Heliyon 8 (2022) e09633

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Review article

An overview of Human Action Recognition in sports based on Computer Vision

Kristina Host *, Marina Ivašić-Kos

Faculty of Informatics and Digital Technologies; Center for Artificial Intelligence and Cybersecurity, University of Rijeka, Radmile Matejčić 2, Rijeka, 51000, Croatia

ARTICLE INFO

Keywords:
Machine learning
Human Action Recognition
Action systematization
Sports dataset
Human action recognition in sports
Sport

ABSTRACT

Human Action Recognition (HAR) is a challenging task used in sports such as volleyball, basketball, soccer, and tennis to detect players and recognize their actions and teams' activities during training, matches, warm-ups, or competitions. HAR aims to detect the person performing the action on an unknown video sequence, determine the action's duration, and identify the action type. The main idea of HAR in sports is to monitor a player's performance, that is, to detect the player, track their movements, recognize the performed action, compare various actions, compare different kinds and skills of acting performances, or make automatic statistical analysis.

As an action that can occur in the sports field refers to a set of physical movements performed by a player in order to complete a task using their body or interacting with objects or other persons, actions can be of different complexity. Because of that, a novel systematization of actions based on complexity and level of performance and interactions is proposed.

The overview of HAR research focuses on various methods performed on publicly available datasets, including actions of everyday activities. That is just a good starting point; however, HAR is increasingly represented in sports and is becoming more directed towards recognizing similar actions of a particular sports domain. Therefore, this paper presents an overview of HAR applications in sports primarily based on Computer Vision as the main contribution, along with popular publicly available datasets for this purpose.

1. Introduction

Artificial Intelligence (AI) is one of the largest computer science branches aiming to build systems that can perform tasks for which some kind of intelligence is required. In other words, the aim is to build machines that can behave like a human, think like a human, and be able to make decisions on their own [1]. A field of AI dealing with processing visual data such as images and videos in the way humans do, extracting useful information from them, and understanding their content is called Computer Vision (CV) [2].

CV is continuously growing due to the large amount of visual data obtained, such as security systems, traffic cameras, and people uploading them online daily [3]. More than 3 billion images are uploaded every day on social networks like Instagram and Facebook. In addition, hundreds of hours of videos are uploaded on YouTube, perhaps the largest search engine with videos [4]. These data can be used to create various datasets for CV implementations. Better hardware with greater computing power and easily accessible open-source machine learning algorithms tested and applied in multiple applications contributes to the exponential growth of CV. Thanks to popular libraries like Tensorflow [5] and OpenCV [6], Facebook [7], and Microsoft libraries [8], it is no longer necessary to start CV projects from scratch, which in turn leads to less time spent on implementations. Some of the main tasks of CV are to recognize whether there is an object or action in the image or video (validation), which category it belongs to (classification), where it is located (detection), and which pixels belong to the object (segmentation) [9].

Researchers have focused on more complex action recognition tasks guided by successful image classification and object detection in images and videos. The task of action recognition is to recognize which human actions are performed, in which sequence of frames, in which time interval, and where a person acting is located in the scene. In [10], the authors declare that the term human action in CV research "ranges from the simple limb movement to joint complex movement of multiple limbs and the human body." Also, they consider this process dynamic and point out that the process is usually conveyed in a video lasting a few seconds.

* Corresponding author.
E-mail address: kristina.host@inf.uniri.hr (K. Host).

https://doi.org/10.1016/j.heliyon.2022.e09633
Received 29 April 2021; Received in revised form 28 February 2022; Accepted 26 May 2022

Finally, in [11], the authors refer to the action as "simple motion patterns usually executed by a single person and typically lasting for short durations of time."

Here we consider that human action refers to a predetermined set of physical movements performed by a person in a time to complete a given task. In some cases, an object should be used to complete a task, or interaction with other persons is needed. Therefore, there is a different level of complexity of action. Some simple actions can be recognized in a single image or frame. Some complex actions take place over a more extended period, requiring a sequence of a larger number of consecutive frames (video sequence) to recognize. For example, a simple action, such as *standing* or *raising a hand*, could be identified using a single frame (image).

A longer sequence of video frames is needed to determine which action occurred for a more complex action like a *long jump* consisting of different physical movements such as *running, jumping, landing,* and *standing* up that last for a particular time. The complexity of the action can also increase if there is an interaction with one or more objects, for example, *hitting a ball with a baseball bat*, which involves a person using one object (baseball bat) to perform the action on another object (ball).

Furthermore, some human actions may involve more than one person, e.g., two people *wrestling*. According to [11], these actions are called activities and are defined as a "complex sequence of actions performed by several humans who could be interacting with each other in a constrained manner." The activities are typically characterized by much longer temporal durations. Despite interaction with multiple persons, some activities can include simultaneous interactions with objects. For example, *passing a ball involves* two persons and an object being exchanged (ball). Researchers often use the terms "action" and "activity" as equal, but we consider these terms relevant and want to clarify the differences in our paper. When referring to Human Action Recognition (HAR), we consider both the actions and activities.

Mainly, Human Action Recognition aims to detect the person performing the action, or multiple persons performing activities on an unknown video sequence, determine the duration of the action, and identify its type. It is a complex task that includes the detection of a person on image or video, the location of the action in time and space on video sequence, and finally, the recognition of the action. To make it all possible, HAR is located at the interaction of various AI fields, such as CV and Machine Learning (ML), together with Image processing (Figure 1).

HAR problems can vary widely, considering various applications and diverse data selection, so no single approach suits all the challenges one may encounter. Moreover, possible applications belong to many domains. For example, applications can be used to analyze the patient's recovery or detect unexpected events such as the fall of an older person in the home or detect suspicious behavior during CCTV surveillance,

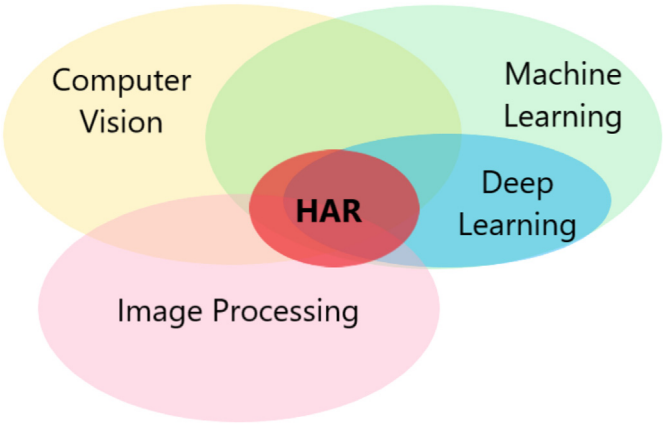

Figure 1. Human action recognition (HAR) is part of different Computer Science fields.

analyze player behavior for various human-computer interactions, and the like.

