Machine Learning for Human Intelligence
2023, Vol. 1, Issue 2

Research Article

Activity Recognition Using Deep Learning for Video Surveillance

Aleena Tariq[1], Muhammad Umair Ahmad[2], and Haris Ali[3,⋆]

[1,2,3]Department of Computer Science, National College of Business Administration and Economics, 59030, Multan, Punjab, Pakistan.; Email: aleenatariq34@gmail.com,umairahmad93@gmail.com, harisali337@gmail.com
⋆Corresponding author: Haris Ali (harisali337@gmail.com)

Article History

Academic Editor:
Dr. Muhammad Sajid
Submitted: January 21, 2023
Revised: May 21, 2023
Accepted: September 1, 2023

Keywords:
Activity Detection; LSTM Networks; Video Monitoring; Convolutional Networks

Abstract

Contemporary security frameworks have progressively integrated automated surveillance systems to oversee both public and private settings. Conventional methods that depend solely on human operators for video surveillance are susceptible to errors and inefficiencies, resulting in significant time and resource expenditures. This study explored the capabilities of sequential modeling architectures for time-series analysis, given that deep convolutional frameworks have primarily focused on static image interpretation tasks. The study created extensive, end-to-end, trainable deep architectures featuring task-specific recurrent convolutional structures for visual understanding. We utilized these frameworks, which demonstrated enhanced efficacy in human behavior detection, to develop a specific model for anomaly detection in surveillance footage. The methodology utilized Convolutional Neural Networks for feature extraction from sequential frame inputs. The study established a classification system that distinguishes between normal and abnormal actions, facilitating accurate categorization of identified anomalies. The performance assessment utilizing the UCF50 dataset achieved remarkable accuracy of around 93%. This performance surpassed other methods, including ConLSTM, when assessed on the same datasets.

1 Introduction

Understanding human behavioral patterns is crucial in various fields, including health monitoring, fitness tracking, remote observation, wearable technology, transportation management, targeted marketing, and security applications. Monitoring daily routines facilitates the estimate of caloric expenditure and the formulation of tailored food recommendations [1]. Likewise, recognizing fall-risk tendencies in senior populations might initiate suitable treatments to avert accidents. Traditional machine learning techniques for behavioral identification rely on manual feature engineering and selection. This method is resource-intensive, requires specialist knowledge, and may yield features that do not satisfy performance standards. Recently, deep learning techniques have proven to be advantageous in minimizing the need for manual feature engineering. These algorithms independently acquire pertinent information from data, reducing human intervention and possibly improving performance.

Deep learning architectures, known as deep neural networks, are artificial neural systems with numerous hidden layers. Research literature categorizes deep learning models into three types: supervised learning, unsupervised learning, and hybrid techniques [2]. Recent years have thoroughly

examined recurrent neural networks (RNNs) in perceptual applications, yielding varying degrees of success. Nonetheless, RNNs encounter a considerable constraint referred to as the "vanishing gradient" problem. This issue arises when transmitting faults across prolonged temporal sequences becomes progressively challenging. A category of models was developed to tackle this difficulty, integrating memory-cell-like neural gates [3]. These models facilitate the preservation, modification, or resetting of state flow by integrating hidden states with nonlinear dynamics. Although these models have shown competence in numerous tasks, their practical significance was recently underscored in studies focused on the prolonged training of voice recognition and language translation models. This highlights the capability of these models to capture long-range connections and attain enhanced performance in intricate tasks, as illustrated in Figure 1.

