IJE TRANSACTIONS A: Basics Vol. 37 No. 04, (April 2024) 662-674

International Journal of Engineering

Journal Homepage: www.ije.ir

Umpire's Signal Recognition in Cricket Using an Attention based DC-GRU Network

A. Dey*[a], S. Biswas[a], L. Abualigah[b,c,d,e]

[a] *Department of Computer Science and Technology, Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India*
[b] *Computer Science Department, Prince Hussein Bin Abdullah Faculty for Information Technology, Al al-Bayt University, Mafra, Jordan*
[c] *Department of Electrical and Computer Engineering, Lebanese American University, Byblos, Lebanon*
[d] *Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, Jordan*
[e] *MEU Research Unit, Middle East University, Amman, Jordan*

PAPER INFO

Paper history:
Received 13 August 2023
Received in revised form 02 November 2023
Accepted 04 November 2023

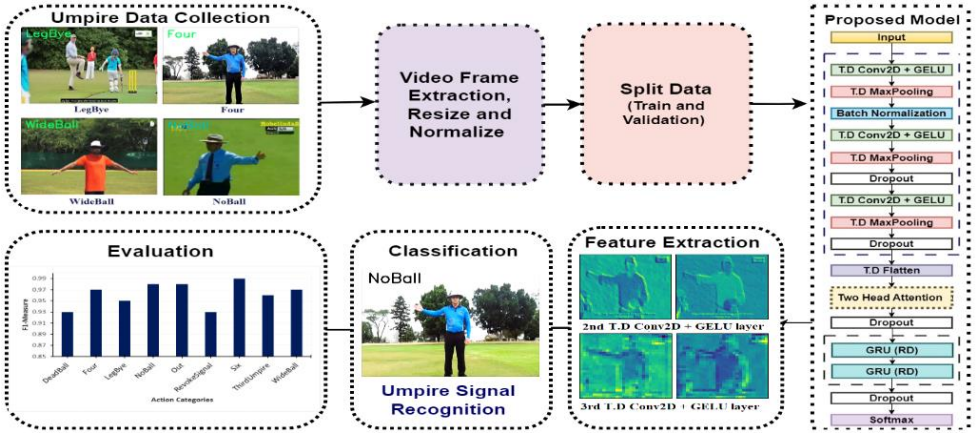
Keywords:
Action Recognition
Umpire Signal
Cricket Game
Attention
DC-GRU
Video Sequences

ABSTRACT

Computer vision has extensive applications in various sports domains, and cricket, a complex game with different event types, is no exception. Recognizing umpire signals during cricket matches is essential for fair and accurate decision-making in gameplay. This paper presents the Cricket Umpire Action Video dataset (CUAVd), a novel dataset designed for detecting umpire postures in cricket matches. As the umpire possesses the power to make crucial judgments concerning incidents that occur on the field, this dataset aims to contribute to the advancement of automated systems for umpire recognition in cricket. The proposed Attention-based Deep Convolutional GRU Network accurately detects and classifies various umpire signal actions in video sequences. The method achieved remarkable results on our prepared CUAVd dataset and publicly available datasets, namely HMDB51, Youtube Actions, and UCF101. The DC-GRU Attention model demonstrated its effectiveness in capturing temporal dependencies and accurately recognizing umpire signal actions. Compared to other advanced models like traditional CNN architectures, CNN-LSTM with Attention, and the 3DCNN+GRU model, the proposed model consistently outperformed them in recognizing umpire signal actions. It achieved a high validation accuracy of 94.38% in classifying umpire signal videos correctly. The paper also evaluated the models using performance metrics like F1-Measure and Confusion Matrix, confirming their effectiveness in recognizing umpire signal actions. The suggested model has practical applications in real-life situations such as sports analysis, referee training, and automated referee assistance systems where precise identification of umpire signals in videos is vital.

***doi**: 10.5829/ije.2024.37.04a.08*

Graphical Abstract



*Corresponding Author Email: arnabdeycs@gmail.com (A. Dey)

Please cite this article as: Dey A, Biswas S, Abualigah L. Umpire's Signal Recognition in Cricket Using an Attention based DC-GRU Network. International Journal of Engineering, Transactions A: Basics. 2024;37(04):662-74.

1. INTRODUCTION

Umpires play a critical role in cricket, ensuring fair play and making important decisions throughout a match. They serve as impartial arbiters, upholding the game's rules and regulations. Their judgments not only ensure fairness but also boost player confidence, enhance the spectator experience, and contribute to the smooth functioning of cricket tournaments. Cricket umpire signals play a vital role in maintaining fairness and making accurate decisions during matches (1). The use of computer vision techniques to recognize and analyze these signals in videos has practical applications across various aspects of cricket. Identifying cricket umpire signals using deep learning can provide valuable insights (2) to players, coaches, and spectators. By integrating umpire signal recognition with broadcast technologies, real-time graphics and overlays can be generated, improving the visual presentation of the game and helping viewers understand the decision-making process. Recognizing umpire signals from video sequences is complex, but advancements in deep learning techniques offer promising solutions. The proposed Two-Head Attention based Deep Convolutional Gated Recurrent Unit (DC-GRU) network shows promise in accurately identifying cricket umpire signals. The proposed model's lightweight design makes it suitable to run on devices with low memory requirements. This means that even general-purpose desktops or laptops can effectively run the proposed model without the need for extensive computational resources. Furthermore, the proposed model can robustly detect and interpret umpire signals. This technology aims to enhance decision-making accuracy and minimize human errors in crucial match situations. In cricket, umpires play a pivotal role in making critical on-field decisions, communicating through unique hand signals and gestures. From sports analysis to referee training and automated referee assistance systems, the applications are diverse and have the potential to significantly enhance the quality of cricket matches, benefiting players, officials, and spectators alike. Recognizing the umpire's signals in cricket is challenging due to similarities and complexities in postures and gestures. One significant obstacle is the lack of publicly accessible datasets designed explicitly for classifying umpire's signals. Due to the absence of an existing video dataset for umpire's signals, we have taken the initiative to create a new video dataset comprising 1179 videos showcasing umpire's signals. The model under consideration has undergone training using a custom dataset that we created. The proposed DC-GRU Attention model mainly classifies the videos into nine umpire signal classes: DeadBall, Four, LegBye, NoBall, Out, RevokeSignal, Six, ThirdUmpire and WideBall.

This research offers the following notable contributions: a) Development of a new video dataset for Cricket Umpire Action (CUAVd) with proper annotations. b) Recognition of Umpire Signals in videos is carried out using the proposed Two-Head Attention based DC-GRU network. c) The proposed DC-GRU Attention model has been evaluated over three standard related benchmark datasets, namely Youtube Actions, HMBD51, and UCF101 action datasets during experimentations, resulting in a high recognition performance of 93.82%, 83.67%, and 92.65%, respectively. d) The effectiveness of the proposed approach is assessed through rigorous evaluation using various well-known classifiers.

