Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH

Open Access

Recognizing sports activities from video frames using deformable convolution and adaptive multiscale features

Lei Xiao[1], Yang Cao[2], Yihe Gai[1], Edris Khezri[3*], Juntong Liu[4] and Mingzhu Yang[1]

Abstract

Automated techniques for evaluating sports activities inside dynamic frames are highly dependent on advanced sports analysis by smart machines. The monitoring of individuals and the discerning of athletic pursuits has several potential applications. Monitoring individuals, detecting unusual behavior, identifying medical issues, and tracking patients within healthcare facilities are examples of these applications. An assessment of the feasibility of integrating smart real-time monitoring systems across a variety of athletic environments is provided in this study. Motion and activity detection for recording sporting events has advanced due to the need for a large amount of both real-time and offline data. Through the use of deformable learning approaches, we extend conventional deep learning models to accurately detect and analyze human behavior in sports. Due to its robustness, efficiency, and statistical analysis, the system is a highly suitable option for advanced sports recording detection frameworks. It is essential for sports identification and administration to have a comprehensive understanding of action recognition. An accurate classification of human activities and athletic events can be achieved through the use of a hybrid deep learning framework presented in this study. Using innovative methodologies, we conduct cutting-edge research on action recognition that prioritizes users' preferences and needs. It is possible to reduce the error rate to less than 3% by using the recommended structure and the three datasets mentioned above. It is 97.84% accurate for UCF-Sport, 97.75% accurate for UCF50, and 98.91% accurate for YouTube. The recommended optimized networks have been tested extensively compared to other models for recognizing athletic actions.

Keywords Action recognition, Sports activities, Deformable convolution network, Hierarchical multiscale deformable, Attention module

Introduction

In the domain of machine comprehension of sports, the significance of advanced automated methodologies for evaluating activities in motion frames cannot be overemphasized [1]. Techniques capable of detecting human movement and recognizing sport events have a diverse array of possible applications. The applications encompass a range of functions, such as monitoring individuals, detecting atypical or questionable conduct, investigating medical ailments, and tracking patients' movements within medical facilities [2, 3]. Various methodologies, including the examination of cinematic portrayals of actual events, have been suggested as a means of utilizing

*Correspondence:
Edris Khezri
edris.khezri@qiau.ac.ir
[1] Chengdu Technological University, Chengdu 610000, China
[2] Xindu Research Institute of Educational Science, Chengdu 610000, China
[3] Department of Computer Engineering, Boukan Branch, Islamic Azad University, Boukan, Iran
[4] Department of Physical Education, Chengdu University of Information Technology, Chengdu 610000, China

video surveillance for the purpose of monitoring sport events up to this point [4, 5]. Empirical evidence indicates that the perception of bodily motion can transcend a singular modality, encompassing the movement of different body segments, such as the upper and lower extremities [6]. This phenomenon has been seen in multiple individuals who have demonstrated satisfactory performance. Relying just on static photos is inadequate in capturing the whole intricacy of a dynamic sequence of events. The examination of human body movements captured in individual frames, together with their interactions with the surrounding environment, is crucial for the identification and comprehension of complete activities within video images [7].

The increasing popularity of video analysis has led to a surge in research efforts focused on utilizing video processing techniques for monitoring sports performance. This field of study focuses on enhancing individuals' overall welfare and existence standard. The topic of activity monitoring has frequently been conceptualized as a process of activity recognition in prior research [8, 9].

Sport activity identification algorithms evaluate video frames from statistical data, including factors such as frame rate per second and real-time monitoring in dynamic contexts [10, 11]. A multitude of databases exist to record the diverse range of activities individuals engage in throughout a day. Shifts may be categorized into two distinct classes: regular, anticipated shifts, and irregular, unforeseen shifts, also called anomalous shifts [12, 13]. The UCF Sports dataset [14] provides a diverse range of authentic activities performed in a natural setting, establishing its significance as one of the pioneering datasets for action recognition. The whole UCF Sports dataset is subjected to analysis, whereby the employed methodologies and their corresponding statistical discoveries are thoroughly examined. The study findings indicate that some action detection algorithms perform exceptionally when applied to sports footage. This talk delves into an exploration of many of these methods. Nevertheless, there are still significant study gaps that have not been adequately addressed as a result of the constraints inherent in current studies. Furthermore, the comprehensive examination of sports motions has proven challenging due to the distinctive attributes inherent in each specific activity. Numerous physical activities require the recognition of individuals in densely populated environments. Moreover, activities like leaping, plunging, swaying, and striking can undermine precision to a certain extent.

Artificial intelligence (AI) is a very interesting topic within computer science, with its primary objective being the development of intelligent machines. The primary objective is to develop computers that possess the cognitive abilities and independent decision-making capabilities necessary to do tasks traditionally attributed to human beings [15]. Computer Vision [16], is a specialized domain within artificial intelligence that focuses on the examination and interpretation of visual data, such as images and videos. Its primary objective is to replicate humans' cognitive skills to extract valuable information and comprehend the underlying content of these visual inputs. Feature extraction and classification algorithms play a crucial role in the analysis of data gathered from sport events, regardless of the specific type of data being considered. Based on the referenced literature [17], it has been shown that neural networks (NNs) and support vector machines (SVMs) serve as proficient main classifiers inside systems that depend on human feature extraction. Significant progress has been made in the domain of image categorization through the utilization of convolutional neural networks (CNNs), which draw inspiration from the hierarchical structure of the human visual cortex [18]. Feature extraction and classification allow CNN to autonomously extract distinctive characteristics from training videos. The analysis of action representations and the extraction of pertinent components has the potential to significantly enhance the enhancement of sport-related activities.

Computer vision presents several challenges, especially in the domain of human activity identification. Historically, frame-by-frame video analysis has been used to monitor and record athletes' movements in competitive environments. There is a need for frameworks that are more context-specific, accurate, and efficient, along with the inclusion of supplemental data, images, and video frames. Deep learning is widely recognized in this domain for its exceptional efficacy and significant computational capabilities. CNNs, sometimes referred to as automated techniques, have been increasingly employed in various research endeavors aimed at the recognition of human behavior in video data. However, there are numerous limitations that affect these systems' capacity to properly handle a significant volume of video frames in real-time. Current research is now investigating the potential applications of human activity recognition in many team and individual sports, including handball, volleyball, hockey, baseball, soccer, basketball, and other related fields. The identification of an athlete's movements is a significant difficulty in light of the dynamic nature of sporting events. These events are characterized by high speeds and the execution of several maneuvers by players as they navigate and transition between positions [19].

The primary objective of this study is to examine the feasibility of using real-time monitoring techniques across several sports. The proposed approach relies heavily on the utilization of a network of video

sensors strategically placed in diverse and ever-changing surroundings. To facilitate tracking, our focus will be on the creation of multispectral control videos. The demand for substantial quantities of real-time and offline data has stimulated the emergence of innovative concepts in motion and activity recognition pertaining to recorded sprot events. Our methodology for identifying human behavior in sports surpasses conventional instructional approaches by including deformable learning techniques. In a movie activity recognition related sport, low-resolution frames are preferrable for presenting several frames in a movie activity recognition related sport, despite their difficulty in recognizing individuals or objects. For storage or network transfer, low-resolution images can be compressed better. Using our proposed structure and sport-related action recognizer, we propose a novel feature extraction and learning methodology. As a result of its robustness, efficiency, and statistical analysis, the technique is a very viable contender for improving activity identification in athletic events of the same type. Detailed analysis of a few of the significant findings from the investigation follows.

1) A comprehensive understanding of action recognition is required for accurate identification and efficient administration of athletic activities. This paper presents a hybrid deep learning system that accurately identifies and classifies human activities and athletic events.
2) We provide innovative and user-centric methodologies for our cutting-edge research on action recognition. By incorporating a learning mechanism, error rates can be reduced to less than 3% using the recommended design.
3) There is a difference between the method described in this study and the conventional scaling methodologies commonly used in deep CNN designs. In contrast, it uses a hybrid scaling strategy to ensure consistency in its architecture. For improved generalization, CNN architectures can be modified with arbitrary dimensions, such as width, height, and resolution.
4) Tests have been conducted to evaluate the resilience of the proposed optimized networks in comparison to other sport-related action recognizer models. A deformable convolution network is used to enhance feature extraction and classification using adaptive multiscale feature generators and improved attention mechanisms. In most approaches to human action recognition (HAR), high-quality frames are emphasized. As a result, alterations to image dimensions or resolution may lead to error in recognition.

The applied framework for our research endeavors is based on an organizational structure. In the second section, a summary of pertinent research is provided. In Sect. 3, we provide a novel feature extraction and learning methodology that leverages our proposed structure and sport-related action recognizer. The experimental findings derived from the proposed methodology for video frame analysis are documented in Sect. 4. The research conclusion is then accompanied by a concise overview of its principal issues, which are further expounded upon in Sect. 5.

Related work

When many participants, often representing conflicting teams, are present on the playing field, the situation gets more intricate. It is imperative to monitor and appreciate these individuals' activities. Furthermore, players can simultaneously perform many tasks, albeit with a delay, ensuring each athlete does a distinct activity at the same time. Challenges associated with the recognition of multiple objects in the presence of occlusion and crowded environments [20] manifest when individuals in a scene exhibit movement relative to the camera, obstruct each other's visibility, enter or exit the camera's visual range, and similar scenarios occur.

In the study [1] conducted by the authors, a comparison is made between CNN, Multilayer Perceptron (MLP), and LSTM models. These models are evaluated using a custom dataset specifically created to differentiate various behaviors observed in handball scenarios. Furthermore, the temporal aspect of the data is considered throughout the analysis. The dataset's foundation consists of over three thousand annotated videos, encompassing many categories such as dribbling, passing, shooting, throwing, and others. Given the considerable variation in the duration required to complete different activities, the researchers examined the resulting results by manipulating the number of input frames. They used strategies such as frame reduction. The researchers determined that the MLP-based model yielded the most favorable findings. Additionally, they observed that increasing the number of frames could enhance action recognition accuracy in this particular case.

HAR may be categorized by deep learning-based techniques, including but not limited to automatic feature extraction from frames, description, and classification. Currently, the prevailing methods used are deep learning-based techniques [21]. Although deep learning eliminates human feature extraction, a substantial dataset is still necessary for the learning process to estimate several parameters in the hidden layers. This network learns to recognize labeled actions [22]. CNN and Long Short-Term Memory (LSTM) are extensively employed

as models for Deep Learning-based HAR in sports [23]. Two-stream CNNs have been utilized in sports applications due to their ability to handle multiple streams of input and process specific network layers separately [24]. This approach has been demonstrated in several sports-related studies [25, 26]. Moreover, studies often integrate many methodologies when considering the temporal aspect of human action recognition [27].

The use of feedback connections in recurrent neural networks (RNNs) enables the effect of past time step input activations on the present input output, in contrast to the employment of feedforward connections in conventional neural networks (CNNs). RNNs include a notable attribute that renders them very suitable for modeling sequences, such as video frames. Long-Short-Term Memory (LSTM), RNNs, and Long-term temporal convolutions can be used for simulating time sequences [28].