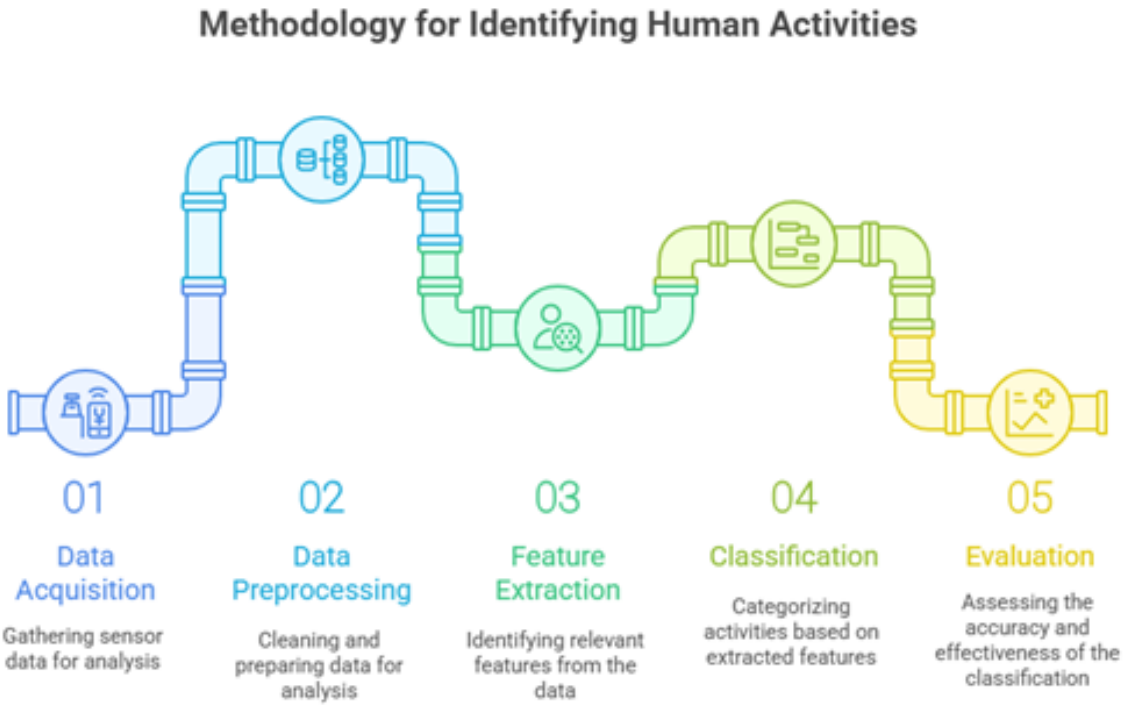

Figure 1: Method for Identifying Human Activity.

This investigation explores the effectiveness of modeling long-term recurring convolutional networks for visual time-series data. We propose that these long-term recurrent neural networks can outperform static or fixed temporal models in visual tasks, particularly when sufficient training data is available for model development and learning. This study demonstrates that LSTM-based architecture models provide a novel, end-to-end approach for mapping pixel-level visual data to natural language descriptions at the sentence level [4]. These models not only improve detection of traditional video activity problems but also enhance description generation from visual examples, bridging the gap between traditional graphical models and language understanding. This study implements the design in three test environments to validate the proposed approach. Initially, we apply it to instructional video identification systems incorporating complex temporal relationships. By directly integrating visual convolutional deep LSTM networks, this study observes improvements of approximately 4% on standard benchmark datasets. This improvement is significant, even though labeled video activity datasets may not exhibit highly complex temporal dynamics regarding 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. 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 to tackle this problem have encountered overfitting issues, necessitating a reduction in learning parameter size [5]

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 findings on the proposed dataset confirm the superiority of data-driven learning methodologies over state-of-the-art handcrafted features. Utilizing larger datasets enables us to fully exploit data-driven methodologies and advance the limits of human activity recognition [6]. The principal aim of this research is to investigate and enhance methodologies for establishing a foundation for action recognition utilizing video data. Although deep learning has been employed by major corporations to ascertain client preferences and improve product offers, there is currently no framework or methodology specifically developed for identifying actions in commercial settings through video data. This work proposes to concentrate on activity recognition to develop an automated security system that is readily available to individuals in need [7, 8].

The research aims to establish a framework for activity recognition that employs a model proficient in extracting activity data and assessing habitual behaviors through diverse component properties. Deep learning methodologies will be deployed for activity recognition using these features, and the UCF50 dataset will be employed to alleviate the financial strain on human activity recognition systems and minimize losses resulting from security breaches. The research seeks to provide precise and automated activity recognition through the development of this framework and the application of deep learning methodologies. This framework will enhance security systems by offering a more efficient and effective method for identifying and addressing possible security concerns.

2 Related Work

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 [9, 10]. 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. With the advancements in the Internet of Things (IoT), the concept of smart environments is gaining significant attention as it offers a range of benefits, such as healthcare monitoring, assistance with daily tasks, energy management, and enhanced safety. Smart environments are equipped with multiple sensors and actuators that enable the monitoring of various parameters like door openings, room lighting levels, temperature, humidity, and more [11].