2. RELATED WORKS

In this section we delve into the realm of prior research endeavors concerning the domain of action recognition. Conventional machine learning (ML) based action recognition systems typically follow a three-step process: feature extraction using manually crafted feature descriptors, feature representation, and feature classification using suitable ML algorithms. In their research, Shi et al. (3) presented a new local spatiotemporal descriptor known as gradient boundary histograms (GBH). The authors showed that the GBH descriptor surpasses other gradient-based descriptors in representing both local structure and motion. The paper focuses on action recognition and introduces the GBH descriptor as a spatiotemporal feature. They have attained 63.2% accuracy on HMDB51 and 86.6% on UCF101 dataset. Challenges associated with handcrafted-based action recognition methods include time-consuming feature selection, labor-intensive processes, and difficulties in determining suitable features (4). To overcome the shortcomings and challenges of handcrafted-based methods, researchers turned to deep learning to develop effective and innovative approaches for cutting-edge action recognition systems based on videos. Deep learning techniques directly analyze videos to identify human actions, using an end-to-end approach. The spatiotemporal and the two-stream networks stand out among deep learning techniques. Xin et al. (5) introduced an adaptive recurrent-convolutional hybrid (ARCH) network. Their approach effectively handles variations in the spatial and temporal domains, as well as intra- and inter-class diversities. This architecture incorporates Temporal-Spacial fusion-Convolutional Neural Networks (CNN) to gather local static and dynamic information, and Recurrent neural network (RNN) focus on global sequence pattern modeling. The seamless connectivity between local feature extraction and global pattern modelling enhances the network's adaptability to actions with varying speeds and durations. Simonyan and Zisserman (6) presented a two-stream CNN model comprising a spatial stream, responsible for

processing individual video frames to capture spatial information, and a temporal stream, dedicated to capturing temporal information within the video sequences. Xiong et al. (7)further extended this concept by proposing a transferable two-stream CNN technique that combined motion and spatial features. The model was trained on the UCF101 dataset, yielding satisfactory results. Wang et al. (8) have implemented action recognition by employing Temporal Segment Network (TSN). This approach uses unique segment-based sampling and aggregation techniques to mimic the patterns and relationships that exist over a longer period of time in a temporal sequence. TSNs divide the input video into multiple segments and independently classify each segment. The classification scores obtained from all segments are then combined to generate the final output. Li et al. (9) introduced a new architecture named VideoLSTM, designed specifically for learning action sequences from videos. VideoLSTM addresses the distinctive characteristics of video data by incorporating convolutions to leverage spatial relationships within images, and it integrates a shallow CNN to capture motion information and produce attention maps-based motion maps. Ge et al. (10) introduced CNN-LSTM with an attention model to recognize human actions. GoogleNet is used to extract features from video sequences. A spatial transformer network is then applied to focus on important regions by transforming the feature maps. The convolutional LSTM module captures sequential information for action classification. Redundant features extracted by GoogleNet are reduced using temporal coherence analysis during training, maintaining high accuracy. Minhas et al. (11) introduced a method to classify shots in field sports videos utilizing AlexNet CNN. The proposed approach achieved 94.07% accuracy. In another study, Rafiq et al. (12) presented a method to classify sports videos and video summarization utilizing transfer learning. By employing an AlexNet CNN-based approach on a relatively smaller dataset exclusively comprised of cricket scenes, they attained an impressive accuracy of 99.26%. Sanchez-Caballero et al. (13) have proposed a real-time method to detect human actions. Their method utilizes a 3D-CNN that can automatically extract spatio-temporal patterns from unprocessed depth sequences. The 3DFCNN classifies the activities based on the spatial and temporal data from depth sequences, without the need to identify people's identities, ensuring privacy is maintained. Savadi Hosseini et al. (14) have introduced a hybrid deep learning architecture that merges the multiple GRU layers with a two-stream inflated CNN network to address action recognition challenges. By combining the strengths of gated recurrent unit (GRU) layers and the two-stream inflated 3D Convolutional neural network (3DCNN), the proposed hybrid architecture enables the model to process video data, extract both local and global features, and improve its understanding of the temporal and spatial characteristics present in the video content. Kavimandan et al. (15) presented a methodology aimed at enhancing the recognition accuracy of actions utilizing only one camera in multi-camera environments. They suggested a modified bag-of-visual-words method, employing Support Vector Machine (SVM) to detect and categorize human actions. Foysal et al. (16) put forward a CNN based model designed to categorize six distinct cricket shots. Using their curated dataset, they have attained good results. In our earlier research concerning the the identification of workout-related actions from images (17), we have achieved a validation accuracy of 92.75% on the WAId dataset prepared by us, and an accuracy of 89% on the Sports Image dataset using the proposed WorkoutNet architecture. In another research (18), we successfully identified diverse human interactions from image data by implementing the AdaptiveDRNet, enhanced with a multi-level attention mechanism. Wu et al. (19) have provided a survey that delves into the realm of video-based sports action recognition, shedding light on the existing datasets and methodologies in this field. Li et al. (20) introduced a framework that tackles the challenge of selecting important spatial parts and modelling temporal motion in action recognition. Their approach involves selecting the most discriminative spatial parts within video clips and effectively modelling temporal motion by incorporating bidirectional temporal information across multiple layers of an LSTM model. Hussain et al. (21) introduced a convolution-free approach that successfully overcomes previous challenges by accurately encoding relative spatial information. The proposed method employs a pre-trained Vision Transformer to extract features at the frame level. These extracted features are subsequently fed into a multilayer LSTM network, enabling the capture of long-range dependencies within videos. Ravi et al. (22) proposed a new image dataset called SNOW, aimed at detecting umpire poses in cricket. The study's main objective was to identify and categorize four important umpire events: WIDE, OUT, NO BALL and SIX. To extract relevant features, the researchers employed pre-trained CNN models like InceptionV3 and VGG19 to extract features from the umpires pose images and then fed them into a linear SVM classifier. Ahmad et al. (23) proposed a technique to recognize human actions by employing a combination of CNN and Bidirectional-GRU (BD-GRU). The approach involves two main steps. Firstly, they utilized CNN to extract deep features from the frame sequences in human activity videos. Secondly, to capture the temporal dynamics within the sequence of frames, the researchers introduce Bi-GRU. The deep and informative features extracted from the frame sequence are fed into the Bi-GRU, which learns the temporal

motions in both forward and backward directions at every time step. In addition, Wickramasinghe (24) conducted a comprehensive review of the diverse applications of machine learning within the realm of cricket. Reddy and Santhosh (25) propose a two-stream spatial CNN method for recognizing human actions in long video datasets. The first stream extracts spatial information from the RGB frames, while another stream incorporates graph-based visual saliency (GBVS) maps produced using the GBVS technique. The outputs of these spatial streams are concatenated utilizing various feature fusion techniques such as sum, max, average, and product. Pan et al. (26) have proposed a sensor based approach for sports referee training. Several recent studies have placed emphasis on enhancing the detection of cricket events (27, 28) and the summarization of cricket videos (29, 30). Nandyal and Kattimani (31) prepared a new dataset named SNWOLF, designed specifically to identify umpire poses in cricket matches. The proposed method focuses on detecting umpire stances from cricket video frames and classifying them into the six events namely WIDE, LEG BYE, FOUR, SIX, NO BALL and OUT using CNN based framework. They have achieved an accuracy of 98.20% on their prepared SNWOLF data.

The majority of studies in the domain of image or video-based action recognition tend to primarily focus on the recognition of everyday actions (14) or the classification of sports events (32). However, due to the absence of a publicly available standardized video dataset for umpire's signal in cricket, we have taken the initiative to develop our own comprehensive Cricket Umpire Action Video Dataset (CUAVd). This dataset serves as a benchmark resource for evaluating and advancing the field of umpire signal recognition in cricket, providing researchers with a valuable tool for further exploration and development.

3. PROPOSED METHOD

The method commenced by gathering videos featuring umpire actions and subsequently trimming them based on the action performed. And then data augmentation is applied. The frame extraction function receives a video path as input and extracts frames from the video. It utilizes the Video Capture object to read the video, determines the total number of frames within the video, and calculates the interval at which frames will be added to the frames list. The frames will be extracted based on a specified sequence length. It then iterates through the frames, resizes them to 64 X 64 height and width, normalizes them by dividing by 255, and appends the normalized frames to the frames list. Finally, it releases the Video Capture object and returns the frames list containing the extracted frames. Next, the dataset is divided into training and validation set in 75:25 ratio. The block diagram illustrating the sequential steps involved in the proposed method is depicted in Figure 1. This study introduces a Two-Head Attention based Deep Convolutional GRU (DC-GRU) network for umpire's signal recognition in cricket.

