, Fast RNN for label generation.

In [2] uses ResNet a pre-trained CNN architecture for features extraction and a deep bi-LSTM network,which consists of two LSTM networks, one of which reads the input image frame and processes it in the forward direction, and the other of which reads the input image and processes it in the backward direction.

[3] uses a basic CNN model for the activity recognition and a LSTM network model for the dataset and then develops a CNN LSTM hybrid model for the classification and prediction of their 6 activities. It is then further extended to a ConvLSTM model where in there are convolutional layers in the LSTM.

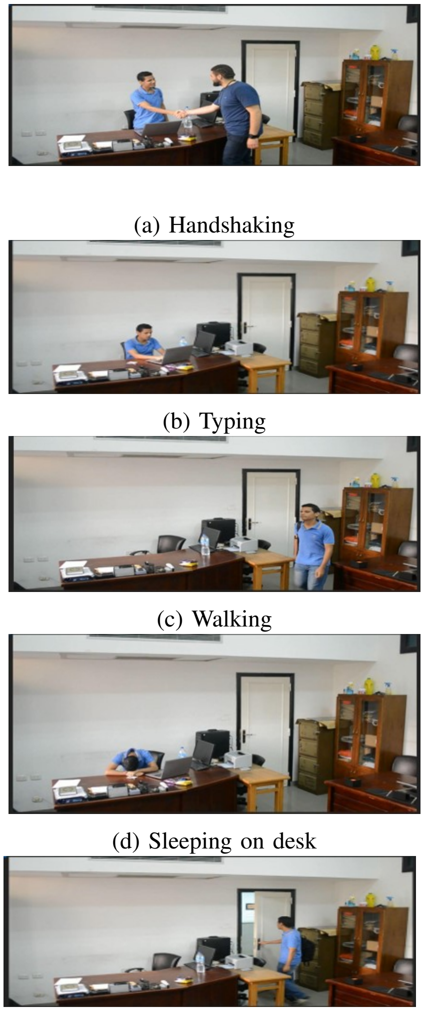
[4] describes activities as a combination of feature attributes known as Category Feature Vectors(CFV).Gaussian Mixture Models(GMM) are in turn used to represent the combination of CFVs of an activity. It uses a Confident Frame Based Recognition and creates a global model and local model to classify the activity.

In [5], Single Shot Detectors are used to detect multiple classes in a single frame.The model can detect normal and abnormal activities very accurately.

[7] tests both single-frame CNN as well as VGG16 to tell apart normal behaviours from suspicious ones. In the end, single-frame CNN proved to perform better for HAR applications.

III. Dataset Collection

Since this work is the first ever work that mainly deals with the activity recognition in the office environment collecting a suitable office video dataset for human activity recognition research is an important task that requires a considerable amount of effort and resources. While some publicly available datasets exist, they may not cover the full range of activities that occur in typical office environments. One approach to overcome this issue is to collect additional data from online sources. Online platforms such as YouTube and Vimeo contain a vast amount of videos that potentially include office-related activities. However, collecting and annotating videos from online sources requires careful consideration of copyright and privacy issues. Moreover, the quality and consistency of the videos in the dataset must be ensured to avoid bias and ensure the robustness of the recognition model. Therefore, while online video sources can provide a valuable resource for expanding office video datasets, careful curation and ethical considerations are necessary to ensure the reliability and usefulness of the resulting dataset.