Taekwondo, a discipline classified as an individual sport, has been identified as a domain in which the HAR methodology has been used [29]. In accordance with this approach, the researchers limited their investigation of tennis player behavior to two specific actions, namely left and right swings. The challenge arises mostly when the player character's height is limited to 30 pixels, particularly in frames designed for distant viewing. The development of tennis action categorization via transudative transfer learning was described by [30]. The authors utilized Histogram of Oriented Gradient (HOG) 3D features to depict activities in their study. To facilitate transfer, they developed two distinct techniques: one included re-weighting features, while the other involved translating and scaling features. The researchers engaged in an analysis of a proprietary dataset, which was not publicly accessible, in order to classify theatrical performances into three distinct categories: "serve," "hit," or "non-hit." The focus of recent scholarly investigations has been on the identification of actions at a micro-level within the domain of tennis. The Inception neural network, which has been trained using a distinct dataset, represents films as sequences of features. In classification, three-layered LSTM networks are trained. The THETIS dataset is used to evaluate the models.

In order to achieve automated recognition of ten fundamental badminton motions, a group of studies [31] integrated a sensor chip into a badminton racquet. The most favorable outcomes were achieved using AFEB-AlexNet, whereby a distinct module for adaptive feature extraction was proposed, followed by AlexNet, and lastly, LSTM networks.

The Siamese Spatio-temporal Convolutional neural network was employed by the author of [32] to successfully classify 20 distinct table tennis stroke movements, exhibiting little variability across the different categories. They compute Optical Flow using RGB images obtained from the TTStroke-21 dataset. This dataset comprises 129 videos, lasting 94 h, capturing table tennis matches. The dataset was submitted to the MediaEval Challenge in 2020. In the proposed Siamese network design, data fusion occurs within a fully connected layer following the application of three spatial-temporal convolutions. The study [33] proposed a multi-stage deep neural network pipeline for stroke type detection in table tennis, utilizing spatial-temporal properties. This pipeline utilizes many methodologies, including RGB image-based, Optical Flow-oriented, pose-oriented, and region-of-interest analysis, to create predictions about the final class. Each stage of the pipeline focuses on different factors to get diverse results. The optimal prediction is derived from the TTStroke-21 dataset through the fusion of outcomes collected at each stage. Ultimately, RGB images integrated with Optical Flow-based methodologies produce the most accurate results. The SoccerNet dataset, [33], has been specifically curated for activity identification in online soccer videos. Goals, cards, and substitutions are among the methods used to categorize annotated activities. The collection has 6637 distinct activity instances. Feature extraction was performed using a 3D CNN, an Inflated 3D CNN (I3D CNN) [35], and a Residual Network (ResNet). The study utilized several deep learning models, including Custom CNN, SoftDBOW, NetFV, NetVLAD, and NetRVLAD, in addition to employing diverse pooling approaches such as mean pooling and max pooling [36].

In the study conducted by [37], action recognition techniques were employed to condense extensive soccer footage into concise summaries. By employing a LSTM network trained on extracted soccer features, as well as a ResNet based on 3D-CNN, the researchers successfully achieved recognition of the five distinct actions (centerline, corner-kick, free-kick, goal action, and throw-in) outlined in their designated dataset, Soccer5.

The authors propose a deep learning method for fine-grained action detection in soccer. The objective was to determine the effectiveness of a player in stopping a soccer ball, based on the analysis of 132 soccer training videos. The dataset used for this study consisted of 2543 annotated instances of ball-stopping actions. This methodology was discussed in detail in [38]. It is pertinent to consider the variability in human-object interaction movements due to the indistinguishable nature of the motions and scenery involved in these two activities. This study proposes the utilization of an object-level trajectory-based cascaded deep network, which combines a YOLOv3 network for detection with an LSTM-driven network for classification.

In order to derive high-level semantic attributes from soccer data, the researchers in [39] conducted fine-tuning of several action recognition models and developed a temporal detection module based on transformers to identify the desired events. The study [40] conducted a study on action recognition in basketball, aiming to discern noteworthy individuals and events depicted in game video. The suggested model has been trained using a dataset exclusive to the organization. Its purpose is to detect and classify eleven distinct types of action or occurrence. Additionally, the model automatically allocates "attention" to the persons responsible for the observed incidents.

The study [41] devised a method called LSTM-DGCN for recognizing basketball players' actions. This methodology is built upon RNNs and deep graph convolutional networks (DGCNs) specifically designed to process skeletal data. Northwestern Polytechnical University researchers developed a comprehensive dataset called NPU RGB + D, which consists of RGB image data and depth data. This dataset was specifically built to capture 12 complex activities, encompassing 32 different atomic actions, performed by basketball players. The collection comprises 2169 videos, equivalent to 75,000 frames, encompassing RGB frame sequences, depth maps, and skeleton coordinates. The dataset utilized in their study has a high level of competitiveness, while the experimental results demonstrate that their technique outperforms current state-of-the-art action recognition systems.

Moreover, the study [42] pertains to volleyball's identification as a group activity. The LSTM model represents sequential activity. In order to obtain a comprehensive understanding of an activity, an additional LSTM model aggregates data at the individual level. By employing a two-step methodology, the researchers successfully constructed a temporal framework for characterizing specific activities. Subsequently, they merged these individual representations to discern collective activity patterns.

In the context of recognizing hockey activities, [43] proposes to employ the pre-trained VGGNet-16, which is a transfer learning model based on deep learning techniques. The authors constructed the hockey dataset by extracting footage from the International Hockey Federation archives and the video sharing site YouTube. The four most prevalent occurrences that they were aware of are free hits, goals, long corners, and penalty corners.

In their publication, the researchers in [44] introduced a transformer network architecture that was utilized to identify National Hockey League (NHL) players based on on-screen jersey numbers observed in broadcast video. Further investigation and refinement of this network will enable it to accurately identify and classify human behaviors. In [45], the authors provide a well-curated dataset consisting of six distinct pitch categories, which exhibits high quality. The researchers utilized a two-stream inflated 3D convolutional neural network to discern several categories of baseball pitches using recorded broadcasts as input data. The authors of reference [46] constructed a dataset by employing multimodal Kinect sensors and cameras to assess the signals and discern the behaviors of the players on the baseball field. The researchers employed a multimodal LSTM model for this purpose. This study proposes 10 different ways to improve the efficacy of baseball pitching and batting techniques. Prior to games or practices, players' health is evaluated by left and right stretches, lunges to the side, and deep squats. The initial results of the proposed model indicate that a baseball coach has the potential to utilize multimodal sensor data to assess a potential player's behavior both during games and in non-game situations.

In the study conducted by [47], HOG features were seen in combination with SVM, KNN, and AlexNet models. In the study conducted by the authors in [48], the OpenPose skeleton key points [49] were utilized and included in a LSTM network as a set of distinctive characteristics. In a subsequent study, a method for discerning between gaming and stroke was introduced, employing a scene recognition algorithm [50].

The study [51] proposed an architectural design for a deep convolutional neural network (DCNN) that is both computationally efficient and straightforward, while also including multiscale processing, with the aim of enhancing human activity identification. Through careful and systematic network design, the researchers significantly increased the network scope and complexity. The recommended architecture was evaluated by rigorous testing on publicly accessible datasets such as UCF101, HMDB51, TV-HI, YouTube, IXMAS, and UCF sports datasets.

Methodology

This methodology is based on a comprehensive framework, as shown in Fig. 1. Using CNN is required for the extraction of characteristics from video.

In the current version of the focus module, large and small feature maps can be handled. Consequently, CNN networks include Deformable Convolution Networks (DCN) in order to represent deformations at various scales.

Preparing step

The To identify sporting events from surveillance footage shot in a variety of settings, a deep learning network is used. In addition to improved accuracy, deep learning systems also have greater capacity to handle

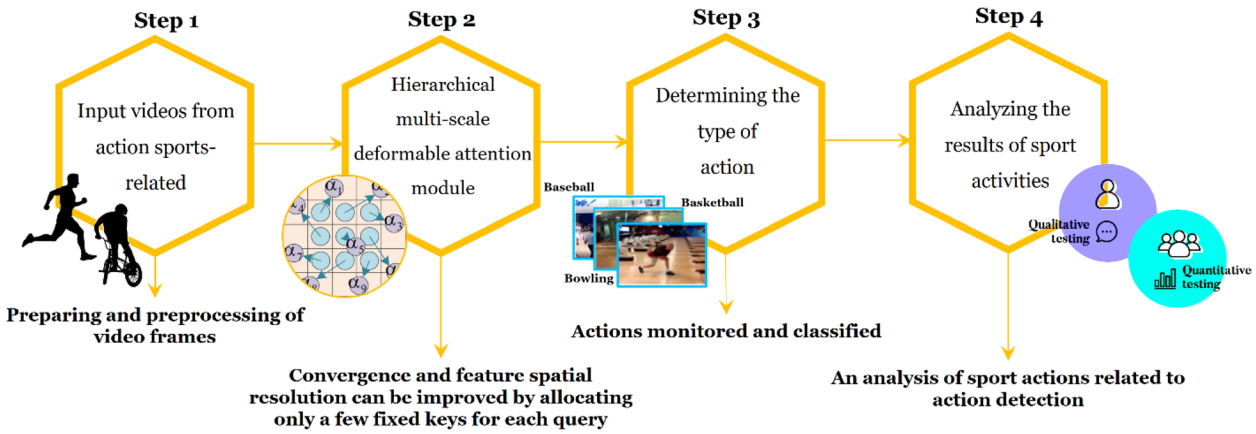

Fig. 1 The diagram provided serves as an illustrative example of the format used for presenting the approach

large datasets. Many different conventional and experimental methods exist for capturing moving images. In the pre-processing phase, frames of recorded content are deleted one by one. Each movie has its own subfolder, which is refreshed with new screenshots. Extracting frames from the video file and saving them as JPEGs is the first step in creating still images. Images at 100%, 70%, and 50% of the original frame resolution were all processed using the bilinear method. Images are compressed by reducing their dimensions (or resolutions) quickly. The speed and precision with which it reduces frames are two strong arguments in favor of its application.

Action videos are generated randomly. Video compression algorithms that use frame sampling reduce the computational burden of processing unnecessary data. Depending on the dataset parameters, the number of shot segments in various films may vary. By selecting a single frame randomly from the available alternatives, the number of images in each section may be minimized. Although the vast majority of an event may be captured in frames, only a few key details are often recorded. Our dynamic image sampling procedure is shown in Fig. 2.

