Alexandria Engineering Journal (2023) 73, 771–779

HOSTED BY

ELSEVIER

Alexandria University

Alexandria Engineering Journal

www.elsevier.com/locate/aej
www.sciencedirect.com

ORIGINAL ARTICLE

Optimized deep learning-based cricket activity focused network and medium scale benchmark

Waqas Ahmad [a,b], Muhammad Munsif [a], Habib Ullah [c], Mohib Ullah [b], Alhanouf Abdulrahman Alsuwailem [d], Abdul Khader Jilani Saudagar [d], Khan Muhammad [e,*], Muhammad Sajjad [a,b,*]

[a] *Department of Computer Science, Islamia College Peshawar, 25000 Peshawar, Pakistan*
[b] *Department of Computer Science, Norwegian University of Science and Technology, 2815 Gjøvik, Norway*
[c] *Faculty of Science and Technology, Norwegian University of Life Sciences, Gjøvik, Norway*
[d] *Information Systems Department, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia*
[e] *VIS2KNOW Lab, Department of Applied Artificial Intelligence, School of Convergence, College of Computing and Informatics, Sungkyunkwan University, 03063 Seoul, South Korea*

Received 19 December 2022; revised 6 April 2023; accepted 25 April 2023
Available online 22 May 2023

KEYWORDS

Activity recognition;
Cricket sports activities;
Convolutional neural network;
Sequence learning;
Spatiotemporal network

Abstract The recognition of different activities in sports has gained attention in recent years for its applications in various athletic events, including soccer and cricket. Cricket, in particular, presents a challenging task for automatic activity recognition methods due to its closely overlapped activities such as cover drive, and pull short, to name a few. Existing methods often rely on hand-crafted features as the limited availability of public data has restricted the scope of research to only the significant categories of cricket activities. To this end, we proposed a cricket activities dataset and an intuitive end-to-end deep learning model for cricket activity recognition. The data is collected from online sources and pre-processed through cleaning, resizing, and organizing. Similarly, an intuitive deep model is designed with a combination of time-distributed 2D CNN layers and LSTM cells for extracting and learning the spatiotemporal information from the input sequences. For benchmarking, we evaluated the model on our cricket datasets and four standard datasets namely UCF101, HMDB51, YouTube action, and Kinetics. The quantitative results show that the proposed model outperforms different variants of recurrent neural networks and achieved an accuracy of 92%, recall of 91%, and F1 score of 91%. Our code and dataset is publicly available for further research on https://drive.google.com/file/d/1c9qcAz4q00qvx4yFA3pSudWFczm1cWUL/view?usp=sharing.

[*] Corresponding authors.

Peer review under responsibility of Faculty of Engineering, Alexandria University.

https://doi.org/10.1016/j.aej.2023.04.062

1. Introduction

In sports, activity recognition plays a significant role to track the performance of the individual player as well as the entire team [1–4]. Activity recognition can be applied to a wide range of athletic events. Each event has its own distinct dynamics, which may differ according to various circumstances. Some sports are driven by simple activities such as walking and sitting, while others are driven by complex activities like an overhead smash in Tennis [5]. Event detection in sports videos is a very challenging and difficult task because of abrupt motion, similar outfits of the players, varied viewpoints, and camera movement [6]. In sports videos, the player's posture changes from moment to moment [7]. For example, physical movement is a simple process of hitting the ball in baseball and cricket. The complete shot of a player is made up of the movements of various body parts, including the legs, arms, head, and entire body working together in coordination. Therefore, an event is a combination of multiple activities that can be categorized into low, medium, and high-level activities or complex activities as shown in Fig. 1.

Various researchers have been contributing towards activity recognition in various sports including but not limited to cricket, badminton, volleyball, and football. A number of paradigms are proposed to efficiently classify the activities [8]. In a nutshell, it can be divided into two groups namely traditional machine learning (ML) [9] and advanced deep learning (DL) activity recognition methods. ML techniques are based on two main steps i.e., (i) feature embedding - feature descriptors selection (ii) and an appropriate classification algorithm for classifying the underlying activity. For the feature descriptor, the local and global features can be used to extract spatial and temporal information for the input sequences [10]. Handcrafted feature extractors are most of the time domain-specific and specifically designed for specific types of tasks. For example, Zhao et al.[11] proposed a key frames-based descriptor that extracts key points. In some cases, these extractors generate similar feature maps for two different activities, making the representation of various activities difficult. To process complicated datasets, hybrid techniques combine various features such as motion, background, and histogram of oriented (HOG) and pass them to the prediction module. However, the high computational complexity in terms of long-time videos and real-time response in continuous video streaming is challenging. To target specifically cricket activities Karmaker et. al. [12] introduced a strategy that predicts the shot type played by players. They used camera motion parameters for calculating the trajectories and then used these calculated trajectories to classify the performed shot into two classes including cover drive and pull shot. In addition, they used a 3D MACH kernel to train a model to recognize four types of shots based on including square-cut, flick, off-drive, and hook. Angle ranges for the final prediction were calculated by utilizing an optical flow features vector by looking for various thresholds for a shot. Noorbhai et. al. [13] analyzed the back lift of the batsmen, helps in the categorization of the position of the bat and helps in player performance analysis. Similarly, Arora et al. [14] proposed an algorithm for ball detection and tracking based on the histogram of gradients (HOG) and support vector machine for classification. Yeole et al. [15] created a strategy for monitoring wickets in order to assist the third umpire in making a run-out decision. Further, Chowdhury et al. [16] come up with a strategy for diving the bowling crease into two sections. They used the image subtraction method to differentiate between a legal ball and a no-ball. Irrespective of the moderate success of traditional approaches, they have several drawbacks like high time complexity and being sensitive towards changes in data such as orientation, illumination, and position of the objects [17].

