Analysis of Deep Learning Approaches for Video Classification of Human Activities: A Comparison of ConvLSTM and LRCN Architectures

Mužinić Mia
University of Split
Faculty of Science
Split, Croatia
mmuzinic@pmfst.hr

Abstract

In recent years, deep learning techniques have shown remarkable success in video classification tasks. The research focuses on evaluating and comparing different state-of-the-art deep learning architectures, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), in terms of their performance and accuracy in classifying human activities from video data. The aim of this research is to compare metrics of ConvLSTM and LRCN deep learning architectures for video classification of human activities on the UCF50 dataset. ConvLSTM combines convolutional and LSTM layers, capturing spatial and temporal dependencies, while LRCN integrates a pre-trained CNN with an LSTM. Performance evaluation using accuracy, precision, recall, and F1 score shows competitive results for both architectures. ConvLSTM excels in capturing temporal dependencies, while LRCN demonstrates better spatial feature extraction. Computational analysis reveals that ConvLSTM has higher computational requirements due to increased parameters and operations. These findings provide insights for selecting appropriate architectures based on specific video classification requirements and computational constraints. In the specific context mentioned, ConvLSTM demonstrated an accuracy of 77.05%, and the LRCN model achieved an accuracy of 87.70% in accurately identifying the activity being performed.

Keywords: video classification, human activity recognition, deep learning, convolutional neural networks, recurrent neural networks

1 Introduction

In recent years, computer vision has witnessed remarkable advancements, enabling machines to interpret and understand visual information in a way that closely resembles human perception [12]. In the field of computer vision, video classification and human activity recognition (HAR) pose fascinating and challenging problems that have attracted significant research attention. With the proliferation of video data from diverse sources such as surveillance cameras, wearable devices, and social media platforms, there is a growing need to develop automated systems capable of understanding and interpreting human actions and activities in videos.

This research provides an overview of the major contributions made by some of the famous researchers in this area, such as Fei-Fei Li, and Karpathy Andrej et al. [20], Jitendra Malik, and Andrew Zisserman et al. [35]. Other notable researchers in computer vision area include Kaiming He [11, 15], and Jorge Luis Reyes Ortiz [4, 3] who distinguished himself with his work in the field of human activity recognition. Also, key challenges and open research questions in video classification and human activity recognition are highlighted, including the need for more robust and accurate models, the development of multimodal datasets, and ethical considerations related to privacy and security.

Video classification refers to the task of categorizing videos into predefined classes or labels, based on the content and context of the visual information presented. This enables machines to identify and understand the activities and events captured in the video stream. On the other hand, HAR focuses specifically on recognizing and analyzing human actions or activities within videos, aiming to infer the type of activity being performed and potentially infer higher-level contextual information. The problem of video classification poses several challenges due to the inherent complexity of videos. Unlike static images, videos are temporal sequences of frames, capturing dynamic and evolving scenes over time. This temporal dimension introduces additional intricacies, such as motion, temporal dependencies, and temporal context, that need to be effectively modeled and analyzed. Moreover, videos often contain various spatial and tem-

poral variations, making it challenging to develop robust algorithms. Spatial variations refer to differences in appearance, scale, viewpoint, and object occlusions, while temporal variations arise from variations in speed, duration, and timing of activities. Additionally, videos can contain multiple activities or events occurring simultaneously or sequentially, requiring the identification and understanding of complex interactions between objects and actors. [47, 26, 30, 36]

To tackle these challenges, researchers have explored a variety of approaches, ranging from traditional machine learning techniques to more recent deep learning methods. Deep learning, in particular, has revolutionized the field by leveraging convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to effectively capture spatial and temporal information from video sequences, leading to significant advancements in video classification and HAR [14, 52]. Due to the complexity of behaviors and the absence of contextual information, existing methodologies that use a random capture from a video or picture to classify human activities for HAR might result in incorrect classifications. To increase accuracy, deep learning models must be developed utilizing video data to extract time and space-related information. This work aims to solve this problem by developing hybrid models, ConvLSTM and LRCN for reliable HAR using video datasets. Previously mentioned two architectures were chosen due to their popularity and frequent use in activity recognition research. This analysis will aid researchers and practitioners in making informed decisions when selecting the architecture that best suits their specific needs and requirements. However, despite the progress made, there are still open research questions and practical challenges in video classification and HAR. Some important aspects to consider in this context are the ability of a system to perform well even when there are changes in lighting conditions or the viewpoint, dealing with situations where certain parts of an object are blocked or hidden, managing and working with large amounts of data, finding ways to tackle the requirement of having labeled training data, and creating algorithms that can process information quickly and efficiently in real-time.

In the framework of video classification - human activity recognition, the following research questions are the focus of this researching:

1. What are the most effective deep learning architectures for video classification of human activities in the latest state-of-the-art?

2. How can different types of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) be optimally combined to improve video classification results?

3. How do the ConvLSTM and LRCN architectures compare in terms of accuracy and performance in classifying human activities in videos?

Based on an analysis of the existing literature and knowledge gaps, these study issues were chosen. Finally, this paper assesses the current state of the art of human activity recognition within video classification. Taking into account the practical recommendations and insights gathered from previously covered articles on this topic, the research aims to offer valuable insights for effective recognition and categorization of human activities in videos. Furthermore, it provides opportunities about potential future advancements and prospects in this field of study.

2 Literature review

In order to collect relevant articles an essential first step in performing a comprehensive literature review is choosing relevant studies. It entails a meticulous and thorough procedure of finding pertinent papers that address the research concerns. To guarantee that the chosen studies are of the highest caliber and closely relevant to the study goals, the inclusion and exclusion criteria for the paper selection process must be carefully considered. In order to identify significant contributions published in prestigious journals and conference proceedings, a thorough literature search was conducted using the Science Citation Index Expanded (SCIE) and the Social Sciences Citation Index (SSCI) via the Web of Science service and the IEEE Xplore database. The search was focused on relevant fields such as title, abstract, and keywords of publications indexed in the Web of Science Core Collection and IEEE Xplore. A special set of search keywords was used, combining the three fragments with the AND operator. The first fragment contained keywords related to the human activity recognition. If these terms were not found in the papers' topics, the term motion was included in the search term, so the term (Human Activity Recognition* OR Human Actions Recognition* OR Human Motion Recognition* OR Human Estimation* OR Activity Recognition*) was used as the first fragment of the search phrase. To cover the video classification, the second fragment of the search phrase was composed as follows: (Video Classification* OR Video-based* OR Video*). To avoid all papers that do not contain deep learning methods, a third fragment (Deep Learning* OR Neural Networks*) is included. Finally, the unique query was established in which the first three fragments were formed as stated at the beginning of this subsection and joined with an AND operator: (Human Activity Recognition* OR Human Actions Recognition* OR Human Motion Recognition* OR Human Estimation* OR Activity Recognition*) AND (Video Classification* OR Video-based* OR Video*) AND (Deep Learning* OR Neural Networks* OR Deep Neural Network*). The list of total 2,606 results from Web of Science Core Collection and 2,176 results from IEEE Xplore obtained after exclusion