In action videos, random sampling keeps the number of frames to a minimum. By avoiding unnecessary data processing, frame sampling reduces video volume and saves computing resources. In the dataset, the number of shot segments varies from video to video. For each subsection, a random frame is selected to reduce the number of usable images. The video captures depict distinct events despite the lack of contextual information. The procedure for dynamically sampling photographs is shown in Fig. 2. Processing time can be reduced by only encoding the frames that change during a video. Different movies will have different shot segments depending on the dataset specifications. Randomly selecting one image from each section reduces the number of images in each section. There are few details missing from a film that can capture the majority of an event. The figure below illustrates how our dynamic image sampling method works. Random frames are generated for action videos. Frame sampling compresses video and avoids unnecessary processing

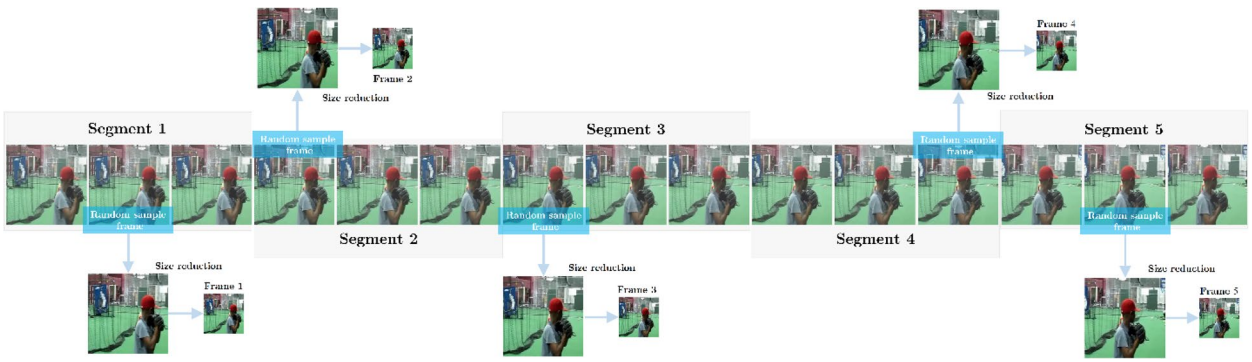

Fig. 2 Shows a randomly selected sample of baseball players. Shrinkage involves selecting frames randomly and assembling them to create reduced-sized movies through a stochastic process

time. Depending on the parameters of the dataset, different films feature different numbers of shot segments. By randomly selecting one frame from the pool of available frames, each segment's image count is reduced. A majority of an event may be captured on video with only a few key details. Our approach to dynamically sampling images is shown in Fig. 2.

Learning procedure

The first phase entails the classification of the sport-related activities portrayed in the video frames. The aforementioned classification is employed for the purpose of identifying instances of type of action in sport within the video frames. This allows for the accurate and timely detection of sport-related activities in the video frames. This is a crucial step, as it allows for the subsequent extraction of the temporal and spatial features of the sport-related activities. The extracted features are then used for the recognition of the sport-related activities.

Deformable convolution model

At the outset, x and y are designated as the initial and final feature maps, respectively. In this context, the symbol p_0 denotes the feature map output coordinate. The following are the template coordinates (p_n) for the convolutional kernel. Mathematical formulations for convolution can be found in the literature when conventional two-dimensional kernels are utilized [52].

$$y(p_0) = \sum_{p_n \in T} x(p_n + p_0).w(p_n) \tag{1}$$

The point that has been sampled in this particular example is denoted by coordinates of the form summation of p_n and p_0. The weight parameter $w(p_n)$ of the convolutional kernel is defined by its template T. In the present context, our focus is directed towards the convolutional kernel size of 3×3 due to its frequent utilization. In the given scenario, the observed pattern can be denoted using the symbol $T_{3\times3}$:

$$T_{3\times3} = \{(-1,-1),(-1,0),....,(0,0),(0,1),...,(1,-1)\} \tag{2}$$

In the default template T, the values of point pairs are exclusively integers. The concept of deformable convolution involves the incorporation of an offset parameter into the T transformation matrix, thereby extending the conventional convolution operation:

$$y(p_0) = \sum_{p_n \in T} x(\Delta p_n + p_n + p_0).w(p_n) \tag{3}$$

Consider allowing a decimal value to represent the variable p_n. Additionally, let $\{\Delta p_n \mid n=1, 2,..., N\}$be equal to the cardinality of the set obtained by taking the absolute value of the set T, N. To enhance ease, we can simplify Eq. (4) by denoting the product of p_n and p_0 as P_n.

$$y(p_0) = \sum_{p_n \in T^*} x(p_n + p_0).w(P_n) \tag{4}$$

This statement suggests the existence of N = |T∗| and T∗ = {P_0, P_1, …, P_N}. The value of P_n is a numerical representation rounded to two decimal places, thereby necessitating consideration of this particular characteristic. The inclusion of decimal sample points (P_n + p_0) is considered superfluous when the positions on the feature map x only consist of integer values. In order to compute the missing value for the decimal sample points, an interpolation function is employed, which may be expressed as:

$$x(p) = \sum_q x(q).G(q,p) \tag{5}$$

The subsequent procedures demonstrate the methodology employed in calculating the value of P_n for each provided x-coordinate q, utilizing the interpolation kernel function G(.,.).

$$\frac{\partial y(p_0)}{\partial \Delta P_n} = \sum_{P_n \in T^*} \left[\frac{\partial y(P_n + p_0)}{\partial P_n}.w(P_n) \right] \tag{6}$$

in which,

$$\frac{\partial y(p_0)}{\partial \Delta P_n} = \sum_{P_n \in T^*} \left(w(P_n).\sum_q \left[x(q) \frac{\partial G(q, P_n + p_0)}{\partial P_n} \right] \right) \tag{7}$$

The application of deformable convolution by CNN-Block presents a viable alternative to conventional convolution for the purpose of extracting and comprehending geometric modifications.

Feature representation

Analysis of static images taken during a dynamic athletic event reveals intricate and unforeseen patterns. In this study, we provide a novel multi-scale deformable attention module that effectively identifies crucial regions and characteristics. Sports activity in the research region is influenced by factors such as complexity, resemblance to baseline circumstances, and small spatial patterns. In order to include accurate spatial information, the convolution approach utilizes a filter with a predetermined receptive field size. The process of extracting several hierarchical representations from convolutional networks employing filters of a single size has similar challenges, as discussed in previous studies [53, 54]. When doing a comprehensive assessment of many sport activities simultaneously, it is imperative to efficiently collect and analyze visual data. This phenomenon is particularly

evident when comparing individuals with typical abilities. Individuals are continuously exposed to a multitude of signals that are similarly complex. This study successfully employed a multi-scale deep convolutional network (DCN) to discern distinct patterns of human behavior in various athletic scenarios. The objective was to identify significant associations between successive video frames.

This study used an analysis approach informed by [52]. Due to their ability to accurately identify body despite geometric deformations, including changes in size, posture, and perspective, deformable convolutional networks have proven to be advantageous. To do this, the offset, convolution, and sample location are constantly adapted. The DCN demonstrates a high level of precision in capturing and reconstructing the intricate and diverse properties present within the receptive field. A comprehensive neural network may perform a range of tasks, including but not limited to applying multi-scale deformable attention modules, utilizing feature maps of different dimensions, and conducting pattern classification to identify the type of sport. Figure 3 shows the intricate network architecture used to identify sports within video frames.

Figure 4 illustrates the potential benefits of employing a flexible attention module at different scales to mitigate gradient vanishing. It is ensured that the set order is maintained throughout the procedure. This design utilizes a multiscale deformable attention module extraction technique to facilitate hierarchical attention maps. Figure 5 illustrates the structural design of the Head, whereas Fig. 4 depicts the configuration of a multiscale deformable attention module.

The multi-scale DCN employs three kernels of varying sizes and integrates them by concatenating the resultant feature maps. Moreover, a 1×1 convolution layer is employed to establish a residual link. Smaller kernel sizes have been shown to be effective for edges and other video frames, but much larger kernel sizes are necessary for capturing diverse properties. To enhance video frame collection in athletic events, it is possible to use data from both low and high frequencies. Subsequently, 1×1 convolutional layer are utilized to standardize the multi-scale input, and the network is trained utilizing a soft-plus activation function.

Ultimately, a comprehensive cartographic representation was produced, highlighting distinctive attributes within different frequency categories. Initially, the hierarchical features undergo normalization using the L2-norm technique. Subsequently, an attention map is employed to construct the ultimate feature map. The previously described feature map is next submitted to sequential processing, wherein each component is individually inspected. The initial layer of the convolutional neural network possesses the ability to detect rudimentary characteristics such as the boundaries and outlines of video frames. Conversely, the deeper layers of the network have the capability to identify more intricate patterns.

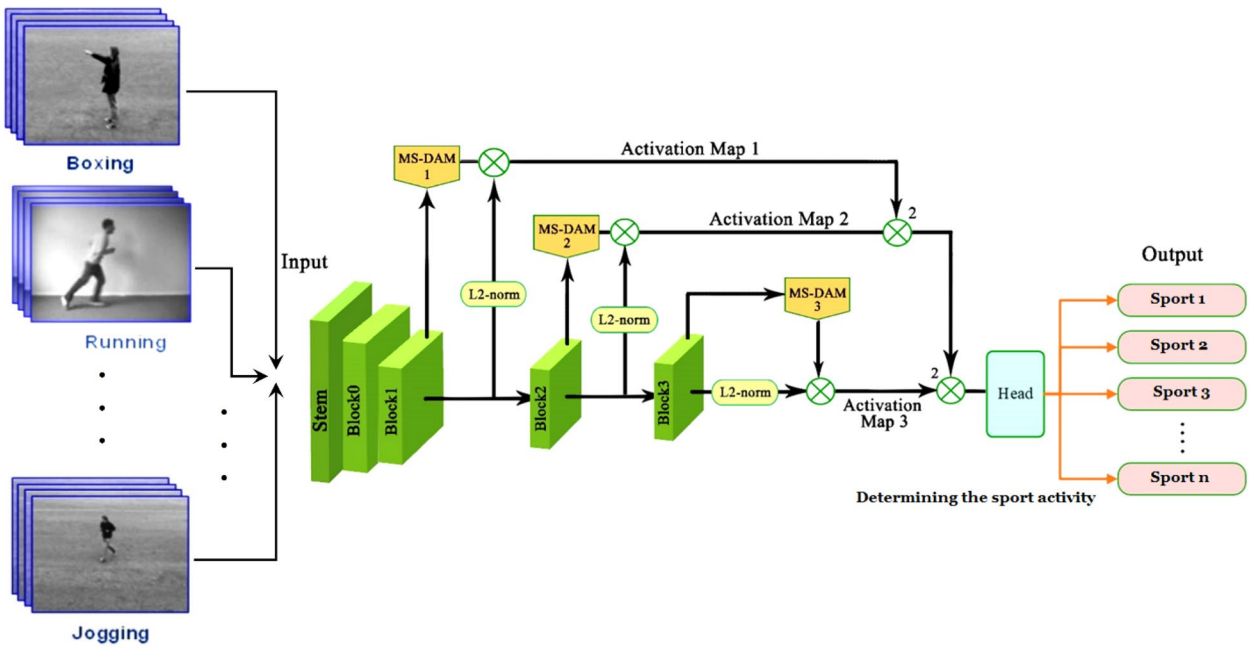

Fig. 3 In our study, we combine hierarchical feature maps with a multiscale deformable attention module to create a network architecture. The purpose of this research is to develop methods for more accurately identifying and locating sporting events within video clips