To handle the limitations and challenges of traditional methods, researchers developed DL-based methods [18–20]. The DL techniques are able to learn effectively and represent high-level visual features and then classify videos [21,20] in an end-to-end fashion. Convolutional Neural Network (CNN) is a popular architecture that often alters the parameters based on the information and uses convolutional operations to learn the best features [22–24]. In the realm of DL, Feichtenhofer et al. [25] proposed simple CNN features and fused temporal and spatial features to recognize activities. Similarly, Tu et al. [26] developed a CNN model consisting of multi streams that are able to learn human-related features for recognizing various activities. Ijjina and Mohan [27] developed a hybrid strategy for learning multiple features for activity recognition by fusing various features [28]. However

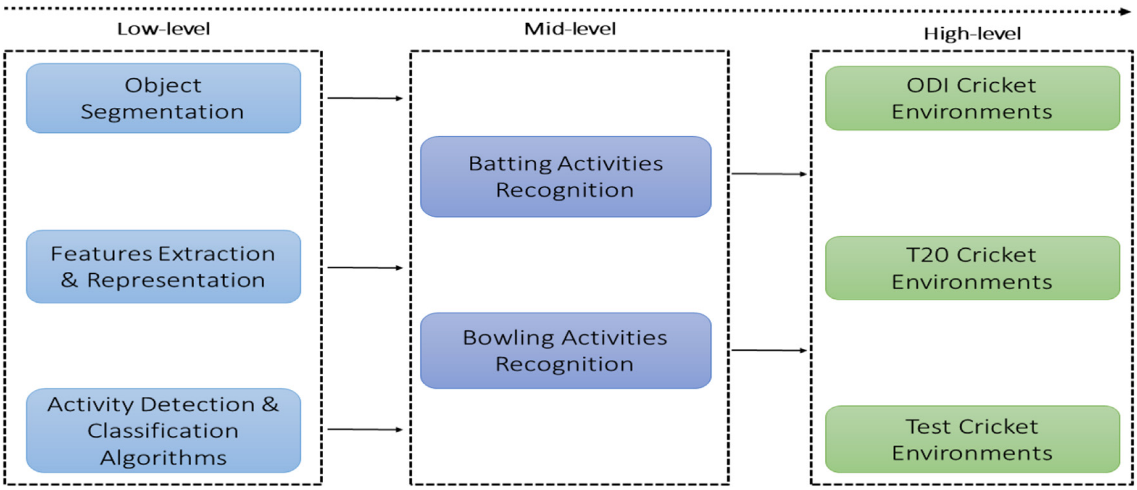

Low-level
Mid-level
High-level
Object Segmentation
Features Extraction & Representation
Activity Detection & Classification Algorithms
Batting Activities Recognition
Bowling Activities Recognition
ODI Cricket Environments
T20 Cricket Environments
Test Cricket Environments

Fig. 1 Levels of Recognition. Different levels of activity recognition are presented namely, low-level, mid-level, and high-level.

one of the basic limitations of this technique is that they are only able to extract short-term information. To extract long-term spatiotemporal cues from video sequences Dai et al. [29] and Gammulle et al. [30] used a two-stream network. Furthermore, Tsunoda et al. [31] presented a centralized deep neural network to classify player activities in sports based on the player's location in the frames. They categorized the sports activity recognition framework into three stages including feature extraction, dictionary learning, and video classification. Most of the contemporary methods achieved better performance. However, they relied on limited data and used hand-crafted descriptors or a separate feature extraction module that has high computational complexity. Due to the unavailability of a public dataset in cricket, less attention has been given to recognizing different activities like batting (hitting a ball with a cricket bat). To this end, we proposed a dataset that is publicly accessible to researchers consisting of complex cricket activities. We also proposed an intuitive deep model that consists of three stages, namely pre-processing, feature extraction, and classification. In the first step, keyframes are selected and resized to the standard size of 100 × 100 for effective model training. In the second step, the resized key-frames are used as input to the CNN, and the CNN extracts spatial features using distributed CNN layers. Each convolutional layer applies a set of learned filters to the input key-frame to capture specific patterns, and the resulting feature maps are passed to the next layer for further processing. The output of the CNN is a set of high-level features represented in a way that is useful for activity recognition. Afterwards, the extracted features vector is fed to a long short-term memory (LSTM) to effectively classify the sports activity from a video sequence. In a nutshell, the contributions of this study are summarized as follows.

- We have compiled a comprehensive cricket dataset encompassing five complex activities. The dataset is fully annotated and is available to the research community.
- We have designed an intuitive CNN-LSTM time-distributed model for the recognition of complex cricket activities, specifically batting.
- An extensive experiment has been performed with our dataset and four benchmark datasets. Our model was evaluated against existing variants of recurrent neural networks (RNNs) using accuracy, recall, and F1 score as metrics. The quantitative results indicate that our model achieves superior results.

The rest of the paper is structured in the following order. Section 2 provides a detailed description of the proposed method. The dataset description, experimental results, and discussion are given in Section 3. Section 4 gives future directions and final remarks that conclude the paper.