was browsed to inspect the titles of the papers and eliminate the items that do not belong in the scope of this research. The search was limited to articles, proceeding papers, and review articles written in English, resulting in 2,591 remaining articles from WoS and conferences, journals early access, and articles resulting in 2,158 from IEEE Xplore, but with open access only left us 158 articles. After checking for duplicates, 346 articles were excluded, which left 2,654 research papers in total. Studies that did not investigate HAR approaches were excluded from further screening, which left 1,058 articles. Also, reading the titles left 436 articles. After checking the keywords in abstracts of the remaining publications left 60 most relevant articles. These articles reported on the performance of various deep learning architectures for HAR in video classification using different datasets and evaluation metrics. Also, relevant articles were searched on Google Scholar, the internet search engine for scientific literature, which is today the first online indicator of citations. I searched by keywords as (Computer Vision, Human Activity Recognition, Video Classification, Deep Learning Approaches, CNNs, RNNs) and selected articles based on the title, abstract and the number of citations.

3 Selected research papers

This section reviews the most related works for video classification and human activity recognition tasks.

The earliest studies on human activity recognition (HAR) were published in the late 1990s. [13, 2]. Turaga et al. (2008) suggested a subtle distinction between "action" and "activity" connotations, where the activity recognition methods were categorized according to their degree of activity complexity [40]. In the literature, a number of additional methods and taxonomies for HAR have been put forth [46]. One such technique is the use of handcrafted features [8], whereby features from video frames are extracted for activity recognition, such as histograms of oriented gradients (HOG) [22] and motion history images (MHI) [51], and fed into machine learning classifiers like support vector machines (SVMs) [3, 28, 33] or random forests (RFs) [5].

Although manually created feature-based approaches, such as sensor-based activity recognition [48, 7], are simple to understand and use, they have drawbacks like high computational costs, a lack of robustness to changes in lighting, and poor accuracy. First of all, most of the commonly used feature extractors are developed based on a specific dataset, and the feature extractors are often database-biased. They do not have general-purpose feature extraction abilities. And secondly, creating a handcrafted feature-based human activity recognition system required careful feature engineering. Thus, handcrafted feature engineering works are labor-intensive and time-consuming, severely hindering the development of related technologies. Nevertheless, even if their extraction process is fully automated, these so-called "handcrafted" features are especially designed to be optimal for a specific task. Thus, despite their high performances, these approaches main drawback is that they are highly problem-dependent. In the last few years, there has been a growing interest in approaches, so-called deep models, that can learn multiple layers of feature hierarchies and automatically build high-level representations of the raw input [38, 14, 52]. They are therefore more generic since the feature building process is fully automated. Therefore, this research is based on deep learning architectures.

In a study by Zhu et al. (2016) [55], both handcrafted and learning-based approaches for action recognition were examined. The authors first evaluated the limitations of the handcrafted methods and then briefly discussed the emergence of deep learning techniques in HAR, till 2016. Another survey conducted by Herath, Harandi, and Porikli et al. (2017) [17] explored similar research areas. They began with the pioneer of human activity recognition (HAR) techniques, which was a handcrafted-feature-based approach that eventually evolved into deep learning-based methods. Unlike previous surveys, this study comprehensively presented deep learning methods, aligning them with HAR datasets, which was a crucial aspect missing in earlier surveys. However, it should be noted that this survey also only included literature up until 2016. In another survey conducted by Koohzadi and Charkari et al. (2017) [24], the role of deep learning in image and video processing for HAR was investigated. The overall approach was categorized into five types of models: supervised-deep generative, supervised-deep discriminative, unsupervised deep, semi-supervised deep, and hybrid models. One notable aspect highlighted in this survey was the discussion of the benefits, tips, and tricks for choosing a deep learning model specifically for HAR within the aforementioned five categories. The authors also explored deep learning approaches to spatio-temporal representation, which involved incorporating time as the third dimension in traditional 2D image processing. Reining et al. (2019) [32] conducted a thorough review of the existing literature on human activity recognition (HAR) in the fields of production and logistics. They provided a comprehensive overview of the latest HAR methods, including statistical pattern recognition and deep learning architectures. A summary of the deep learning video classification approach is visually presented in the article by Rehman et al. (2021) [31]. Also, Ullah. H et al. (2021) [43] uses existing research on video-based human activity recognition to review state-of-the-art deep learning architectures from 2015 to 2020 in terms of different methods, challenges, and problems. Authors explored various deep learning techniques available for HAR, challenging researchers to build robust models and using state-of-the-

art datasets used for evaluation. The research consists of a description of the review process, definitions of the research questions, and a report of the results for each question, with included recent deep neural network architecture approaches and an overview of some recent research. This review also had some limitations, including a lack of analysis of the latest state-of-the-art approaches.

Given that the goal of this research is to compare the metrics of different deep learning architectures, specifically ConvLSTM and LRCN architectures, below are presented relevant articles in which these architectures are explained in more detail with the already existing results achieved by these architectures on different datasets. The Convolutional LSTM (ConvLSTM) network was introduced in the work by Shi et al. (2015) [34], and Luo et al. (2017) [27]. ConvLSTM incorporates convolutional operations in its recurrent layers [45, 53, 54]. ConvLSTM allows for feature extraction and temporal modeling to be performed simultaneously, which makes it an effective architecture for HAR in video classification. Several studies have reported high accuracy rates using ConvLSTM for HAR in video classification, such as the work by Yue-Hei et al. (2015) [54]. A hybrid design that combines CNNs and RNNs is called LRCN, on the other hand. In LRCN, the RNN is used to model the temporal connections between video frames while the CNN is used to extract spatial characteristics from video frames. Numerous studies have demonstrated the efficacy of LRCN in HAR for video classification, reporting high accuracy rates. As an example of one of them, Donahue et al. (2015) [10] suggested end-to-end trainable class of architectures for visual recognition and description, and used LRCN to obtain 83.5% accuracy on the UCF101 dataset. ConvLSTM and LRCN differ mostly on how their architecture is designed, as it is mentioned earlier. While LRCN combines CNNs and RNNs, ConvLSTM enables the integration of convolutional operations into RNNs. Their accuracy, efficiency, and generalizability are all impacted by this architectural difference. For example, Donahue et al. (2015) found that while the ConvLSTM variation utilized in their research doesn't have long-term memory, LRCN is able to capture long-term dependencies in the video sequence utilizing recurrent connections. [10]

A new term, transformers have recently become the subject of study into video classification area [29]. Transformers can analyze the video frames concurrently, which is computationally more efficient and can capture long-term dependencies in the video, in contrast to typical deep learning algorithms. Treating the video frames as a series of tokens and then applying the transformer architecture to this series is one method of using transformers for video classification. A CNN is used to first encode each video frame into a feature vector, and then the transformer architecture treats these feature vectors as tokens. For a variety of video classification tasks, such as action identification and human position estimate, this method has been proven to be successful. However, more research is needed to realize the full potential of transformers for video classification tasks.