(a) Handshaking

(b) Typing

(c) Walking

(d) Sleeping on desk

(e) opening/closing door

Fig. 1: Dataset Snapshots

IV. Methodology

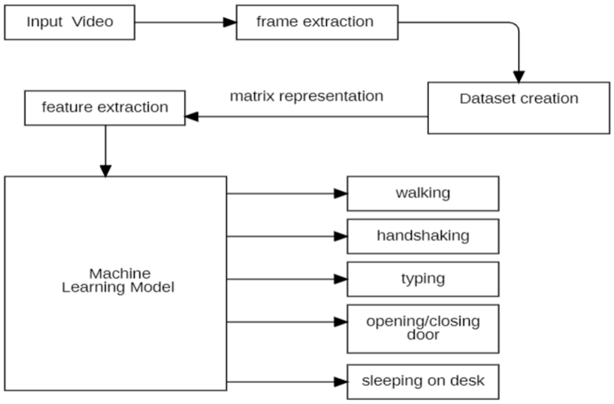


Fig. 2: Model Architecture

A. Data Visualisation

Data visualization is a crucial tool for analyzing and interpreting large amounts of data, and can be especially useful for tasks such as human activity recognition from video. In the context of office surveillance, a common goal is to detect and classify activities such as walking, sleeping on desk, typing, handshaking, and opening/closing door.In this first step, the video data is visualized by showing one frame from each class along with labels to get an idea about what the dataset. A dataset containing five different folders for each activity is read using CV2.VideoCapture function of cv2 module, a random video is chosen from every activity folder of dataset, for a chosen video first frame is extracted using function video_reader() ,each frame is resized with defined imaged height and width and displayed with their associated labels written on them.

B. Frame Extraction and Preprocessing

Frame extraction and preprocessing are crucial steps in many computer vision applications, including human activity recognition from video. In the context of video processing, a video is essentially a sequence of frames, each of which can be processed independently to extract visual features. Frame extraction involves capturing each individual frame from the video stream, typically at a fixed frame rate. Once the frames are extracted, they are preprocessed to enhance their quality and reduce noise. One common preprocessing step is to normalize the pixel values by dividing them by the maximum value (usually 255 for 8-bit images) to obtain values between 0 and 1. This step helps to improve the performance of the subsequent feature extraction and classification algorithms. By normalizing the pixel values, the image data is made more consistent and easier to process, while also reducing the impact of any variations in lighting or color. Overall, frame extraction and preprocessing are important steps in the video processing pipeline, and can have a significant impact on the performance of downstream tasks such as human activity recognition.

C. Dataset Creation

A dataset is created in a suitable form for the way a model expects the input for training. This function will iterate through all the classes mentioned in the CLASSES_LIST(which contains all 5 activity class names) and will call the function frame_extraction() mentioned above on every video file of the selected classes. This will return a list of resized and normalized frames for each video, which is then appended to the main features list. Along with extracting features for each video, a class label is added to the label list. A video_file_path list is also created, containing a path to each video file present on the local disk.

As a whole, this function will return a features list, with each element in the list containing a 3D matrix (number_of_frames*image_height*image_width) of each video, and each element in the 3D matrix being a RGB value of a pixel. This will also return a labels list containing the class label of each video. The feature list and labels are converted into numpy arrays before they are returned.Returned labels numpy array is converted into one-hot-encoded vectors using the Keras to_category method. In this project, there are 5 classes, namely: typing, walking, sleeping on a desk, opening and closing doors, and handshaking. When these 5 classes are converted into one-hot-encoded vectors, numerical values from 0 to 4 are given for each class name.

D. Splitting data into Train and Tests Set

Splitting data into training and test sets is a common practice in machine learning and data science. The purpose of this step is to divide the available data into two separate sets, one for training a model and one for testing its performance. The training set is used to fit the model to the data, while the test set is used to evaluate the model's performance on new, unseen data.The general approach is to randomly split the available data into two sets, with a larger portion assigned to the training set and a smaller portion assigned to the test set. The typical split ratio used in this project is around 70-30, but this can vary depending on the size of the dataset and the complexity of the model being trained. It is important to ensure that the split is representative of the underlying distribution of the data, so that the model is not biased towards any particular subset of the data.

Once the data is split, the model is trained on the training set, and its performance is evaluated on the test set. The test set should be held out until the very end of the model development process, to avoid any bias in model selection or hyperparameter tuning. In addition, cross-validation can be used to further evaluate the model's performance on different subsets of the training data, and to assess its generalization ability. Overall, splitting data into training and test sets is a critical step in machine learning, and can help to ensure that the resulting model is accurate, reliable, and generalizable to new data.