2. Methodology

The graphical representation of the proposed model is given in Fig. 2. The complete information on the model is tabulated in Table 1. The proposed model architecture consists of various layers including Time distributed 2D-CNN, Max pooling (MP), and fully connected layers. The model accepts a sequence of frames and is analyzed by the CNN network that extracts spatial features from the input frames. The LSTM layers are utilized to capture the temporal features and dependencies between the frames. The model includes multiple convolutional layers with 64 feature maps each. An activation function, specifically the rectified linear unit (ReLU) function, is used in each convolutional layer to introduce non-linearity to the output of each layer. The outputs of the last max-pooling layer in the CNN are flattened and fed to the LSTM layers, which capture the temporal dependencies in the sequence of frames. Flattening the output of the max-pooling layer means that the spatial dimensions are reduced, resulting in a 1-dimensional vector that can be fed to the LSTM layer. The layer uses a kernel size of 3 × 3 to perform convolution on the input frames. The 2 × 2 jump refers to the stride length or step size of the convolutional kernel, which determines how many pixels the kernel moves between each convolution operation. To prevent the loss of information at the boundaries of the input frames, the layer uses the same padding technique. Padding involves adding extra pixels around the edges of the input frames to create a larger frame before performing the convolution. This ensures that the convolutional kernel can process the pixels at the edges of the frames without missing any information. To reduce the size of the feature maps, each max-pooling layer contains a kernel size of 4 × 4, meaning it takes a sub-region of 4 × 4 pixels from the input feature map along with a filter jump or stride length of 2 × 2 means that the sub-region moves 2 pixels in both the horizontal and vertical directions between each pooling operation and as above the same padding strategy is utilized. As shown in Table 1, the second convolutional layer produces 32 feature maps as output by 3 × 3 kernel size, 2 by 2 stride, and a ReLU function as activation for each kernel. The technique is replicated on the feature maps created by the first layer by the second layer and the 2D max-pooling layer with a distinguishing filter size of 4 × 4 and a stride size of 2 × 2 considered. In addition, the third CNN layer in the model has 64 feature maps, which are similar to the previous two CNN layers. This layer also uses the same padding technique to prevent the information from being skipped at the input frame boundaries while convolution through a kernel size of 3x3 along with a stride size of 2x2. After the convolution, the output feature maps are passed through a max-pooling layer with a filter size of 2 × 2 and a stride of 2 × 2 to reduce their size. The same padding strategy is utilized in this max-pooling layer as well. These strategies are repeated in the third layer to enhance the useful features in the video sequence.

3. Experimental analysis and results

All the experiments were carried out in a python virtual environment consisting of Keras along the back-end of TensorFlow-GPU installed on the personal computer. The computer is equipped with an NVIDIA GeForce GTX 3060 GPU and 12 GB of RAM. Further, we used CUDA toolkit 9.2 and cuDNN v7.0. Details of the environment specification are presented in Table 2. The categorical cross-entropy loss function is used to measure the error between the predicted values and actual values. During the training phase, the Adam optimizer is used to optimize the weights and biases of the neural network model. The batch size is set to 16, and the number of epochs is set to 60.

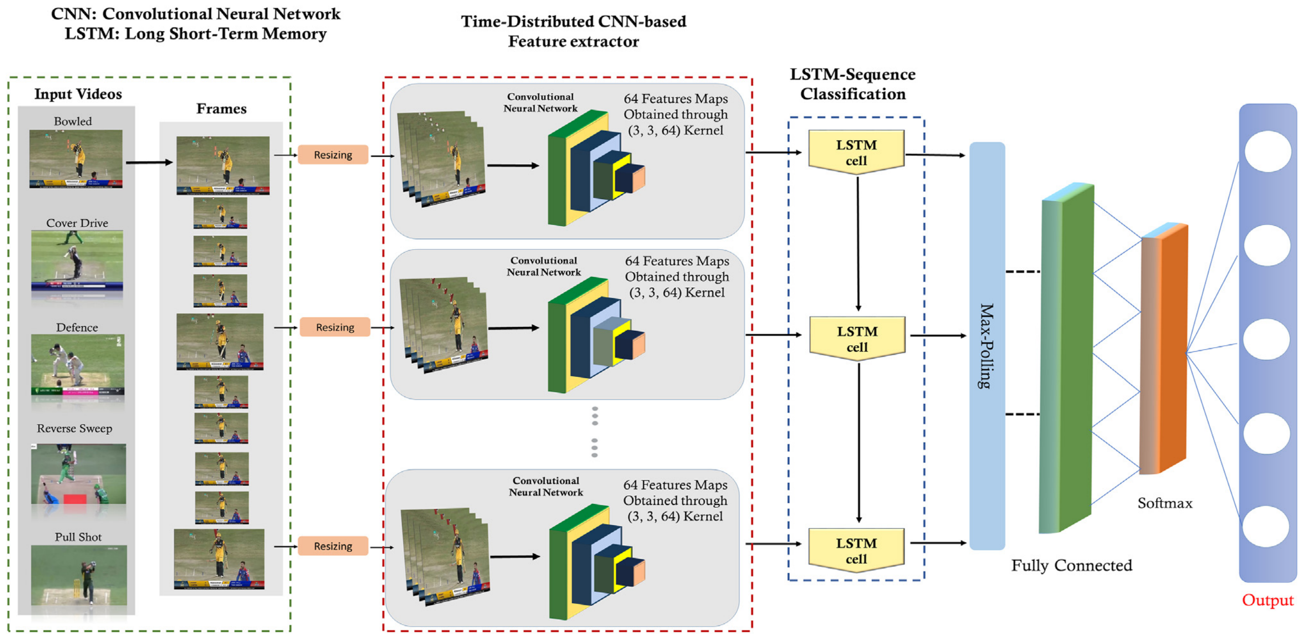


Fig. 2 An overall structure of the proposed model. Firstly, the keyframes are extracted from the sports videos and fed to the CNN model for features extraction. Then, the extracted features are fed to the LSTM model for effective sports activity recognition.