In addition to architectural approaches, other factors, such as the size and quality of the dataset, the pre-processing techniques used, and the choice of optimisation algorithm, also impact the performance of HAR in video classification. For instance, importance of having a large and diverse dataset for training deep neural networks is reported by Simonyan and Zisserman (2014) [35]. Furthermore, the availability of large-scale annotated datasets such as UCF101 [37], HMDB51 [25], ActivityNet-200 [16] and Kinetics [21] has facilitated the development and evaluation of HAR in video classification methods [6]. These datasets contain thousands of video clips that span different human activities and are commonly used to benchmark the performance of HAR in video classification methods. In addition to the accuracy of HAR in video classification, other factors such as model interpretability, computational complexity, and energy efficiency are becoming increasingly important considerations in practical applications. For instance, real-time activity recognition in video streams requires low-latency and energy-efficient models, which may require sacrificing some accuracy. Recent research has proposed lightweight deep learning models and compression techniques that aim to strike a balance between accuracy and efficiency, such as the work by Agarwal et al. (2020) [1] that proposed a lightweight deep learning model for HAR in video classification.

4 Used Deep Learning Methods

In this section, a more detailed overview of the architectures used in this work is given, as well as an overview of recent literature with experimental results.

Convolutional Neural Networks (CNNs) are a type of deep learning model widely used in computer vision tasks. They excel at processing visual data, such as images or videos. CNNs employ a mathematical operation called convolution, which examines small portions of an input image at a time, allowing them to capture local patterns and features effectively. Unlike traditional fully connected networks, CNNs have shared weights and hierarchical architectures that enable them to extract hierarchical representations from the input. Through a series of convolutional layers, pooling layers, and activation functions, CNNs learn to recognize complex patterns and objects, making them powerful tools for tasks like image classification, object detection, and image segmentation.
Recurrent Neural Networks (RNNs) are a type of neural network designed to process sequential data by considering the temporal dependencies among inputs. Unlike feedforward networks, RNNs have a feedback connection that allows them to maintain an internal state,

enabling information to persist over time. This architecture enables RNNs to handle variable-length sequences and capture long-range dependencies. The key component of an RNN is the recurrent layer, which processes each input while incorporating information from previous inputs through recurrent connections. This allows RNNs to model sequential patterns and make predictions based on the context of previous inputs. RNNs find applications in various domains, including natural language processing, speech recognition, and time series analysis. To summerise, Recurrent neural networks (RNNs) are used to describe the temporal structure of video sequences, whereas convolutional neural networks (CNNs) are frequently employed for feature extraction from video frames [41, 20, 49, 35, 39, 19, 39] **Long Short-Term Memory (LSTM)** is a type of recurrent neural network (RNN) that specializes in processing sequential data. Unlike regular RNNs, LSTMs have a unique memory cell that allows them to capture and retain information over longer sequences. They are like smart memory banks that learn to remember important things while filtering out irrelevant details. LSTMs address a common problem in regular RNNs called the vanishing gradient, which makes it difficult for the network to remember long-term dependencies. By using special gates, LSTMs selectively store, update, and retrieve information, making them highly effective in tasks involving time series data, language processing, and anything where context and long-range dependencies matter. [41, 49, 42]

Given that the previous section Related Works presented related works detailing various deep learning approaches used in human activity recognition video classification, as well as ConvLSTM and LRCN architectures, such as Ullah et al. (2021) [43] which includes relevant articles and techniques up to 2020, I decided for modifying the search strategy with the same queries, in which I added the fourth fragment (CNN-LSTM OR ConvLSTM OR LRCN) and the condition that the articles are from 2020 to 2023, which left 32 articles from WoS and IEEE Xplore databases. After checking the titles and abstracts, 6 most relevant papers are left.

4.1 ConvLSTM architecture

ConvLSTM (Convolutional Long Short-Term Memory) is a powerful combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. It combines the spatial feature extraction capabilities of CNNs with the ability of LSTMs to capture temporal dependencies in sequential data. ConvLSTM models excel in tasks that involve both spatial and temporal information, such as video analysis. In ConvLSTM, the input goes through convolutional layers, similar to CNNs, which extract spatial features at each time step. These features are then fed into LSTM cells, allowing the model to capture and learn temporal correlations between consecutive frames. This fusion of convolutional and LSTM layers enables ConvLSTM to effectively analyze video data and perform tasks like action recognition and video prediction.

Convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and fully connected layers are sequentially combined to form the proposed ConvLSTM network. Pre-calculating skeleton coordinates from the image or video sequence using human detection and posture estimation is done by the acquisition system, shown in Figure 1. The innovative guided features are built using the raw skeleton coordinates and their distinctive geometrical and kinematic properties in the ConvLSTM model [50]. Authors in this article presented experimental results with the accuracy of the proposed ConvLSTM is 98.89% on the KinectHAR dataset, which is higher than the accuracy of CNNs and LSTMs, which are 93.89% and 92.75%, respectively. The suggested method has been put through real-time testing, and it has been discovered to be unaffected by poses, camera facing, people, clothes, etc.

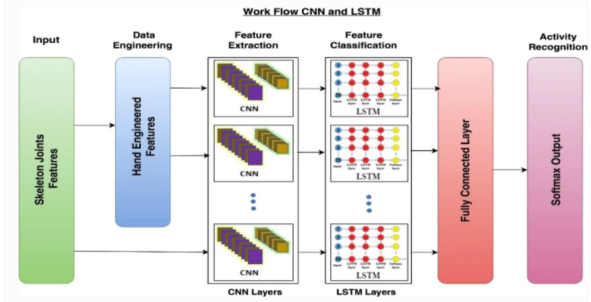

Figure 1: Skeleton-based human activity recognition using ConvLSTM and guided feature learning, Yadav, Santosh Kumar and Tiwari, Kamlesh and Pandey, Hari Mohan and Akbar, Shaik Ali, 2022. [50]

Uzzaman et al. (2022) [44] presented results of ConvLSTM model trained on UCF50 and HMBD51 datasets, with various of combinations activation functions and same dropout, where accuracy varies from 0.5687% to 0.8197%.

Khater et al. (2022) [23] compared two ConvLSTM architectures on several datasets, for instance KTH, and Weizmann dataset with different human activities and obtained accuracy results from 47.66% recognizing UCF Sports Action to recognizing jogging, running, and walking activities, and 98.9% for recognizing boxing, clapping, and waving activities. As shown, performance decreases, compared to boxing, clapping, and waving, with activities jogging and running in UCF Sports, both activities are confused with each other and with walking because those activities have similar features.