Furthermore, they allow users to 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 [12]. Examples of such approaches include various models and sparse coding approaches, as well as Bag of Words methods. 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 systems. Among these systems, computer vision technologies utilized in surveillance systems stand out for reducing the need for manual monitoring and enhancing people's safety, such as in community security and crime prevention [13]. This research leverages a deep learning network incorporating RNN and LSTM 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 safety applications.

This study utilizes the LSTM 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 [14, 15]. The researchers drew inspiration from previous studies 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. Additionally, score fusion is performed between CNN and LSTM channels [16]. This fusion technique demonstrates superior performance compared to combining multiple LSTM channels, thanks to the complementary nature of the CNN and LSTM models. The proposed approach is evaluated on standard datasets, yielding cutting-edge results.

3 Materials and Methods

Depending on the specifics of the task at hand and the dataset's characteristics, preprocessing techniques can vary. It is essential to carefully consider which preprocessing techniques are most appropriate for a particular video classification task. Preprocessing is an important step in video classification since it improves the classifier's performance by preparing the data for analysis. For video classification, common preprocessing techniques are applied [17]. During preprocessing, video files are read from the dataset. Video frames are scaled to specific dimensions to facilitate computations. Additionally, the data is normalized to lie between 0 and 1. To accelerate convergence during model training, pixel values are divided by 255. We scale the frames to 64 $\times$ 64 to improve accuracy. To achieve better results, we increase the frame size to 64 $\times$ 64, though this raises computational costs. The sequence length is provided to the LSTM, specifying how many video frames are given to the model in a particular sequence. The more sequences there are, the larger the network and the longer it will take to train [18]. Convolutional Neural Networks 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 [19]. 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. The study investigates different methods, such as utilizing recovered region data as input to locate and examine the behavior of objects within the video as shown in Figure 2.

3.1 ConLSTM Architecture

The initial strategy is implemented using a combination of ConLSTM cells. Convolutional processes are incorporated into LSTM network versions called ConLSTM cells [20]. This LSTM does not work with 1D data but only 3D data. It is built on complex processes [21]. Recurrent layers from the Keras ConLSTM2D framework are used to build the model. The ConLSTM2D layer additionally considers the kernel's size and the number of filters required to apply convolutional operations [22]. 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 employ MaxPooling3D layers to shrink the size of the frames and eliminate unnecessary computations, dropout layers to prevent the model from overfitting the data as shown in Figure 3. 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 progresses 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 feature maps. The network has more features since there are more filters. It gives the network wider scope and allows for more precise training [23, 24, 25]. 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 reduce the scope of the feature maps and increase the precision of the predictions, max pooling is added repeatedly with increasing filters.

3.2 Recurrent Convolutional LSTM Network

LSTMs were developed to describe temporal sequences. The method is applied following data preprocessing, which removes unwanted, missing, and null signal values [26]. The LSTM provides a solution