Table 1 Details of the CNN-LSTM model.

Layer	Kernels	Size	Padding	Jump	A-Function	Output Maps
TD (Con2D)$_1$	16	3×3	same	2×2	Relu	10, 100, 100, 16
TD (Max-Pooling2D)$_1$	1	4×4	-	2×2	-	10, 25, 25, 16
TD (Con2D)$_2$	32	3×3	same	2×2	Relu	10, 25, 25, 32
TD (Max-Pooling2D)$_2$	1	4×4	-	2×2	-	10, 25, 25, 16
TD (Con2D)$_3$	64	3×3	same	2×2	Relu	10, 6, 6, 64
TD (Max-Pooling2D)$_3$	1	2×2	-	2×2	-	10, 3, 3, 64
TD (Con2D)$_4$	128	3×3	same	2×2	Relu	10, 3, 3, 128
Time-Distributed (Max-Pooling2D)$_4$	1	2×2	-	2×2	-	10, 1, 1, 128
LSTM	-	-	-	-	-	None, 10, 128
Time-Distributed (Flatten)	-	-	-	-	-	49408

Table 2 System specifications and hardware/software configurations for developing the proposed method.

Hardware	Model (version)
PC Type	i5-10400 CPU@2.90 GHz
Processor	NVIDIA Ge-Force GTX 3060
Operating system	Windows 10 Pro
RAM	16 GB
CUDA toolkit	11.3
CUDNN	v7.0
Python version	3.7
TensorFlow	2.8.0

3.1. Datasets

3.1.1. Our Cricket Dataset

The cricket video dataset is collected from YouTube and cricket-info websites. The dataset includes 722 videos that represent different classes of batting activities, including pull shot, bowled, reverse sweep, defence, and cover drive as shown in Fig. 3. We present the details of our dataset in Table 3. The videos were all recorded at the same frame rate of 30 and had the same background, ground, pitch, and spectator accommodation. Each class contains 150 videos, and the shortest video in the dataset has 56 frames. The dimensions of each frame are 840 x 480. We used five different classes from each video. The video editor tool was used to extract short video clips of 1 to 3 s in duration for each of these categories. This process was repeated for each video, resulting in a dataset of short video clips that can be used for various purposes such as recognition, analysis, etc. This approach of creating a dataset of short video clips is common in ML tasks such as action recognition or gesture recognition. By using short video clips, the amount of data required for recognition is reduced, and it becomes easier to train ML models. The resulting dataset can be used for a range of applications such as sports analysis, surveillance, or human–computer interaction. The action tags in the dataset are not in a numerical format. We used one hot encoding as the encoding method. This is because there were very few zero values in the dataset, and hot-encoding encoding is a suitable technique for converting categorical variables into numerical values when there are few unique categories. Next, frames were extracted from videos in the dataset, and these frames were split into training and validation sets. The split was made in such a way that 75% of the frames were used for training the DL model, while the remaining 25% were used for validation or testing.

3.1.2. UCF101 Dataset

UCF101 is a challenging dataset representing realistic activities performed in real life and is unique compared to other datasets depicting activities performed by actors. UCF101 is a data collection of realistic activities in videos taken from YouTube with 101 activity categories. UCF101 is the most complex data set to present, with 13320 videos from 101 activity classes and huge variations in camera motion, object appearances and position, object scale, perspective, background clutter, illumination variation, and so on. Because the majority of available activity recognition data sets are not realistic and are performed by performers, UCF101 intends to inspire further activity recognition research by learning and exploring new realistic activity categories. We only consider videos from this dataset that resemble cricket activities. These are basketball dunk, cricket bowling, table tennis shot, tennis swing, and volleyball spiking. Some of the videos present different illumination conditions, viewpoints, and poses. We present the details in Table 4.

Table 3 Our cricket dataset, which consists of 723 videos divided into five classes. We also provide the number of videos for each class.

