

Enhancing Human Activity Analysis in Video Surveillance with Recurrent Neural Networks

Ali Ahmad Sabir¹, Ujala Saleem¹, Naeem Aslam¹, Abdul Rehman², Muhammad Sajid^{3*} and Muhammad Fuzail¹

¹Department of Computer Science, NFC Institute of Engineering and Technology Multan, Pakistan.

²Department of Computer Science, Lahore Garison University, Lahore, 54000, Pakistan.

³Air University Islamabad, Multan Campus, Pakistan.

*Corresponding Author: Muhammad Sajid. Email: msajid@aumc.edu.pk

Received: March 22, 2023 Accepted: May 29, 2023 Published: June 05, 2023

Abstract: Recently, numerous security systems have been implemented to enhance security in public and private spaces. However, relying solely on human monitoring of surveillance cameras can lead to errors and missed events, making it inefficient and time-consuming. To address this issue, this research explored the effectiveness of recurrent models in investigations involving sequences, as deep convolutional network models have primarily dominated image interpretation challenges. This study developed a group of end-to-end trainable, deep, and task-specific recurrent convolutional architectures for visual understanding. These models excelled in detecting human activities and were utilized to create a model specifically designed for identifying unusual events in security camera data. This study approach employed a Convolutional Neural Network to extract important features from each frame in the input sequence. Additionally, this study implemented a classification mechanism capable of distinguishing between common and abnormal behaviors, enabling the system to categorize each detected aberration accurately. To evaluate the performance of the proposed model, this study utilized the UCF50 dataset and achieved an impressive accuracy of approximately 93%. This accuracy surpassed other models, such as ConLSTM, when tested on the same dataset.

Keywords: Human Activities Recognition; Long Short-term Memory; Surveillance; CNN.

1. Introduction

Understanding human behavior is crucial in various applications, including healthcare monitoring, exercise tracking, remote monitoring, wearable technology, traffic planning and control, targeted advertisement, and safety [1]. For example, tracking a person's daily activities makes it possible to determine their calorie consumption and provide personalized recommendations for a healthy diet. Similarly, detecting patterns indicative of a risk of falling in an elderly individual can trigger the appropriate support to prevent accidents. Traditional machine-learning techniques have been used to identify human behavior but rely on manually designing and selecting relevant features. This process is time-consuming and requires specialized knowledge and the resulting features may not always meet expectations. In recent years, deep learning algorithms have emerged as a promising approach to alleviate the burden of manual feature engineering [2]. These algorithms can automatically learn relevant features from data, reducing the need for human intervention and potentially improving performance.

Deep learning networks, also known as deep neural networks, are artificial neural networks with multiple hidden layers. According to [3] researchers, deep-learning models can be categorized into three

groups: supervised learning, unsupervised learning, and hybrid methods. Over the years, recurrent neural networks (RNNs) have been extensively studied in perceptual applications, yielding varying degrees of success. However, RNNs suffer from a significant drawback known as the "vanishing gradient" problem. This problem arises when back-propagating errors through a long sequence of time steps becomes increasingly challenging. A class of models was introduced in [4] to address this issue, which incorporated memory-cell-like neural gates. These models allow the state to flow unchanged, updated, or reset by combining hidden states with nonlinear dynamics. While these models have shown effectiveness in various tasks, their actual value was recently demonstrated in studies involving extensive learning of speech recognition [5] and language translation models. This highlights the potential of these models in capturing long-range dependencies and achieving superior performance in complex tasks, as shown in figure 1

This study explores the effectiveness of modeling long-term recurring convolutional networks for visual time-series data. We argue that these long-term recurrent neural networks (RNNs) can outperform static or flat temporal models in visual tasks, especially when ample training data is available for model development and learning. This study demonstrates that models based on LSTM (Long Short-Term Memory) architecture provide a novel, end-to-end approach to mapping pixel-level visual data to natural language descriptions at the sentence level. These models not only improve the detection of traditional video activity problems but also enhance the generation of descriptions from visual examples, bridging the gap between traditional graphical models and language understanding [6]. This study implements the design in three test environments to validate the proposed approach. Firstly, we apply it to instructional video identification systems incorporating intricate temporal relationships. By directly integrating visual convolutional deep LSTM networks, this study observes improvements of approximately 4% on standard benchmark datasets [7]. This improvement is significant, even though the labeled video activity datasets may not exhibit highly complex temporal dynamics regarding the captured actions or activities. Overall, our study highlights the potential of long-term recurrent convolutional networks in modeling visual time-series data, showcasing their advantages in various visual tasks and paving the way for more advanced applications in the field.