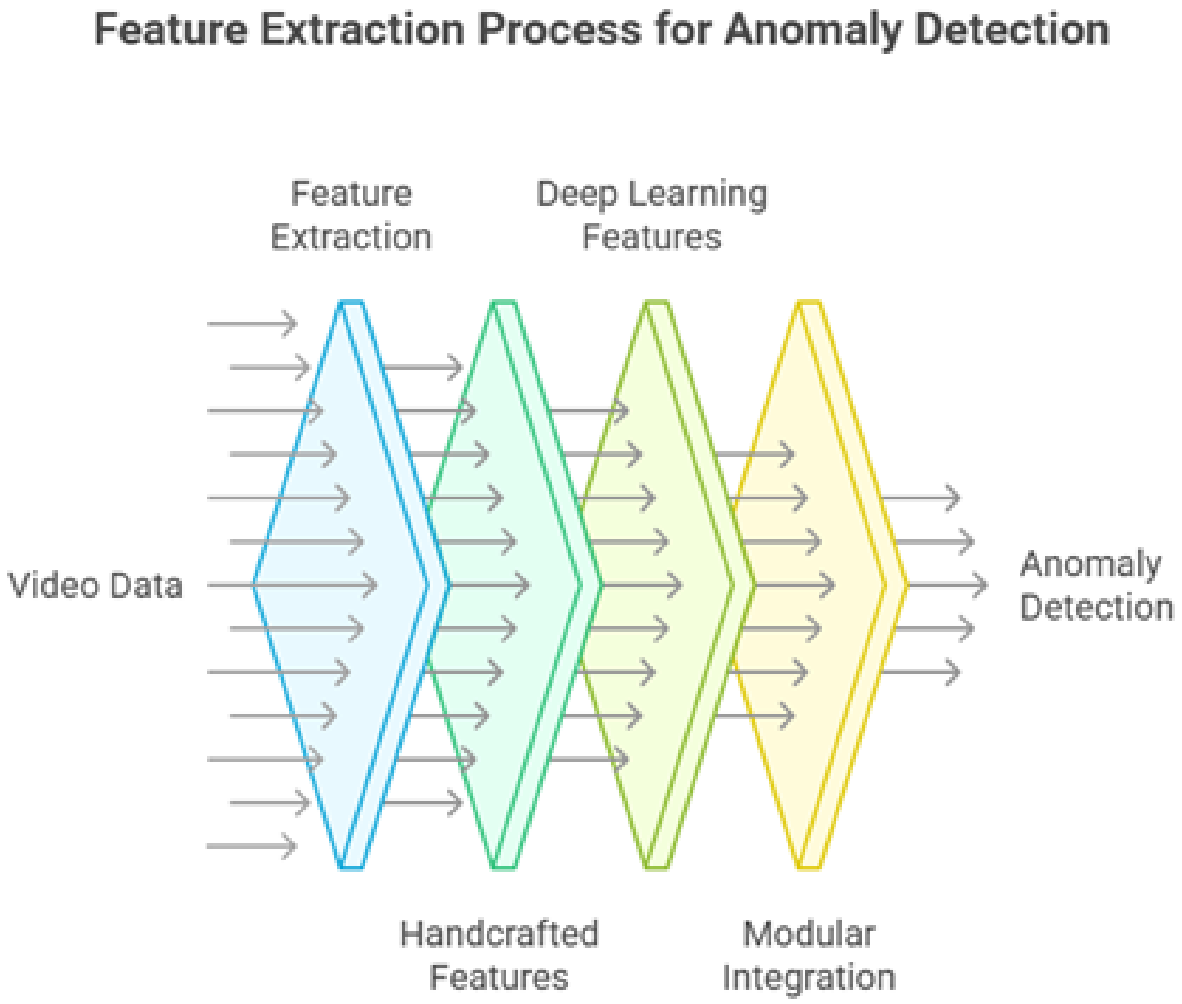

Figure 2: Feature Extraction framework.

by including a memory cell to reliably encode knowledge at each stage. The memory cell is controlled by an input gate, forget gate, and output gate. The input that is read during categorization is monitored by these gates [27], as given in Equation 1.

$$z_m^k = f\left(\sum_{vN_k} X c_m^k + b_m^k\right) \tag{1}$$

Since we're working with video, we focus on many-to-one LSTM networks because we want to transmit many frames through the network before receiving an action prediction. As discussed, LSTM works best with data sequences, whereas CNN is excellent for image classification [28]. Although we've discussed various methods for classifying images and identifying actions, none of them could provide accurate predictions alone as can be calculated using Equations 2 - 7. Convolutional networks extract frames from videos, and the LSTM network uses this output to perform action recognition as shown in Figure 4.