Utilizing the proposed DC-GRU Attention model with Callback functions, we train the umpire dataset, extract video frame features, and classify umpire signals with the last layer of the proposed model namely the Softmax layer. The proposed approach effectively analyze both spatial and temporal features extracted from video data, resulting in accurate action predictions. The following subsections explain: A) Training dataset preparation, B) DC-GRU Attention: Model Architecture, C) Loss Function and Optimizer Details, and D) Importance of Attention Mechanism in Umpire Signal Recognition.

3. 1. Training Dataset Preparation Initially, we collected the umpire signal dataset and used data augmentation to enhance the model's performance by increasing the training data and reducing overfitting. We have implemented random transformations as part of video data augmentation, including zooming within a range of 0.2 factor, rotation range of 20 degrees, adjusting brightness within a range of 0.5 to 0.9 and applying a contrast range of 1.5. To generate augmented video data, we have utilized the Moviepy python library. Data augmentation, through random transformations applied to the training videos, generates additional training data. This technique perform effectively on unseen data.

3. 2. DC-GRU Attention: Model Architecture The proposed Two-Head Attention based DC-GRU Model comprises of a total of 16 layers, including Time Distributed (T.D) Conv2D layers, T.D max pooling layers, T.D dropout layers, T.D Flatten layer, Two Head Attention mechanism, GRU layers and a dense output layer. The proposed model utilized for implementing the proposed work to recognize umpire signals in cricket is depicted in Figure 2. Time-distributed (T.D) layers give more importance to temporal dynamics in video based umpire signal recognition by applying convolution and pooling operations across the time dimension.

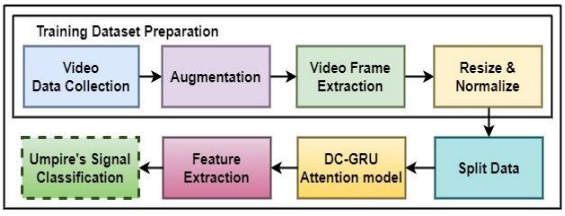


Figure 1. Overview of the Proposed Method

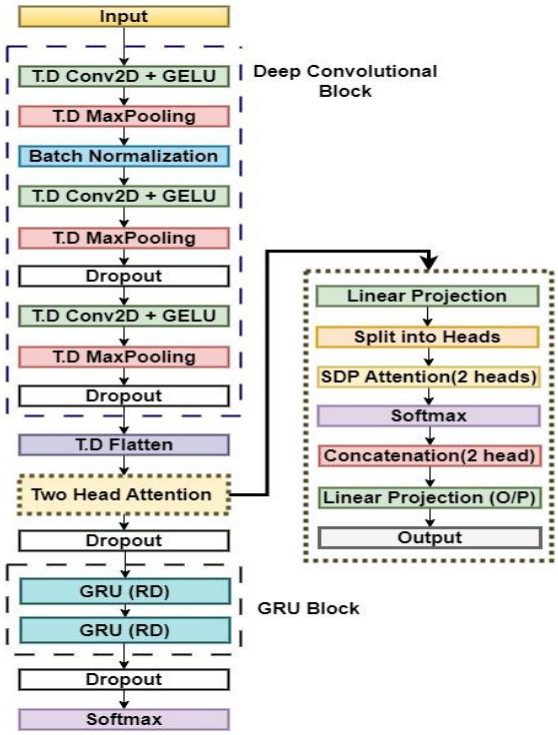


Figure 2. Proposed DC-GRU Attention Model

The breakdown of the model architecture are as follows:

3. 2. 1. Input Layer The input layer accepts a sequence of video frames with a shape of (SeqLength, FrameHeight, FrameWidth, nChannels), where SeqLength represents the number of frames considered as a sequence, FrameHeight and FrameWeight are the dimensions of each frame and nChannels are the number of channels. Here, the SeqLength, FrameHeight, FrameWidth are taken as 15, 64, 64 and 3 respectively.

3. 2. 2. Deep Convolutional Block This block applies several convolutional layers to each frame in the input sequence, capturing spatial features from the frames. This branch uses three sets of convolutional layers, with 32, 64, and 64 filters, respectively. In the first block, the input sequences of video frames are processed through three layers. The initial Time distributed (T.D) Conv2D+GELU layer (Layer 1) applies 32 filters with a 3x3 kernel size to capture local features in each frame. A TimeDistributed (T.D) wrapper allows each frame in the sequence to be treated independently. The T.D Conv2D layer is mathematically represented using Equation 1. Here, I (t-i) denotes the input sequence at time step (t-i), W_i denotes the convolutional kernel at time step i and Y(t) represent the output at time step t.

$$Y(t) = \sum_i W_i * I(t-i) \quad (1)$$

Following this, a MaxPooling2D layer (Layer 2) with a pool size of 4x4 reduces spatial dimensions, emphasizing key features while reducing computation. Each T.D convolutional layer is followed by an Gaussian Error Linear Unit (GELU) activation function and max pooling to reduce the spatial dimensions. The GELU activation (GA) utilized in this study is mathematically represented using Equations 2 and 3, where u denote the input to the activation function. GA offers a smooth transition from linear to non-linear behavior, addressing vanishing gradient problems while providing superior performance.

$$GA\ (u) = 0.5\ u\ (1 + \tanh(CT)) \quad (2)$$

$$CT = \left(\sqrt{\frac{2}{\pi}}\ (u + 0.044715\ u^3)\right) \quad (3)$$

BatchNormalization (Layer 3) is used to normalize the batches. The subsequent layers continues feature extraction with deeper layers. Another T.D Conv2D+GELU layer (Layer 4) applies 64 filters, maintaining the 3x3 kernel size for local feature extraction. Subsequently, a MaxPooling2D layer (Layer 5) further compresses feature maps spatially, improving computational efficiency. A Dropout (0.15 rate) layer (Layer 6) is applied for regularization, randomly deactivating portions of the network during training to prevent overfitting. In this block, the complexity of the features is increased. And the third T.D Conv2D+GELU layer (Layer 7) applies 64 filters to capture more intricate patterns. MaxPooling2D (Layer 8) further reduces spatial dimensions, while Dropout (0.15 rate) (Layer 9) maintains regularization. Then T.D flatten layer (Layer 10) is added to reshape the 2D feature maps into a 1D vector format.

3. 2. 3. Two Head Attention Block After feature extraction, the sequence of flattened features enters an attention mechanism (Layer 11). The proposed model for recognizing umpire actions in video sequences incorporates a pivotal element known as the scaled dot product mechanism with two attention heads. In a two-head attention mechanism, two distinct sets of query, key, and value vectors facilitate simultaneous focus on different sequence parts. Each set acts as an individual head, allowing the model to learn distinct relationships and extract various sequence features. Computation occurs in parallel across these two heads, and their results are usually combined before progressing through subsequent model layers. This method significantly improves the model's capability to capture diverse sequence relationships and patterns.