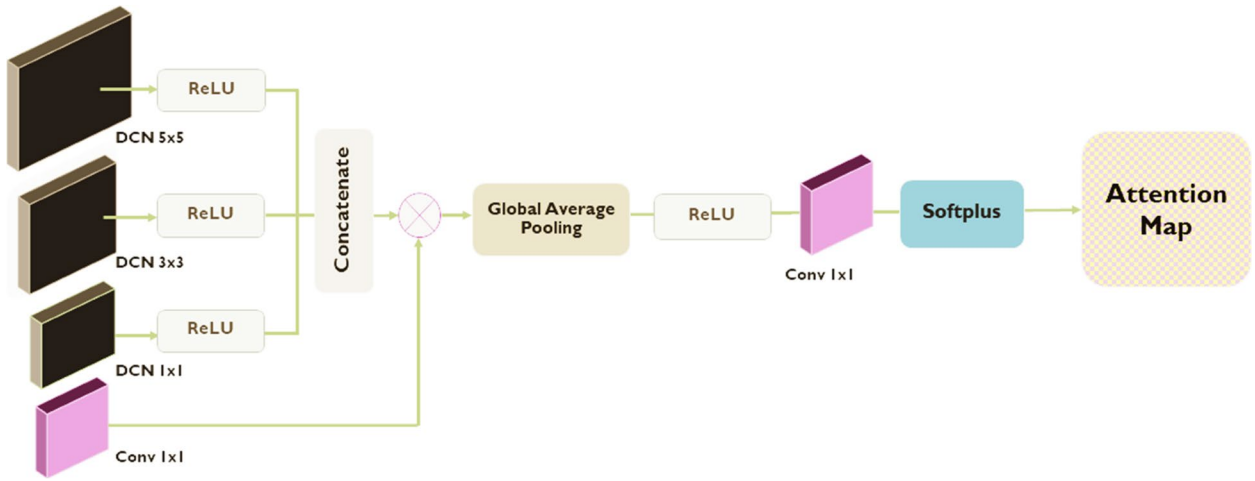

Fig. 4 An intricate network of a number of DCN components makes up the deformable attention module that can work at various scales, as shown in the diagram below

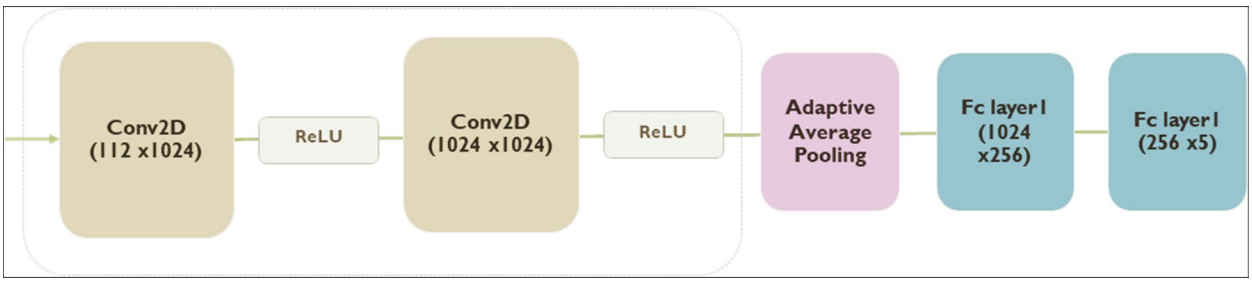

Fig. 5 The graphic presented illustrates the anatomical structure of the Head part

Edge computing

In order to enhance comprehension of human activities, a framework has been devised that encompasses a scenario for the acquisition of data pertaining to sports. This is done through the process of action recognition. In order to advance the discussion on this subject, we employed the recommended deep learning network in combination with an edge computing video analysis approach. The first proposition was the introduction of a theoretical framework for edge computing, aimed at mitigating power consumption and enhancing processing velocity in interconnected devices. The proposed instructional methodology utilizes real-time video frames obtained in low-resolution environments, which are subsequently analyzed inside a sample data analysis system to generate forecasts for human performance across several sports disciplines. Figure 6 shows the planned human activity related sports edge computing architecture.

Edge computing and interconnected devices have the advantage of enhancing the security of individuals' personal information. Users' privacy will not be affected when using this platform for activity data analysis. Within the framework of edge computing architecture, real-time applications encompass the expeditious and accurate transmission of data across cloud systems, as well as the immediate accessibility of user data via remote monitoring systems. To conduct a more comprehensive analysis of this phenomenon, we integrated real-time data processing into our investigation. In this particular case, data aggregation is facilitated by HUBs, which can comprise different nodes. The data set and flow module is responsible for gathering and distributing the outcomes of real-time activity analysis conducted on video footage of athletic events generated within the edge computing environment. This particular component facilitates the effective transmission of real-time data while minimizing data loss.

Results

In this section, we will conduct an analysis of the outcomes, with a particular emphasis on the implementation components of the study strategy. The study starts by doing a comprehensive inspection of the video frames.

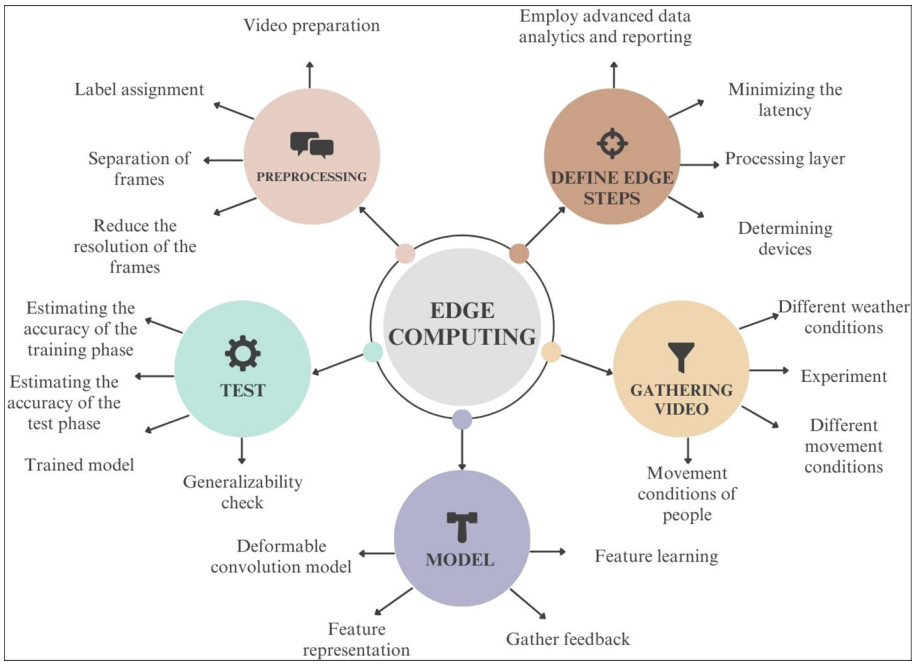

Fig. 6 The figure shows the proposed framework for edge computing in sport-related human activities

Dataset

The datasets employed in this investigation include UCF Sports [14], UCF50 [55], and YouTube (UCF11) [56]. The UCF Sports dataset, created by the Action Recognition Society, encompasses a diverse range of sports-related activities across many sporting disciplines. Events of this nature are often covered by broadcast networks such as the BBC and ESPN. Figure 7 illustrates a selection of stock footage sources, including BBC Motion Gallery and Getty Images, from which the video sequences were obtained. The collection under consideration comprises 150 sequences, each having a resolution of 720×480 pixels. The diverse range of places and perspectives encompassed within this compilation offers a rich array of potential activities. This data is intended to encourage further investigation into the field of action recognition in natural settings through the public availability of this dataset. Since the dataset was created, it has been utilized in a diverse range of situations. This includes action identification, action localization, and saliency detection, among others. In Fig. 7, we display a sample received frame from the UCF Sports dataset containing various activities.

The UCF50 dataset [55] exhibits a diverse range of human behaviors as a result of the multitude of camera motions, object views and locations, object scale, perspective, cluttered backdrop, and ambient lighting conditions. There exist several sets of movements that exhibit similarities, such as an individual performing the piano four times from significantly distinct perspectives.

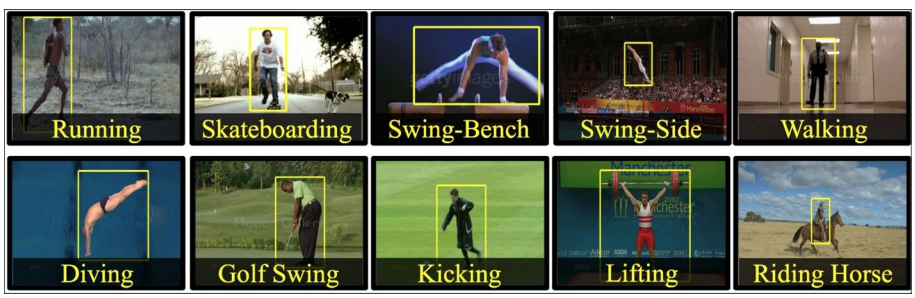

Fig. 7 The sample received frames from the dataset "UCF Sports" containing videos of diving (14 videos), swinging (18 videos), kicking (20 videos), lifting (6 videos), riding horses (12 videos), running (13 videos), skateboarding (12 videos), swing-bench (20 videos), swing-side (13 videos), and walking (22 videos)

The YouTube dataset [56] has 11 unique activity categories, including strolling with a dog, volleyball spiking, trampoline leaping, tennis swinging, soccer juggling, horseback riding, golf swinging, diving, shooting, biking/cycling, swinging, and basketball. The dataset provided presents notable difficulties arising from substantial variations in camera motion, object attributes such as visual features, body position, and size, differing perspectives, complex backdrops, and a wide range of lighting conditions, among other contributing variables. The films have been systematically classified and arranged into 25 distinct categories, with each category containing a minimum of 4 action portions. The video clips within a certain group demonstrate common features, such as the inclusion of the same actor, similar background settings, a consistent perspective, and other relevant qualities. The videos are encoded in Microsoft MPEG-4 format. To access the files, install a suitable codec. One example is the K-lite Codec Pack, which provides a compilation of codecs that may be employed for this particular objective.

Implementation details

The In both the training and testing phases, the Keras framework was combined with the TensorFlow library. Furthermore, a GeForce RTX 3070 graphics card equipped with 4GB of GDDR5 RAM was used in the experiment. Our testing involved comparing the proposed attention module with state-of-the-art models for image recognition tasks. Convolutional networks, such as RegNetY, exhibit compact and efficient design spaces. In addition to depth, initial width, and gradient, we have collected several other key attributes. To optimize the model's depth, breadth, and input size, EfficientNet uses a compound scaling technique. Following their initial training on ImageNet, both models were fine-tuned using proprietary data. The findings from iterations 2–4 of each model were also incorporated into the primary research. A comparison of the two models was conducted to determine their similarities and differences.

In terms of class distribution, the brain activity dataset showed significant variability. Due to its cost-effectiveness as a machine learning method, focused loss was used as the loss function in this study. By assigning higher importance to instances with less data, cost-sensitive learning reduces disparities between groups. Cases with increased complexity in their categorization are given greater importance by the focused loss function. Identifying the ideal combination of Alpha = 1 and Beta = 0 minimized the overall loss in this experiment. Overlapping athletic activities may compromise the accuracy of video frame measurements. In order to address the challenge of amalgamating several frames of sports video, the concentrated loss function was employed. The goal was to mitigate conflicts among socioeconomic groups. Rectified Adam (RAdam) was the optimization approach used in our study. In addition, cosine decay principles were used to establish learning rate schedules. Within a range of 0.00001 to 0.0002, the training rates were constrained. The training batches consisted of 16 individuals, and the iterations totaled 1000. Validation of the results was performed using a 5-fold cross-validation method. For evaluation purposes, the collection includes a variety of example images. A maximum of 20% of the available information can be used for assessment, leaving the remaining 80% for training. Using K-fold cross-validation (CV) with a value of 5, 80% of the training and validation data are used for model training, while the remaining 20% are used for evaluating the model's validity.