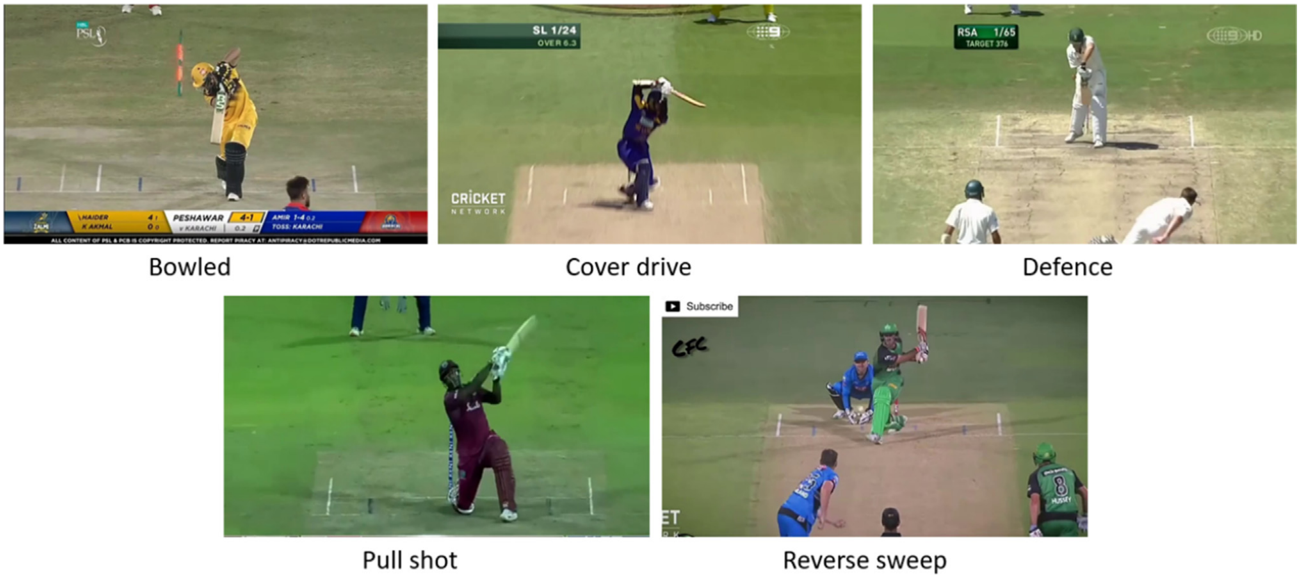
Classes	No. of Videos	Duration (s)	Frame/s
Bowled	150	3:00	30.57
Cover Drive	150	3:00	30.82
Defence	150	3:00	30.97
Pull Shot	150	3:00	30.50
Reverse Sweep	122	3:00	30.91

Table 4 UCF101 Dataset. We provide the details in terms of classes, number of videos, and frame rate.

Classes	No. of Videos	Frame/s
Basketball Dunk	131	25.00
Cricket Bowling	139	26.33
Table Tennis Shot	140	25.00
Tennis Swing	166	29.70
Volleyball Spiking	116	27.00

3.1.3. HMDB51 dataset

The HMDB51 dataset represents human interactions with various physical interactions, facial movements, and body movements. This dataset is significantly challenging because the clips in each category are grouped with different lights for different topics, with 4–6 clips per item performing the same process in different poses and contexts. This HMDB51 dataset includes realistic videos from many sources, like movies and YouTube videos. The dataset contains 6,849 short videos from 51 activity classes (such as ”jump,” ”kickball,” ”laugh,” and golf”), with at least 101 clips in each category. We only con-



Bowled

Cover drive

Defence

Pull shot

Reverse sweep

Fig. 3 Our cricket dataset, consisting of five classes: bowled, cover drive, defence, pull shot, and reverse sweep.

Table 5 HMDB51 dataset. We provide the details regarding classes, the number of videos, and the frame rate.

Classes	No. of Videos	Frame/s
Ball-dribble	145	30.00
Golf	105	30.00
Kickball	128	30.00
Swing baseball	143	30.00
Throw	102	30.00

Table 6 Kinetics dataset. We provide the details regarding classes, the number of videos, and the frame rate.

Classes	No. of Videos	Frame/s
HorseRace	127	27.00
Fencing	111	28.33
Punch	105	27.00
PushUps	106	29.70
BaseballPitch	123	29.00

sider the classes: ball-dribble, golf, kickball, swing-baseball, and throw. The original-evaluation scheme uses two different training/test splits. We provide the details of the selected classes in Table 5.

3.1.4. Kinetics dataset

The Kinetics dataset is a high-quality, large-scale dataset for detecting human behavior in videos. This dataset comprises 400 human activity classes each with at least 400 short video clips. Each clip is about 10 s long and is labeled with a corresponding activity class. The activities are human-centered and include a variety of classes, including human-object interactions like playing an instrument and interactions between people like shaking hands. We consider the five sports classes: horse race, fencing, punch, pushUps, and baseball pitch. The dataset details are in Table 6.

3.1.5. YouTube dataset

The YouTube dataset, which encompasses 11 activity classes, is highly demanding, as it exhibitswide variability in camera movement, object appearance, pose, object scaling, perspective, background noise, and lighting conditions. Videos within each group have certain features in common. We only consider classes representing biking, diving, trampoline-jumping, soccer juggling, and swing, as listed in Table 8.

3.2. Performance Evaluation

We used the F1-score, recall, precision, accuracy, and confusion matrix to analyze the proposed model's performance. Mathematically, the performance metrics are defined as:

$$\textbf{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$\textbf{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\textbf{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\textbf{F}-\textbf{Score} = \frac{2 * Recall * Precision}{Recall + Precision} \tag{4}$$

Accuracy refers to the proportion of samples that are classified correctly out of the total number of samples. Fig. 4 reports accuracy and loss during training and validation stages using the cricket dataset. The specific choice of 60 epochs for training the network is based on experimentation with the dataset and model, where the number of epochs was adjusted to achieve the best balance between training time and model accuracy. Furthermore, the choice of 60 epochs was based on prior experience with similar tasks or datasets. However, it is important to note that the optimal number of epochs can vary depending on the complexity of the model, the size of the dataset, and other factors. Therefore, it is often necessary to experiment with different numbers of epochs to find the best choice for a particular task. As can be seen, our CNN-LSTM model learns the motion patterns efficiently without facing overfitting and underfitting.