The dimentionality of the model (d_model) is set to 128, which defines the size of hidden layers of the Two-

Head attention. Here's a breakdown of this Two-Head Attention mechanism:

1. Utilizing Query, Key, and Value Vectors with Two Attention Heads: The scaled dot product (SDP) mechanism employs three distinct sets of vectors: Queries (R), Keys (K) and Values (U). In the specific context of the umpire action recognition model, these vectors are derived from the flattened features that are extracted by the preceding convolutional layers of the model. Linear Projections for R, K and U are represented using Equation 4. Here, I denote the input sequence, h is the number of attention heads (two here), and W_R, W_K, W_U are learnable weight matrices specific to R, K, V and it is computed for h=1, and h=2.

$$R, K, U = I.W_{R_h}, I.W_{K_h}, I.W_{V_h} \quad (4)$$

2. Dot Product and Scaling with Two Attention heads: This mechanism proceeds by computing the dot product between the Query and Key vectors. Attention scores (AS) are computed using the expression given in Equation 5. This computation quantifies the similarity between various elements in the sequence. To maintain the dot product within manageable bounds, it scales down the values by the square root of the dimension of the K vectors (dim_k) as shown in Scaled AS expression represented in Equation 6. AS_h is computed for h=1,2.

$$AS_h = R_h K_h^T \quad (5)$$

$$Scaled\ AS_h = \frac{Q_h K^T}{\sqrt{dim_K}} \quad (6)$$

3. Attention Weights (AW) with Two Attention Heads: Subsequently, the scaled dot product mechanism applies a softmax function to the SDP, producing the attention weights (AW) as represented in Equation 7. These weights signify the significance and relevance of each element within the sequence relative to the others. Higher attention weights denote that a given element bears greater importance in the context of the task at hand. AW_h is computed for h=1, and h=2.

$$AW_h = Softmax\ (\ Scaled\ AS_h) \quad (7)$$

4. Weighted Summation with Two Attention Heads: Finally, the mechanism computes a weighted sum of the Value vectors (U) by leveraging the attention weights (AW) represented using Equation 8. This weighted summation effectively captures vital sequence information, allowing focused attention on crucial segments while disregarding noise. AO represents the Attention Output for h=1, and h=2, resulting in concatenated weighted sum outcomes.

$$AO_h = WS_h = (AW_h).U_h \quad (8)$$

In the realm of umpire action recognition, this mechanism equips the model with the ability to discern which frames or moments in the video sequence hold the utmost relevance for identifying specific umpire actions. By directing its attention towards the pertinent segments of the video data, the model can furnish more precise predictions. The utilization of self-attention mechanisms, such as the scaled dot product attention with two attention heads, emerges as a potent tool for modeling intricate relationships within sequential data. Scaled Dot Product (SDP) Attention layer computes attention scores between the feature vectors from different frames within the sequence. This operation applies attention to the features, assigning higher importance to elements that have higher attention scores. This step results in attention-enhanced features that are more focused on the relevant parts of the sequence. This process is followed by a Dropout layer with 0.18 dropout rate to promote regularization.

3. 2. 4. GRU Block Two GRU layers (Layers 13 and 14) with 128 units each and recurrent dropout (0.23) process the attention-enhanced features, capturing temporal dependencies among the frames in the video sequence. The GRU, a variant of the recurrent neural network (RNN), excels in capturing temporal dependencies and sequence-related information. It operates by processing the current input and the hidden state inherited from the preceding node. This process results in the generation of output, and the updated hidden state is subsequently transmitted to the subsequent node. GRU only needs one unit to complete forget and selecting memory operations. The updated equation of GRU can be represented as depicted in Equations 9 and 10, where X_t is the external input vector, the update gate is denoted as U and reset gate as rs, the hidden state from the previous node (hs^{t-1}) are fed through GRU, the updated hidden state (hs^t) which is then passed to the next node, the parameters for two matrices W and U, bias vector b.

$$\mathrm{hs}^t = (1 - U) \odot hs^{t-1} + U \odot hs' \quad (9)$$

$$hs' = \tanh(W_h\ X_t + (rs_t * U_{hs} hs_{t-1}) + b_{hs}) \quad (10)$$

In this context, the initial GRU layer, comprising 128 units, is harnessed for the purpose of capturing temporal dependencies embedded within the sequence. Subsequently, a second GRU layer, also featuring 128 units, is employed to derive the ultimate output sequence. In the model, the initial GRU layer is configured to return sequences rather than a single output. This empowers the model to grasp temporal nuances spanning across frames while preserving the output sequence. The second GRU layer does not return sequences, so it returns a single output. Then a dropout layer (Layer 15) with 0.23 dropout rate is applied.

3. 2. 5. Output Layer Dense layer with softmax activation (Layer 16) is used to produce the final output probabilities for the nine umpire's signal classes. This layer maps the concatenated features to the corresponding class probabilities. The sum of those probabilities, expressed in decimal form, equals 1. In this study, we are dealing with a multi-class problem. During training, the model employs the Categorical cross-entropy loss function, serving as the metric to gauge the error the model endeavors to reduce.

The proposed Attention based DC-GRU model offers significant utility in recognizing umpire's signals within video sequences, leveraging the distinct strengths of its constituent layers. Time-Distributed (T.D) Conv2D layers perform spatial feature extraction, using 2D convolutional filters that slide across frames, capturing spatial patterns, edges, and textures. As the filters convolve over the frames, they generate feature maps. Each feature map signifies the activation patterns of a distinct filter across various spatial positions within individual frames. These feature maps capture spatial information such as local patterns, object boundaries, and spatial relationships within the frames. The time-distributed convolutional layers are responsible for extracting spatial features from each frame within the sequence. They operate on each frame independently within the sequence. The feature map(s) derived from the three T.D Conv2D+GELU layers when applied to a sample frame containing the 'No Ball' umpire signal are highlighted in Figure 3. Multiple layers hierarchically abstract spatial features, ranging from low-level patterns to complex spatial structures.

Shared weights across frames ensure consistency over time, crucial for recognizing umpire signals. This is particularly useful for tasks that require understanding how spatial patterns evolve across a sequence. The integration of an attention mechanism further enhances the model's ability to focus on relevant frames where signals are likely to occur for making predictions. On the other hand, the recurrent layers, particularly the GRU components, specialize in capturing temporal dependencies and context across frames present within the sequence. The combination of convolutional layers, attention mechanism, and recurrent layers allows the model to extract both spatial features from individual frames and temporal features from the entire sequence. The proposed Two-Head Attention based DC-GRU architecture is particularly effective in modeling short-term dependencies and intricate temporal dynamics. By combining these layers, the proposed model achieves a comprehensive understanding of umpire signals, encompassing both the complex temporal evolution of these signals. Its ability to learn intricate spatial and temporal features, coupled with its capacity to focus on relevant segments, positions it as a valuable tool for automating the recognition of umpire's signals, thereby enhancing the efficiency and accuracy of sports match analysis and decision-making. With a total number of trainable parameters of 297,097 and a model size of approximately 1.13 MB, the proposed model is suitable for running on low-memory requirement devices like general purpose desktops or laptops.

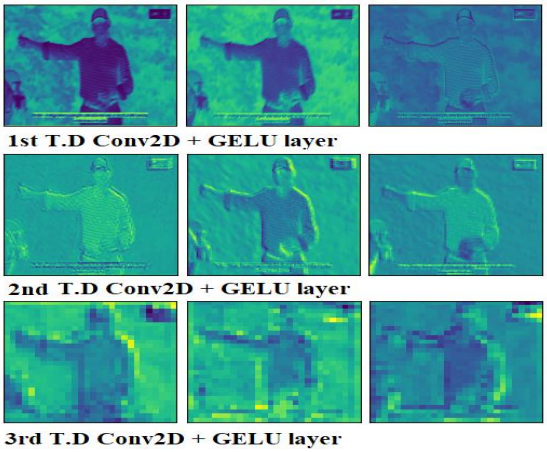


Figure 3. Feature map results of 'No Ball' umpire signal

3. 3. Loss Function and Optimizer Details In the proposed work, we are dealing with a multi-class problem. To measure the training error, we utilized the Categorical Cross-Entropy (CCE) loss function. This function minimizes the disparity between expected and actual probability distributions, computing the error as shown in Equation 11 for each instance.

$$Loss_{cce} = \sum_{c=1}^{9} t_c . \log f(U\,S_c) \quad (11)$$

Here, t_c denotes the ground truth label of c^{th} class, $f(U\,S_c)$ denotes the softmax probability value for c^{th} class, and c signifies the overall count of scalar values contained within the model's output. The proposed model utilizes the Nadam optimizer with initial learning rate of 0.01 to minimize the error function. The Nadam optimizer (33) combines elements of two popular optimization methods, namely Nesterov Accelerated Gradient (NAG) and Adam optimizer.