To investigate researchers' interest in the HAR topic, we examined search statistics for similar or related issues using the Google Trends application [12]. The application shows the total number of searches of given keywords in the Google Search Engine through the desired period and for the desired location. However, the number of searches does not necessarily correlate with the number of publications in the field, given that one person can search the same keywords more than once, but it certainly shows there is an interest. The main keywords we used to investigate the interest in HAR were "Human action recognition" and "Human activity recognition." However, since researchers often write only *action/activity recognition* implying that it refers to humans, we also used the keywords "Action recognition" and "Activity recognition." In this case, the search results show all actions/activities that can be recognized, such as animal activity monitoring [13]. Still, such papers are rarer and do not significantly affect the display of trends and popularity of searches related to action recognition in humans. Worldwide search results from 2005 to 2021 are shown in Figure 2., in terms of the proportion of searches of all topics on Google. The graph shows that HAR is frequently searched and that the interest has been continuous throughout the years.

Besides, Figure 3 indicates the number of publications on HAR from 2005 to 2021 retrieved using the web Dimensions application [14] for the keywords "Human Action Recognition," which should appear in the title or abstract of the publication. There is a growing trend of research on the topic of HAR throughout the observed period.

In this overview, the focus is on implementing HAR in sports, analyzing different applications in various sports, and considering specific goals for which it is necessary to recognize the actions of athletes. The complexity of athletes' actions can vary significantly from basic exercises such *as running* and *jumping* to more complex techniques such as *blocking* in volleyball, *dunking a ball* in basketball, and more complex activities involving multiple athletes interacting, such as *crossing* in handball. With this in mind, we proposed a new systematization of actions to simplify the differentiation of actions by complexity, determine the appropriate set of sports scenes, and more correctly compare the performance and capabilities of HAR algorithms.

The main contribution of this paper is to propose a new systematization of sports actions for HAR and to give an overview of computer vision-based HAR implementations in various sports. We were primarily focused on sports with a ball and two players, such as tennis and badminton, or team sports, such as volleyball, basketball, hockey, and soccer, and the description of popular publicly available sports datasets with visual data. Although most applications for HAR in sports focus on RGB visual data, researchers use other types of data, such as depth-based and skeleton-based data [15] or Optical Flow-based data [16]. Furthermore, despite the vision-based data, data obtained from sensors [17] can be used for HAR in sports, so such papers are mentioned in the overview if considered relevant when using different ML and DL methods. Still, the focus of this overview is primarily on articles with RGB visual data. When using sensors, the researchers mostly analyze the trajectory obtained by the sensors, so it is important to place sensors on the position of each joint of interest, like in [18].

Many papers deal with HAR in general (e.g., [19, 20, 21, 22]), mostly focusing on everyday activities, simple actions from different sports, and other implementations. Everyday activities include, for example, walking, shaking hands, sitting down, eating, etc., whereas simple actions from different sports include biking, horse riding, running, etc. In [23], the authors present a comprehensive overview of significant developments towards recognizing human actions, from discussing classical methods that use handcrafted features to deep learning oriented. In [20], the authors categorized the HAR methods into two main categories, unimodal (e.g., stochastic, rule-based) and multimodal (e.g., behavioral, affective), according to the nature of sensor data they employ, and then analyzed them. In [19], the authors considered action recognition for

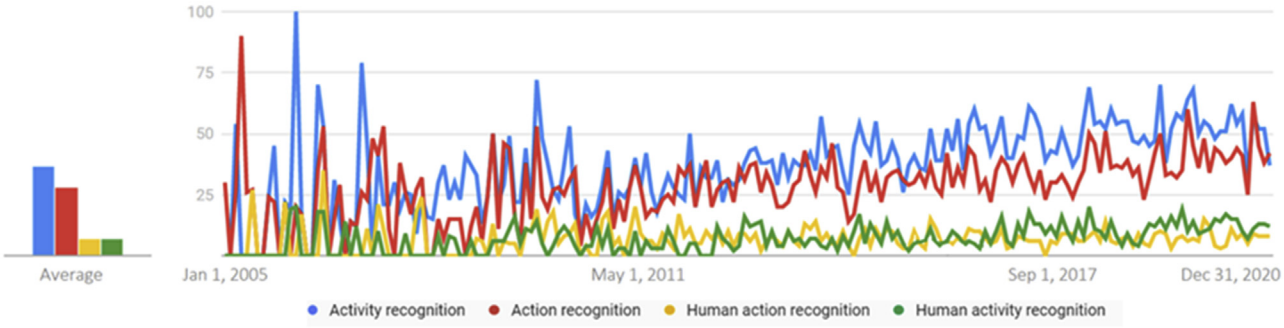

Figure 2. Interest worldwide over time (2005–2021).

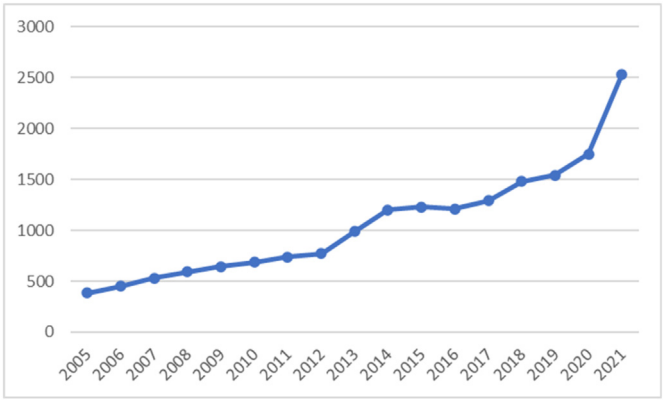

Figure 3. Number of publications over time (2005–2021).

some general actions and discussed different challenges, such as intra- and inter-class variations, cluttered background and camera motion, insufficient annotated data, action vocabulary, and uneven predictability. They use techniques common for different kinds of actions, such as decomposing complex actions at high levels of the hierarchy into a combination of actions at a lower level and dealing with uneven predictability because some actions are instantly predictable, and others need more frames to be observed. These challenges are even more pronounced in sports, so authors often try to improve existing methods or develop new ones to reduce and overcome them in the context of sports actions. Also, there are other challenges in sports such as more demanding scenes, multiple players on the field, occlusion, actions that differ very little, etc. For example, in [24], the authors provided an overview of existing commercial technologies for the sports domain. Still, they concluded that there is no tool that can automatically recognize and classify sports activities for meaningful analysis. They also discussed the implementation of deep learning for sports actions, but in their survey, there are not many papers focusing on specific actions for a particular sport (which is what we are focusing on), as the research is more focused on understanding the role of deep learning.

To include as many useful articles as possible in our overview, we investigated review articles on HAR in Computer Vision using Google Scholar, IEEE, Scopus, arXiv, and other sources. We then analyzed and selected articles that are specific examples of research done for a particular sport. We found out that most reviews focus on a set of general activities, so our idea was to focus on sports, particularly sports with a ball, to help other researchers in the field get an overview and ideas about what can be done with HAR methods in a particular sport, how far previous research has come, and to make it easier to elaborate further work. Therefore, for a more detailed examination of articles related to sports, we have used specific keywords in the sources mentioned above, such as human action, action recognition, human action classification, and the like, along with the name of the sport, such as tennis, badminton, etc.