The dataset's confusion matrix, shown in Fig. 5, reflects differences between actual and predicted labels. Three activities – bowled, cover drive, and pull shot – show an accuracy of 97%. The reverse sweep class shows lower accuracy equal to 79% due to its complex nature. Precision (Eq. (2)) is the number of accurate outputs generated by the model. We present precision, recall, and f1-score in Table 7 considering the cricket dataset. The recall (Eq. (3)) is the percentage of positive classes predicted correctly by our model out of all positive classes. Because it may be difficult to compare two models that have low precision and high recall, we utilized the F1 score (Eq. (4)) to assess the two measures simultaneously. If the recall equals the accuracy, the F-score is maximized. As can be seen, the higher values for all metrics indicate the effectiveness of the proposed method. We also present the performance in terms of validation accuracy for the UCF101, the HMDB51, the Kinetic, and the Youtube action datasets in Table 9. Our proposed method achieves a validation accuracy equal to 90.03% on the UCF101 dataset. Considering the HMDB51 dataset, we achieve a validation accuracy equal to 89.10%. In fact, our method can learn frame-by-frame changes regardless of viewpoint, pose, and subject. We achieve 86% accuracy and 71.10% accuracy on the kinetic dataset and the Youtube dataset, respectively. In order to demonstrate the robustness of the proposed method using the cricket dataset, we compare the results of our method with reference methods in the literature. The methods are: SimpleRNN [32], Con3D [33], ConLSTM2D [34], Bi-directional LSTM [35], and Gated Recurrent Unit [36]. We present the validation accuracy in Table 10 for the reference methods and our proposed method. Our method outperforms all the reference methods by achieving an accuracy equal to 92.65% considering the cricket activity recognition videos.

4. Conclusion and future scope

In this work, we proposed a cricket activities dataset consisting of five classes (Bowled, cover drive, defence, pull shot, and reverse sweep). The data was pre-processed via cropping and resizing and then organized in relevant classes. Further, an end-to-end time-distributed CNN-LSTM model is developed, where the time-distributed CNN is used for frame-level feature extraction and processed with LSTM layers to learn sequence patterns and make ultimate predictions in the final stage.

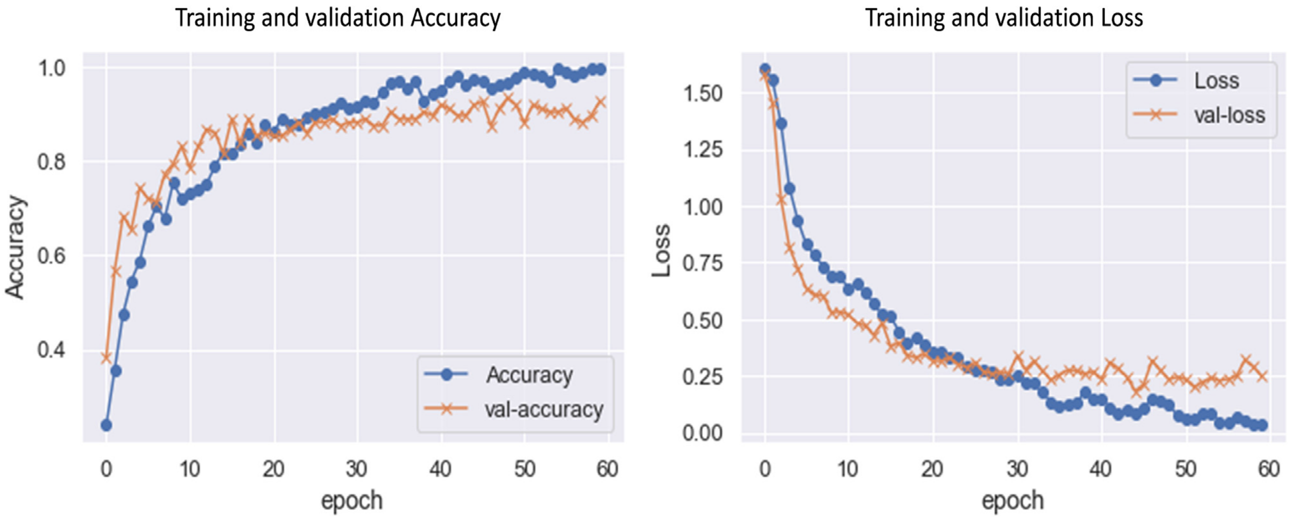


Fig. 4 The CNN-LSTM model. The left graphs show training and validation accuracy highlighted in orange and blue, respectively. The right graphs show training and validation losses highlighted in orange and blue, respectively. We set the batch size equal to 16 and the number of epochs equal to 60.