4.2 LRCN architesture

LRCN (Long-term Recurrent Convolutional Networks) is a hybrid architecture that combines Convolutional

Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. It aims to leverage both spatial and temporal information for tasks like video understanding. LRCN starts by using CNNs to extract spatial features from individual frames of a video. These features are then fed into LSTM layers, which capture temporal dependencies across frames and model the overall video context, as it is shown in Figure 2. By combining the strengths of CNNs in visual feature extraction and LSTMs in capturing sequential patterns, LRCN achieves robust video analysis, enabling tasks such as action recognition, video captioning, and video classification. [44, 18, 9]

Uzzaman et al. (2022) [44] presented results of LRCN model trained on UCF50 and HMBD51 datasets, with various of combinations activation functions and dropouts, where accuracy varies from 0.1967% to 0.9344%.

LRCN is an RNN-based network that was developed to investigate whether it works effectively on data of the RNN family as sequence input data (video, streaming data). Each frame in input data is sequentially applied to the CNN to extract features, and then applied to the LSTM. Jeon et al. (2023) [18]

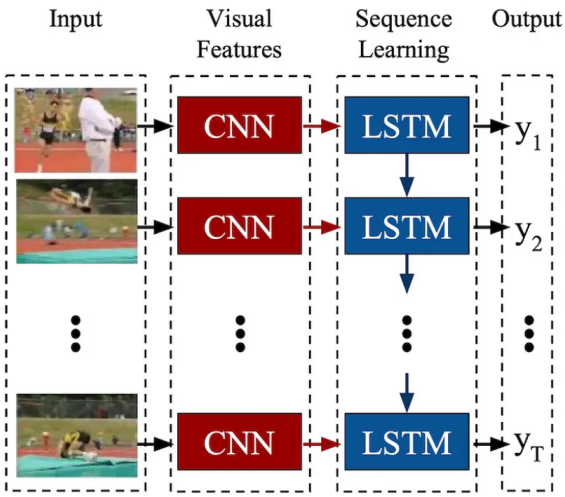

Figure 2: Donahue et al. (2016) Long-term Recurrent Convolutional Networks for Visual Recognition and Description [10]

In summary, HAR in video classification is an active research area with numerous approaches proposed in the literature. ConvLSTM and LRCN are two popular deep learning architectures that have shown promise in recognising human activities from video data. While both architectures have their strengths and limitations, the choice of architecture depends on the specific application and the available resources. Further research could focus on developing new architectures that address the limitations of existing approaches or exploring the combination of multiple architectures. Other factors such as the size and quality of the dataset, the preprocessing techniques used, and the choice of optimization algorithm also impact the performance of HAR in video classification.

5 Experimental Setup

This section presents the dataset building procedure, and experimental results using the deep learning architectures, i.e., ConvLSTM, and LRCN. The proposed model has been tested on a UCF50 dataset.

5.1 Dataset

Due to time and resource limitations, the UCF50 dataset was chosen for this researching since it offers a wide variety of human activity videos. The choice to utilize UCF50 was made based on its accessibility, fit for the study's goals, and capacity to analyze the dataset within the project's computational parameters. Utilizing Google Colaboratory, a cloud-based platform that allowed for effective use of computer resources and enabled team collaboration, the tests and analysis were carried out.

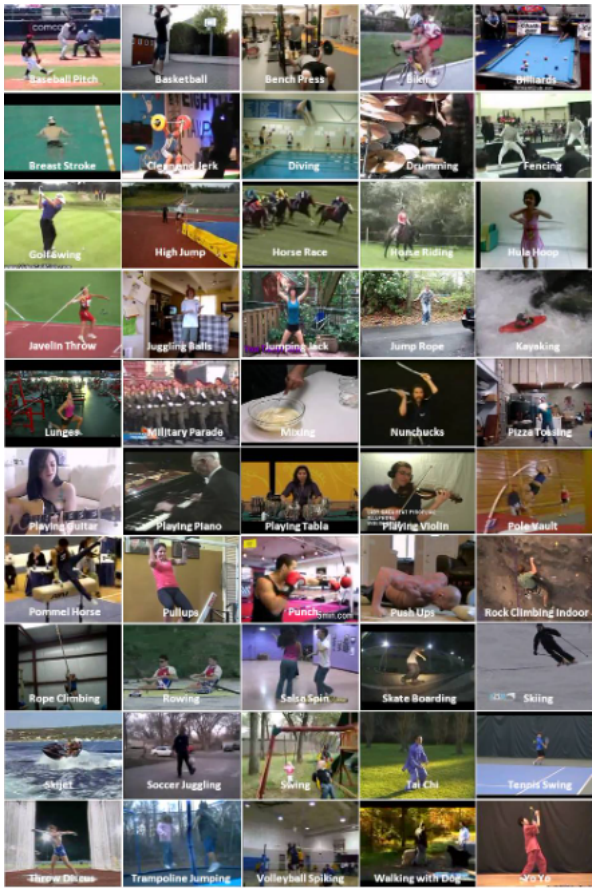

Figure 3: UCF50 dataset

UCF50 is an action recognition dataset which contains:

- 50 Action Categories consisting of realistic YouTube videos

- 25 Groups of Videos per Action Category
- 133 Average Videos per Action Category
- 199 Average Number of Frames per Video
- 320 Average Frames Width per Video
- 240 Average Frames Height per Video
- 26 Average Frames Per Seconds per Video

Many existing action recognition datasets lack realism as they are often artificially staged by actors. However, this dataset aims to address this limitation by providing the computer vision community with a collection of realistic videos sourced from YouTube. This dataset was specifically designed to be challenging, as it encompasses a wide range of variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, and illumination conditions. By incorporating these real-world factors, this dataset offers a more authentic representation of the complexities involved in action recognition tasks.
The UCF50 dataset contain activities as: Jumping Jack, Fencing, Pole Vault, Baseball Pitch, Horse Riding, Billiards Shot, Biking, Diving, Drumming, Golf Swing, Clean and Jerk, Biking, Horse Race, Volleyball Spiking, Lunges, Rock Climbing Indoor, Javelin Throw, Playing Violin, High Jump, Playing Tabla, Military Parade, Skiing, Skateboarding, Swing, Jump Rope, Soccer Juggling, Basketball Shooting, Salsa Spins, Yoga, Ski Jet, Playing Piano, Tennis Swing, Rope Climbing, Walking with a dog, Trampoline Jumping, Bench Press, Playing Guitar, Pull Ups, Hula Hoop, Rowing, Breaststroke, Tai Chi, Pizza Tossing, Nun chucks, Mixing Batter, Drumming, Kayaking, Yo-Yo, Pommel Horse, and Push Ups. Such information may be recorded by an external camera, an on-board computer system, or a number of security cameras, shown in Figure 3.