The following chapter introduces the basic concepts of HAR in sports. In Chapter 3, a new systematization of HAR actions is proposed. In Chapter 4., the available sports datasets are described. In Chapter 5., an overview and discussion of HAR implementation considering sports with two players and team sports, along with various methods used and different challenges, is presented. The article concludes with a conclusion and a plan for future research in the field of HAR.

2. HAR in the sports domain

Recognition of human action in sports refers to using computer vision methods to detect players or recognize athletes' actions or activities. Players' or athletes' activities can be monitored during warm-ups and fitness training, specific training for a sport, matches or competitions. Therefore, HAR can be used to monitor athletes' performance (e.g., running), compare the various actions performed by different players or multiple executions of one player (e.g., backhand in tennis, serve in volleyball) to help practice a technique, or to improve the style and the like. In addition, using HAR, some automated statistical analysis of a sports match or individual athlete performances can be provided.

Nowadays, researchers are investigating various types of sports where HAR can be implemented: the individual ones, like skiing or swimming, the ones played with two players, such as tennis, and the ones played in teams, such as basketball, soccer, baseball, hockey, volleyball, handball, etc.

As a rule, sports scenes are dynamic scenes in which athletes perform several actions using various techniques while moving and changing positions, so recognizing an athlete's actions is a difficult task.

In team sports, the scene is even more complex because it involves many players on the field, often from rival teams, who need to be monitored and recognize their actions in parallel. In addition, players can perform various activities simultaneously, but with a time lag, so that each athlete performs a separate action at some point in a specific way. Often, players change their position, distance, and angle towards the camera, cover one another, enter and exit the camera's field of vision, and the like, which causes problems such as multi-object recognition with occlusion and cluttered scenes [25].

The benefits of HAR in sports can be multiple. For example, HAR can help players and coaches improve players' performance [26], physiotherapists prevent player injuries, and journalists collect statistics like the number of shots on goal [24].

HAR in sports is a typical supervised learning task on sports data, as shown in Figure 4. It begins with collecting and annotating data for a task of interest, preprocessing such as removing digital noise, and extracting features. Feature extraction in traditional machine learning (ML) techniques was manual (marked with dashed borderlines in Figure 4.), but with the advent of deep learning (DL), it was automated [15].

Methods based on hand-crafted features involve two main steps: feature extraction and dimensionality reduction, if needed, before

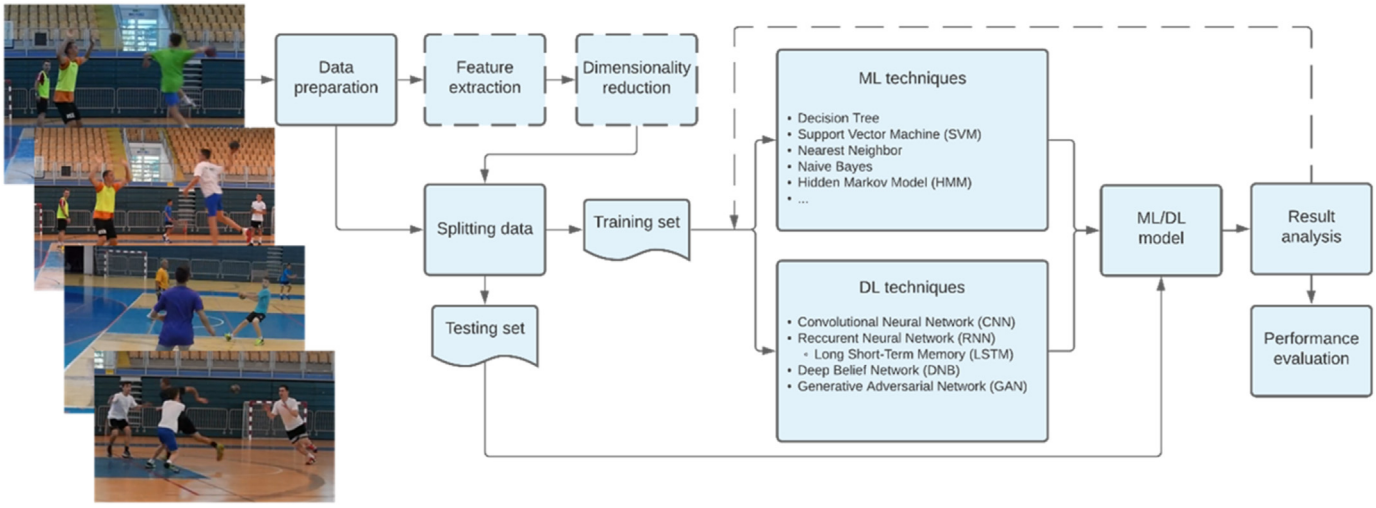

Figure 4. The workflow of the process of Human action recognition.

classification. Feature extraction can be described as a pre-processing part to remove the redundant part from the data. Features are divided into low-level and high-level features. The key points for low-level features are corners, edges, or contours. High-level features should include domain knowledge to get structured information related to the action being taken [27]. Various features, such as Optical Flow for extracting motion information, and Histogram of Oriented Gradients (HOG), can be used for HAR in sports (e.g., [28, 29, 30, 31]). Optical Flow is usually applied to a series of images, i.e., video frames, that have a short time step between them. The main idea is to calculate the velocity for points within the frames and estimate where points could be in the next image (video frame). The direction and length of arrows represent the direction and magnitude of Optical Flow at each point, as shown in Figure 5. There are many different approaches to Optical Flow, such as the Lucas and Kanade method, which can be found in [32].

The feature extraction step in HAR may allow identifying the object movement in the scene, but these descriptors do not explain the actions. Therefore, classification with ML techniques such as Support Vector Machine (SVM) is used in many cases thanks to its fast execution (e.g., [33, 34]). Other ML methods that authors use less frequently are K-nearest neighbor (KNN) [35], K-Means algorithm [36], and similar.

Actions in HAR can also be classified using different DL-based techniques, including automatic feature extraction from images, description, and classification. Lately, the DL-based methods have been used the most [37]. Although DL eliminates the manual feature extraction phase still needs a large data set to learn a large number of hidden layer parameters and more computing resources and accelerators such as graphics processing units (GPU) [38] or tensor processing units (TPU) [39]. For example, a neural network that learns to classify labeled actions will represent the input data in its hidden layers that can be used as features to represent such data [24]. The most popular models being used in DL-based implementation of HAR in sports are the Convolution Neural Networks (CNN), along with Long Short-Term Memory (LSTM) [40].

A CNN is an artificial neural network based on convolution operation, considered a two-step end-to-end classifier. At the beginning of the network, the convolutional layers are used for feature extraction and the end (fully connected, dense) for classification. One such network that represents a breakthrough in image classification and has inspired new deep learning methods, such as VGGNet [41] and I3D CNN [42], is the AlexNet [43]. The architecture of a simple CNN is represented in Figure 6. The main idea is that convolutional layers perform convolutional operations with multiple trainable filters. The convolutional results are fed to neurons with nonlinear activation and then pooled to reduce the input dimension.