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

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

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

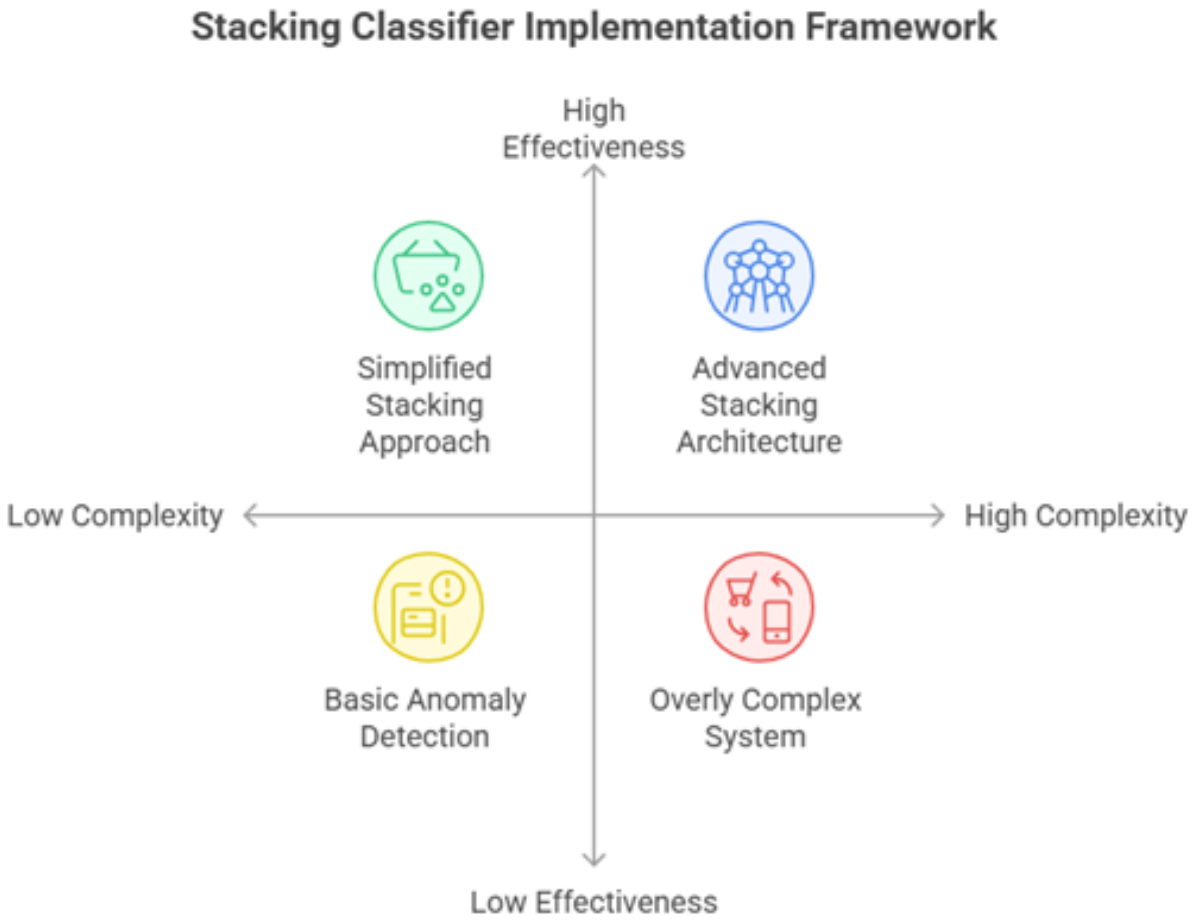

Figure 3: Structure for Stacking Classifiers.

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

$$C_t = m_t \odot C_{t-1} + r_t \odot y_t \tag{6}$$

$$h_t = n_t \odot \tanh C_t \tag{7}$$

The convolutional neural network learns spatial information, while the LSTM learns temporal information [29]. 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 problem can extract spatial information from the video's frames [30]. A robust model is produced because the system acquires spatiotemporal properties during end-to-end training. We work with a TimeDistributed wrapper layer, which enables us to apply the same layer to each video 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) as shown in Figure 5. This is especially useful because it allows feeding the entire video into the model in a single operation [31].

4 Experimental Findings and Analysis

Based on the results, the RCLSTM model appears to have performed exceptionally well for a small number of classes. As a result, we tested 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 [32]. The UCF50, which contains numerous unusual, illegal, and hostile