5.2 Evaluation framework

The human activity recognition model is built using the steps listed below:

1. Download and Visualize the Data with its Labels
2. Preprocess the Dataset
3. Split the Data into Train and Test Set
4. Implement the ConvLSTM Approach
   - 4.1. Construct the Model
   - 4.2. Compile & Train the Model
   - 4.3. Plot Model's Loss & Accuracy Curves
5. Implement the LRCN Approach
   - 5.1. Construct the Model
   - 5.2. Compile & Train the Model
   - 5.3. Plot Model's Loss & Accuracy Curves
6. Test the Best Performing Model on YouTube videos

To construct the ConvLSTM model, the **ConvLSTM2D** recurrent layers provided by Keras were utilized. Figure 4 illustrates the structure of these ConvLSTM2D layers. The configuration of the ConvLSTM2D layer included the specification of the number of filters and kernel size, which are essential for performing the convolutional operations. The output of these layers was then flattened and passed to a Dense layer with softmax activation, producing the output probabilities for each action category. The ConvLSTM2D layer in Keras is designed to handle 2D spatio-temporal inputs, typically in the form of video data. It extends the standard LSTM cell by replacing the matrix multiplications in the LSTM cell with convolutional operations. This allows the ConvLSTM2D layer to effectively capture spatial and temporal dependencies in video data, enabling it to learn complex patterns and make predictions based on both local and global context. This makes ConvLSTM2D a popular choice for tasks such as action recognition, video prediction, and video generation.

MaxPooling3D layers were incorporated to reduce the dimensions of the frames and minimize unnecessary computations. MaxPooling3D is often used in CNN architectures for video analysis, where it helps to extract the most relevant spatio-temporal features. It is typically applied after convolutional layers to downsample the feature maps and reduce the spatial dimensions of the data. This downsampling facilitates subsequent layers to focus on higher-level representations and reduces the model's sensitivity to small variations in the input data. Dropout layers were also added to mitigate overfitting, ensuring the model generalizes well to unseen data. The architecture was intentionally kept simple, with a small number of trainable parameters. This decision was influenced by the fact that the model only needed to handle a specific subset of the dataset, making a large-scale model unnecessary.

In summary, this approach effectively captures the spatial relationships within individual frames and the temporal relationships across different frames, making it suitable for video classification tasks. The ConvLSTM architecture accommodates 3-dimensional inputs (width, height, and number of channels), whereas a basic LSTM can only handle 1-dimensional inputs. Consequently, an LSTM alone is inadequate for modeling spatio-temporal data, whereas the ConvLSTM's convolutional structure is designed precisely for that purpose.