This network in Figure 6 is an example of two-dimensional convolution, but 3D convolution can often be used for feature extraction in action recognition (e.g., [44, 45, 46]). A single 3D convolution can simultaneously capture the Spatio-temporal information of video action behaviors, while a single 2D convolution requires some help in capturing time information [47]. If the input data for action consists of multiple frames, 2D convolution accumulates all the results from different video frames. It outputs one image opposite the 3D convolution, which preserves the temporal information in all different video frames and generates an output volume, a collection of frames [48], as demonstrated in Figure 7. Like the pooling layers in 2D-CNNs, 3D pooling produces one output pixel for the pixel of the same color performing along with the adjacent video frames.

Also, CNN can be provided with multiple streams of data, and some layers of the network can be processed separately [49], so there are applications in sports with two-stream CNNs of different dimensions (e.g., [50, 51]). Furthermore, researchers often apply a fusion of different approaches to consider the time dimension for recognizing human actions [52].

Unlike the CNNs, which have feedforward connections, Recurrent neural networks (RNNs) [53] have relationships that feed activations from an input in a previous time step back into the network to affect the

Figure 5. Example of two adjacent frames in a video and the corresponding Optical Flow field [16].

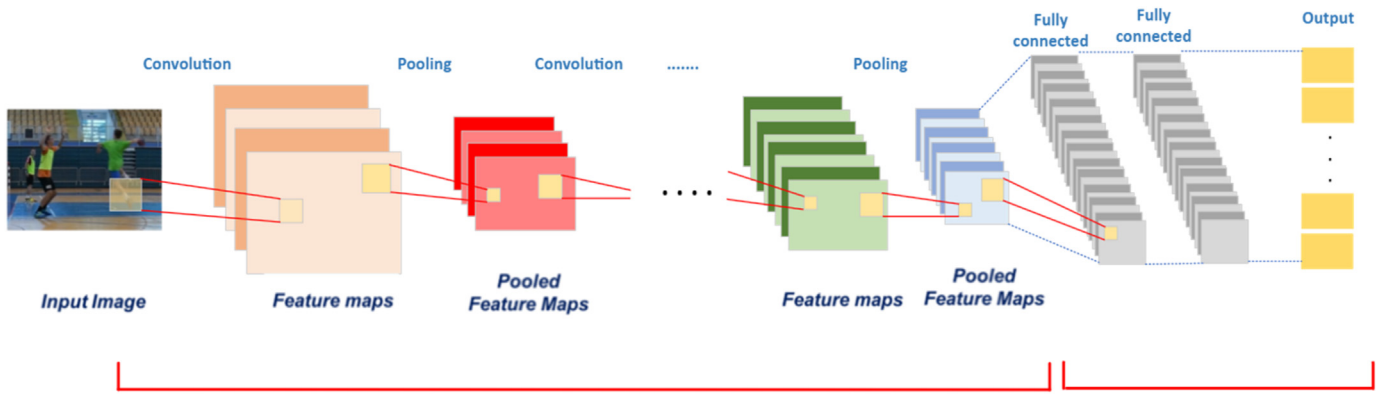

Figure 6. A simple CNN architecture.

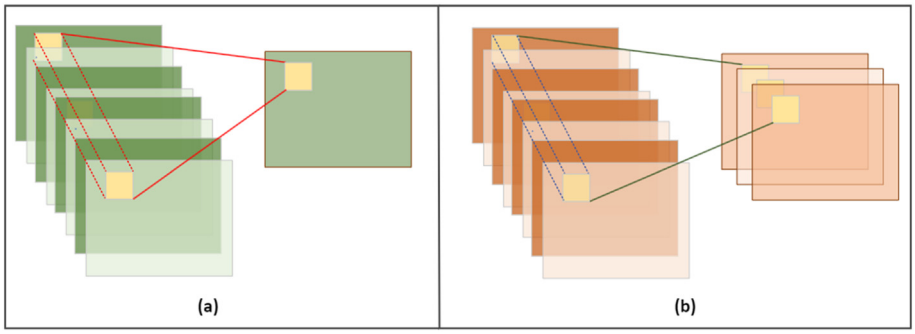

Figure 7. A comparison of the 2D convolution (a) and 3D convolution (b).

output for the current input. This property makes RNNs suitable for modeling sequences, such as video frames in action recognition. One type of RNN designed for modeling temporal sequences is called LSTM [40]. LSTMs consist of special units called memory blocks located in the recurrent hidden layers and contain memory cells. The memory cells store the temporal state of the network thanks to the self-connections. Different types of multiplicative units, called gates, are used to control the flow of information. The input gate controls the flow of input activations into the memory cell, and the output gate controls the output flow of cell activations into the rest of the network. The forget gate causes adaptive forgetting or resetting of the cell's memory by scaling its internal state before adding it as input to the cell via the cell's self-recurrent connection [16] (see Figure 8).

Such sequential flow does not make use of today's GPUs very well (designed for parallel computation), so the Transformer neural network architecture [54] that employs an encoder-decoder architecture is introduced to speed up the process. The difference from the RNNs is that the input sequence can be passed in parallel, i.e., there is no concept of the time step for the input, so all the data are passed simultaneously. Also, transformers use a self-attention mechanism (multi-head attention) that can better capture long-term dependencies than RNNs. They are typically pretrained using pretext tasks on large-scale unlabeled datasets to avoid manual annotations. The transformers were designed for natural language processing, but the authors expanded their application to visual data such as images [55] and videos (e.g., [56, 57]). In [57], the authors created a Video Action Transformer Network on top of a 3D CNN

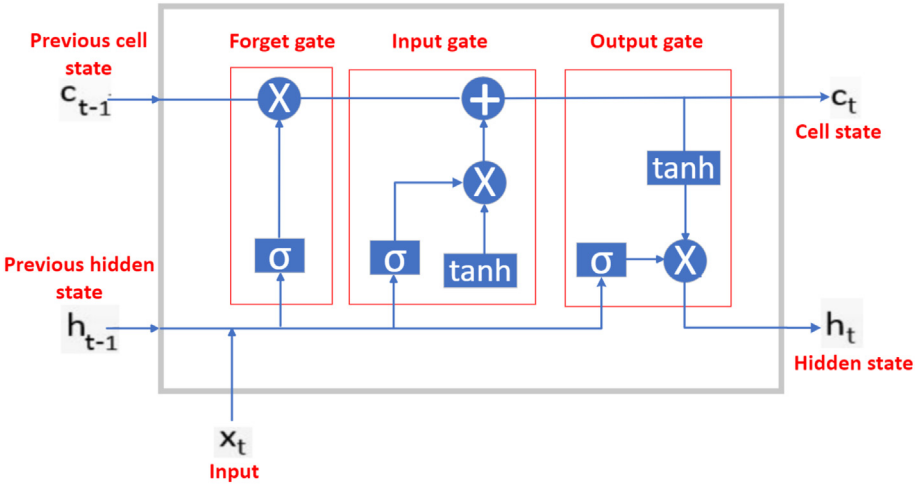

Figure 8. A representation of a simple LSTM.

representation for action localization and action classification. In their method, the Action Transformer architecture, to recognize human action, learns to attend to relevant regions of the person of interest and their context, such as other people and objects in the scene (Figure 9). Each head computes a clip embedding, which focuses on different parts like the face, hands, and the other people to recognize that the person of interest is 'holding hands' and 'watching a person.'