E. Machine Learning Models

1) CovnLSTM Model: ConvLSTM is a type of neural network that combines the convolutional layers with the Long Short-Term Memory (LSTM) architecture. An LSTM network is designed specifically to operate on a data sequence since it generates an output while accounting for all of the inputs that came before. Because of a phenomenon known as the “Vanishing gradient problem”, recurrent neural networks (RNNs), which are what LSTMs are in actuality, are known to be inefficient for handling long-term dependencies in input sequences. ConvLSTM is a kind of RNN (recurrent neural network) used for spatio-temporal prediction that incorporates convolutional structures in both input-to-state and state-to-state transitions. By using the inputs and previous states of its local neighbors, the ConvLSTM predicts the future state of each cell in the grid. Using a convolution operator in the state-to-state and input-to-state transitions makes this simple to accomplish.

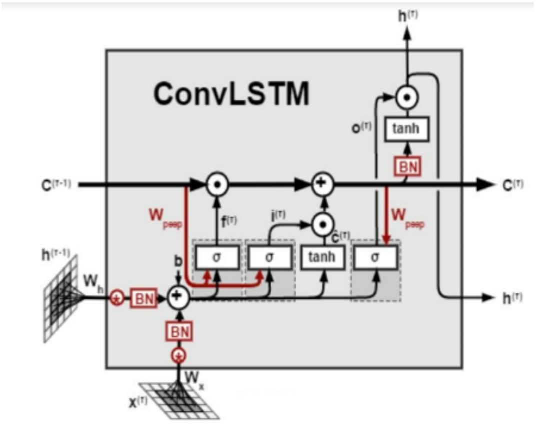


Fig. 3: ConvLSTM Cell Architecture

A ConvLSTM cell is a type of LSTM network that incorporates convolutional processes. It is an LSTM with embedded convolution, which enables it to recognise spatial aspects of the data while taking into consideration the temporal relationship. This method efficiently captures the geographical relationship between the individual frames and the temporal relationship between them for video categorization. Because of this convolution structure, the ConvLSTM may accept input in three dimensions (width, height, and number of channels). Methodology The Keras ConvLSTM2D recurrent layers is used to build the model. The amount of filters and kernel size required for applying the convolutional operations are also taken into account by the ConvLSTM2D layer. After being flattened, the layers’ output is passed to the Dense layer, which uses softmax activation to output the probabilities for each action category. Additionally, we’ll employ Dropout layers to stop the model from being overfitted to the data and MaxPooling3D layers to limit the size of the frames and eliminate pointless computations. The design is straightforward and just includes a few trainable parameters. This is due to the fact that we are only working with a small portion of the dataset, which does not call for a complex model.

In total the model has 54,605 tainable parameters.The model has trained with categorical cross entropy loss function and adam optimizer with batch size of 4 for 20 epochs. With these parameters the model had good training accuracy of 94.47%. When it came to testing the model had an accuracy of 60.68% with a loss of 1.466. So the model turned out to be largely overfit.

2) Gated Recurrent Convolutional Neural Network[GRCNN]: The goal of human activity recognition for office surveillance is to identify the activity performed by an employee in real-world office videos. To address this issue, a novel architecture called Gated RCNN (GR-CNN) is proposed, which is inspired by a newly published universal image classification model called Recurrent Convolutional Neural Network. Its primary building block, the Gated Recurrent Convolutional Layer, is created by including gates in the Recurrent Convolution Layer of the RCNN. Gates manages the RCL’s context modulation to balance repetitive information and feedforward information. The model consists of several Gated Recurrent units as an input layer. The main purpose of these layers is to extract the features from the input frames of the office video. GRUs are pretty similar to LSTM cells, but it only has three gates, unlike LSTM, and it doesn’t keep track of the internal state of the cell. The data that is kept in an LSTM recurrent unit’s internal cell state is incorporated into the gated recurrent unit’s hidden state. The next Gated Recurrent Unit receives this group of data.