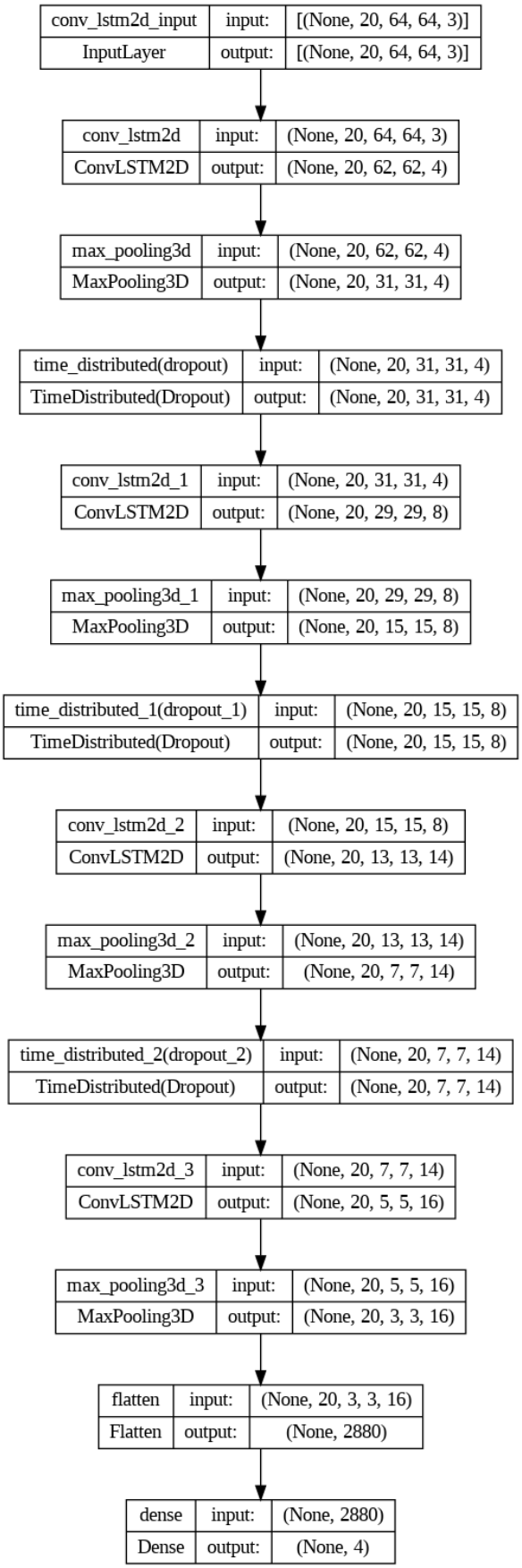

Figure 4: ConvLSTM model

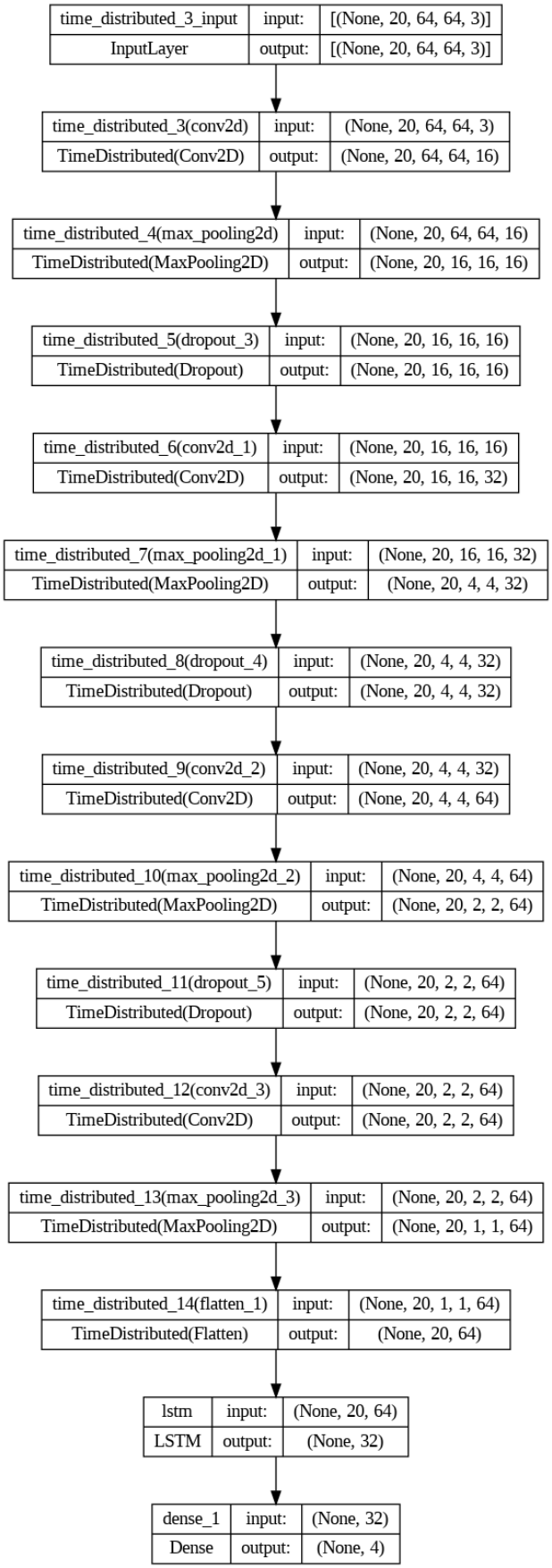

Figure 5: LRCN model

LRCN Approach was implemented by combining Convolution and LSTM layers in a single model. Another similar approach involved using a CNN model and LSTM model trained separately. The CNN model was utilized to extract spatial features from the frames in the video, and a pre-trained model could be employed for this purpose, which could be fine-tuned to address the specific problem. Subsequently, the LSTM model utilized the features extracted by the CNN to predict the action being performed in the video.

The **TimeDistributed** wrapper layer was also employed, which facilitated the application of the same layer to every frame of the video independently. This allowed a layer, around which it was wrapped, to handle input of shape (number of frames, width, height, number of channels), whereas the original layer's input shape was (width, height, number of channels). This capability proved highly advantageous as it enabled the model to process the entire video as a single input.

To implement the LRCN architecture, time-distributed Conv2D layers were utilized, followed by MaxPooling2D and Dropout layers. The features extracted by the Conv2D layers were subsequently flattened using the Flatten layer and fed into an LSTM layer. The output from the LSTM layer was then utilized by a Dense layer with softmax activation to predict the action being performed, Figure 5.

6 Experimental Results

To assess the performance of ConvLSTM and LRCN architectures, commonly used evaluation metrics such as accuracy, precision, recall, and F1 score (explained in more detail below) were employed. Classification report is a summary of the performance of a classification model that provides key evaluation metrics such as precision, recall, and F1 score for each class.

These metrics are crucial because they help the accuracy and effectiveness of the model in predicting the correct class labels and provides insights into the model's strengths, and weaknesses for different classes. High precision indicates a low false positive rate, which is important in scenarios where misclassifications can have significant consequences. High recall signifies a low false negative rate, ensuring that the model can capture as many positive instances as possible.

By analyzing these metrics in the classification report, we can gain insights into the model's strengths and weaknesses. It helps us understand how well the model is performing and provides guidance for potential improvements, such as fine-tuning the model's parameters or adjusting the classification threshold. Ultimately, the classification report allows us to evaluate and enhance the model's performance for better decision-making and problem-solving.

1. Precision: Percentage of correct positive predictions relative to total positive predictions. Represents the ability of the model to accurately identify positive instances out of all the instances it predicted as positive.

$$Precision = \frac{TP}{TP + FP}$$

2. Recall: Percentage of correct positive predictions relative to total actual positives. Measures the model's ability to identify all the actual positive instances from the dataset.

$$Recall = \frac{TP}{TP + FN}$$

3. F1 Score: A weighted harmonic mean of precision and recall, providing an overall assessment of the model's performance. The closer to 1, the better the model.

$$F1score = 2 * \frac{precision * recall}{precision + recall}$$

Where TP, FP, TN, FN represents:

- True positives: data points labeled as positive that are actually positive
- False positives: data points labeled as positive that are actually negative
- True negatives: data points labeled as negative that are actually negative
- False negatives: data points labeled as negative that are actually positive

Table 1: Classification report - ConvLSTM model

	precision	recall	f1-score
WalkingWithDog	0.76	0.54	0.63
TaiChi	0.90	0.70	0.79
Swing	0.78	0.76	0.77
HorseRace	0.71	0.97	0.82

Table 1. displays the classification report for the ConvLSTM model based on choosed predicted classes WalkingWithDog, TaiChi, Swing, and HorseRace that were previously trained.

Table 2: Classification report - LRCN model

	precision	recall	f1-score
WalkingWithDog	0.70	0.79	0.75
TaiChi	0.90	1.00	0.95
Swing	0.90	0.85	0.88
HorseRace	0.97	0.87	0.92

Table 2. displays the classification report for the LRCN model based on choosed predicted classes WalkingWithDog, TaiChi, Swing, and HorseRace that were previously trained.

Accuracy represents the number of correctly classified data instances over the total number of data instances.

$$Accuravy = \frac{TN + TP}{TN + FP + TP + FN}$$

Total accuracy is a metric that measures the overall correctness (in percentage) of classification across all classes in a dataset. It provides an indication of how well the model performs in classifying the data, taking into account all categories. On the other hand, total validation accuracy is a metric specifically used during the validation phase of model training to assess its performance on unseen data. It helps evaluate the model's ability to generalize and make accurate predictions on new, unseen examples. The main difference between total accuracy and total validation accuracy lies in the datasets on which they are measured. These metrics provide valuable insights into the overall performance of the model, indicating how well it is able to classify data across all classes and its generalization capabilities.

ConvLSTM exhibited superior accuracy in capturing temporal dependencies. It leverages the power of both convolutional layers and LSTM (Long Short-Term Memory) cells to effectively model and analyze sequential data. In the specific context mentioned, ConvLSTM demonstrated an accuracy of 70.49% in identifying the activity being done.

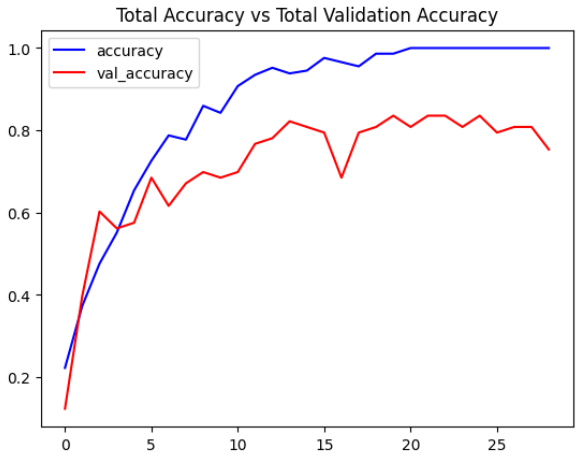

Figure 6: ConvLSTM model accuracy

On the other hand, LRCN (Long-term Recurrent Convolutional Networks) demonstrated stronger capabilities in extracting spatial features. It combines convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to capture both spatial and temporal information from the input data. In the mentioned scenario, the LRCN model achieved an accuracy of 88.52% in accurately identifying the activity being performed.