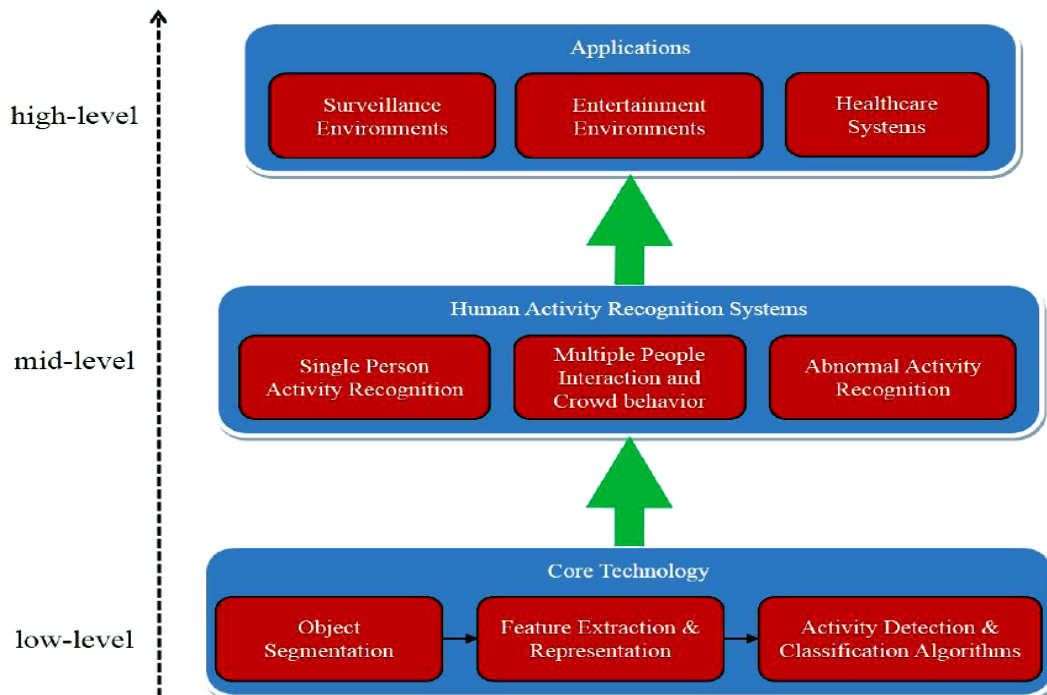


Figure 1. Broad Method for Identifying Human Activity

One critical challenge in addressing this issue is the scarcity of video samples, which hampers the ability to effectively employ sophisticated data-driven learning techniques. Previous attempts [8] to tackle this problem have encountered issues with overfitting, necessitating a reduction in the size of the learning parameters. Significantly larger sample sizes are required for these techniques to generalize and achieve superior performance on testing data. However, the proposed dataset provides a solution by offering

sufficient data for data-hungry approaches like deep learning techniques to excel. This enables the research community to advance in 3D human activity research. Our test results on the suggested dataset validate the superiority of data-driven learning techniques compared to state-of-the-art hand-crafted features [9]. By leveraging larger datasets, we can leverage the full potential of data-driven approaches and push the boundaries of human activity recognition. The primary objective of this research is to explore and improve techniques for building a foundation for action recognition of customers based on video data. While deep learning has been utilized by large businesses to determine customer preferences and enhance product offerings, there currently lacks a framework or method designed explicitly for recognizing customer activities in a shopping center using video data. Therefore, this study aims to focus on customer activity recognition to provide an automated security system that is easily accessible to those in need [10].

To achieve this goal, the research will develop a framework for activity recognition that utilizes a model capable of extracting customer activity information and analyzing their routine behaviors based on various component features. Deep learning strategies will be employed for activity recognition using these features, and the UCF50 dataset will be utilized to reduce the financial burden on the human activity recognition (HAR) system and mitigate losses caused by theft [11]. The research aims to enable accurate and automated recognition of customer activities in a shopping center by developing this framework and employing deep learning techniques. This will contribute to advancing security systems by providing a more efficient and effective means of identifying and responding to potential security threats.

2. Related Work

HAR, or Human Activity Recognition, plays a vital role in people's everyday lives due to its remarkable ability to derive deep insights into human behavior directly from raw camera inputs. It is a crucial component in various applications such as video surveillance, home behavior analysis, and signal recognition [12]. Two fundamental types of HAR exist video-based HAR and sensor-based HAR. Sensor-based HAR primarily relies on motion data collected from intelligent sensors such as accelerometers, gyroscopes, Bluetooth devices, and sound sensors. In contrast, video-based HAR involves analyzing videos or images captured by cameras, specifically focusing on human motion. Everyday usage of smartphones has made them the most indispensable items in our lives, and as technology advances, they continually enhance their ability to meet customer expectations and demands. To enhance the capabilities of these devices, designers make hardware modifications by incorporating new components and modules. Built-in sensors are ubiquitous in nearly all smartphones as they are crucial in expanding their functionality and environmental awareness [13]. With the advancements in the Internet of Things (IoT), the concept of a smart home is gaining significant attention as it offers a range of benefits, such as healthcare monitoring, assistance with daily tasks, energy management, and enhanced safety [14]. A smart home is equipped with multiple sensors and actuators that enable the monitoring of various parameters like door openings, room lighting levels, temperature, humidity, and more.