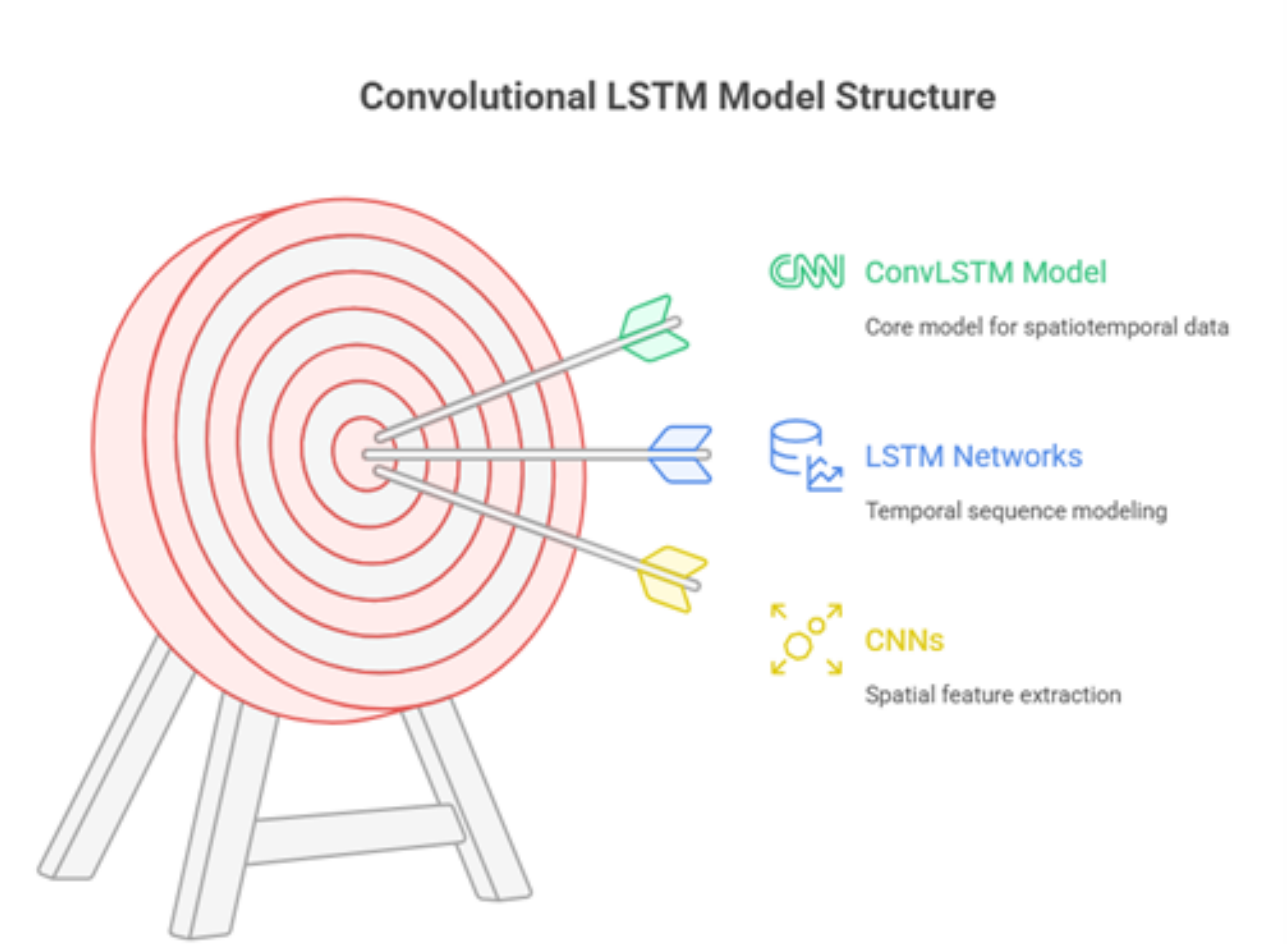

Figure 4: ConLSTM Model.

behaviors captured on video in public spaces, including streets, stores, and schools, is used in this research to implement the recommended model. The fact that this dataset was created from actual events that could happen anywhere at any moment was a deciding factor. These unusual behaviors can also have serious 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 background and surroundings. UCF50 is a diverse collection of human behaviors. It consists of fifty action classes, with videos from each class organized into separate groups that have some shared characteristics. The realistic UCF50 Action recognition dataset was utilized as given in Table 1. We evaluate the suggested

Table 1: Dataset Specifications

Dataset Characteristics	**UCF50 Values**
Mean Videos per Action Class	100
Mean Frame Count per Video	199
Mean Video Frame Height	320
Mean Video Frame Width	240
Mean Frame Count per Video	240
Mean Frame Rate per Video	26

system's action recognition performance technique on various example videos as given in Table 2. The table compares the results of our testing regarding the initial weight and optimizer types and whether data augmentation was used [13]. As a result, we expanded our dataset and ran tests using Adam as