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t}+\epsilon}\left(\beta_1 \frac{m_t}{1-\beta_1} + \frac{(1-\beta_1)\nabla L(\theta_t)}{1-\beta_1^t}\right) \quad (12)$$

In Equation 12, θ_{t+1} denotes the parameter updated at the next time step, t+1, while θ_t represents the parameter at the current time step, t. The symbol η corresponds to the learning rate, determining the step size of the parameter updates. Furthermore, $\hat{v}_t$ signifies the exponentially weighted average of the squared gradients, offering insights into both the direction and magnitude of the parameter updates, ϵ serves as a minuscule value essential for averting division by zero, β_1 is the first momentum decay term, m_t is the exponentially weighted

average of the gradients, and $\nabla L(\theta_t)$ represents the gradient of the loss function w.r.t the parameter at time step t.

The Nadam optimizer exhibits several advantages compared to other optimizers such as Adam. Firstly, it incorporates NAG, which enables faster convergence by including momentum in the gradient descent process. This momentum helps the optimizer to navigate areas with steep gradients more efficiently, leading to quicker convergence and improved optimization. Secondly, Nadam combines the benefits of NAG with the adaptive learning rate capabilities inherent in the Adam optimizer. This dynamic learning rate adjustment feature enables the optimizer to fine-tune the learning rates for individual parameters, guided by their gradients and past updates.

3. 4. Importance of Attention Mechanism in Umpire Signal Recognition Attention mechanisms are of paramount importance when it comes to the identification and recognition of umpire signals, such as "FOUR", "SIX", "NO BALL", "WIDE", and others in the context of cricket matches. The diverse array of umpire signals involves distinct visual representations, each conveying specific game events. For instance, signaling a "FOUR" entails raising both arms with all fingers extended, while a "SIX" is typically indicated by both arms raised above the head. The ability to recognize and differentiate these unique visual cues is fundamental in comprehending the dynamics of the ongoing match. Umpire signals sometimes involve subtle hand movements, which can be challenging to detect, particularly when the frame includes other players or objects. Attention mechanisms help the model to concentrate on the umpire's hand or the pertinent area, improving recognition. They enhance robustness to lighting, camera angles, and clutter for effective operation in various conditions. They also allocate computational resources efficiently, reducing processing costs. In complex cases with gesture combinations, attention mechanisms excel at identifying components and their sequence, ensuring accurate signal interpretation. Thus, attention mechanisms are indispensable in the recognition of umpire signals in cricket matches for dynamic focus, capturing dependencies, enhance robustness to variations and ensuring context-aware recognition, making them invaluable for diverse and context-dependent umpire signals in cricket game.

4. EXPERIMENTAL RESULTS

This section provides a comprehensive overview of the dataset, conducts a thorough performance evaluation, and finally analyses the results. The study was conducted on the Google Colab platform without utilizing any GPU. This research has involved a comprehensive investigation of relevant literature and the exploration of various well-established deep learning models. In order to evaluate the overall classification performance of our approach, we have reported the training and validation accuracy, along with the F1-score.

4. 1. Dataset Details Several publicly accessible datasets that are related to the proposed work include HMDB51, UCF50, Youtube Actions and UCF101. These datasets provide valuable resources for conducting research and analysis in the area.

We have created a unique dataset called Cricket Umpire Action Video dataset (CUAVd)[1] by gathering videos from different social media platforms, cricket tournaments and some YouTube videos. Once the data collection phase was completed, we manually identified and extracted the specific segments of the videos that showcase umpire actions. These videos were then organized into separate directories based on different categories to ensure proper labelling. The dataset primarily focuses on umpire signals performed by various umpires during cricket matches. Figure 4 illustrates sample video frames showcasing the LegBye, Four, WideBall, and NoBall umpire signals extracted from the umpire signals video dataset.

In the CUAVd dataset, various umpire signals in cricket have been given by various umpires. Within the proposed dataset, there are 9 distinct categories encompassing various umpire actions, and it comprises a total of 1179 RGB videos. The umpire's signal categories are DeadBall, Four, LegBye, NoBall, Out, RevokeSignal, Six, ThirdUmpire and WideBall. The training dataset is collected from various Cricket tournaments, TV episodes

Figure 4. Sample video frames of CUAVd dataset

[1] **CUAVd :** https://sites.google.com/view/cuavd/home

and Youtube videos. Data collection is a vital step in the proposed work for maintaining the integrity of the research due to the unavailability of various umpire action videos.

4. 2. Performance Evaluation Python libraries such as TensorFlow-Keras, Callbacks, and Optimizers have been employed to construct the models on the Google Colab platform. These libraries offer a wide array of functionalities and tools that facilitate creating and training deep learning models. In addition, the Sklearn and Matplotlib libraries have been used for analysis. Sklearn offers various machine learning algorithms and evaluation metrics, while Matplotlib provides visualization capabilities, allowing for the creation of informative and visually appealing plots. Figure 5 exhibits the accuracy curve for the Two-Head Attention based DC-GRU model under consideration, with a training accuracy of 96.17% and a corresponding validation accuracy of 94.38%.

The model's performance did not significantly improve over several consecutive epochs, so the training process was halted at 24 epochs using callback functions.

The proposed study has undergone evaluation using several advanced benchmark deep learning models on the CUAVd dataset. Table 1 showcases the accuracy and F1-Measure achieved by these models, namely VidLSTM [9], CNN-LSTM with Attention [10], 3DFCNN [13], BD-LSTM [20], the DC-LSTM, the DC-GRU model (without attention), and our proposed model. Through extensive experimentation and analysis, it was observed that our proposed DC-GRU Attention model outperformed all the baseline models considered in this study. Moreover, the BD-LSTM [20], 3DFCNN [13], DC-GRU model (without attention), and CNN-LSTM with Attention model [10] also displayed noteworthy performance in classifying Cricket Umpire action videos (CUAVd), achieving validation accuracies of 91.82%, 89.69%, 93.86%, and 90.35% respectively.

TABLE 1. Experimental Results with Various deep learning models on our CUAVd dataset

Model	Train Accuracy	Validation Accuracy	F1-Score
VidLSTM [9]	94.20%	87.40%	0.87
CNN-LSTM + Attention [10]	93.57%	90.35%	0.90
3DFCNN [13]	94.61%	89.69%	0.89
BD-LSTM [20]	95.36%	91.82%	0.91
DC-LSTM	94.85%	91.43%	0.90
DC-GRU	95.69%	93.86%	0.94
Proposed Model	96.17%	94.38%	0.96

The classification performance report depicted in Figure 6 indicates that the proposed model demonstrates strong performance, achieving an impressive Average F1-Score of 0.96.

In Figures 7 and 8, the confusion matrices for both the DC-GRU model and our proposed model when tested on our CUAVd dataset is shown. Notably, our proposed model outperforms the DC-GRU model (without Attention) in terms of classification accuracy.

The Receiver Operating Characteristic (ROC) curve for the proposed DC-GRU Attention model on the CUAVd dataset is depicted in Figure 9. Notably, all umpire signal categories achieve a perfect AUC of 1, indicating ideal performance. This is evident in intersecting ROC curves, showcasing optimal performance for each category. The Attention-based DC-GRU Model demonstrates outstanding classification performance, with each color on the ROC curve representing a specific umpire signal class, offering valuable insights into its proficiency at different classification thresholds.