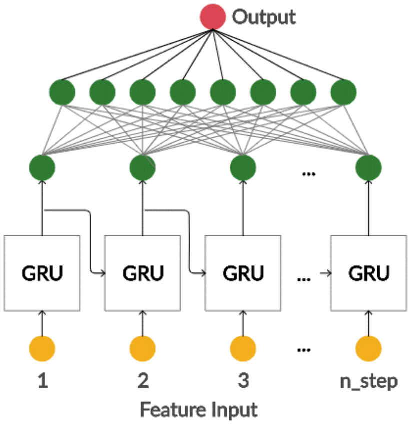


Fig. 4: GRCNN Architecture

Update Gate determines how much of the past must be transmitted into the future. It is comparable to an LSTM recurrent unit’s Output Gate.Reset Gate chooses how much of the prior information to erase. It is comparable to how the Input Gate and Forget Gate work together in an LSTM recurrent unit. The current memory gate is similar to how the Input Modulation Gate is a component of the Input Gate, it is incorporated into the Reset Gate and is used to make the input zero-mean and introduce some nonlinearity into the

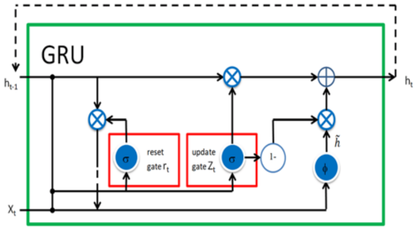


Fig. 5: GRU Cell

input. Reducing the influence that earlier information has on the information that is being transferred into the future is another justification for making it a component of the Reset gate. GRU cells take the current input and the previous hidden state as input vectors. The values for three different gates are calculated by performing element-wise multiplication between the relevant vector and the corresponding weights for each gate to determine the parameterized current input and previously hidden state vectors for each gate. Apply the "relu" activation function to the parameterized vectors for each gate element. The output of one cell will be fed as input to the next cell in the same layer and also to the cells in the next layer connected to it. To generate the labels for the given input office video, the GRU layers feed the output to fully connected dense layers. Each neuron in the dense layer receives information from all the neurons in the preceding layer. Softmax activation function is used in dense output layers.

In total the model has 99,909 tainable parameters.The model has trained with sparse categorical cross entropy loss function and adam optimizer with batch size of 64 for 50 epochs. With these parameters the model had good training accuracy of 98.73%. When it came to testing the model had an accuracy of 82.67% with a loss of 0.5674.

3) Long Term Recurrent Convolutional Network[LRCN]: LRCN or Long Term Recurrent Convolutional Network, refers to a class of models that can handle both spatial and temporal features. It combines the CNN and LSTM layers into a single model. Many different problems involving time-varying visual input or sequential outputs can be solved using LRCN models. The end-to-end model architecture provided by LRCN is simple to implement and train. The convolutional layers are used for spatial feature extraction from the frames, and the extracted spatial features are fed to the LSTM layer at each time step for temporal sequence modeling. This way, the network learns spatiotemporal features directly in an end-to-end training, resulting in a robust model. Input is passed into time distributed convolutional layers where the features are learnt, then into the max pooling layers , where the input size is reduced, and finally into the dropout layers to prevent overfitting. There are 4 layers of these. After this, the data is flattened and given to a 32 unit LSTM. The output of the LSTM is given to the dense layer, which uses the softmax activation function. Finally, this gives the probability of each activity, and the maximum one is selected.e input video is divided into a fixed number of frames, and CNN takes each frame and applies kernels, also known as filters, on each frame and generates feature maps. These feature maps are used to check if a particular feature is present in that location or not. Initially, there are a limited number of feature maps, but as we go deeper into the network, the number of feature maps increases and the size of the feature maps reduces, respectively, without losing any important information. This is done using pooling operations.