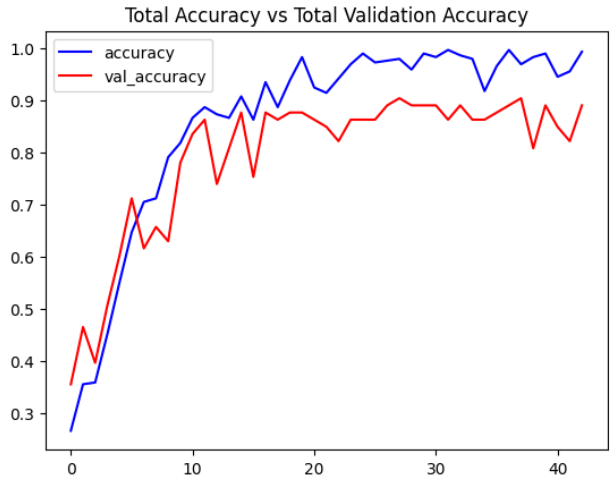

Figure 7: LRCN model accuracy

The confusion matrices were also examined to evaluate the architectures' ability to differentiate between similar activities. The experimental findings demonstrated competitive performance by both architectures in classifying human activities. The confusion matrix is a table that provides a comprehensive summary of the performance of a classification model. It presents the predicted and actual class labels, allowing us to analyze the model's performance in terms of true positives, true negatives, false positives, and false negatives. It is a valuable tool for evaluating the accuracy and reliability of a classification model and provides insights into the types of errors it makes. By examining the confusion matrix, we can identify patterns, assess the model's strengths and weaknesses, and make informed decisions for improving its performance.In Figures 8 and 9, the confusion matrices for the ConvLSTM and LRCN models are displayed.

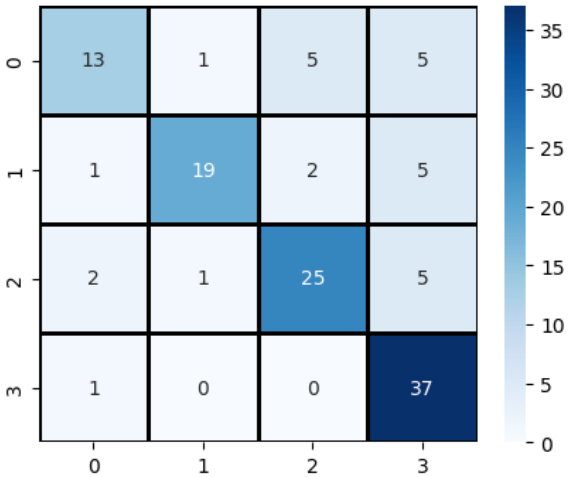

Figure 8: ConvLSTM model confusion matrix

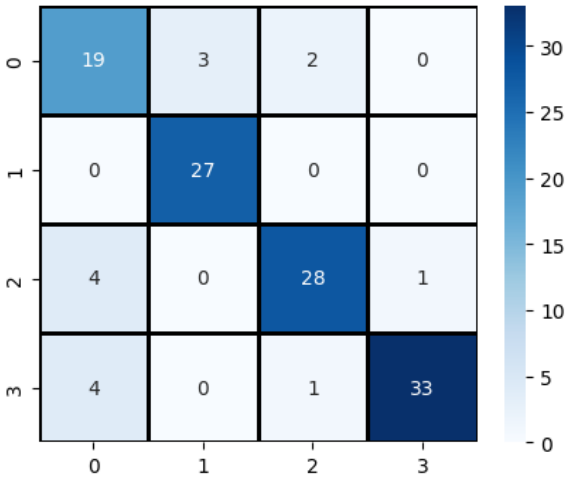

Figure 9: LRCN model confusion matrix

Additionally, a computational efficiency analysis was conducted to compare the computational demands of the two architectures. The experiments revealed that ConvLSTM incurred higher computational costs due to its larger number of parameters and operations compared to LRCN. Since the LRCN model outperformed the ConvLSTM model in terms of performance, we can say LRCN is the best model for identifying human activities.

In summary, this research paper presents a comparative investigation of ConvLSTM and LRCN architectures for video classification of human activities using the UCF50 dataset. The results indicate promising outcomes for both architectures, each showcasing its own strengths and trade-offs in capturing spatial and temporal information. These insights can assist researchers and practitioners in selecting the appropriate architecture for video classification tasks, considering their specific requirements and computational limitations.

Conclusion and Future Work

In this experimental research paper, a comparative analysis of ConvLSTM and LRCN architectures was conducted for video classification of human activities using the UFC50 dataset. The obtained results demonstrated the effectiveness of both architectures in accurately recognizing and categorizing human activities. The particular application and the available resources have an impact on the architecture decision, even though each architecture has advantages and disadvantages. The findings revealed that the LRCN architecture achieved an impressive accuracy of 87.70%, surpassing the ConvLSTM architecture, which achieved an accuracy of 77.05%. The superior performance of the LRCN model can be attributed to its ability to effectively extract spatial features from video frames, complemented by the temporal modeling capabilities of recurrent neural networks. These results emphasize the significance of architecture selection when dealing with video classification tasks. The strong performance of the LRCN architecture suggests that it may be better suited for capturing the complex spatial relationships present in human activity videos, contributing to its higher accuracy.

While this research provides valuable insights into the comparative performance of ConvLSTM and LRCN architectures on the UFC50 dataset, there are several areas for further investigation and improvement. Firstly, the dataset used in this study, UFC50, offers a limited scope of human activities. Expanding the evaluation to larger and more diverse datasets encompassing a wider range of activities would provide a more comprehensive assessment of the architectures' capabilities. Additionally, exploring different variations and configurations of ConvLSTM and LRCN architectures could lead to improved results. Fine-tuning hyperparameters, adjusting network depths, or incorporating pre-trained models could potentially enhance the performance of both architectures. Furthermore, investigating other deep learning architectures specifically designed for video classification, such as two-stream networks or 3D convolutional networks, could offer alternative approaches and potentially yield even higher accuracies. Finally, conducting an extensive analysis of misclassified samples and error analysis would provide valuable insights into the limitations and potential areas of improvement for both architectures. By addressing these aspects, future research can further advance the field of video classification of human activities and potentially lead to more accurate and robust models.