Transformers are achieving a better result than the methods mentioned above and are becoming a new, not yet sufficiently researched trend for HAR in sports. However, it can be assumed that more and more researchers will take this path to recognize action in sports, as the authors in [58] did.

3. Categorization of sports actions for HAR

Sports actions can be of various complexity, so inspired by human actions' categorization in [59], a new categorization for actions and sports activities is proposed in this paper (Figure 10).

According to [59], there are four levels of complexity of human activities: gestures, actions, interactions, and group activities. The gesture is considered basic movement involving some motions of the hands, fingers, or some part of the body, like stretching an arm. Unlike gestures described by mere movements, the action is composed of multiple motions organized temporally. It has an inherent goal to be achieved, e.g., for walking, one should lift legs and move arms to maintain balance and move. Interactions involve two or more persons and sometimes objects, e.g., two persons fighting or a basketball player throwing a ball. Finally, group activities are performed by groups composed of multiple persons using objects, e.g., a group of persons running.

Our proposed method is elaborated in more detail and explicitly modified for HAR application in sports. Also, actions and activities are distinguished so that actions include one player and activities several players. Certain rules predefine the actions and activities in sports. Thus, it requires an expert in a field with knowledge of the predetermined set of physical movements that must be performed to be classified as an action/activity. Of course, there can be differences in the execution of an action/activity that reflect a player's style. Still, once the predefined movements have been performed, the actions and activities can be categorized and recognized, regardless of how widely the players move their hands, whether they take larger steps or jump higher than the others.

We propose a new systematization of human actions related to sports that consists of three main categories: individual actions, joint activities, and team activities. These actions and activities can be intended (e.g., passing a ball in soccer, shooting at the goal frame, etc.) and unintended (e.g., an accidental fall or contact, a kick, a hand play in soccer, etc.). Both types can be recognized in HAR in sports, where for the first one, the goal is to determine what action was performed and whether it was performed correctly, and for the second whether an unexpected event or incident occurred. When analyzing the unintended actions and activities, it has been observed that they are mostly individual actions and joint activities. However, there are also examples of unintended team activities, such as the bench penalty in ice hockey when a team has more than the allowed number of players on the ice at the same time.

When only one player is involved in the action that needs to be recognized, it is called **individual action**. Individual actions can be divided into:

(1) gestures, the mere movement of different body attributes, mostly hands, which are used, for example, by referees to signal different kinds of fouls such as "out" in soccer and various kinds of penalties in hockey, or by players to signal for example the need for "time out" (equivalent to the gesture category from [59]);
(2) simple action, an action that consists of person movements with different body attributes simultaneously and is more complicated than the gesture itself (most similar to the action category from [59]). It can be categorized as:
   a. simple action while standing, simple action without moving across the field, such as squatting, plie and fondue in ballet, dodging in boxing, etc.
   b. simple action while moving, simple action consisting of successive movements of the person moving across the field, that is a cyclic repetition of simple action, for example, a step, that turns into walking or running
(3) interaction with objects, an action that involves objects, like throwing or catching a ball (most similar to the interaction category from [59], but in their category, the interaction can also be with other people, which we have separated in a new one called joint activity);
(4) complex action, an action that involves one or more simple actions often combined with interactions with objects, like a jump-shot in basketball or a serve in tennis. The jump-shot in basketball consists of simple actions combined with interaction with an object, that is, one jump straight vertical (simple action) while simultaneously holding the ball in place with one hand until the shutting towards the basket with the other (interaction with object). This jump-shot can be made from movement, for example, while dribbling (simple action + interaction with object). The serve in tennis is an example of complex action consisting of two interactions with objects. First, the player throws a ball in the air with one hand (interaction with object), then swings his hand and smashes the ball with a racket (interaction with object).

Actions involving at least two players, but not the whole team, performing a combination of one or more simple actions, interactions with objects, or/and complex actions are referred to as **joint activities**. Joint activity is most similar to the interaction category from [59], but in their

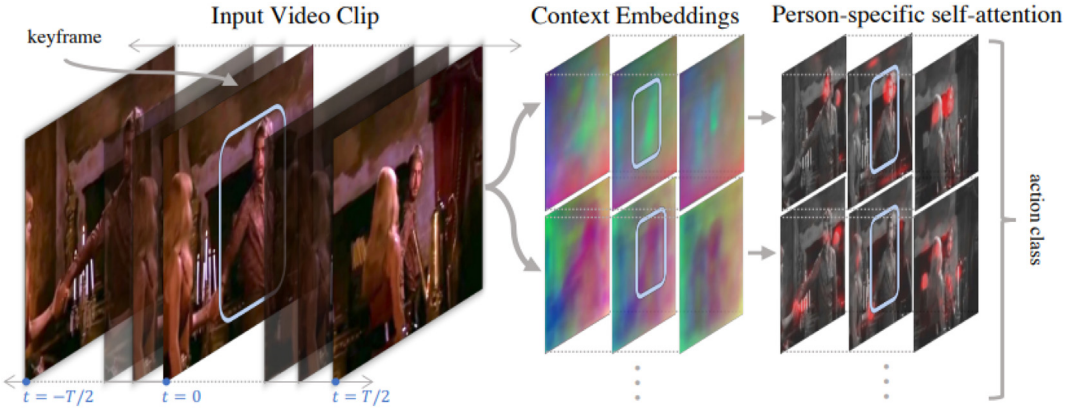

Figure 9. Action Transformer architecture learns to attend to relevant regions of the person of interest and their content [57].