Our proposed model has been evaluated on various video-based action recognition datasets using established performance metrics, generating promising results. To

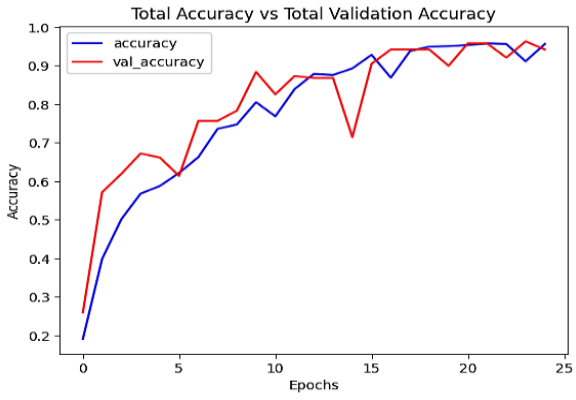


Figure 5. Model Accuracy Trend on CUAVd Dataset

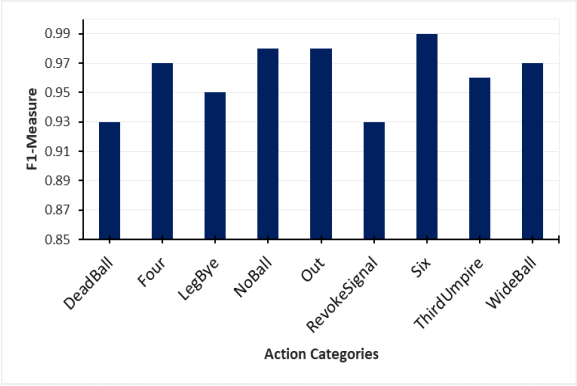


Figure 6. Classification accuracy of proposed model on CUAVd

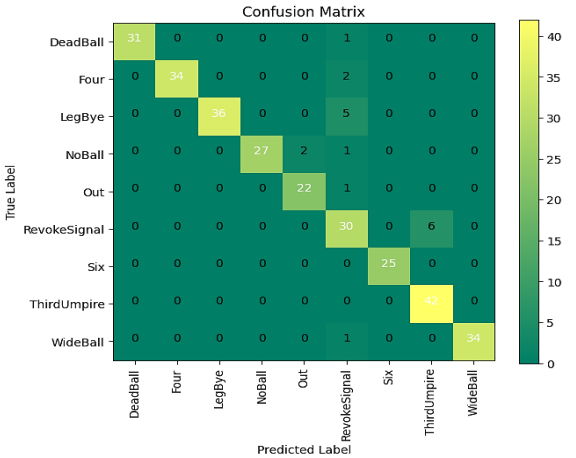


Figure 7. Confusion Matrix of DC-GRU model on CUAVd

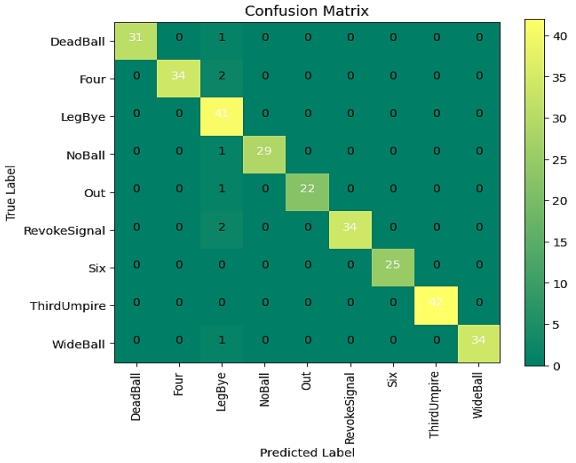


Figure 8. Confusion Matrix of Proposed model on CUAVd

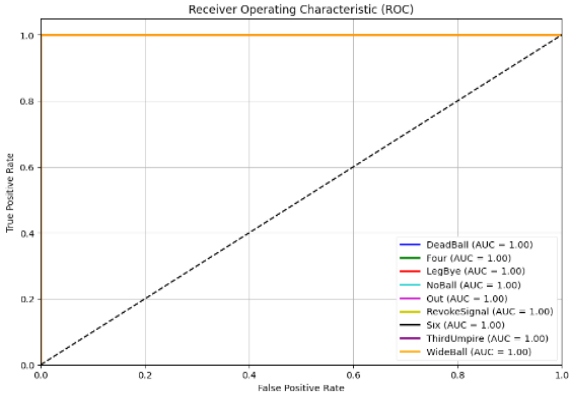


Figure 9. ROC curve of proposed model on CUAVd

thoroughly assess the effectiveness of our proposed approach, we computed F1-score, precision and recall values across different related action datasets. A detailed summary of these values can be found in Table 2. Notably, our proposed model performs well when tested on well-known datasets, namely Youtube Actions, HMDB51, UCF50 and UCF101. The precision and recall scores achieved by our technique are well-balanced across the benchmark datasets, highlighting a reduced number of both the true and false negatives. Additionally, the proposed Two-Head Attention based DC-GRU model achieved impressive F1-measures of 0.83, 0.93, 0.94, and 0.91 for the HMDB51, UCF50, Youtube Actions, and UCF101 respectively. The F1-score obtained illustrates its effectiveness, further reinforcing its superiority in action recognition tasks.

The performance of our proposed DC-GRU Attention model was further assessed by introducing random noise to our CUAVd dataset, as illustrated in Table 3. The model's performance showed a minor decline with the introduction of noise in the data, achieving an F1-score of 0.91. In contrast, the DC-GRU model without attention attain F1-score of 0.88.

Confusion matrix of Youtube Actions and UCF50 dataset utilizing proposed model is depicted in Figures 10 and 11, respectively. It shows remarkable performance.

The efficacy of the proposed DC-GRU Two-Head Attention model is evaluated by comparing it with a range of existing standard approaches. Table 4 presents the outcomes of this comparison, showcasing the performance of different deep learning-based models on related benchmark action datasets, namely HMDB51, YouTube Actions, and the UCF101 dataset.

The model proposed in this study demonstrated an impressive accuracy rate of 93.82% when evaluated on the YouTube Actions dataset. Our method demonstrated

TABLE 2. Evaluating the Proposed Model's Performance on various Existing Action Datasets

Dataset	Action Classes	Net Precision	Net Recall	Net F1-Score
Youtube Actions	11	0.94	0.95	0.94
HMDB51	51	0.83	0.84	0.83
UCF50	50	0.93	0.94	0.93
UCF101	101	0.92	0.91	0.91
Our CUAVd Dataset	9	0.97	0.95	0.96

TABLE 3. Experimental Results with Effect of random Noise on our CUAVd dataset

Model	Train Accuracy	Validation Accuracy	F1-Score
DC-GRU	90.39%	88.26%	0.88
Proposed Model	92.57%	90.42%	0.91

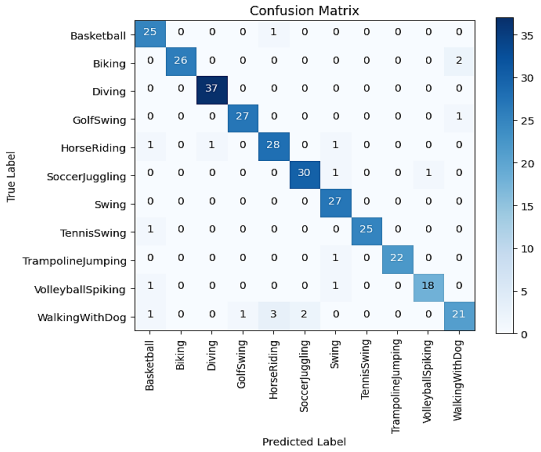


Figure 10. Confusion Matrix of Youtube Actions dataset utilizing Proposed model

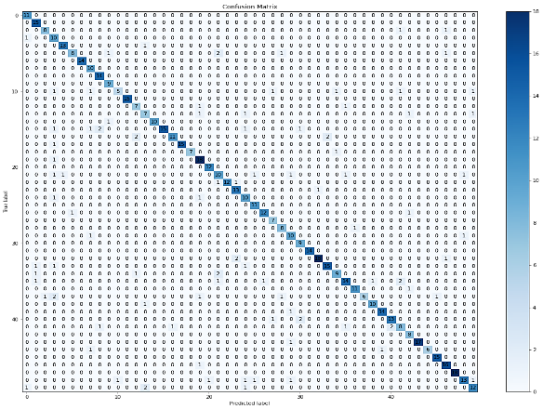


Figure 11. Confusion Matrix of UCF50 dataset using Proposed model

TABLE 4. Comparative Analysis of Various Benchmark Video based Action Recognition methods