Furthermore, it lets users control devices such as heating systems, blinds, lighting fixtures, and home appliances. Current research efforts predominantly concentrate on developing methods to adapt feature representations through learning to focus on relevant areas in human detection. Examples of such approaches include the model and sparse coding proposed by [15] and the Bag of Words approach. The progress in deep learning algorithms, the availability of vast amounts of data, and the computational power of modern computers have significantly contributed to advancing Human Action Recognition (HAR) systems. Among these systems, Poolview stands out as a machine vision technology utilized in surveillance systems to reduce the need for manual monitoring and enhance people's safety, such as in community security and crime prevention [16]. This research leverages a deep learning network incorporating RNN (Recurrent Neural Network) and LSTM (Long Short-Term Memory) architectures to classify various activities based on dynamic video motion, particularly in sporting events. The findings of this study have implications for performance evaluation and sports safety. Please note that you are permitted to print or copy portions of the study for personal or educational use as long as proper attribution is given and not used for profit or commercial purposes.

This study utilizes the LSTM (Long Short-Term Memory) model to effectively capture long-term contextual information in the temporal domain, leveraging its powerful modeling capabilities. The LSTM model is enriched with a range of spatial domain variables. The researchers drew inspiration from previous

studies [17], where class scores from multiple sources were combined. In this research, class score fusion is applied across various LSTM channels that process different types of features [18]. Additionally, score fusion is performed between the CNN (Convolutional Neural Network) and LSTM channels. This fusion technique demonstrates superior performance compared to combining multiple LSTM channels, thanks to the complementary nature of the CNN and LSTM models [19]. The proposed approach is evaluated on the NTU RGB+D Dataset, yielding cutting-edge results [20].

3. Materials and Methods

Depending on the specifics of the work at hand and the dataset's characteristics, preprocessing techniques can alter. It is essential to carefully consider which preprocessing techniques are most appropriate for a certain video classification task. Preprocessing is an important step in categorizing movies since it improves the classifier's performance by preparing the data for analysis. For classifying films, common preprocessing techniques include [21]. During preprocessing, a video file is read from the dataset. Video frames are scaled to a specific width and height to make calculations easier. Additionally, the data is adjusted to lie between 0 and 1. To speed up convergence during the model's training, pixel values are divided by 255. We will scale the frames to 64 X 64 to improve accuracy. To get better results, we must increase the frame size to 64 X 64 because doing so raises the cost of computing. Give the sequence length to the LSTM. It specified how many video frames were given to the model in a particular sequence. The more series there are, the larger the network and the longer it will take to train.

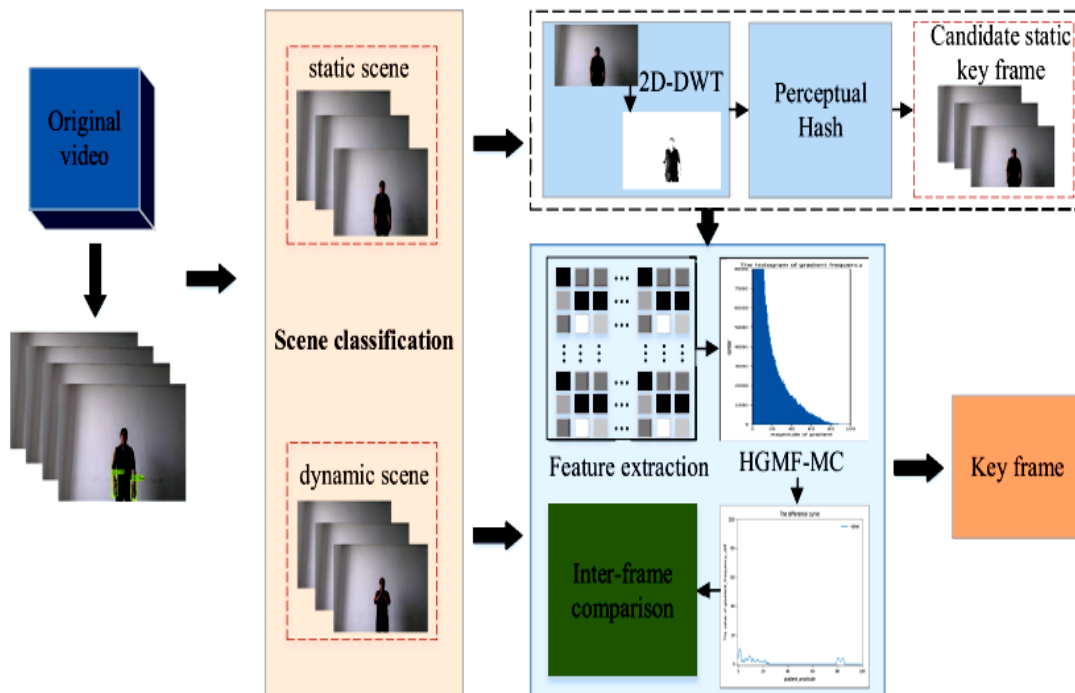


Figure 2. Feature Extraction framework

Convolutional Neural Networks (CNNs) excel in handling image data and image classification tasks, while LSTM techniques are well-suited for processing sequential data. However, both CNN and LSTM methods can categorize videos and address challenges like activity recognition as shown in figure 2. In this study, we explore various approaches TensorFlow employs to classify videos. Traditionally, video surveillance systems heavily rely on human analysis. However, this study focuses on developing highly autonomous systems that can analyze, process, and handle video inputs without human intervention. The system automatically analyzes, processes, and treats suspected events captured in the video footage as depicted in figure 3. The study investigates different methods, such as utilizing recovered region data as input to locate and examine the behavior of objects within the video.