Evaluations

The presence of several categories within the provided confusion matrix indicates that further tests were conducted to validate cross-validation correctness. Each of the four modified structures retained an same total number of parameters as their respective original versions. The evaluation criteria exhibit a consistent pattern across all designs, with few or negligible discrepancies seen in test outcomes. In comparison to other models, the model under evaluation demonstrates a higher likelihood of accuracy. When employed within the appropriate framework, the suggested methodology yields optimal outcomes. In contrast to previous comparable models, this deep convolutional learning model has a reduced parameter count and a more streamlined computing implementation. The model indicated above has improved generalization skills in data tests, and it demonstrates successful learning from a smaller training dataset than rival models. The enhanced resilience of the model in mitigating overfitting is attributed to the reduction in the number of variables, resulting in heightened accuracy. Each component's performance is assessed separately. The majority of the activities seen in the datasets included in the research had an inherent athletic component. Our strategies consistently yield enhanced experimental efficiency. In this work, a technique was designed to perform a multi-class categorization assignment that incorporates sports-related behaviors. The approach achieved accuracy values of 97.84% for UCF sport, 97.75% for UCF50, and 98.91% for YouTuble (UCF11).

The presence of several categories within the provided confusion matrix indicates that further tests were conducted to validate cross-validation correctness. Each of the four modified structures retained a same total number of parameters as their respective original versions. The evaluation criteria exhibit a consistent pattern across all designs, with few or negligible discrepancies seen in test outcomes.

In comparison to other models, the model under evaluation demonstrates a higher likelihood of accuracy.

When employed within the appropriate framework, the suggested methodology yields optimal outcomes. In contrast to previous comparable models, this deep convolutional learning model has a reduced parameter count and a more streamlined computing implementation. All actions related to sports have been taken into account, leading to the presentation of the mean estimations for accuracy, recall, precision, F-score, and Kappa over five K-fold cross-validation cycles in Table 1.

The identical criteria are often employed for the analysis of sports data obtained from recorded films filmed in diverse locations. The overall construction quality of the vehicle surpasses that of its primary competitors. The validity of our studies was assessed by employing a 5-fold cross validation (CV = 5) technique. The extensive range of video frames facilitates the achievement of high experimental repeatability. Despite the presence of imitation and an uneven distribution of frames in our study and dataset, we successfully addressed challenges related to imbalanced data and overfitting through

Table 1 A comparison of experimental results between our suggested technique and similar structures is presented in the table below. Analyzing video frames from many scenes, the analysis focuses on three distinct activities within the context of sports

Dataset	Resolution	Model	Accuracy	Recall	Precision	F-score	Kappa
UCF sport	50%	CNN	95.32%	95.05%	94.97%	95.01%	73.99%
		CNN + Attention	96.07%	95.63%	95.59%	95.61%	78.17%
		DCN + Attention	96.21%	95.76%	96.02%	95.89%	78.95%
		Proposed	96.98%	96.81%	96.87%	96.79%	83.25%
	70%	CNN	96.17%	95.77%	95.99%	95.88%	78.73%
		CNN + Attention	96.21%	95.92%	95.94%	95.92%	78.94%
		DCN + Attention	96.67%	96.14%	96.40%	96.26%	81.48%
		Proposed	97.16%	97.07%	96.90%	96.98%	84.24%
	100%	CNN	96.47%	96.11%	96.29%	96.19%	80.38%
		CNN + Attention	96.65%	96.39%	96.31%	96.34%	81.37%
		DCN + Attention	97.20%	96.90%	97.22%	97.05%	84.46%
		Proposed	97.84%	97.85%	97.91%	97.87%	87.99%
YouTube (UCF11)	50%	CNN	96.63%	96.04%	96.16%	96.10%	79.63%
		CNN + Attention	98.09%	97.51%	97.59%	97.55%	88.44%
		DCN + Attention	98.28%	97.68%	97.72%	97.70%	89.58%
		Proposed	98.51%	97.97%	98.05%	98.01%	90.96%
	70%	CNN	96.49%	95.92%	95.99%	95.95%	78.75%
		CNN + Attention	98.16%	97.73%	97.66%	97.69%	88.87%
		DCN + Attention	98.46%	97.92%	97.93%	97.93%	90.69%
		Proposed	98.77%	98.19%	98.23%	98.21%	92.56%
	100%	CNN	97.06%	96.50%	96.49%	96.50%	82.24%
		CNN + Attention	98.36%	97.88%	97.84%	97.86%	90.08%
		DCN + Attention	98.49%	97.94%	97.98%	97.96%	90.89%
		Proposed	98.87%	98.38%	98.36%	98.37%	93.15%
UCF50	50%	CNN	93.43%	93.06%	93.01%	93.03%	60.24%
		CNN + Attention	95.48%	95.03%	94.96%	94.99%	72.65%
		DCN + Attention	96.71%	96.20%	96.21%	96.19%	80.12%
		Proposed	97.62%	97.00%	97.04%	97.02%	85.63%
	70%	CNN	94.13%	93.62%	93.69%	93.66%	64.46%
		CNN + Attention	95.86%	95.39%	95.35%	95.36%	74.93%
		DCN + Attention	96.80%	96.23%	96.27%	96.25%	80.67%
		Proposed	97.70%	97.22%	97.26%	97.24%	86.09%
	100%	CNN	94.23%	93.80%	93.79%	93.79%	65.10%
		CNN + Attention	96.58%	95.99%	96.06%	96.03%	79.29%
		DCN + Attention	96.92%	96.39%	96.40%	96.39%	81.37%
		Proposed	97.73%	97.16%	97.16%	97.15%	86.25%

the implementation of our suggested approach. It is important to bear in mind, however, that certain video frames depict elements unrelated to sports. Therefore, it is imperative to establish a rigorous specification standard.

The aforementioned findings provide evidence for the soundness and reliability of the 5-fold cross validation technique. Furthermore, the study demonstrated that the implementation of this method not only mitigated the impact of potential biases but also enhanced the level of accuracy. The video stills are from a dataset characterized by classes exhibiting low accuracy in their categorization. The assessment was conducted using three models as described in Table 1: a CNN, a CNN with attention module, a DCN with module, and attention with a unique multi-scale deformable attention module. The aforementioned classes were assessed for their prevalence in accordance with the norms of our study, utilizing these models. Due to the inherent decreased accuracy of lower resolution frames compared to higher resolution frames, the likelihood of generating false positives and false negatives is increased. The number of cases in the database is limited. Nevertheless, when employed in the assessment of sports-related actions, the aforementioned methodology has potential.

The labeling of each frame is determined by random sampling from the original labels associated with each video. To alleviate computational load [57, 58], we opted to examine one of the three frames in an arbitrary manner. Furthermore, we considered alternative scenarios when determining overall accuracy by randomly selecting frames from a range of 2, 3, 4, …, and 10 frames. The selection of one of the three frames yielded the highest accuracy level, as seen in Table 2. Out of the potential frames numbered 2, 3, 4, …, and 10, only one frame is retained and documented in Table 2.

Furthermore, the accuracy and computational complexity of frames were assessed by evaluating frames with lower resolutions and varying the number of selected frames simultaneously. As the number of removed frames grows, there is a decline in the correlation between the retrieved characteristics and the video sequence.

The decline in accuracy is not particularly noteworthy. Various methodologies can be employed to dynamically determine the optimal video frames. The utilization of various methodologies in conjunction with the process of video preparation and frame selection can, however, be a time-consuming endeavor. To ensure the accuracy of the retrieved characteristics, it is important to select the optimal frames. Automating this process can help to reduce the time and effort required.

Discussion

The proposed methodology has the capability to effectively categorize action recognition tasks by utilizing video frames obtained from a diverse range of individuals in various locations and circumstances. This enables the detection and diagnosis of a broad spectrum of disorders associated with human actions. Hence, a more efficient framework is employed to assess, analyze, and

Table 2 The table presents an evaluation of the accuracy of various frame sequences when both resolutions and selected frames are simultaneously lowered. This reduction mitigates computational complexity. Reducing resolutions and frames simultaneously does not result in a significant decrease in accuracy diversity

No. Selected frames	UCF sport			YouTube (UCF11)			UCF50		
	Resolution	Accuracy	Computational complexity	Resolution	Accuracy	Computational complexity	Resolution	Accuracy	Computational complexity
One of two frames	100%	97.84%	High	100%	98.90%	High	100%	97.75%	High
One of three frames	90%	97.70%	High	90%	98.81%	High	90%	97.70%	High
One of four frames	80%	97.43%	Moderate	80%	98.73%	Moderate	80%	97.65%	Moderate
One of five frames	70%	97.03%	Moderate	70%	98.65%	Moderate	70%	97.61%	Moderate
One of six frames	60%	96.90%	Moderate	60%	98.42%	Moderate	60%	97.57%	Moderate
One of seven frames	50%	96.81%	Low	50%	98.38%	Low	50%	97.50%	Low
One of eight frames	40%	96.55%	Low	40%	98.29%	Low	40%	97.40%	Low
One of nine frames	30%	96.50%	Low	30%	98.19%	Low	30%	97.25%	Low
One of ten frames	20%	96.44%	Low	20%	98.08%	Low	20%	97.11%	Low

classify behaviors shown at athletic events. In this analysis, we will deconstruct each technique and elucidate the underlying rationale, afterwards presenting the resulting consequences.

Robustness

The UCF-sport dataset includes ten main types of sporting-related video classification. Diving (Div), swinging (Swi), kicking (Kic), lifting (Lif), riding horses (Rid-ho), running (Run), skating (Ska), running-bench (Run-ben), swing-side (Swi-si), and walking (Wal) videos are just some of the many subcategories inside each main category. Some of them are quite similar to other types of human behavior. The results of the algorithm are depicted in descending rosettes in Fig. 8 for three different video quality levels. Confusion matrices for three conditions, including the original video with all frames, a 30% reduction of resolution with 1 chosen frame from 5 frames, and a 50% reduction of resolution with 1 chosen frame from 10 frames, are displayed from left to right, indicating that while the frame size has not changed, the output accuracy varies slightly from the original resolution in addition to reducing the selected frames.

Despite the drop in frame resolution, there is minimal disparity between the two outputs. There is a relatively low degree of standard variance among them. Despite the comprehensive range of sport-related courses offered by the prescribed curriculum, it has developed the ability to differentiate among them and articulate these distinctions using a predetermined framework. Hence, it can be observed that among the many action categories in sports, more than ten exhibited an accuracy rate of 97% or above, while five of them showed an accuracy rate of 98% or above. Also shown in Fig. 9 are 11 distinct activity categories from the YouTube dataset, including walking with a dog (St-dog), volleyball spiking (Vo-sp), trampoline leaping (Tra-le), tennis swinging (Te-swi), soccer juggling (So-jug), horseback riding (Ho-rid), golf swinging (Go-swi), diving (Div), shooting (Sho), biking/cycling (Bik), swinging (Swi), and basketball (Bas).