Model	HMDB51	Youtube Actions	UCF101
GBH+BoW [3]	63.2%	83.40%	83%
ARCH [5]	58.2%	---	85.30%
CNN Two Stream [6]	76%	88.46%	84.65%
TSN [8]	71%	---	94.90%
VidLSTM [9]	56.40%	---	92.20%
CNN-LSTM + Attention [10]	67.10%	93.48%	92.50%
3DFCNN [13]	72.54%	89.18%	87.41%
3DCNN+ GRU [14]	81.87%	86.64%	---
BD-LSTM [20]	70.40%	91.80%	94.20%
ViT+LSTM [21]	73.71%	---	96.14%
Bi-GRU [23]	71.89%	93.28%	91.79%
Proposed Model	**83.67%**	**93.82%**	**92.65%**

high classification scores on both the YouTube Actions dataset and HMDB51 dataset, yielding remarkable accuracies. The ViT+LSTM [21] achieved the highest accuracy of 96.14%, and TSN [8] model achieved an accuracy of 94.90% on UCF101 dataset. Out of all the models evaluated, the proposed DC-GRU Attention model secured exceptional accuracy on the UCF101 dataset, ranking among the top three performers with an impressive accuracy of 92.65%. For the YouTube Actions dataset, our proposed model achieved the topmost accuracy of 93.82%, followed by the CNN-LSTM with Attention [10] and Bi-GRU [23] and BD-LSTM [20]. In terms of the HMDB51 dataset, the ViT+LSTM [21], 3DFCNN [13], CNN Two-Stream Fusion [6] and 3DCNN+GRU [14] achieved accuracies of 73.71%, 72.54%, 76% and 81.87% respectively.

However, ARCH [5] and GBH+BoW [3] achieved accuracies of 58.2% and 63.2%, respectively, on the HMDB51 dataset. In comparison, our proposed technique achieved 83.67% accuracy on HMDB51, which surpassed all others, securing the top position in Table 4. The proposed model has undergone rigorous testing with multiple video samples, demonstrating its ability to accurately classify umpire signals by correctly labelling the videos using the trained model. The classification result on a realistic video sample with the 'Six' umpire signal category is indicated in Figure 12.

The experimental findings indicate that the proposed model excels in accurately detecting the umpire's signal from the videos. So, the proposed model can also be utilized to detect the umpire signals from the video sequences in real-life.

5. CONCLUSION AND FUTURE SCOPE

Cricket umpire's signal recognition in videos is a captivating and rapidly evolving field within computer vision. The ability to automatically identify and classify umpire signal actions in cricket matches has immense practical applications. It contributes to fair decision-making, enhances the game flow, and provides valuable

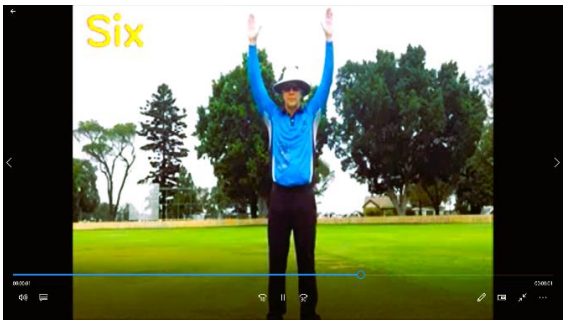


Figure 12. Test Video of 'Six' umpire signal recognized using proposed model

insights for players, coaches, and spectators. The proposed Two-Head Attention based Deep Convolutional GRU framework has shown great promise in accurately recognizing and classifying umpire signal actions. With a high accuracy rate of 94.38% and a Net F1-Score of 0.96, this model stands out as an effective solution. This framework can serve as a virtual trainer for novice umpires, assisting them in practising their cricket umpire signal actions effectively. A significant contribution of this research is the creation of the Cricket Umpire Action Video dataset (CUAVd), which provides a comprehensive collection of videos showcasing the umpire's signals in cricket. This dataset enables the training and evaluation of the proposed DC-GRU Attention model. The model's lightweight design ensures efficient performance even on devices with limited memory, enabling its widespread usage. The proposed approach has been extensively tested and evaluated using three well-established benchmark action datasets, consistently showcasing remarkable recognition performance. Moreover, the effectiveness of the proposed model has been confirmed through successful testing on a diverse range of sample videos, accurately identifying and labelling the umpire signals.

The future scope of this research lies in refining the proposed model, expanding the dataset size to encompass a wider range of umpire signal actions, generalizing to other sports, integrating with sports technology, collaborating with sports organizations, and addressing ethical considerations. Further optimization of the model's architecture and network combinations can enhance its performance.

6. REFERENCES

1. Oslear D. Wisden's The Laws Of Cricket: Random House; 2010.
2. Naik BT, Hashmi MF, Bokde ND. A comprehensive review of computer vision in sports: Open issues, future trends and research directions. Applied Sciences. 2022;12(9):4429. 10.3390/app12094429
3. Shi F, Laganiere R, Petriu E, editors. Gradient boundary histograms for action recognition. 2015 IEEE Winter Conference on Applications of Computer Vision; 2015: IEEE. 10.1109/WACV.2015.152
4. Kong Y, Fu Y. Human action recognition and prediction: A survey. International Journal of Computer Vision. 2022;130(5):1366-401. 10.1007/s11263-022-01594-9
5. Xin M, Zhang H, Wang H, Sun M, Yuan D. Arch: Adaptive recurrent-convolutional hybrid networks for long-term action recognition. Neurocomputing. 2016;178:87-102. 10.1016/j.neucom.2015.09.112
6. Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems. 2014;27. 10.48550/arXiv.1406.2199
7. Xiong Q, Zhang J, Wang P, Liu D, Gao RX. Transferable two-stream convolutional neural network for human action recognition. Journal of Manufacturing Systems. 2020;56:605-14. 10.1016/j.jmsy.2020.04.007
8. Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, et al. Temporal segment networks for action recognition in videos. IEEE transactions on pattern analysis and machine intelligence. 2018;41(11):2740-55. 10.1109/TPAMI.2018.2868668
9. Li Z, Gavrilyuk K, Gavves E, Jain M, Snoek CG. Videolstm convolves, attends and flows for action recognition. Computer Vision and Image Understanding. 2018;166:41-50. 10.1016/j.cviu.2017.10.011
10. Ge H, Yan Z, Yu W, Sun L. An attention mechanism based convolutional LSTM network for video action recognition. Multimedia Tools and Applications. 2019;78:20533-56. 10.1007/s11042-019-7404-z
11. Minhas RA, Javed A, Irtaza A, Mahmood MT, Joo YB. Shot classification of field sports videos using AlexNet Convolutional Neural Network. Applied Sciences. 2019;9(3):483. 10.3390/app9030483
12. Rafiq M, Rafiq G, Agyeman R, Choi GS, Jin S-I. Scene classification for sports video summarization using transfer learning. Sensors. 2020;20(6):1702. 10.3390/s20061702
13. Sanchez-Caballero A, de López-Diz S, Fuentes-Jimenez D, Losada-Gutiérrez C, Marrón-Romera M, Casillas-Perez D, et al. 3dfcnn: Real-time action recognition using 3d deep neural networks with raw depth information. Multimedia Tools and Applications. 2022;81(17):24119-43. 10.1007/s11042-022-12091-z
14. Savadi Hosseini M, Ghaderi F. A hybrid deep learning architecture using 3d cnns and grus for human action recognition. International Journal of Engineering, Transactions B: Applications,. 2020;33(5):959-65. 10.5829/ije.2020.33.05B.29
15. Kavimandan PS, Kapoor R, Yadav K. Human action recognition using prominent camera. International Journal of Engineering, Transactions B: Applications,. 2021;34(2):427-32. 10.5829/ije.2021.34.02b.14
16. Foysal MFA, Islam MS, Karim A, Neehal N, editors. Shot-Net: A convolutional neural network for classifying different cricket shots. Recent Trends in Image Processing and Pattern Recognition: Second International Conference, RTIP2R 2018, Solapur, India, December 21–22, 2018, Revised Selected Papers, Part I 2; 2019: Springer. 10.1007/978-981-13-9181-1_10
17. Dey A, Dutta A, Biswas S, editors. Workoutnet: A deep learning model for the recognition of workout actions from still images. 2023 3rd International Conference on Intelligent Technologies (CONIT); 2023: IEEE. 10.1109/CONIT59222.2023.10205926
18. Dey A, Biswas S, Le D-N. Recognition of Human Interactions in Still Images using AdaptiveDRNet with Multi-level Attention. International Journal of Advanced Computer Science and Applications. 2023;14(10).
19. Wu F, Wang Q, Bian J, Ding N, Lu F, Cheng J, et al. A survey on video action recognition in sports: Datasets, methods and applications. IEEE Transactions on Multimedia. 2022. 10.1109/TMM.2022.3232034
20. Li W, Nie W, Su Y. Human action recognition based on selected spatio-temporal features via bidirectional LSTM. IEEE Access. 2018;6:44211-20. 10.1109/ACCESS.2018.2863943
21. Hussain A, Hussain T, Ullah W, Baik SW. Vision transformer and deep sequence learning for human activity recognition in surveillance videos. Computational Intelligence and Neuroscience. 2022;2022. 10.1155/2022/3454167
22. Ravi A, Venugopal H, Paul S, Tizhoosh HR, editors. A dataset and preliminary results for umpire pose detection using SVM classification of deep features. 2018 IEEE Symposium Series on