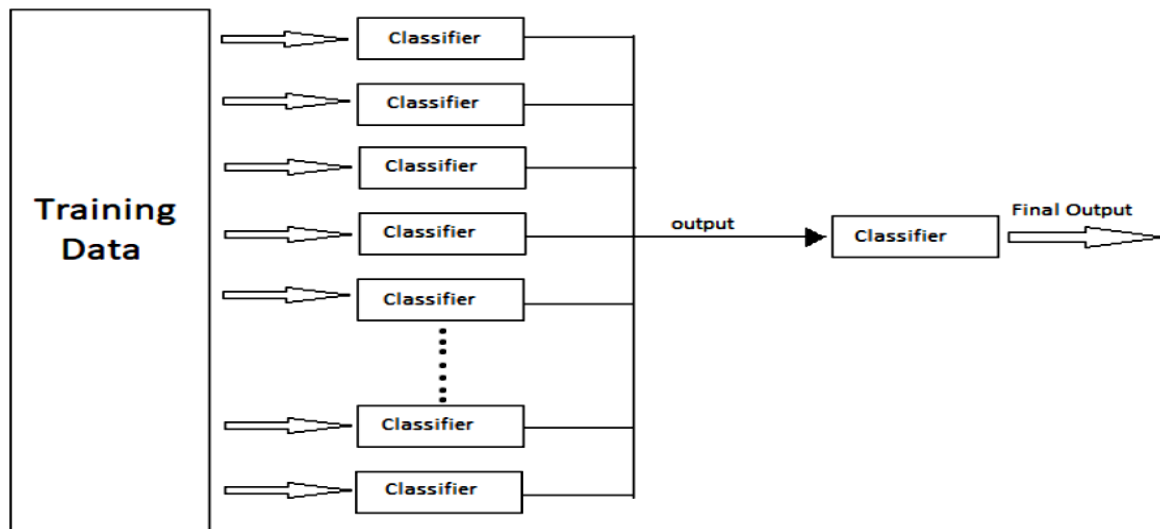


Figure 3. Structure for Stacking Classifiers

3.1 ConLSTM Model

The initial strategy will be implemented in this phase using a combination of ConLSTM cells. Convolutional processes are incorporated into LSTM network versions called ConLSTM cells. It is LSTM with CNN included in the proposal, permitting it to differentiate between input items having a spatial and temporal component. This method efficiently captures the temporal and geographic links between individual frames to classify films. This convolution structure allows the ConLSTM to receive input in 3 ways (width, height, and the number of channels). An LSTM as shown in figure 4, cannot independently represent spatiotemporal data since only one input dimension can be used with simple LSTM. This LSTM does not work with 1D data but only 3D data. It is built on complicated processes [22].

Recurrent layers from the Keras ConLSTM2D framework will be used to build the model. The ConLSTM2D layer additionally considers the kernel's size and the number of filters required to apply convolutional operations. The dense layer receives the output from the layers after they have been flattened and utilizes SoftMax activation to calculate the probability for each action category. Additionally, we'll employ MaxPooling3D layers to shrink the size of the frames and get rid of needless computations, dropout layers to prevent the model from overfitting the data, and both. The straightforward design has a few trainable parameters. This is because a very insignificant dataset slice is being used, which does not require a complicated model [23].