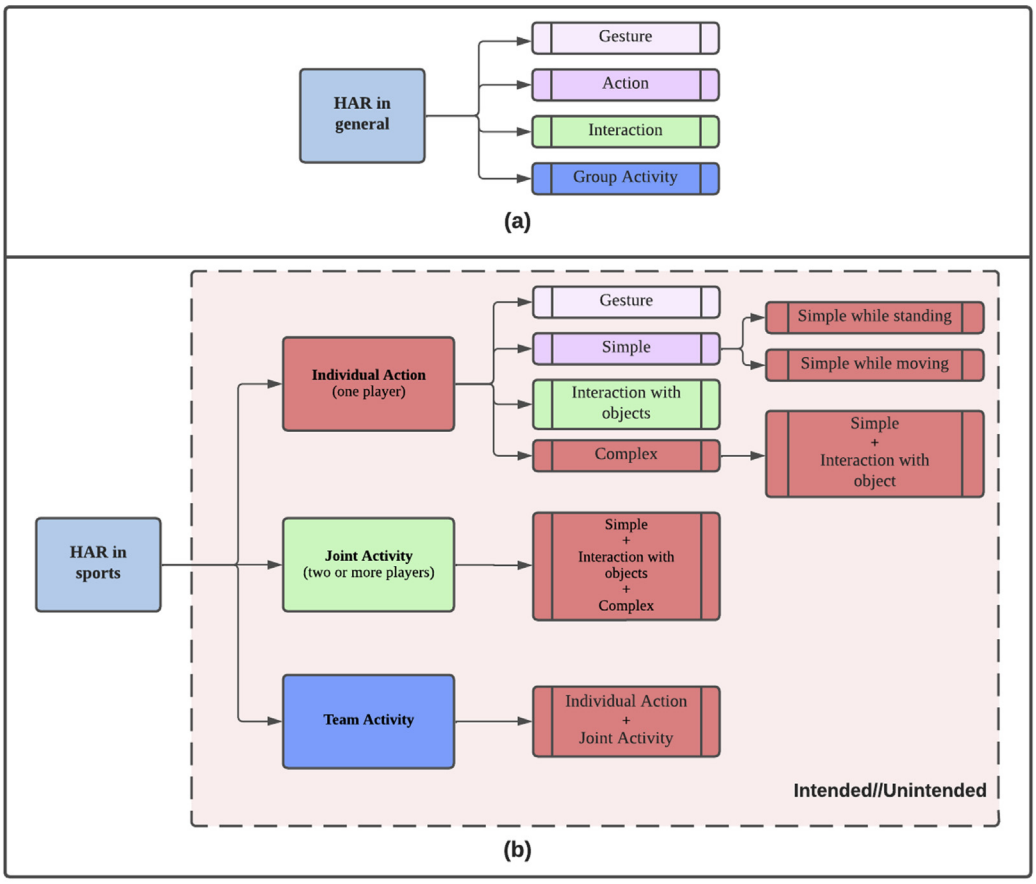

Figure 10. HAR categorization in general [59] (a) vs. HAR categorization in the sports domain (b). In the HAR in sports categorization, novel categories are presented in red shades, and with the same color as in [59] are represented the categories that can be marked as equivalent.

categorization, the interaction can also be with objects, which we have separated into a new individual action called interaction with objects. We have also presented the subdivision of which sports actions the joint activity may consist of. An example of joint activity with a combination of two simple actions is a front kick in taekwondo, while an example of a combination of two interactions is passing a ball between players in basketball, where one player throws a ball, and the other catches it (the players can additionally be running at the same time, so there are two more simple actions). Furthermore, an example of the combination of interaction with the object and complex action is when in volleyball, one player is setting the ball in the air, and the other is doing a spike (making a few steps, jumping, and hitting the ball with a hand).

Actions that are strategic and involve the whole team are called **team activities.** The team activity corresponds to the group activity category from [59], but with a detailed subdivision of what actions and activities in sports it may consist of. It includes various combinations of one or more individual actions or/and joint activities. An example of a team activity that consists of more individual actions simultaneously is the rotation of team players in a clockwise manner in volleyball. In contrast, a combination of individual and joint activities is an attack in soccer, where all the players from a team are running forward (individual actions) toward the goal of the opponents, simultaneously passing the ball between each other (joint activities), and eventually shooting at the goal frame (individual).

The main difference between [59] and our categorization, despite the elaboration and branching out in detail the individual actions in sports based on their complexity (the complex action mostly differs from the [59] categorization because it consists of more simple actions combined with interactions with objects, which was not considered in [59]), is that we pointed out that the number of players also plays a big role in defining

Figure 11. Example of consecutive frame for a complex action in handball: jump-shot.

the complexity of actions and activities because, in sports, it is not the same if only one player is performing an action, or if there are several, or if there is a whole team where approximately 6–15 players can be observed simultaneously.

For individual sports, individual actions are typical and often interact with objects, for example, in golf, there is an interaction with a golf club and a ball. In contrast, in sports with two players, such as tennis, badminton, and table tennis, there is mostly interaction between players, most rivals, and player interaction with the object—e.g., with racket and ball. All the combinations of actions and activities can be involved in team sports.

Here are more examples of actions and activities focusing on handball to clarify the novel categorization of human actions that can be recognized in sports. Gestures are mostly used by referees to signal the start of the game, fouls, and similar. Other actions, whether intended or not, are performed by players. A simple action in handball is a player running on the field, interaction with an object is shooting the ball into the goal frame, the complex is dribbling, which consists of simultaneous running (simple action) and hitting the ball on the floor (interaction with object). Figure 11 shows consecutive frames from a video sequence containing a complex action of the handball game, a jump-shot, consisting of more sequential actions starting from left to right.

The jump-shot is a complex action that consists of simple actions combined with interaction with an object, that is, a player doing three big steps (simple action), then a jump (simple action), and simultaneously throwing a ball at the goal frame (interaction with object), and finally landing (simple action).

An example of joint activity is a pass action. It is a combination of two-player activity and ball interaction in which both players stand or move (simple action) and alternately throw or catch the ball (interaction with object). Another example of joint activity is the cross-activity in which three players participate where one of them passes the ball at a certain time (interaction with object), and all of them switch their positions and direction of moving while running (simple actions).

An example of a team activity would be defense, where, for example, all team players stand or move either forward or backward, or from left to right, shifting weight from one leg to the other while raising or outstretching their arms to prevent the opponent from kicking the ball toward the goal, or passing it to a teammate (simple actions).

As shown, a player's action in a team sport, such as handball, can be analyzed on an individual level of performance (e.g., Figure 12, red bounding box, the player in white jersey is throwing a ball) and be categorized as individual action, but also can be observed for a small group performing an activity (e.g., Figure 12, green bounding box, two players in white jerseys passing a ball) and be categorized as a joint activity, or for the whole team (e.g., Figure 12 blue bounding box, four players in white jerseys performing an attack) and be categorized as a team activity.

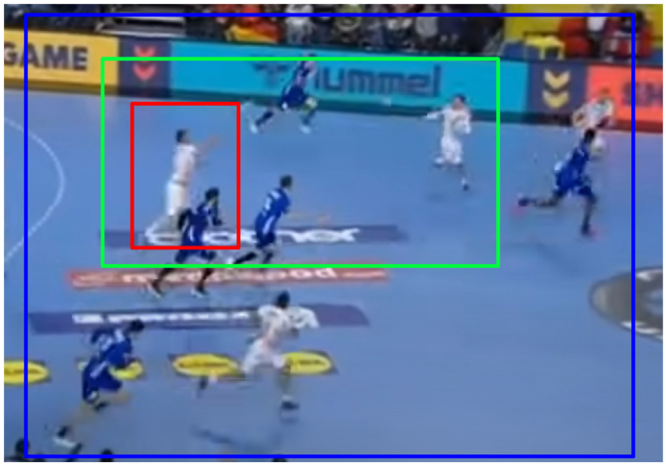

Figure 12. Example of HAR categorization in handball: red – individual action (throwing), green – joint activity (passing), blue – team activity (attack).