Computational Intelligence (SSCI); 2018: IEEE. 10.1109/SSCI.2018.8628877

23. Ahmad T, Wu J, Alwageed HS, Khan F, Khan J, Lee Y. Human Activity Recognition Based on Deep-Temporal Learning Using Convolution Neural Networks Features and Bidirectional Gated Recurrent Unit With Features Selection. IEEE Access. 2023;11:33148-59. 10.1109/ACCESS.2023.3263155
24. Wickramasinghe I. Applications of machine learning in cricket: a systematic review. Machine Learning with Applications. 2022;10:100435. 10.1016/j.mlwa.2022.100435
25. Reddy P S, Santhosh C. Multimodal Spatiotemporal Feature map for Dynamic Gesture Recognition from Real Time Video Sequences. International Journal of Engineering, Transactions B: Applications,. 2023;36(8):1440-8. 10.5829/ije.2023.36.08B.04
26. Pan T-Y, Tsai W-L, Chang C-Y, Yeh C-W, Hu M-C. A hierarchical hand gesture recognition framework for sports referee training-based EMG and accelerometer sensors. IEEE Transactions on Cybernetics. 2020;52(5):3172-83. 10.1109/TCYB.2020.3007173
27. Ahmad W, Munsif M, Ullah H, Ullah M, Alsuwailem AA, Saudagar AKJ, et al. Optimized deep learning-based cricket activity focused network and medium scale benchmark. Alexandria Engineering Journal. 2023;73:771-9. 10.1016/j.aej.2023.04.062
28. Das S, Mahmud T, Islam D, Begum M, Barua A, Tarek Aziz M, et al. Deep Transfer Learning-Based Foot No-Ball Detection in Live Cricket Match. Computational Intelligence and Neuroscience. 2023;2023. 10.1155/2023/2398121
29. Shingrakhia H, Patel H. SGRNN-AM and HRF-DBN: a hybrid machine learning model for cricket video summarization. The Visual Computer. 2022;38(7):2285-301. 10.1007/s00371-021-02111-8
30. Raval KR, Goyani MM. A survey on event detection based video summarization for cricket. Multimedia Tools and Applications. 2022;81(20):29253-81. 10.1007/s11042-022-12834-y
31. Nandyal S, Kattimani SL. Cricket Event Recognition and Classification from Umpire Action Gestures using Convolutional Neural Network. International Journal of Advanced Computer Science and Applications. 2022;13(6). 10.14569/IJACSA.2022.0130644
32. Siddiqui HUR, Younas F, Rustam F, Flores ES, Ballester JB, Diez IdlT, et al. Enhancing Cricket Performance Analysis with Human Pose Estimation and Machine Learning. Sensors. 2023;23(15):6839. 10.3390/s23156839
33. Zhang Q, Zhang Y, Shao Y, Liu M, Li J, Yuan J, et al. Boosting Adversarial Attacks with Nadam Optimizer. Electronics. 2023;12(6):1464. 10.3390/electronics12061464

Persian Abstract

چکیده

بینایی کامپیوتر کاربردهای گسترده ای در حوزه های مختلف ورزشی دارد و کریکت، یک بازی پیچیده با انواع رویدادهای مختلف، از این قاعده مستثنی نیست .تشخیص سیگنال های داور در طول مسابقات کریکت برای تصمیم گیری منصفانه و دقیق در گیم پلی ضروری است .این مقاله مجموعه داده ویدیوی اکشن داور کریکت (CUAVd) را ارائه می‌کند، یک مجموعه داده جدید که برای تشخیص وضعیت‌های داور در مسابقات کریکت طراحی شده است .از آنجایی که داور دارای قدرت قضاوت های مهم در مورد حوادثی است که در زمین رخ می دهد، هدف این مجموعه داده کمک به پیشرفت سیستم های خودکار برای تشخیص داور در کریکت است .شبکه GRU کانولوشن عمیق مبتنی بر توجه پیشنهادی، اقدامات سیگنال داور مختلف را در دنباله‌های ویدیویی به دقت شناسایی و طبقه‌بندی می‌کند .این روش به نتایج قابل‌توجهی در مجموعه داده‌های CUAVd آماده‌شده و مجموعه‌های داده در دسترس عموم، یعنی HMDB51 ، Youtube Actions و UCF101 ، دست یافت .مدل توجه DC-GRU اثربخشی خود را در گرفتن وابستگی های زمانی و تشخیص دقیق اقدامات سیگنال داور نشان داد .در مقایسه با سایر مدل‌های پیشرفته مانند معماری‌های سنتیCNN ، CNN-LSTM با توجه، و مدل3DCNN+GRU ، مدل پیشنهادی به طور مداوم از آنها در تشخیص اقدامات سیگنال داور برتری داشت .در طبقه‌بندی صحیح ویدیوهای سیگنال داور، دقت اعتبار بالای94.38 ٪ را به دست آورد .این مقاله همچنین مدل‌ها را با استفاده از معیارهای عملکرد مانند F1-Measure و Confusion Matrix ارزیابی کرد و کارایی آنها را در تشخیص اقدامات سیگنال داور تأیید کرد .مدل پیشنهادی دارای کاربردهای عملی در موقعیت‌های واقعی مانند آنالیز ورزشی، آموزش داور، و سیستم‌های خودکار کمک داور است که در آن شناسایی دقیق سیگنال‌های داور در ویدئوها حیاتی است.