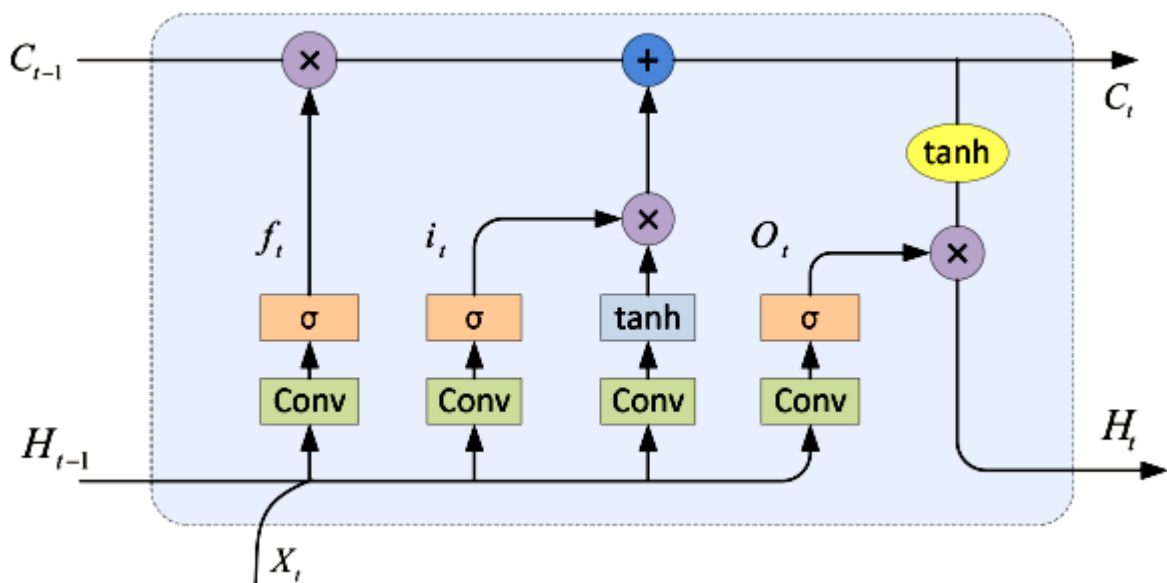


Figure 4. ConLSTM Model

Layers are added to the ConLSTM2D model to construct a cell for a particular network. As more filters are used, a network is more capable of learning. As one digs deeper into the network, more filters are added, including 2, 4, 8, 14, and 16. The CNN image's grid is being expanded with more and more feature maps. The network has more features since there are more filters. Processing is accelerated by using pooling layers to shrink the size of the feature map after each convolutional layer has been added. It gives the network wider scope and allows for more precise training. We employ a variety of image frames because of the network, so Maxpooling3D is used. The size is often reduced by half whenever pooling is employed. There are four convolutional layers added to the network. To lessen the scope of the feature maps and upsurge the precision of the predictions, max pooling is added repeatedly with increasing filters.

3.2. Recurrent Convolutional LSTM

Using features extracted by CNN, a straightforward algorithm has been developed to recognize human activity. A series of inputs are then sent to an LSTM to detect human activity.

$$z_m^k = f\left(\sum_{vN_k} Xc_m^k + b_m^k\right) \quad (1)$$

An RNN-like structure is the LSTM architecture. RNNs' long-range dependencies and memory backup make LSTMs more accurate and efficient than conventional RNNs [24]. LSTMs were developed to describe temporal sequences. The method is applied following data preprocessing, which removes unwanted, missing, and null signal values. The LSTM provides a solution by including a memory cell (C) to reliably encode knowledge at each phase. The memory cell y_t is controlled by an input gate r_t , forget gate m_t , and output gate. The input that is read during categorization is monitored by these gates. Additionally, these control gates help the LSTM send the information through to an unsupervised deep-learning hidden state without changing the output as shown in figure 5. The definition of the LSTM's gates and updating at time t is:

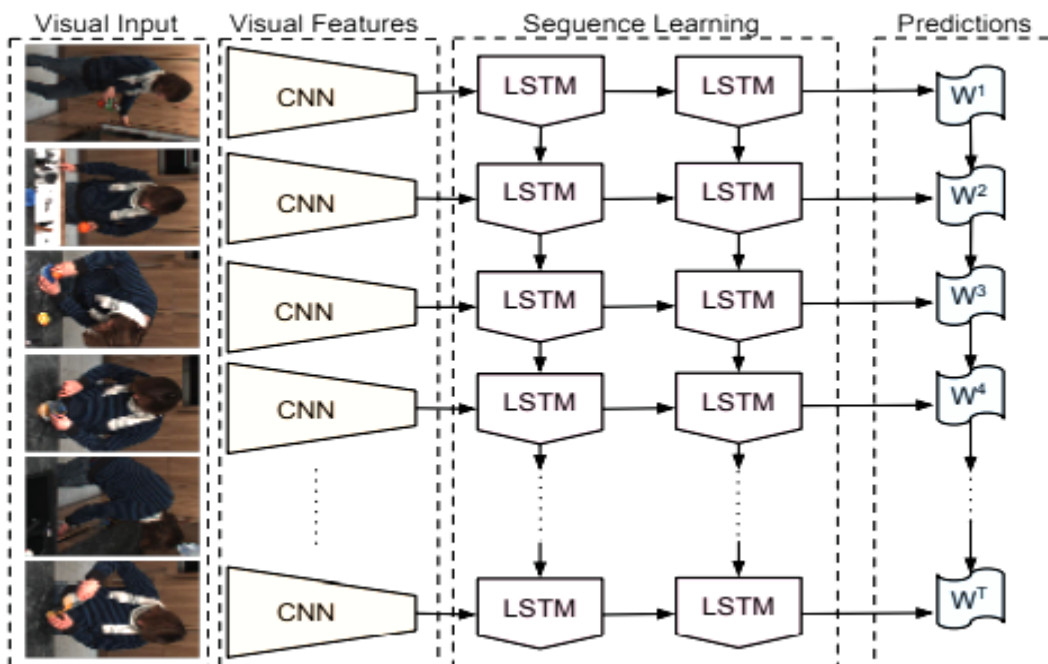


Figure 5. RCLSTM

$$r_t = S(Z_{xr}X_t + Z_{hr}h_{t-1} + Z_{vr}V_t + b_r) \quad (2)$$

$$m_t = \sigma(Z_{xm}X_t + Z_{hm}h_{t-1} + Z_{vm}V_t + b_m) \quad (3)$$

$$n_t = \sigma(Z_{xn}X_t + Z_{hn}h_{t-1} + Z_{vn}V_t + b_n) \quad (4)$$

$$y_t = \tanh(Z_{xc}X_t + Z_{hc}h_{t-1} + Z_{vc}V_t + b_c) \quad (5)$$

$$C_t = m_t \odot C_t - 1 + r_t \odot y_t \quad (6)$$

$$h_t = n_t \odot \tanh C_t \quad (7)$$

We use one input, a feature vector from the series of video frames we feed to LSTM, which only requires one block of data to train a series of captions instead of the video captioning model.

$$j_t = \frac{1}{h} \sum_{i=1}^h \bar{Z}_i f_i \quad (8)$$

$$\bar{P}_t = \bar{Z}^T \tanh(\bar{Z}_h h_t + M_h R_f + b_h) \quad (9)$$

$$C_t = \text{softmax}(\bar{P}_t) \quad (10)$$

The output probabilities from the SoftMax classification layer are then shown in C_t format.