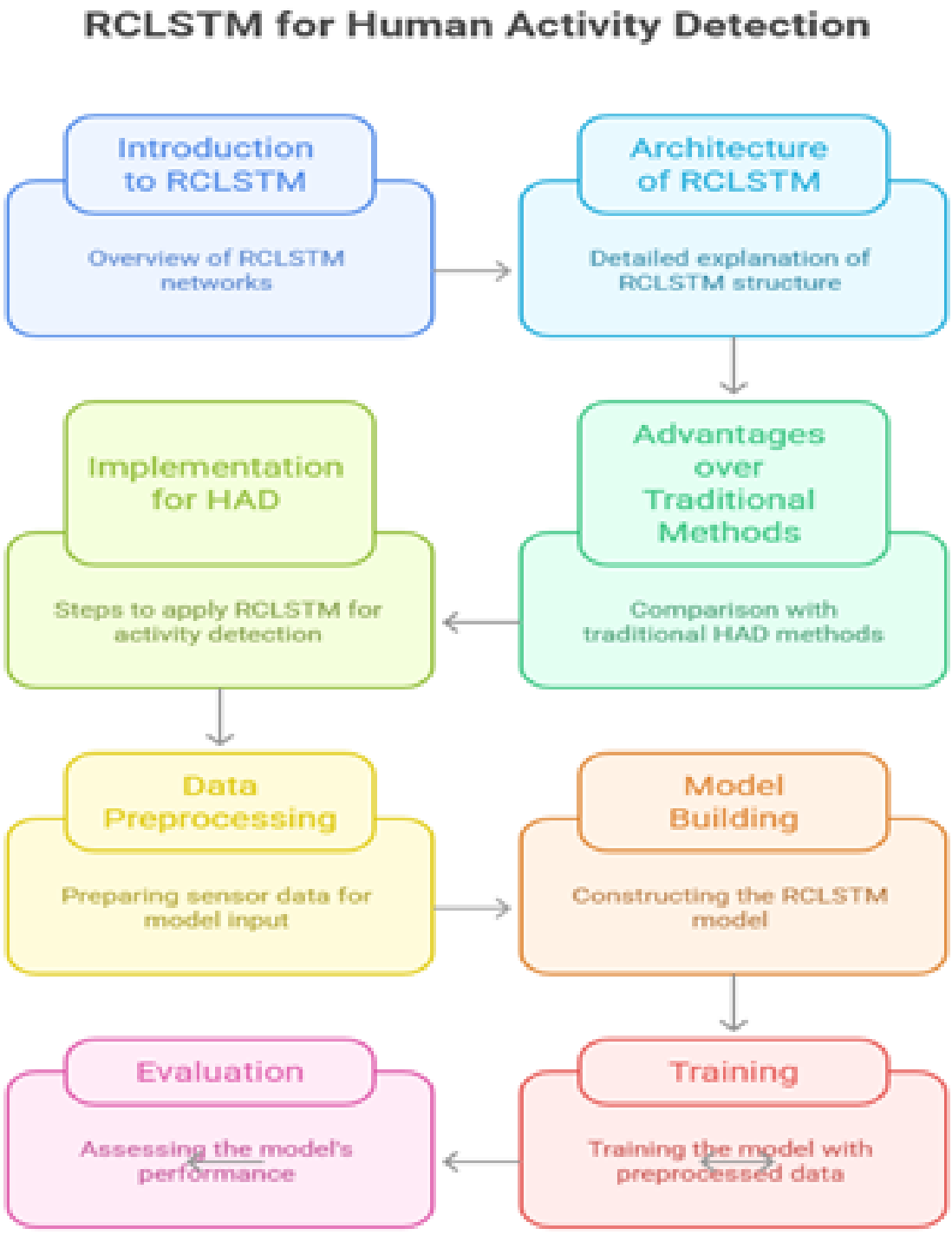

Figure 5: ConLSTM Model.