It should be emphasized that recognizing an action in sports is more complicated if focused on multiple players. For example, suppose you want to recognize the action of a single athlete performing the long jump. In that case, you need one or more cameras positioned on that athlete, and it is easier to prepare the input data for the HAR process, unlike when you have a team and want to detect a joint activity. In this case, you need to take more steps to pre-process the data, such as detecting and tracking all the players to find out which players are active and performing the activity that can be recognized at the end. For team activity, you also need to analyze the behavior of multiple players, for example, from the bird's eye view or with multiple cameras.

4. HAR datasets

Labeled data are required to train and evaluate the HAR methods. For the task of HAR in general, traditionally available public data sets are KTH [60] and Weizmann [61], which contain various actions, some of which can be related to the sports domain. They were introduced at the beginning of the 21st century. Compared to today's datasets, they contain a small number of action classes and a modest number of samples recorded in laboratory conditions. The KTH data set consists of only six classes (walking, running, boxing, waving, and clapping) performed by 25 people several times in four different scenarios. In contrast, sets such as HACS [62] and Kinetics 700–2020 [63] have significantly more classes, significantly more data, and were recorded in realistic conditions. E.g., the Kinetics set is a large dataset (with 400/600/700 human action classes, depending on the version) containing manually tagged videos downloaded from YouTube. The HACS (1.5M) is a large dataset for identifying and temporally localizing human actions collected from web videos. There are 6,273 times more samples in HACS than, for example, in KTH (2,391). For HAR in the sports domain, there are six public action datasets that include different actions from different sports: UCF Sports Action Data Set [27], Olympic Sports Dataset [64], Sports-1M [65], FineGym [66], SVW [67], and MultiSports [68].

4.1. UCF Sports Action Data Set

The UCF Sports Action Data Set, shown in Figure 13, consists of different labeled actions collected from various sports typically featured on broadcast television channels, together with their bounding box annotations of the humans shown in yellow. The dataset contains 150 sequences with a resolution of 720 × 480 and a frame rate of 10fps. The total duration of the dataset is 958s, and the mean sequence length is about 6.39s. The dataset includes ten actions in different environments (sports hall, sports field, nature, etc.): Diving, Golf Swing, Kicking, Lifting, Riding Horse, Running, Skateboarding, Swing-Bench, Swing-Side, and Walking. The number of sequences is not the same for each class, e.g., the action of lifting contains a minimum of 6 sequences and the action of walking a maximum of 22 sequences. Furthermore, specific actions are short, such as kicking, compared to walking or running, which are relatively longer.

Considering the categorization of sports actions proposed in the previous section, this dataset contains only individual action that needs to be recognized. Also, there are no examples of primitive actions in the images. Examples of simple actions are, for example, walking and running; an example of interaction is kicking, lifting, swinging in golf, and complex activities are, for example, swings and swing benches.

4.2. Olympic Sports Dataset

The current release of the Olympic Sports Dataset contains video sequences of athletes practicing 16 different sports. It contains 50 videos from each of 16 classes: high jump, long jump, triple jump, pole vault,

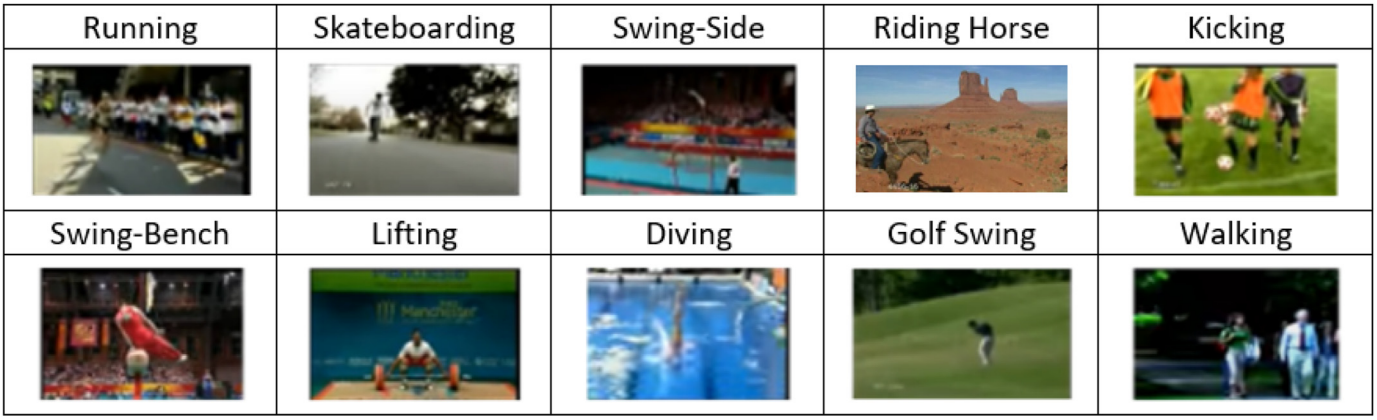

Figure 13. Example of video frames of different actions from the UCF Sports Action Data Set.

discus throw, hammer throw, javelin throw, shot put, basketball layup, bowling, tennis-serve, platform diving, springboard diving, snatch weightlifting, clean and jerk weightlifting and gymnastics vault.

The videos are taken from YouTube and contain realistic scenes. Videos are annotated with Amazon Mechanical Turk's help, and the shots are characterized by camera movements, changing the distance of the object from the camera, and severe occlusions. Figure 14 shows example frames of several action classes of the dataset, from where it is evident that action takes place in the sports hall or sports fields, depending on the observed action. In terms of complexity, all actions in the set are individual actions; precisely, there are interactions like tennis-serve and bowling, and complex actions that prevail, such as high-jump, long-jump, triple jump, etc. For instance, sequences from the long-jump action class show an athlete first standing still in preparation for the jump, followed by running, jumping, landing, and finally standing up.

4.3. Sports-1M dataset

The Sports-1M dataset contains 1,133,158 video URLs for YouTube videos, annotated automatically with 487 sports labels using the YouTube Topics API. The classes are arranged in a manually curated taxonomy that contains internal nodes such as Aquatic Sports, Team Sports, Winter Sports, Ball Sports, Combat Sports, Sports with Animals, and generally becomes fine-grained by the leaf level.

The actions are performed in gyms, swimming pools, sports halls, sports fields, roads, forests, ice rinks, ski slopes, sky, parks, and other places people encounter daily. Some examples of sports scenes and their appropriate labels are shown in Figure 15. There are all types of action in this dataset, but the main focus is on differentiating between sports such as tennis, hockey, swimming, water polo, skiing, etc.

4.4. Sports Videos in The Wild (SVW)

A dataset of Sports Videos in the Wild (SVW) is a challenging dataset captured by amateur users of the Coach's Eye mobile app with their smartphones while practicing a sport or watching a game. The dataset contains 4100 videos and consists of 30 sports categories and 44 actions such as hockey, long jump, pole vault, rowing, running, shot put, figure skating, skiing, soccer, swimming, tennis, volleyball, and others (Figure 16). Due to the unprofessional capturing of the videos, there are challenges with camera vibration and motion, occlusion, viewpoint variation, and poor illumination. Also, due to the imperfect practice of amateur players, there is a different performance of an action in unrepresentative environments, typically outside the sports field (background clutter).