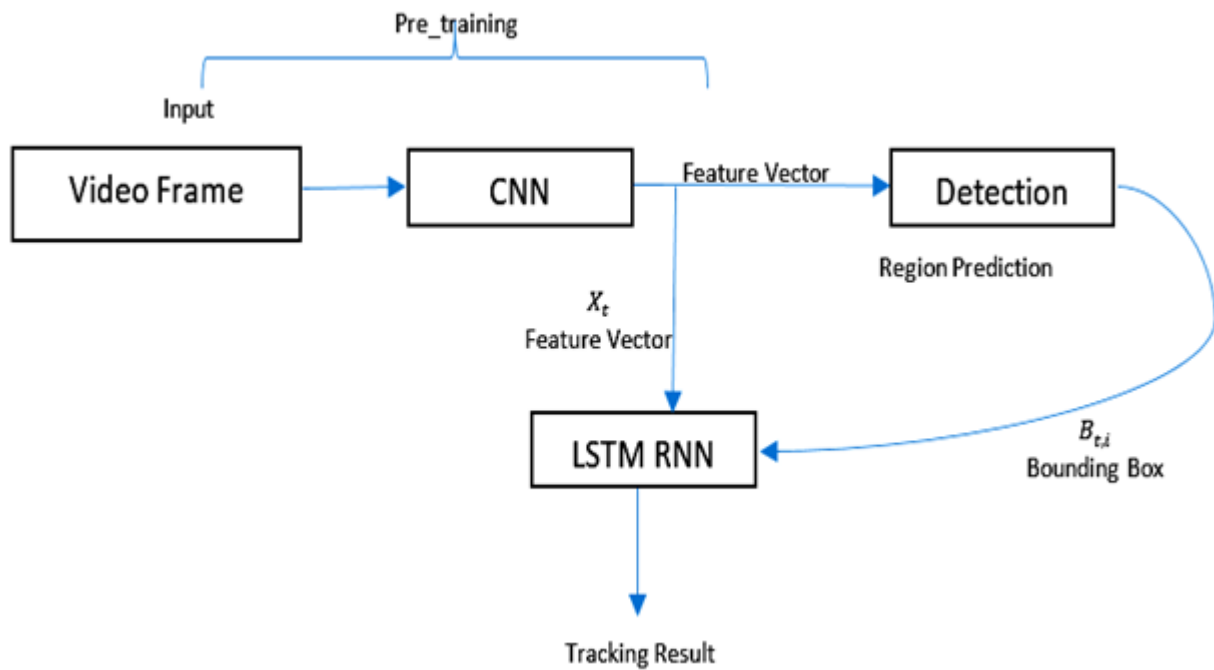


Figure 6. Human Activity Detection Using RCLSTM

$$Loss = -\frac{1}{M} \sum_{a=1}^M \left(\sum_{d=1}^D \sum_{j=1}^{C+F} 1\{x_t^a = k\} \log n_k \right) \quad (11)$$

Since we'll be working with video, we'll only concentrate on many-to-one LSTM networks because we want to transmit many frames through the network before receiving an action prediction. We would therefore be focusing on this. As we've discussed, LSTM works best with a data sequence, whereas CNN is excellent for picture classification. Although we've discussed various methods for classifying images and identifying actions, none of them could provide us with an accurate prediction. Finally, to benefit from both methods and obtain the necessary human actions, we integrated the convolutional neural network and LSTM network. Convolutional networks will extract frames from videos, and the LSTM network will use this output to perform action recognition as depicted in figure 6.

The convolutional neural network will learn spatial information, while the LSTM will learn temporal information. In this stage, the RCLSTM approach will chain the CNN and LSTM layers into a single model. Another related method combines two independently developed models: a CNN model and an LSTM model. Using the CNN model, a pre-trained model adaptable to the issue can excerpt spatial information from the video's frames. The LSTM model can then use the information obtained by CNN to forecast what will occur in the video. The LSTM layer(s) get spatial data from the frames from the convolutional layers at each time step to represent the temporal sequence. A robust model is produced because the system acquires Spatiotemporal properties during end-to-end training. We will work with a time Distributed wrapper layer, which enables us to apply the same layer to each movie frame. If the layer's original input shape is not the desired form, it allows the layer (around which it is wrapped) to accept shape input (number of frames, width, height, and channels) (width, height, and the number of channels). This is especially useful because it allows you to feed the entire video into the model in a single shot.