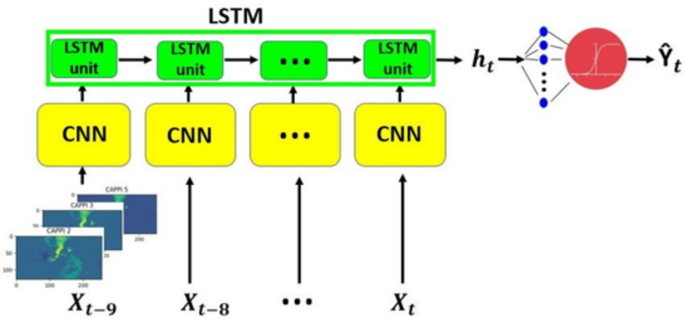


Fig. 6: LRCN Architecture

After the features have been extracted, the feature vector is then given to the sequence of LSTM units. LSTM (long short-term memory) has the advantage of being able to deal with the Vanishing Gradient problem. It is used to discover and identify the temporal information between the frames and predict the corresponding activity. For each frame, LSTM predicts an output and also transfers this information into the next LSTM. This way, the temporal relations are identified. The output probabilities of the LSTMs are averaged, and the one with the highest probability corresponds to the target activity. In total the model has 73,093 tainable parameters.The model has trained with categorical cross entropy loss function and adam optimizer with batch size of 4 for 70 epochs. This model provided the highest accuracy with a training accuracy of 96.72% and testing accuracy of 86.17%.

F. RESULTS AND DISCUSSION

ConvLSTM can capture both the spatial and temporal features of video data, making it a powerful model for analyzing complex human activities.It has been shown to be effective in recognizing human activities from video data, achieving high accuracy rates in various human activity recognition.

The Fig.7 shows the graphs of total accuracy and the validation accurary vs number of epochs, and total loss and validation loss.It clearly explains the model is overfitted i.e it performs well in training (94.47%) but not in testing (60.68%). ConvLSTM is a computationally expensive model, which can require significant computational resources for training and inference. It can also be sensitive to the choice of hyperparameters and architecture design, which can affect its performance and generalization ability. The interpretability of the model is also a challenge, as it can be difficult to understand how the model is making predictions, especially

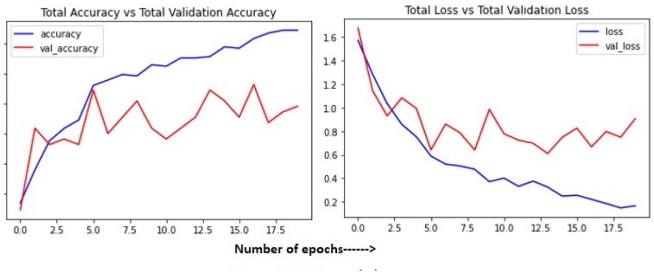


Fig. 7: ConvLSTM Result Graph

in complex scenarios.

Compared to ConvLSTM, GRCNN has several advantages when it comes to human activity recognition (HAR).GRCNN has been shown to achieve higher accuracy than ConvLSTM in several studies on HAR. This is because GRCNN is better able to capture temporal dependencies in the data, and it can learn more complex patterns in the input.GRCNN can be faster to train than ConvLSTM, as it uses fewer parameters and has a simpler architecture. This can be an advantage in real-world applications, where training time is an important factor.GRCNN requires less memory than ConvLSTM, as it does not store as much information about the past states of the network. This can be important when working with large dataset or when running the model on low-power devices.GRCNN is more robust to noisy data than ConvLSTM, as it can learn to filter out irrelevant information and focus on the important features. This can be important in real-world applications where the input data may be noisy or incomplete.

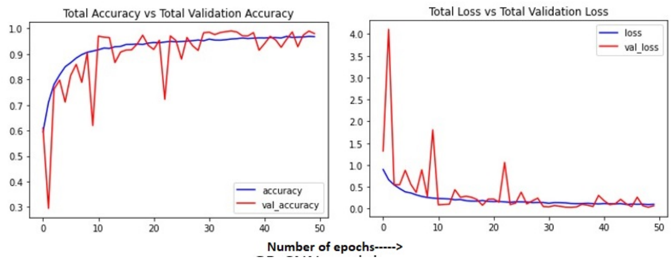


Fig. 8: GRCNN Result Graph

The GRCNN model has a complex architecture and involves a large number of parameters. As a result, it can be computationally expensive to train and evaluate, which can be a drawback in real-time applications.GRCNN has shown improved performance compared to convLSTM model in some cases, the performance gains may not always be significant enough to justify the added complexity and computational cost.Though it is more far better than convLSTM it's not best suitable for real world applications due to its sensitivity to hyperparameters. Like other deep learning models, GRCNN performance is heavily influenced by its hyperparameters, such as learning rate, batch size, and number of layers. Finding the optimal set of hyperparameters can be a challenge, and suboptimal choices could negatively impact model performance.This can be verified with Fig.8 that clearly shows high deviations in validation accuracy and validation loss with total accuracy and total loss when plotted against number of epochs. GRCNN typically requires large amounts of data to train effectively. In some cases, it may be difficult to collect sufficient training data for specific applications, which can limit the usefulness of the model.GR-CNN,as opposed to ConvLSTM, yields better results in both training (98.73%) and testing (82.67%).