an optimizer. The model's learning rate is additionally set to 10^{-4}. The code terminates when the loss function converges since continuing would be nonsensical. Likewise, the epochs are set to 20. If the difference between the loss functions of two succeeding epochs is less than the tolerance value, the method ends, and the result is absolute correctness [23]. We set this sequence length to 20 during our tests. We evaluate the accuracy and compare our proposed method with a 3D convolutional network

Table 2: Experimental Configuration - RCLSTM & ConvLSTM Settings

Configuration Parameters	**RCLSTM Values**	**ConvLSTM Values**
Filter Count	16,32,64	4,8,14,16
Kernel Dimensions	3x3	3x3
Sequence Duration	20	20
Class Count	6	6
Image Dimensions	64x64	64x64
Activation Method	Relu	Tanh
Pooling Type	2D	3D
Recurrent Dropout Rate	0.25	0.20
Optimization Algorithm	Adam	Adam
Training Epochs	70	50

[19]. The evaluation of ConvLSTM revealed an increase in computation time as shown in Table 3. Now, we generated a forecast of human action identification using our proposed RCLSTM model and comparison with other models. We obtained the findings after using the RCLSTM model. Other data that do not contain any abnormal occurrences are classified as "Normal," whereas all previously mentioned unusual event types are combined into one category called "Anomaly." The test classifier shows the possibility that uncommon events will be successfully identified [16].

Table 3: Performance Assessment of the Proposed RCLSTM Model

Model Architecture	**Recall Rate**	**Precision Rate**	**F1 Measure**	**Accuracy Rate**
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
RCLSTM Approach	77	88	83	93

5 Conclusion

In this investigation, we conducted video classification, discussed different approaches, discovered the significance of temporal 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. Our research indicates that utilizing a deep sequence model to learn sequential dynamics can outperform previous methods that solely focused on a deep hierarchy of visual parameters and those that employed a static visual representation of the input while only learning the dynamics of the output sequence. Advanced sequence modeling techniques, such as RCLSTM, are becoming increasingly essential for vision systems addressing problems with sequential structures as the field of computer vision evolves beyond static input and prediction tasks. These methods seamlessly integrate into existing optical identification systems and necessitate no human feature development or input preprocessing, making them a potential alternative for perceptual situations requiring dynamic visual input or sequential outputs. We will resolve these constraints in the future and utilize various databases to identify activities involving multiple individuals. Even though it requires some effort, we will continue to use it in the future with a variety of setups and characteristics.

Supplementary Materials

All relevant data is within the manuscript and its supporting information files.

Funding

This research received no external funding.

Data Availability Statement

Data sharing does not apply to this article as no new data were created or analyzed in this study.

Acknowledgments

We acknowledge that we did not receive any support or funding.

Conflicts of Interest

The authors declare no conflict of interest.

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] Aboah, A. (2021). A vision-based system for traffic anomaly detection using deep learning and decision trees. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4207-4212).

[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).

[25] Patrikar, D. R., & Parate, M. R. (2022). Anomaly detection using edge computing in video surveillance system. *International Journal of Multimedia Information Retrieval*, *11*(2), 85-110.

[26] Wang, G., Wang, Y., Qin, J., Zhang, D., Bao, X., & Huang, D. (2022, October). Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *European Conference on Computer Vision* (pp. 494-511). Cham: Springer Nature Switzerland.

[27] Wang, G., Wang, Y., Qin, J., Zhang, D., Bao, X., & Huang, D. (2022, October). Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *European Conference on Computer Vision* (pp. 494-511). Cham: Springer Nature Switzerland.

[28] Haji, S. H., & Ameen, S. Y. (2021). Attack and anomaly detection in IoT networks using machine learning techniques: A review. *Asian Journal of Research in Computer Science*, *9*(2), 30-46.

[29] Hooshmand, M. K., & Hosahalli, D. (2022). Network anomaly detection using deep learning techniques. *CAAI Transactions on Intelligence Technology*, *7*(2), 228-243.

[30] Wyatt, J., Leach, A., Schmon, S. M., & Willcocks, C. G. (2022). AnoDDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 650-656).

[31] Pirnay, J., & Chai, K. (2022, May). Inpainting transformer for anomaly detection. In *International Conference on Image Analysis and Processing* (pp. 394-406). Cham: Springer International Publishing.

[32] Di Biase, G., Blum, H., Siegwart, R., & Cadena, C. (2021). Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16918-16927).