4. Results and Discussion

Grounded on the results, the RCLSTM model seems to have performed exceptionally well for a small number of classes. As a result, we will test the RCLSTM model on videos at this point. We compare the results of our experimental work with those from earlier methods applied to the UCF50 to determine how well our model performs. The UCF50, which has a lot of strange, illegal, and hostile conduct captured on video in public spaces, including streets, stores, and schools, is used in this research to put the recommended model into effect. The fact that this dataset was created from actual events that could happen anywhere at any moment was a deciding factor. These bizarre habits can also have major negative effects on people and society. It is uncommon in our everyday life to use a handmade dataset in several publications, whether a public dataset or one with specific backdrop and surroundings (for example, the dataset from a hockey game and the dataset from a movie).

CF50 as given in Table 1, is a varied assortment of human behaviors. It consists of fifty action classes, with movies from each class organized into separate groups that have some shared characteristics. For instance, one collection can include four recordings of the same person playing the piano, each shot from a different perspective. In addition to the category of routine events, this dataset also contains longer, straggly surveillance video feeds for various oddities, such as fighting, robberies, explosions, robberies, shoplifting, and traffic accidents. We used 75% of the data for training and 25% for testing in our trials to allow for a fair comparison with other studies in the field. The realistic UCF50 Action recognition dataset will be used.

Table 1. Dataset

Dataset	UCF50
---------	-------

Average Videos per Action Category	100
Average Number of Frames per Video	199
Average video frame height	320
Average video frame width	240
The average number of frames per video	240
Average Frames Per Seconds per Video	26

The initial frame of one random video from each genre will be shown, along with the labels that go with it, in 20 randomly selected categories. The proposed system was tested using a Jupiter Notebook running Windows 10 and equipped with a Core™ i5 processor and 8 GB of RAM. LSTM is used to evaluate data sequences, ConLSTM, an enhanced deep learning technique, to extract CNN features, and the RCLSTM "classification Learner" model to learn action recognition. We evaluate the suggested system's action recognition performance technique on various example movies. Table 2 gives the experimental setup for each model.

Table 2. Experimental Setup - RCLSTM & ConvLSTM Parameters

Parameters	RCLSTM	ConvLSTM
Number of Filters	16,32,64	4,8,14,16
Kernel Size	3x3	3x3
Sequence Length	20	20
Number of Classes	6	6
Image Width and Height	64x64	64x64
Activation Function	Relu	Tanh
MaxPooling	2D	3D
Recurrent Dropout	0.25	0.20
Optimizer	Adam	Adam
No. of Epochs	70	50

The suggested model was applied in our testing using ConvLSTM, a piece of the Keras tool set. We tweaked the model using several hyperparameters to acquire the best results from our studies. Table 3 compares the results of our testing regarding the initial weight and optimizer types and whether data augmentation was used. As a result, we expanded our dataset and ran tests using Adam as an optimizer. The model's learning rate is additionally set to 10^{-4} . The code terminates when the loss function converges since to continue would be nonsensical. Likewise, the epochs are set to 20. If the difference between the loss functions of the two succeeding epochs is less than the tolerance value, the method ends, and the result is absolute correctness. This implies that the loss function is calculated after each epoch. We must also specify how many frames from each video file we will extract. We set this sequence length to 20 during our tests. We evaluate the accuracy and compare our proposed method with a 3D convolutional network. The evaluation of ConvLSTM revealed an increase in computation time. Now, we generated a forecast of the identification of human actions using our fictitious RCLSTM model and comparison with other models is given in table 3. We obtained the findings displayed in figure 7, after using the RCLSTM model. Other

data that do not contain any abnormal occurrences are classified as "Normal," whereas all previously mentioned strange event types are combined into one category called "Anomaly." The test classifier shows the possibility that uncommon events will be successfully identified as shown in figure 8.






Frame Sequence showing an action	Reality	Predictions	Confidence Score
	YoYo	YoYo	99.50
	TaiChi	TaiChi	99.00
	Walking with Dog	Fighting with Dog	75.00
	Basket Ball	Soccer Juggling	56.00
	Billiards	Billiards	96.05

Figure 7. Predicted Human Activity with Confidence Level

Table 3. Evaluation of the Proposed RCLSTM Model's Performance

Model	Recall	Precision	F1 Score	Accuracy
MobileNetv2-LSTM	85	74	76	88
MobileNetv2-BD-LSTM	80	81	75	83
MobileNetv2-Res-LSTM	89	78	84	91
ConvLSTM	78	73	81	90
VGG19	76	74	87	90
Inceptionv3	80	89	83	89
ResNet50v2	80	79	76	84

LRSTM-Proposed Method	77	88	83	93
-----------------------	----	----	----	----

```
1/1 [=====] - 0s 42ms/step
Action Predicted: HorseRace
Confidence: 0.9937543272972107
Moviepy - Building video __temp__.mp4.
Moviepy - Writing video __temp__.mp4
```

```
Moviepy - Done !
Moviepy - video ready __temp__.mp4
```