References

[1] Preeti Agarwal and Mansaf Alam. A lightweight deep learning model for human activity recognition on edge devices. *Procedia Computer Science*, 167:2364–2373, 2020.

[2] Jake K Aggarwal and Quin Cai. Human motion analysis: A review. *Computer vision and image understanding*, 73(3):428–440, 1999.

[3] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L Reyes-Ortiz. Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In *Ambient Assisted Living and Home Care: 4th International Workshop, IWAAL 2012, Vitoria-Gasteiz, Spain, December 3-5, 2012. Proceedings 4*, pages 216–223. Springer, 2012.

[4] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, Jorge Luis Reyes-Ortiz, et al. A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, page 3, 2013.

[5] Pierluigi Casale, Oriol Pujol, and Petia Radeva. Human activity recognition from accelerometer data using a wearable device. In *Pattern Recognition and Image Analysis: 5th Iberian Conference, IbPRIA 2011, Las Palmas*

de Gran Canaria, Spain, June 8-10, 2011. Proceedings 5, pages 289–296. Springer, 2011.

[6] Jose M Chaquet, Enrique J Carmona, and Antonio Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6):633–659, 2013.

[7] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys (CSUR)*, 54(4):1–40, 2021.

[8] Zhenghua Chen, Le Zhang, Zhiguang Cao, and Jing Guo. Distilling the knowledge from handcrafted features for human activity recognition. *IEEE Transactions on Industrial Informatics*, 14(10):4334–4342, 2018.

[9] Pavan Dasari, Li Zhang, Yonghong Yu, Haoqian Huang, and Rong Gao. Human action recognition using hybrid deep evolving neural networks. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

[10] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[12] David A Forsyth and Jean Ponce. *Computer vision: a modern approach*. prentice hall professional technical reference, 2002.

[13] Dariu M Gavrila. The visual analysis of human movement: A survey. *Computer vision and image understanding*, 73(1):82–98, 1999.

[14] Nit Hamirpur. Human activity recognition with deep learning: Overview, challenges & possibilities. 2021.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 961–970. IEEE, 2015.

[17] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. *Image and vision computing*, 60:4–21, 2017.

[18] DaeHyeon Jeon and Min-Suk Kim. Deep-learning-based sequence causal long-term recurrent convolutional network for data fusion using video data. *Electronics*, 12(5):1115, 2023.

[19] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[20] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[21] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[22] Shian-Ru Ke, Hoang Le Uyen Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. A review on video-based human activity recognition. *Computers*, 2(2):88–131, 2013.

[23] Sarah Khater, Mayada Hadhoud, and Magda B Fayek. A novel human activity recognition architecture: using residual inception convlstm layer. *Journal of Engineering and Applied Science*, 69(1):45, 2022.

[24] Maryam Koohzadi and Nasrollah Moghadam Charkari. Survey on deep learning methods in human action recognition. *IET Computer Vision*, 11(8):623–632, 2017.

[25] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[26] Jingen Liu, Jiebo Luo, and Mubarak Shah. Recognizing realistic actions from videos "in the wild". In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1996–2003. IEEE, 2009.

[27] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 439–444. IEEE, 2017.

[28] Mona M. Moussa, Elsayed Hamayed, Magda Bahaa Eldin Fayek, and Heba A. El Nemr. Video based human activity detection, recognition and classification of actions using svm. 2018.

[29] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3163–3172, 2021.

[30] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II 11*, pages 392–405. Springer, 2010.

[31] Atiq Rehman and Samir Brahim Belhaouari. Deep learning for video classification: A review. 2021.

[32] Christopher Reining, Friedrich Niemann, Fernando Moya Rueda, Gernot A Fink, and Michael ten Hompel. Human activity recognition for production and logistics—a systematic literature review. *Information*, 10(8):245, 2019.

[33] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004.

[34] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[35] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[36] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, IEEE International Conference on*, volume 3, pages 1470–1470. IEEE Computer Society, 2003.

[37] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[38] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[39] Zhigang Tu, Wei Xie, Qianqing Qin, Ronald Poppe, Remco C Veltkamp, Baoxin Li, and Junsong Yuan. Multi-stream cnn: Learning representations based on human-related regions for action recognition. *Pattern Recognition*, 79:32–43, 2018.

[40] Pavan Turaga, Rama Chellappa, Venkatramana S Subrahmanian, and Octavian Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video technology*, 18(11):1473–1488, 2008.

[41] Amin Ullah, Jamil Ahmad, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. Action recognition in video sequences using deep bi-directional lstm with cnn features. *IEEE access*, 6:1155–1166, 2017.

[42] Amin Ullah, Khan Muhammad, Javier Del Ser, Sung Wook Baik, and Victor Hugo C de Albuquerque. Activity recognition using temporal optical flow convolutional features and multilayer lstm. *IEEE Transactions on Industrial Electronics*, 66(12):9692–9702, 2018.

[43] Hadiqa Aman Ullah, Sukumar Letchmunan, M Sultan Zia, Umair Muneer Butt, and Fadratul Hafinaz Hassan. Analysis of deep neural networks for human activity recognition in videos—a systematic literature review. *IEEE Access*, 9:126366–126387, 2021.

[44] Muhammad Sajib Uzzaman, Chandan Debnath, Dr Uddin, Md Ashraf, Md Islam, Md Talukder, Shamima Parvez, et al. Lrcn based human activity recognition from video data. *Ashraf and Islam, Md. Manowarul and Talukder, Md. Alamin and Parvez, Shamima, LRCN Based Human Activity Recognition from Video Data.*

[45] Greg Van Houdt, Carlos Mosquera, and Gonzalo Nápoles. A review on the long short-term memory model. *Artificial Intelligence Review*, 53:5929–5955, 2020.

[46] Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2:28, 2015.

[47] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In *Bmvc 2009-british machine vision conference*, pages 124–1. BMVA Press, 2009.

[48] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern recognition letters*, 119:3–11, 2019.

[49] Xuanhan Wang, Lianli Gao, Jingkuan Song, and Hengtao Shen. Beyond frame-level cnn: saliency-aware 3-d cnn with lstm for video action recognition. *IEEE signal processing letters*, 24(4):510–514, 2016.

[50] Santosh Kumar Yadav, Kamlesh Tiwari, Hari Mohan Pandey, and Shaik Ali Akbar. Skeleton-based human activity recognition using convlstm and guided feature learning. *Soft Computing*, pages 1–14, 2022.

[51] Xiaodong Yang, Chenyang Zhang, and YingLi Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1057–1060, 2012.

[52] Chaitanya Yeole and H Singh. Deep neural network approaches for video based human activity recognition. 2021.

[53] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019.

[54] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

[55] Fan Zhu, Ling Shao, Jin Xie, and Yi Fang. From handcrafted to learned representations for human action recognition: A survey. *Image and Vision Computing*, 55:42–52, 2016.