While the frame resolution has dropped in this dataset, the difference between the two outputs remains similar to what was found in the previous dataset. There is a relatively small difference between them. Its curriculum has improved its ability to differentiate and elucidate the many sport-related courses it offers while integrating these distinctions into a cohesive theoretical framework. It was found that five of the action categories examined in sports exhibited accuracy rates above 98%. Also, 98.90% or higher accuracy rates were found in 10 action categories.

There are 50 action categories in the UCF50 data set that have been collected from YouTube: Yo Yo (1), Baseball Pitch (2), Basketball Shooting (3), Bench Press (4), Biking (5), Billiards Shot (6), Breaststroke (7), Clean and Jerk (8), Diving (9), Drumming (10), Fencing (11), Golf Swing (12), Playing Guitar (13), High Jump (14), Horse Race (15), Horse Riding (16), Hula Hoop (17), Javelin Throw (18), Juggling Balls (19), Jump Rope (20), Jumping Jack (21), Kayaking (22), Lunges (23), Military Parade (24), Mixing Batter (25), Nun chucks (26), Playing Piano (27), Pizza Tossing (28), Pole Vault (29), Pommel Horse (30), Pull Ups (31), Punch (32), Push Ups (33), Rock Climbing Indoor (34), Rope Climbing (35), Rowing (36), Salsa Spins (37), Skate Boarding (38), Skiing (39), Skijet (40), Soccer Juggling (41), Swing (42), Playing Tabla (43), TaiChi (44), Tennis Swing (45), Trampoline Jumping (46), Playing Violin (47), Volleyball Spiking (48), Walking with a dog (49), and Motor biking (50). The UCF50 dataset is a complex and challenging research resource due to its

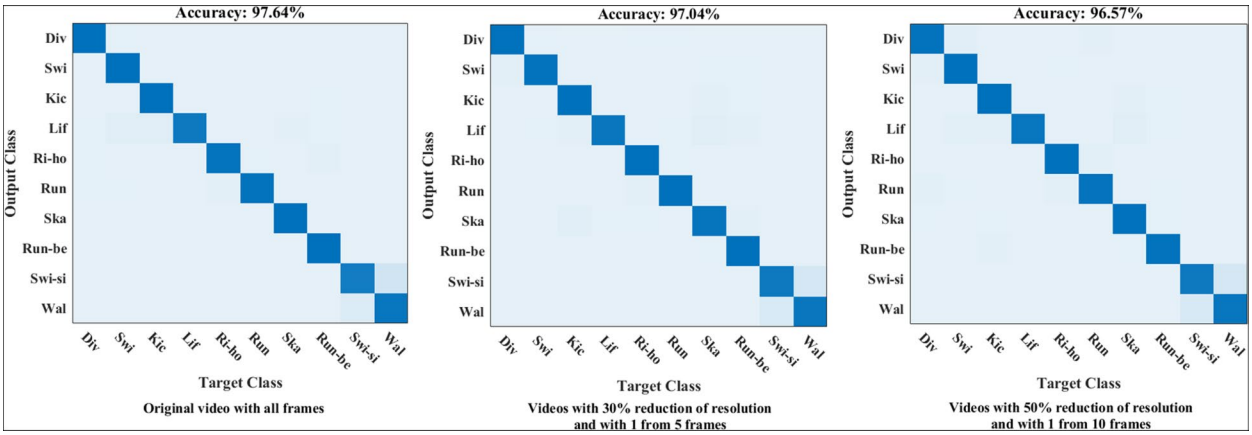

Fig. 8 To reduce computational complexity, we tested three different frame states for UCF-Sport dataset. It shows that the method is robust and recognizes all kinds of sports movements accurately

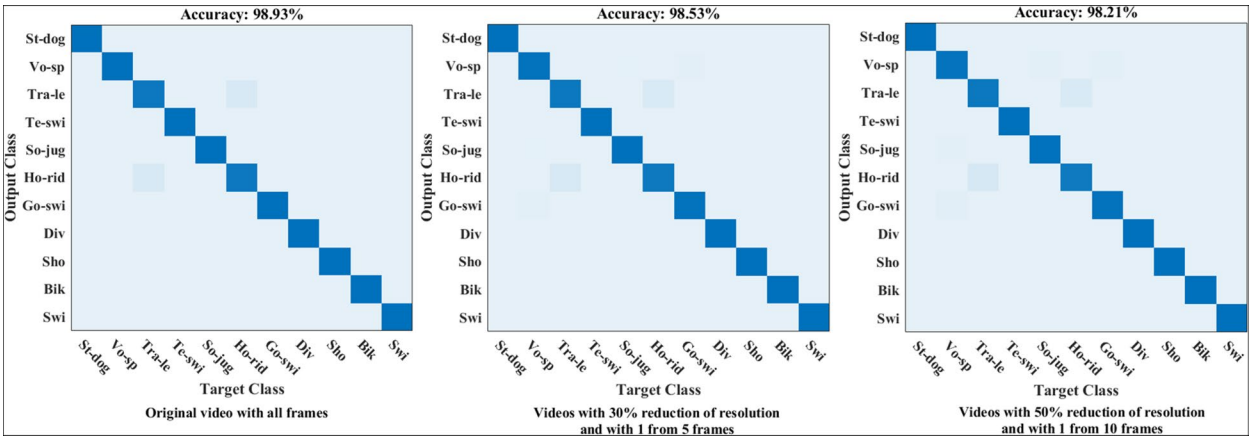

Fig. 9 To reduce computational complexity, we tested three different frame states for YouTube (UCF11) dataset. It shows that the method is robust and recognizes all kinds of sports movements accurately

inclusion of several action classes, which are depicted through persons participating in a diverse array of activities, particularly those related to sports, and utilizing a large variety of items. Figure 10 illustrates that the classification error rate remains relatively constant when the frame size and resolution are reduced from 30 to 50% of their original values. In certain scenarios, a level of accuracy above 97% has been attained, even when video frames are suboptimal.

Effect of features and accuracy dispersion

Using feature selection, we extracted features from each sporting event video at two levels of feature count, i.e., we used mutual information because of its accuracy and speed. As a result of its proven effectiveness and precision, this strategy was chosen. In order to enhance precision, many components were integrated. Conversely, the decision to include a small number of characteristics was motivated by the desire to simplify the model. Having achieved this equilibrium, we were able to maintain maximum efficiency while ensuring accuracy and thoroughness. Nevertheless, the robustness of the models was examined as part of the assessment. The durability of models developed using mutual information technique was examined empirically. Maintaining the model's stability requires an appropriate balance between specificity and complexity. The characteristics associated with each level are therefore included in several designs. By examining their sensitivity, specificity, and accuracy, the models are assessed for stability and robustness. An evaluation of the effectiveness of the suggested technique compared to current state-of-the-art methodologies can be seen in Fig. 11.

Analyzed are different features counts classified as high dimensional and low dimensional. Errors are more likely to occur when the number of characteristics used

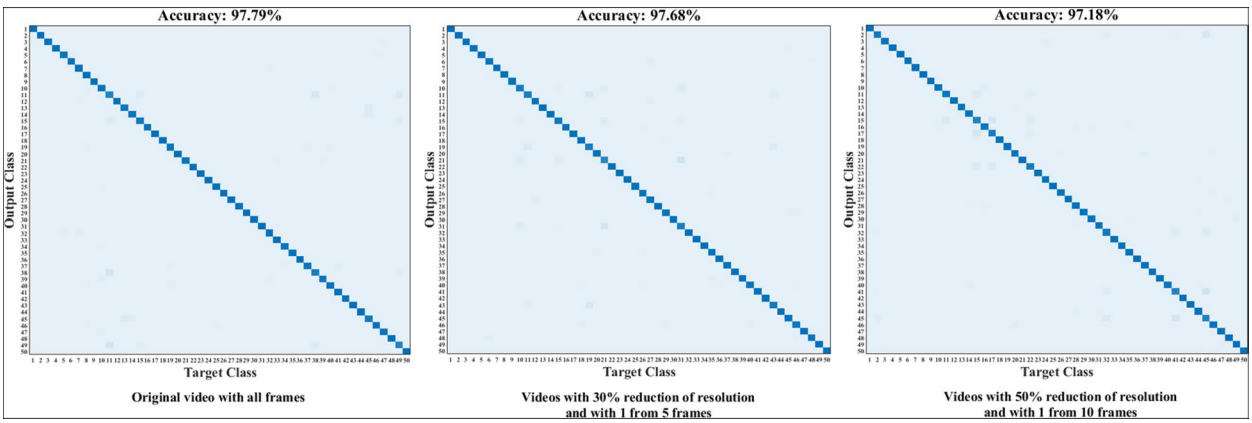

Fig. 10 To reduce computational complexity, we tested three different frame states for UCF50 dataset. It shows that the method is robust and recognizes all kinds of sports movements accurately

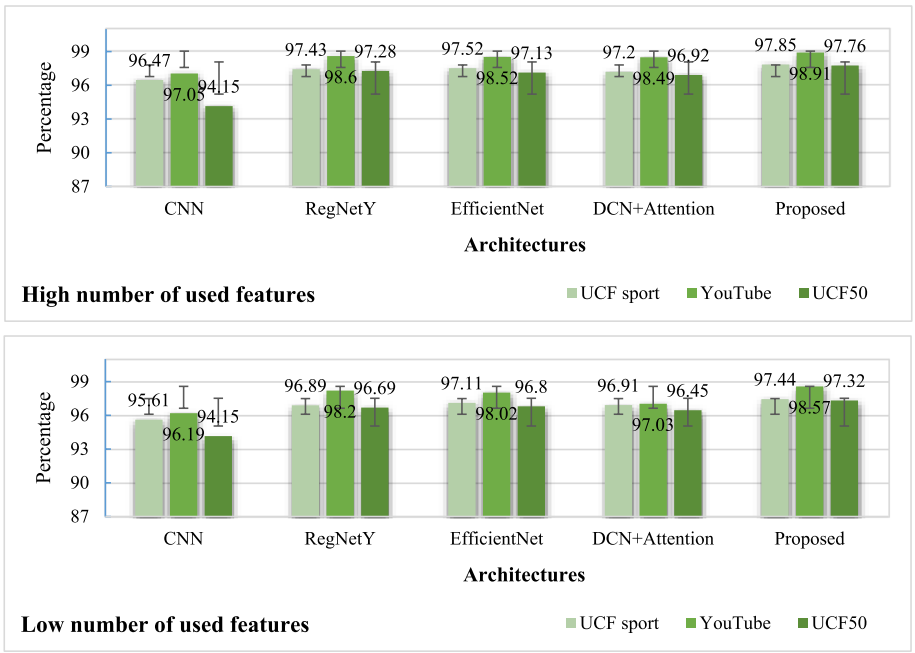

Fig. 11 Three sports-related action recognition datasets with varying numbers of characteristics are used to evaluate the proposed technique against existing alternatives. From what can be seen in the graph, the suggested model has a rather narrow set of requirements. However, when the number of features is decreased, EfficentNet and RegNetY display no variance in accuracy

in model construction decreases. By using a limited number of characteristics, the technique reduces errors. The methodology's capacity to provide a wider array of features for selection may primarily account for the reduction in errors.