Figure 8. Horse Race Prediction with Confidence Level 99%

5. Conclusions

In this study, we carried out the video classification, talked about different approaches, discovered the significance of time frames aspects of the data to improve video classification accuracy, and applied CNN and RNN with enhanced LSTM architectures in TensorFlow to identify human activity in videos by using the temporal and spatial information of the data. We used the OpenCV library to preprocess videos to create an image collection. The RCLSTM model class we developed is flexible enough to be used for a range of vision problems requiring sequential inputs and outputs. Our research demonstrates that learning sequential dynamics using a deep sequence model can beat earlier strategies that just learned a deep hierarchy of visual parameters and techniques that employed a fixed visual representation of the input and only learned the dynamics of the output sequence. Deep sequence modeling methods like RCLSTM are becoming more and more crucial to vision systems for issues with sequential structure as the area of computer vision progresses beyond tasks requiring static input and predictions. Since they integrate readily into current optical identification pipelines and need little to no manually developed features or input preprocessing, these approaches are a promising answer for perceptual instances involving time-varying visual input or sequential outputs. In the future, these limits will be loosened, and different databases would be used to identify activity involving lots of people. Even though it requires a lot of work, we will keep using it in the future with different setups and parameters.

References

1. Clémentin, T. D., Cabrel, T. F. L., & Belise, K. E. (2021). A novel algorithm for extracting frequent gradual patterns. *Machine Learning with Applications*, 5, 100068.
2. Martarelli, N. J., & Nagano, M. S. (2021). How have high-impact scientific studies designing their experiments on mixed data clustering? A systematic map to guide better choices. *Machine Learning with Applications*, 5, 100056.
3. Nikpour, B., Sinodinos, D., & Armanfard, N. (2022). Deep Reinforcement Learning in Human Activity Recognition: A Survey.
4. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625-2634).
5. Park, S. U., Park, J. H., Al-Masni, M. A., Al-Antari, M. A., Uddin, M. Z., & Kim, T. S. (2016). A depth camera-based human activity recognition via deep learning recurrent neural network for health and social care services. *Procedia Computer Science*, 100, 78-84.
6. Vrigkas, M., Nikou, C., & Kakadiaris, I. A. (2015). A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2, 28.
7. WANG, Y., & CHEN, M. (2019). Machine Learning approach to summer precipitation nowcasting over the eastern Alps.
8. Xu, L., Yang, W., Cao, Y., & Li, Q. (2017, July). Human activity recognition based on random forests. In *2017 13th international conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD)* (pp. 548-553). IEEE.
9. Patel, C. I., Garg, S., Zaveri, T., Banerjee, A., & Patel, R. (2018). Human action recognition using fusion of features for unconstrained video sequences. *Computers & Electrical Engineering*, 70, 284-301.
10. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156-3164).
11. Gu, F., Chung, M. H., Chignell, M., Valaee, S., Zhou, B., & Liu, X. (2021). A survey on deep learning for human activity recognition. *ACM Computing Surveys (CSUR)*, 54(8), 1-34.
12. Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2019). Deep learning for sensor-based activity recognition: A survey. *Pattern recognition letters*, 119, 3-11.
13. Bulbul, E., Cetin, A., & Dogru, I. A. (2018, October). Human activity recognition using smartphones. In *2018 2nd international symposium on multidisciplinary studies and innovative technologies (ismsit)* (pp. 1-6). IEEE.
14. Bouchabou, D., Nguyen, S. M., Lohr, C., LeDuc, B., & Kanellos, I. (2021). A survey of human activity recognition in smart homes based on IoT sensors algorithms: Taxonomies, challenges, and opportunities with deep learning. *Sensors*, 21(18), 6037.
15. Gowda, S. N. (2017). Human activity recognition using combinatorial deep belief networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 1-6).
16. Wu, H., Ma, X., Zhang, Z., Wang, H., & Li, Y. (2017). Collecting public RGB-D datasets for human daily activity recognition. *International Journal of Advanced Robotic Systems*, 14(4), 1729881417709079.
17. Wu, D., Sharma, N., & Blumenstein, M. (2017, May). Recent advances in video-based human action recognition using deep learning: A review. In *2017 International Joint Conference on Neural Networks (IJCNN)* (pp. 2865-2872). IEEE.
18. Aldossary, R. S., Almutairi, M. N., & Dursun, S. (2023, March). Personal Protective Equipment Detection Using Computer Vision Techniques. In *SPE Gas & Oil Technology Showcase and Conference* (p. D021S031R001). SPE.
19. Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9, 611-629.
20. Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps: Automation of Decision Making*, 323-350.
21. Sajid, M., Aslam, N., Abid, M. K., & Fuzail, M. (2022). RDED: Recommendation of Diet and Exercise for Diabetes Patients using Restricted Boltzmann Machine.
22. Babiker, M., Khalifa, O. O., Htike, K. K., Hassan, A., & Zaharadeen, M. (2017, November). Automated daily human activity recognition for video surveillance using neural network. In *2017 IEEE 4th international conference on smart instrumentation, measurement and application (ICSIMA)* (pp. 1-5). IEEE.
23. Yang, H., Tian, Q., Zhuang, Q., Li, L., & Liang, Q. (2021). Fast and robust key frame extraction method for gesture video based on high-level feature representation. *Signal, Image and Video Processing*, 15, 617-626.
24. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625-2634).