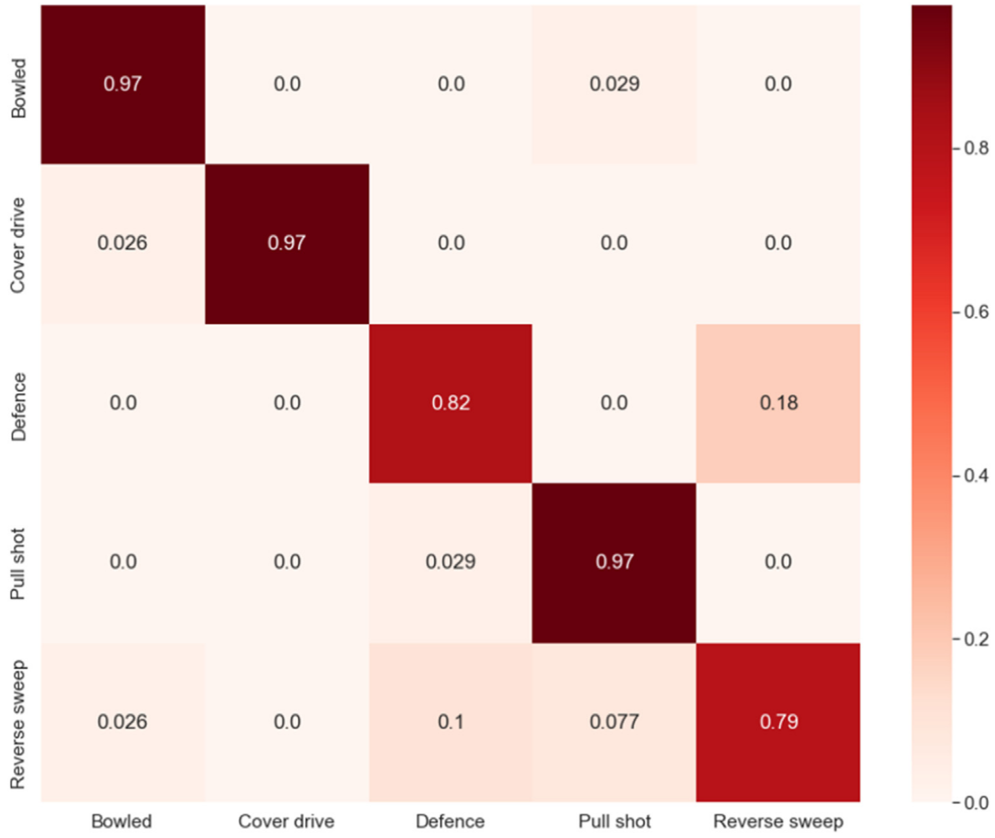


Fig. 5 Confusion matrix. The matrix shows the difference between the actual and predicted labels considering bowled, cover drive, defence, pull shot, and reverse sweep activities. We achieve higher performances for bowled, cover drive, and pull shot activities.

Experiments are performed both on our collected dataset and publicly available datasets. The results showed promising performance of our model in terms of accuracy, recall, precision, and F1 score in comparison with many references as well as baseline techniques used for similar problem-solving. In the future, we would like to extend this work by adding more classes to the current dataset such as fast bowling (fast and spin bowling), Fielder catches, and spin bowling. To achieve better real-time performance, we will conduct experiments by using various attention mechanisms to consider the most

Table 7 Performance of CNN-LSTM. We present the performances in terms of precision, recall, and f1-score using the cricket dataset.

Activities	Precision (%)	Recall (%)	F1-score (%)
Bowled	94	97	96
Cover dive	100	97	99
Defence	84	82	83
Pull shot	89	97	93
Reverse sweep	84	79	82

Table 8 YouTube-Action dataset. We provide the details regarding classes, number of videos, and frame rate.

Classes	No. of Videos	Frame/s
Trampoline-jumping	118	29.00
Soccer-juggling	156	30.00
Biking	145	30.00
Diving	156	29.97
Swing	137	29.40

Table 9 Results of CNN-LSTM under consideration of different datasets.

Datasets	Proposed Method	Accuracy (%)
UCF101	CNN-LSTM	90
HMDB51	-	89
Kinetics	-	86
YouTube-Action	-	71
Cricket	-	92

Table 10 Comparison results with the reference methods considering the cricket Dataset.

Year	Methods	Dataset	Accuracy (%)
2021	SimpleRNN [32]	Cricket	84.71
2022	Con3D [33]	-	86.34
2022	ConLSTM2D [34]	-	73.26
2022	Bi-directional LSTM [35]	-	90.00
2022	Gated Recurrent Unit [36]	-	87.67
2023	**Proposed CNN-LSTM model**	**-**	**92.65**

significant and conspicuous portions of the frame instead processing the whole frame at the later stage of the activity recognition model. Further, to intelligently assess crowd behaviour and dense situations in a stadium, we will combine our proposed method with people counting techniques.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) through the Research Partnership Program No. RP-21-07-02.

References

[1] X. Cui, R. Hu, Application of intelligent edge computing technology for video surveillance in human movement recognition and taekwondo training, Alexandria Eng. J. 61 (2022) 2899–2908.

[2] C. Ladha, N.Y. Hammerla, P. Olivier, T. Plötz, Climbax: skill assessment for climbing enthusiasts, in: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, 2013, pp. 235–244.

[3] R. Montoliu, R. Martín-Félez, J. Torres-Sospedra, A. Martínez-Usó, Team activity recognition in association football using a bag-of-words-based method, Human Move. Sci. 41 (2015) 165–178.

[4] L.N.N. Nguyen, D. Rodríguez-Martín, A. Català, C. Pérez-López, A. Samà, A. Cavallaro, Basketball activity recognition using wearable inertial measurement units, in: Proceedings of the XVI International Conference on Human Computer Interaction, Interacción '15, Association for Computing Machinery, New York, NY, USA, 2015.

[5] M. Ullah, M. Mudassar Yamin, A. Mohammed, S. Daud Khan, H. Ullah, F. Alaya Cheikh, Attention-based lstm network for action recognition in sports, Electronic Imaging 2021 (2021), 302–1.

[6] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, S.W. Baik, Action recognition in video sequences using deep bi-directional lstm with cnn features, IEEE access 6 (2017) 1155–1166.

[7] A. Nadeem, A. Jalal, K. Kim, Automatic human posture estimation for sport activity recognition with robust body parts detection and entropy markov model, Multimedia Tools Appl. 80 (2021) 1–34.