Comparison

Methods that demonstrate similarities and rely on individuals' routine and daily actions sometimes lack accuracy in identifying sports activities. However, many techniques possess the capability of detecting activities with enhanced precision, while simultaneously mitigating computational costs. The methodology presented in this study exhibits enhanced performance compared to traditional approaches to feature extraction, mostly attributed to its automated characteristics. In recent years, there has been a notable increase in the use of novel strategies, such as deep learning techniques, in order to improve the existing performance of action recognition. Despite the significant progress in architectural development, several approaches based on deep learning and convolutional networks have faced difficulties in addressing uncertainty barriers, leading to a decrease in video quality, frame rate, or resolution.

Although deep learning models require powerful hardware in recent years, various algorithms can improve performance. Cloud computing and the Internet of Things (IoT) are examples of methods that can be used and applied to better process video processing [59]. Especially for action recognition data, learning-based processing and systems can be applied to cloud computing, edge processing, and the IoT [60].

Characteristic acquisition, particularly those with discriminative traits that may be adaptively applied across many settings, has significant importance. Video skeletons are used in some cases to extract characteristics. However, the information obtained from skeletal remains is often overlooked, resulting in a methodology that exhibits diminished accuracy and precision. As a result, action characteristics will improve less. This paper introduces a revolutionary methodology for efficiently reducing video frames and dimensions by integrating the attention mechanism, auto-encoder network, and convolutional structure. Table 3 displays a comparative examination of the present findings in relation to those obtained by comparable approaches employed in recent investigations. Moreover, it demonstrates a competitive nature to other action detection systems that have made progress in recent years, by utilizing deep learning methodologies.

Limitations and challenges

Human sports activities have been defined by several people, but few have been successful in overcoming the obstacles. A major challenge in this field is standardizing

Table 3 Our model is compared to various approaches in sport action recognition in order to determine their accuracy levels

Reference	Accuracy			Architecture
	UCF sport	YouTube	UCF50	
Ullah et al. [13]		96.21%	96.40%	Optimized deep autoencoder and CNN
Liu et al. [61]	-	89.7%	93.20%	Hierarchical clustering multi-task learning
Sadanand et al. [62]	95.00%	-	57.9%	High-level representation
Tu et al. [63]	97.53%	-	-	Multi-stream CNN
Afza et al. [64]	99.30%	94.50%	-	Features fusion and weighted entropy-variances
Muhammad et al. [65]	99.10%	98.30%	-	Attention based LSTM network with dilated CNN
Meng et al. [66]	93.20%	89.70%	-	Spatial-temporal convolutional neural network and LSTM
Gammulle et al. [67]	92.20%	89.20%	-	Two stream LSTM
Ijjina et al. [68]	98.90%	94.60%	-	Hybrid deep neural network
Zhou et al. [69]	98.75%	97.60%	-	Density clustering and context-guided Bi-LSTM
Xiong et al. [70]	-	-	96.71%	Two-Stream 3D Dilated Neural Network
Zhang et al. [71]	-	-	60.40%	LSTM and fully-connected LSTM with different attentions
Dai et al. [72]	98.90%	96.90%	-	Two-stream attention-based LSTM
Proposed model	**97.84%**	**98.90%**	**97.75%**	**Deformable convolution and adaptive multiscale features**

video capture of human motion. Time constraints, camera location, weather changes, visual interference, and movement categorization uncertainty are all factors that affect video surveillance systems. A person's position and mobility affect the quality of a video image and the effectiveness of an identification system. Variable weather conditions and intense illumination decreased the accuracy of human action recognition related to sports. Moreover, it may be difficult to assess a performer's ability from recorded footage because of the camera's location. There are several scenarios in which the paradigm can be applied. There is a great need for a large number of training videos. Complexity, time, and low video frame quality make this project challenging. Using video technology, multiple sporting activities can be conducted simultaneously. To overcome this issue, alternative videos could be explored for model training in order to solve the problem of multitasking. In addition, human behavior is complicated and difficult to comprehend. Videos shot in poor conditions tend not to be used as training data for action recognition techniques related to sports. It is also possible for pixel occlusion to be caused by issues with implementation and constraints. Camera movement and perspective distortion can affect sports performance, and when the camera moves, recognition occurs. The operational efficiency of a system is affected by changes in its functionality. Moreover, there are differences between some sports activities, for example, Kabaddi differs from Soccer in some ways, but the automated system may not be able to distinguish between them. It is necessary to distinguish between groups in order to understand human behavior in sports. Detecting human activity becomes more difficult when style, viewpoint, behavior patterns, and clothing differ. A study is being conducted on human-object communication and sports activities. With limited training data, it can be difficult to detect anomalies such as odd physical behavior and abnormal physical activity. Likewise, monitoring and tracking many actions can be challenging.

In order for a system to function optimally, several impediments must be overcome. Among them are slow networks, limited bandwidth, inadequate privacy and security measures, and a shortage of accessible servers. To improve the accuracy of identification across increasingly large and more complex datasets is an important avenue for future investigation in deep learning.

Conclusion

To enhance sports-related activities assessment, we suggest a novel approach that leverages deep learning techniques for feature extraction. Specifically, our method incorporates deformable convolution and adaptive multiscale methodologies. These methodologies effectively capture spatial and deformable qualities in video frames. The final feature map was constructed by integrating multiple hierarchical feature maps using the multi-scale deformable attention module. The aforementioned approach exhibits a high likelihood of successfully detecting prevalent sports-related behaviors. In order to do this, a system was developed that collects and integrates multi-scale features that exhibit sensitivity to various forms of video frame data and temporal fluctuations across multiple situations. The evaluation encompassed the assessment of overfitting, computational cost, and

the general robustness of the model. Once these concerns have been addressed, the model may be used to categorize and monitor action sports. The suggested methodology demonstrates superior performance compared to the current state-of-the-art in the categorization of action activity linked to sports, while simultaneously requiring a reduced level of analysis. The objective of this research is to employ a hierarchical multi-scale deformable attention module in order to establish control groups for human athletic motions under comparable conditions. In order to facilitate the timely detection of sports associated with action-oriented activities and provide support to those impacted by such activities, the development of recognition algorithms for video frames using extensive datasets will be imperative. This method can be generalized and applied to video games as well, to recognize sports activities of characters in video games. Research on this issue will be conducted by the authors in the future. Moreover, taking into account the potential for advancement in deep learning for human activity recognition related sport within an edge computing environment, we aim to explore the future limitations of edge computing.

Authors' contributions

L. X. and Y. C. have proposed the main idea. L. X. and Y. G. have implemented it. E. D. has technically edited the manuscript. E. D. and M. Y. have co-managed the research. J. L. has provided part of the data used.

Funding

No funding was received.

Availability of data and materials

All the data and codes are available through the corresponding author.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 7 October 2023 Accepted: 18 November 2023
Published online: 01 December 2023

References

1. Soomro K, Zamir AR (2015) Action recognition in realistic sports videos. In Computer vision in sports. pp. 181–208. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-09396-3_9
2. Qi W, Wang N, Su H, Aliverti A (2022) DCNN based human activity recognition framework with depth vision guiding. Neurocomputing 486:261–271. https://doi.org/10.1016/j.neucom.2021.11.044
3. Ramasamy Ramamurthy S, Roy N (2018) Recent trends in machine learning for human activity recognition—A survey. 8(4):e1254. https://doi.org/10.1002/widm.1254. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery
4. Wang X, Zheng S, Yang R, Zheng A, Chen Z, Tang J, Luo B (2022) Pedestrian attribute recognition: a survey. Pattern Recogn 121:108220. https://doi.org/10.1016/j.patcog.2021.108220
5. Chen K, Zhang D, Yao L, Guo B, Yu Z, Liu Y (2021) Deep learning for sensor-based human activity recognition: overview, challenges, and opportunities. ACM Comput Surv (CSUR) 54(4):1–40. https://doi.org/10.1145/3447744
6. Kim K, Jalal A, Mahmood M (2019) Vision-based human activity recognition system using depth silhouettes: a smart home system for monitoring the residents. J Electr Eng Technol 14:2567–2573. https://doi.org/10.1007/s42835-019-00278-8
7. Ullah W, Ullah A, Haq IU, Muhammad K, Sajjad M, Baik SW (2021) CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks. Multimedia Tools and Applications 80:16979–16995. https://doi.org/10.1007/s11042-020-09406-3
8. Qi W, Su H, Yang C, Ferrigno G, De Momi E, Aliverti A (2019) A fast and robust deep convolutional neural networks for complex human activity recognition using smartphone. Sensors 19(17):3731. https://doi.org/10.3390/s19173731
9. Singh R, Kushwaha AK, Srivastava R (2023) Recent trends in human activity recognition–A comparative study. Cogn Syst Res 77:30–44. https://doi.org/10.1016/j.cogsys.2022.10.003
10. Ahmad T, Wu J (2023) SDIGRU: spatial and deep features integration using multilayer gated recurrent unit for human activity recognition. IEEE Trans Comput Social Syst. https://doi.org/10.1109/TCSS.2023.3249152
11. Li Y, Liu Y, Yu R, Zong H, Xie W (2023) Dual attention based spatial-temporal inference network for volleyball group activity recognition. Multimedia Tools and Applications 82(10):15515–15533. https://doi.org/10.1007/s11042-022-13867-z
12. Khan AA, Shao J, Ali W, Tumrani S (2020) Content-aware summarization of broadcast sports videos: an audio–visual feature extraction approach. Neural Process Lett 52:1945–1968. https://doi.org/10.1007/s11063-020-10200-3
13. Ullah A, Muhammad K, Haq IU, Baik SW (2019) Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments. Future Generation Computer Systems 96:386–397. https://doi.org/10.1016/j.future.2019.01.029
14. Rodriguez MD, Ahmed J, Shah M (2008) Action mach a spatio-temporal maximum average correlation height filter for action recognition. IEEE conference on computer vision and pattern recognition. pp 1–8. https://doi.org/10.1109/CVPR.2008.4587727
15. Giuggioli G, Pellegrini MM (2023) Artificial intelligence as an enabler for entrepreneurs: a systematic literature review and an agenda for future research. Int J Entrepreneurial Behav Res 29(4):816–837. https://doi.org/10.1108/IJEBR-05-2021-0426
16. Prince SJ (2012) Computer vision: models, learning, and inference. Cambridge University Press. Jun 18
17. Oreifej O, Liu Z (2013) Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 716–723. https://doi.org/10.1109/CVPR.2013.98
18. Yang X, Tian Y (2014) Super normal vector for activity recognition using depth sequences. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 804–811. https://doi.org/10.1109/CVPR.2014.108
19. Host K, Ivašić-Kos M (2022) An overview of Human Action Recognition in sports based on computer vision. Heliyon 1. https://doi.org/10.1016/j.heliyon.2022.e09633
20. Host K, Ivasic-Kos M, Pobar M (2020) Tracking Handball Players with the DeepSORT Algorithm. ICPRAM. pp 593–599. https://doi.org/10.5220/0009177605930599
21. Al-Faris M, Chiverton J, Ndzi D, Ahmed AI (2020) A review on computer vision-based methods for human action recognition. J Imaging 6(6):46. https://doi.org/10.3390/jimaging6060046
22. Rahmad NA, As'Ari MA, Ghazali NF, Shahar N, Sufri NA (2018) A survey of video based action recognition in sports. Indonesian J Electr Eng Comput Sci 11(3):987–993. https://doi.org/10.11591/ijeecs.v11.i3.pp987-993
23. Sak H, Senior AW, Beaufays F (2014) Long short-term memory recurrent neural network architectures for large scale acoustic modeling, In:

Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September
24. Ghislieri M, Cerone GL, Knaflitz M, Agostini V (2021) Long short-term memory (LSTM) recurrent neural network for muscle activity detection. J Neuroeng Rehabil 18:1–5. https://doi.org/10.1186/s12984-021-00945-w
25. Malawski F, Kwolek B (2019) Automatic analysis of techniques and body motion patterns in sport. AGH University of Science and Technology
26. Cai, Neher Z, Vats K, Clausi DA, Zelek J (2019) Temporal hockey action recognition via pose and optical flows. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp 0–0. https://doi.org/10.1109/CVPRW.2019.00310
27. Gu X, Xue X, Wang F (2020) Fine-grained action recognition on a novel basketball dataset. ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp 2563–2567. https://doi.org/10.1109/ICASSP40776.2020.9053928
28. Varol G, Laptev I, Schmid C (2017) Long-term temporal convolutions for action recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1510–1517. https://doi.org/10.1109/TPAMI.2017.2712608
29. Lee J, Jung H, Tuhad (2020) Taekwondo unit technique human action dataset with key frame-based Cnn action recognition. Sensors 20(17):4871. https://doi.org/10.3390/s20174871
30. FarajiDavar N, De Campos T, Kittler J, Yan F (2011) Transductive transfer learning for action recognition in tennis games. 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). pp 1548–1553. https://doi.org/10.1109/ICCVW.2011.6130434
31. Wang Y, Fang W, Ma J, Li X, Zhong A (2019) Automatic badminton action recognition using cnn with adaptive feature extraction on sensor data. In Intelligent Computing Theories and Application: 15th International Conference, ICIC 2019, Nanchang, China, August 3–6, 2019, Proceedings, Part I. pp. 131–143. Springer International Publishing. https://doi.org/10.1007/978-3-030-26763-6_13
32. Martin PE, Benois-Pineau J, Péteri R, Morlier J (2018) Sport action recognition with siamese spatio-temporal cnns: Application to table tennis. 2018 International Conference on Content-Based Multimedia Indexing (CBMI). pp 1–6. https://doi.org/10.1109/CBMI.2018.8516488
33. Aktas K, Demirel M, Moor M, Olesk J, Ozcinar C, Anbarjafari G (2021) Spatiotemporal based table tennis stroke-type assessment. SIViP 15(7):1593–1600. https://doi.org/10.1007/s11760-021-01893-7
34. Giancola S, Amine M, Dghaily T, Ghanem B (2018) A scalable dataset for action spotting in soccer videos. Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp 1711–1721. https://doi.org/10.1109/CVPRW.2018.00223
35. Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp 6299–6308. https://doi.org/10.48550/arXiv.1705.07750
36. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 770–778. https://doi.org/10.1109/2fCVPR.2016.90
37. Agyeman R, Muhammad R, Choi GS (2019) Soccer video summarization using deep learning. IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). pp 270–273. https://doi.org/10.1109/MIPR.2019.00055
38. Xiong J, Lu L, Wang H, Yang J, Gui G (2019) Object-level trajectories based fine-grained action recognition in visual IoT applications. IEEE Access 7:103629–103638. https://doi.org/10.1109/ACCESS.2019.2931471
39. Zhou X, Kang L, Cheng Z, He B, Xin J (2021) Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection. arXiv preprint arXiv:2106.14447. https://doi.org/10.48550/arXiv.2106.14447
40. Ramanathan V, Huang J, Abu-El-Haija S, Gorban A, Murphy K, Fei-Fei L (2016) Detecting events and key actors in multi-person videos. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 3043–3053. https://doi.org/10.48550/arXiv.1511.02917
41. Ma C, Fan J, Yao J, Zhang T (2021) NPU RGB + D dataset and a Feature-Enhanced LSTM-DGCN Method for Action Recognition of Basketball Players. Appl Sci 11(10):4426. https://doi.org/10.3390/app11104426
42. Ibrahim MS, Muralidharan S, Deng Z, Vahdat A, Mori G (2016) A hierarchical deep temporal model for group activity recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. pp 1971–1980. https://doi.org/10.1109/CVPR.2016.217
43. Rangasamy K, As'ari MA, Rahmad NA, Ghazali NF (2020) Hockey activity recognition using pre-trained deep learning model. ICT Express 6(3):170–174. https://doi.org/10.1016/j.icte.2020.04.013
44. Vats K, McNally W, Walters P, Clausi DA, Zelek JS (2022) Ice hockey player identification via transformers and weakly supervised learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3451–3460. https://doi.org/10.48550/arXiv.2111.11535
45. Chen R, Siegler D, Fasko M, Yang S, Luo X, Zhao W (2019) Baseball pitch type recognition based on broadcast videos. InCyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health: International 2019 Cyberspace Congress, CyberDI and CyberLife, Beijing, China, December 16–18, 2019, Proceedings, Part II. pp. 328–344. Springer Singapore. https://doi.org/10.1007/978-981-15-1925-3_24
46. Sun SW, Mou TC, Fang CC, Chang PC, Hua KL, Shih HC (2019) Baseball player behavior classification system using long short-term memory with multimodal features. Sensors 19(6):1425. https://doi.org/10.3390/s19061425
47. Moodley T, van der Haar D (2019) Cricket Stroke recognition using computer vision methods. In Information Science and Applications: ICISA 2019. pp. 171–181. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-15-1465-4_18
48. Moodley T, van der Haar D (2020) Casrm: cricket automation and stroke recognition model using openpose. InInternational Conference on Human-Computer Interaction pp. 67–78. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-49904-4_5
49. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. InProceedings of the IEEE conference on computer vision and pattern recognition. pp. 7291–7299. https://doi.org/10.48550/arXiv.1611.08050
50. Moodley T, van der Haar D (2020) Scene recognition using alexnet to recognize significant events within cricket game footage. In Computer Vision and Graphics: International Conference, ICCVG 2020, Warsaw, Poland, September 14–16, Proceedings 2020 (pp. 98–109). Springer International Publishing. https://doi.org/10.1007/978-3-030-59006-2_9
51. Kushwaha A, Khare A, Prakash O (2023) Micro-network-based deep convolutional neural network for human activity recognition from realistic and multi-view visual data. Neural Comput Appl 35(18):13321–13341. https://doi.org/10.1007/s00521-023-08440-0
52. Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, Wei Y (2017) Deformable convolutional networks. InProceedings of the IEEE international conference on computer vision. pp. 764–773. https://doi.org/10.1109/ICCV.2017.89
53. Jiang G, Lu Z, Tu X, Guan Y, Wang Q (2021) Image super-resolution using multi-scale space feature and deformable convolutional network. IEEE Access 9:74614–74621. https://doi.org/10.1109/ACCESS.2021.3079519
54. Tang H, Xiao B, Li W, Wang G (2018) Pixel convolutional neural network for multi-focus image fusion. Inf Sci 433:125–141. https://doi.org/10.1016/j.ins.2017.12.043
55. Reddy KK, Shah M (2013) Recognizing 50 human action categories of web videos. Mach Vis Appl 24(5):971–981
56. Liu J, Luo J, Shah M (2009) Recognizing realistic actions from videos "in the wild". In IEEE Conference on Computer Vision and Pattern Recognition 2009 Jun 20 (pp. 1996–2003). https://doi.org/10.1109/CVPR.2009.5206744
57. Kong L et al (2022) Time-aware missing healthcare data prediction based on ARIMA model. IEEE/ACM Trans Comput Biol Bioinf. https://doi.org/10.1109/TCBB.2022.3205064
58. Yang Y et al (2022) ASTREAM: data-stream-driven scalable anomaly detection with accuracy guarantee in IIoT environment. IEEE Trans Netw Sci Eng. https://doi.org/10.1109/TNSE.2022.3157730
59. Wang F et al (2021) Edge-cloud-enabled matrix factorization for diversified APIs recommendation in mashup creation. World Wide Web 1–21. https://doi.org/10.1007/s11280-021-00943-x
60. Rezaee K et al (2021) A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. Personal Uniquit Comput 1–17. https://doi.org/10.1007/s00779-021-01586-5
61. Liu AA, Su YT, Nie WZ, Kankanhalli M (2016) Hierarchical clustering multi-task learning for joint human action grouping and recognition. IEEE Trans Pattern Anal Mach Intell 39(1):102–114. https://doi.org/10.1109/TPAMI.2016.2537337

62. Sadanand S, Corso JJ (2012) Action bank: A high-level representation of activity in video. In IEEE Conference on computer vision and pattern recognition. pp. 1234–1241. https://doi.org/10.1109/CVPR.2012.6247806
63. Tu Z, Xie W, Qin Q, Poppe R, Veltkamp RC, Li B, Yuan J (2018) Multi-stream CNN: learning representations based on human-related regions for action recognition. Pattern Recogn 79:32–43. https://doi.org/10.1016/j.patcog.2018.01.020
64. Afza F, Khan MA, Sharif M, Kadry S, Manogaran G, Saba T, Ashraf I, Damaševičius R (2021) A framework of human action recognition using length control features fusion and weighted entropy-variances based feature selection. Image Vis Comput 106:104090. https://doi.org/10.1016/j.imavis.2020.104090
65. Muhammad K, Ullah A, Imran AS, Sajjad M, Kiran MS, Sannino G, de Albuquerque VH (2021) Human action recognition using attention based LSTM network with dilated CNN features. Future Generation Computer Systems 125:820–830. https://doi.org/10.1016/j.future.2021.06.045
66. Meng B, Liu X, Wang X (2018) Human action recognition based on quaternion spatial-temporal convolutional neural network and LSTM in RGB videos. Multimedia Tools and Applications 77(20):26901–26918. https://doi.org/10.1007/s11042-018-5893-9
67. Gammulle H, Denman S, Sridharan S, Fookes C (2017) Two stream lstm: A deep fusion framework for human action recognition. In IEEE winter conference on applications of computer vision (WACV). pp. 177–186. https://doi.org/10.1109/WACV.2017.27
68. Ijjina EP, Mohan CK, Hybrid (2016) Deep neural network model for human action recognition. Applied soft computing. 46:936-52. https://doi.org/10.1016/j.asoc.2015.08.025
69. Zhou T, Tao A, Sun L, Qu B, Wang Y, Huang H (2023) Behavior recognition based on the improved density clustering and context-guided Bi-LSTM model. Multimedia Tools and Applications. 1–8. https://doi.org/10.1007/s11042-023-15501-y
70. Xiong X, Min W, Han Q, Wang Q, Zha C (2022) Action Recognition Using Action Sequences Optimization and Two-Stream 3D Dilated Neural Network. Computational Intelligence and Neuroscience. 2022. https://doi.org/10.1155/2022/6608448
71. Zhang Z, Lv Z, Gan C, Zhu Q (2020) Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions. Neurocomputing 410:304–316. https://doi.org/10.1016/j.neucom.2020.06.032
72. Dai C, Liu X, Lai J (2020) Human action recognition using two-stream attention based LSTM networks. Appl Soft Comput 86:105820. https://doi.org/10.1016/j.asoc.2019.105820

Publisher's Note