LRCN has the ability to capture the long-term dependencies within sequential data due to its ability to remember past inputs and propagate relevant information through time. This allows the LRCN to better model complex temporal dynamics in human activity recognition, which can be especially beneficial in scenarios where activities may overlap or transition into one another.

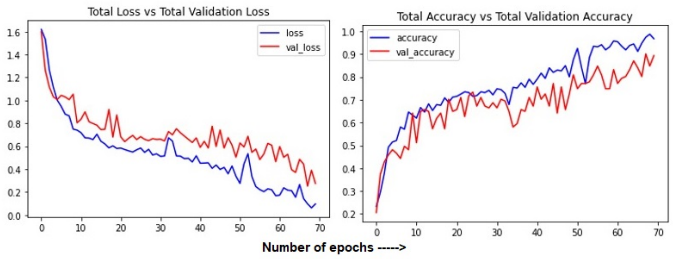
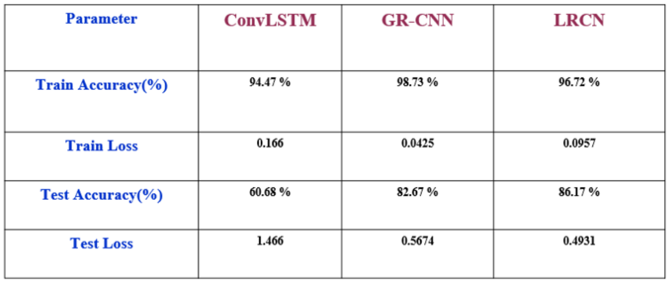


Fig. 9: LRCN Result Graph

The LRCN model is less susceptible to overfitting compared to other two models.This can be verified with Fig.9 where it shows less deviations in validation accuracy and validation loss with total accuracy and total loss when plotted against number of epochs. LRCN typically has a smaller number of parameters than both GRCNN and ConvLSTM, which can lead to less overfitting on smaller datasets. This can make LRCN a more effective option for scenarios where limited data is available.LRCN yields stable results in both training (96.72%) and testing (86.17%) .



Parameter	ConvLSTM	GR-CNN	LRCN
Train Accuracy(%)	94.47 %	98.73 %	96.72 %
Train Loss	0.166	0.0425	0.0957
Test Accuracy(%)	60.68 %	82.67 %	86.17 %
Test Loss	1.466	0.5674	0.4931

Fig. 10: Result comaparison of convLSTM,GRCNN,LRCN Model

From all the observations from three different models we can summarize the result and performance of models in terms of accuracy and loss by the above table.It is very clear that LRCN has better results in terms of accuracy and loss in training and testing.

G. CONCLUSIONS

One of the difficult problems in computer vision in recent years has been the automatic identification of human action and activity. It is crucial for many artificial intelligence applications, including video surveillance, computer gaming, robotics, and human-computer interactions. When researching relevant research on the subject, it was discovered that the feature extraction, action representation, and classification steps of processes can be used to construct a human action recognition system. Among the various use cases of human activity recognition, office activity recognition is considered essential for efficient and safe day-to-day working in the office.

Human activity recognition from video for office surveillance using ML models poses several challenges. One of the main challenges is the collection and annotation of a large dataset that accurately reflects the diversity of activities in an office environment. Another challenge is the design and tuning of the deep learning models to achieve high accuracy while minimizing overfitting. Additionally, the high computational requirements for training and inference of deep learning models can be a challenge, especially when dealing with large datasets. Finally, generalization of the trained models to other similar environments may also pose a challenge. This project carefully considered and planned to ensure the success over these critical challenges.