[8] G. Zhang, L. Zhong, Research on volleyball action standardization based on 3d dynamic model, Alexandria Eng. J. 60 (2021) 4131–4138.

[9] H. Khan, I.U. Haq, M. Munsif, S.U. Khan, M.Y. Lee, Automated wheat diseases classification framework using advanced machine learning technique, Agriculture 12 (2022) 1226.

[10] X. Zhen, L. Shao, Action recognition via spatio-temporal local features: A comprehensive study, Image Vis. Comput. 50 (2016) 1–13.

[11] Y. Zhao, H. Guo, L. Gao, H. Wang, J. Zheng, K. Zhang, Y. Zheng, Multifeature fusion action recognition based on key frames, Concurr. Comput.: Pract. Experience (2021) e6137.

[12] D. Karmaker, A. Chowdhury, M. Miah, M. Imran, M. Rahman, Cricket shot classification using motion vector, in: 2015 Second International Conference on Computing Technology and Information Management (ICCTIM), IEEE, 2015, pp. 125–129.

[13] H. Noorbhai, M.M.A. Chhaya, T. Noakes, The use of a smartphone based mobile application for analysing the batting backlift technique in cricket, Cogent Medicine 3 (2016) 1214338.

[14] U. Arora, S. Verma, S. Sahni, T. Sharma, Cricket umpire assistance and ball tracking system using a single smartphone camera, PeerJ Preprints 5 (2017) e3402v1.

[15] S. Yeole, N. Sharma, Y. Shinde, S. Shaikh, Use of image processing techniques for making run out decision in cricket, Int. J. Eng. Comput. Sci. 3 (2014).

[16] A.E. Chowdhury, M.S. Rahim, M.A.U. Rahman, Application of computer vision in cricket: Foot overstep no-ball detection,

in: 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), IEEE, pp. 1–5.

[17] X.-S. Wei, P. Wang, L. Liu, C. Shen, J. Wu, Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples, IEEE Trans. Image Process. 28 (2019) 6116–6125.

[18] H. Kwon, Y. Kim, J.S. Lee, M. Cho, First person action recognition via two-stream convnet with long-term fusion pooling, Pattern Recogn. Lett. 112 (2018) 161–167.

[19] J. Lee, H. Jung, Tuhad: Taekwondo unit technique human action dataset with key frame-based cnn action recognition, Sensors 20 (2020) 4871.

[20] H. Yasin, M. Hussain, A. Weber, Keys for action: an efficient keyframe-based approach for 3d action recognition using a deep neural network, Sensors 20 (2020) 2226.

[21] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, J. Garcia-Rodriguez, A survey on deep learning techniques for image and video semantic segmentation, Appl. Soft Comput. 70 (2018) 41–65.

[22] M. Munsif, H. Afridi, M. Ullah, S.D. Khan, F.A. Cheikh, M. Sajjad, A lightweight convolution neural network for automatic disasters recognition, 2022 10th European Workshop on Visual Information Processing (EUVIP), IEEE, 2022, pp. 1–6.

[23] T.M. Lee, J.-C. Yoon, I.-K. Lee, Motion sickness prediction in stereoscopic videos using 3d convolutional neural networks, IEEE Trans. Visual. Comput. Graph. 25 (2019) 1919–1927.

[24] M. Munsif, M. Ullah, B. Ahmad, M. Sajjad, F.A. Cheikh, Monitoring neurological disorder patients via deep learning based facial expressions analysis, in: IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, 2022, pp. 412–423.

[25] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1933–1941.

[26] Z. Tu, W. Xie, Q. Qin, R. Poppe, R.C. Veltkamp, B. Li, J. Yuan, Multi-stream cnn: Learning representations based on human-related regions for action recognition, Pattern Recogn. 79 (2018) 32–43.

[27] E.P. Ijjina, C.K. Mohan, Hybrid deep neural network model for human action recognition, Appl. Soft Comput. 46 (2016) 936–952.

[28] C.I. Patel, S. Garg, T. Zaveri, A. Banerjee, R. Patel, Human action recognition using fusion of features for unconstrained video sequences, Computers & Electrical Engineering 70 (2018) 284–301.

[29] C. Dai, X. Liu, J. Lai, Human action recognition using two-stream attention based lstm networks, Appl. Soft Comput. 86 (2020) 105820.

[30] H. Gammulle, S. Denman, S. Sridharan, C. Fookes, Two stream lstm: A deep fusion framework for human action recognition, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2017, pp. 177–186.

[31] T. Tsunoda, Y. Komori, M. Matsugu, T. Harada, Football action recognition using hierarchical lstm, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 99–107.

[32] D. Kollias, S. Zafeiriou, Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset, IEEE Trans. Affect. Comput. 12 (2021) 595–606.

[33] M. Qu, J. Cui, T. Su, G. Deng, W. Shao, Video visual relation detection via 3d convolutional neural network, IEEE Access 10 (2022) 23748–23756.

[34] R. Singla, S. Mittal, A. Jain, D. Gupta, Convlstm for human activity recognition, in: International Conference on Innovative Computing and Communications, Springer, 2022, pp. 335–344.

[35] X. Yin, Z. Liu, D. Liu, X. Ren, A novel cnn-based bi-lstm parallel model with attention mechanism for human activity recognition with noisy data, Scient. Rep. 12 (2022) 1–11.

[36] L. Lu, C. Zhang, K. Cao, T. Deng, Q. Yang, A multichannel cnn-gru model for human activity recognition, IEEE Access 10 (2022) 66797–66810.