This project looks at some of the common activities that exist in an office and tries to classify them. We made use of some of the most popular algorithms that are used in the machine learning and artificial intelligence fields, such as the ConvLSTM, which is a LSTM with convolution layers; GRCNNs, which make use of GRU units for feature extraction; a fully connected dense layer for caption generation; and finally, LRCN, which is a combination of CNN and LSTM, where the CNN is used for feature extraction and the LSTM is used for identifying the temporal relations and classifying the activity.LRCN offers several advantages for human activity recognition from video for office surveillance. Firstly, LRCN has a simpler architecture compared to convLSTM and GRCNN, which makes it faster and easier to train.

Additionally, LRCN can process variable length inputs, making it more versatile in real-world scenarios.Secondly LRCN is less susceptible to overfitting and it has a smaller number of parameters than both GRCNN and ConvLSTM, which can lead to less overfitting on smaller datasets. This can make LRCN a more effective option for scenarios where limited data is available. In conclusion, the above methods have produced satisfactory accuracy and are therefore promising methods that can be improved upon.The choice of model architecture for human activity recognition depends on the specific requirements and constraints of the application. However, based on the analysis of the performance and advantages/disadvantages of the three models, it can be concluded that LRCN is a suitable choice for HAR in office surveillance scenarios.

This work is not only limited to the office environment and can be further expanded to other places, such as government institutions like schools, colleges, hospitals, etc. It can also be used for abnormal activity detection, i.e., surveillance in banks, shops, etc., or even in the monitoring of the elderly in rehabilitation centers or patients in hospitals.

REFERENCES

[1] ANDREI DE SOUZA INÁCIO 1,2, MATHEUS GUTOSKI 1 , ANDRÉ EUGÊNIO LAZZARETTI 1 AND HEITOR SILVÉRIO LOPES. OS-VidCap: A Framework for the Simultaneous Recognition.Graduate Program in Electrical Engineering and Industrial Informatics,Received August 20, 2021, accepted September 19, 2021, date of publication September 29, 2021, date of current version October 12, 2021.

[2] Mihanpour,Akram; Rashti, Mohammad Javad; Alavi, Seyed Enayatallah (2020). [IEEE 2020 6th International Conference on Web Research (ICWR) - Tehran, Iran (2020.4.22-2020.4.23)]Human Action Recognition in Video Using DB-LSTM and ResNet. , (), 133–138. doi:10.1109/ICWR49608.2020.9122304 .

[3] Shiranthika, C., Premakumara, N., Chiu, H.-L., Samani, H., Shyalika, C.,and Yang, C.-Y. (2020). Human Activity Recognition Using CNN and LSTM. 2020 5th International Conference on Information Technology Research (ICITR). doi:10.1109/icitr51448.2020.93107

[4] Weiyao Lin, ; Ming-Ting Sun, ; Poovandran, Radha; Zhengyou Zhang, (2008). 2008 IEEE International Symposium on Circuits and Systems - Human activity recognition for video surveillance. , (), 2737–2740. doi:10.1109/ISCAS.2008.4542023

[5] A. Sunil, M. H. Sheth, S. E and Mohana, "Usual and Unusual Human Activity Recognition in Video using Deep Learning and Artificial Intelligence for Security Applications," 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2021, pp. 1-6, doi: 10.1109/ICECCT52121.2021.9616791.

[6] Chaitanya Yeole, Hricha Singh, Hemal Waykole, Anagha Deshpande, "Deep Neural Network Approachesfor Video Based Human Activity Recognition", 2021, International Journal of Innovative Science and Research Technology ISSN No:-2456-2165

[7] M. R. Siyal, M. Ebrahim, S. H. Adil and K. Raza, "Human Action Recognition using ConvLSTM with GAN and transfer learning," 2020 International Conference on Computational Intelligence (ICCI), 2020, pp. 311-316, doi: 10.1109/ICCI51257.2020.9247670.

[8] Wang, J.,and Hu, X. (2021). Convolutional Neural Networks with Gated Recurrent Connections. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–1. doi:10.1109/tpami.2021.3054614