4.5. FineGym dataset

The FineGym dataset is built on top of 708 h of gymnastic videos hosted on YouTube, where over 95% of these videos are of high resolutions (720P and 1080P). The videos in FineGym are all official recordings of top-level competitions, where the action instances are rich and diverse in terms of viewpoints and poses. Also, unlike other datasets where the background can help distinguish between actions, all instances in FineGym have relatively consistent backgrounds. The dataset is different from the datasets mentioned above because it provides temporal annotations at both actions and sub-actions levels with a three-level semantic hierarchy: event, set, and element (Figure 17).

The authors analyzed four gymnastic routines: balance-beam, uneven-bars, vault, and floor exercise (Figure 18). These routines are identified in each video, and they are decomposed via manually constructed decision trees into sub-actions (e.g., a *balance beam* event is

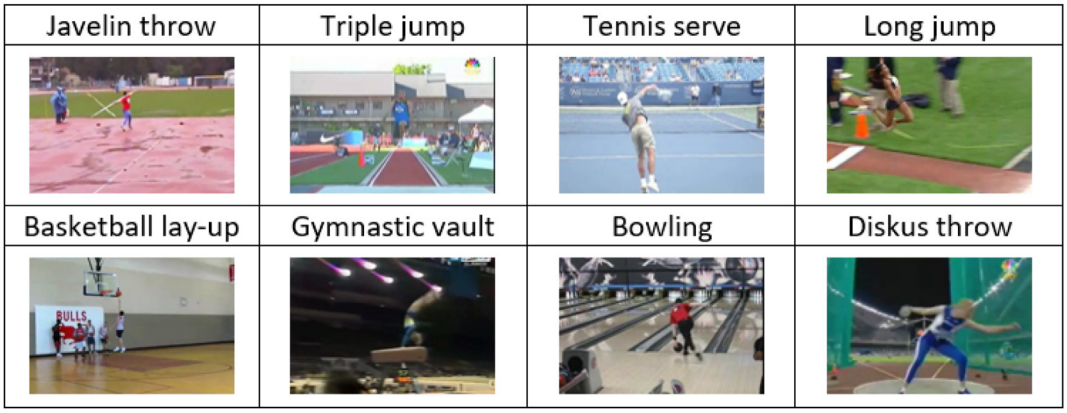

Figure 14. Example of video frames containing different actions from the Olympic Sport Dataset.

Figure 15. Example of video frames containing different actions from the Sports-1M Dataset.

Figure 16. Example of video frames containing different categories from the Sports Videos in The Wild (SVW) dataset.

annotated as a sequence of elementary sub-actions derived from five sets: leap-jumphop, beam-turns, flight-salto, flight-handspring, and dismount, where the sub-action in each set is further annotated with finely defined class labels).

In order to consider all element category combinations and enable easier database replenishment, a total of 530 categories have been defined, where 354 have at least one instance. The number of instances in each element category ranges from 1 to 1, 648, and the total number of samples considering all classes reaches 32 697. Because of missing instances for some element categories, it is more advisable to use a version named Gym288 with 288 classes or a more balanced one, Gym99 with 99 classes.

4.6. *MultiSport*

MultiSport is a new realistic multiperson dataset of Spatio-temporal localized sports actions. The data is collected for four different sports:

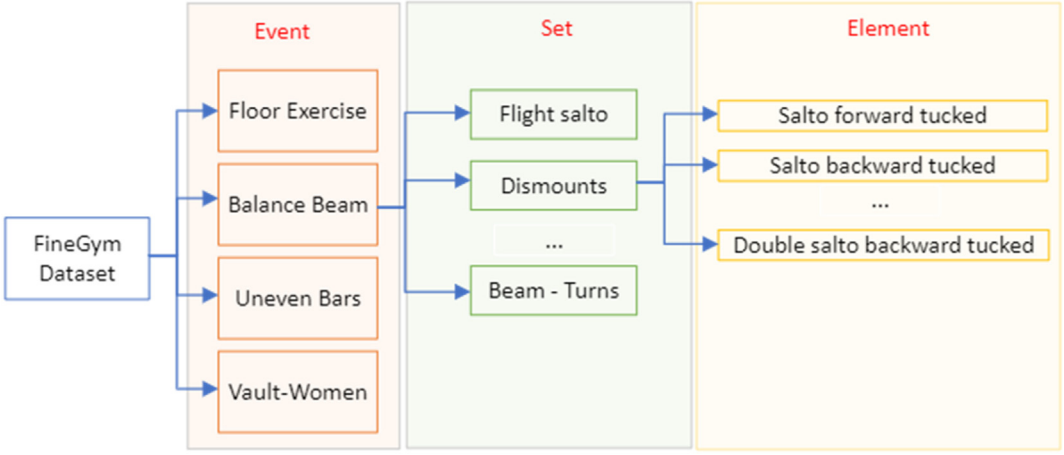

Figure 17. FineGym has a three-level semantic hierarchy (event, set, and element).

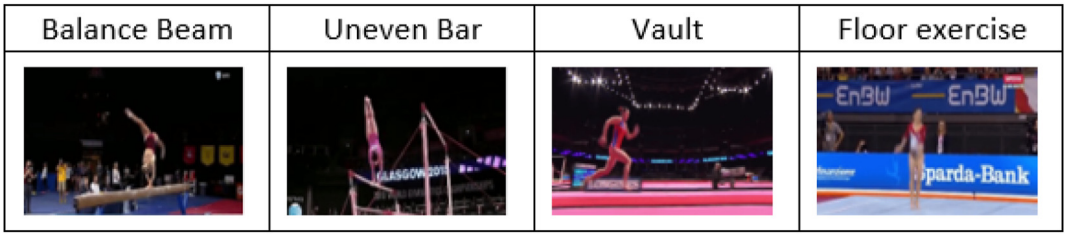

Figure 18. Example of video frames of different sub-actions from the FineGym Dataset.

aerobic gymnastics, basketball, football, and volleyball, where the actions of each sport are fine-grained and annotated, which brings to 66 action classes such as pass, drive, and dribble in basketball (see Figure 19). From 3200 video clips, 37701 action instances are annotated with precise and dense annotations in spatial and temporal domains, representing 902k bounding boxes around players acting. The dataset is large-scale and high-quality but also very challenging due to multiple players in the scene and fine-grained actions that are very similar, and the background is far less characteristic, and it cannot provide sufficient information for fine-grained action recognition.

4.7. Custom datasets

Despite these six datasets being used for recognizing various sports actions such as bowling, swimming, or running, there are numerous videos of sports scenes, matches, and competitions that can be downloaded from the website. However, they are mostly not tagged and prepared for machine learning, so it takes much work to prepare the appropriate sports scenes for machine learning. For this reason, to train a HAR model for a sport of interest, researchers are very often forced to create their own sets because there is no representative data set that could be used as a benchmark. That can be seen in the following chapter, where various HAR implementations are described. Most of them use custom-made datasets, like in [31], where authors created three custom datasets for ballet, tennis, and football to test their method for action recognition.

An example of a custom dataset can be found in [69], where researchers recorded a handball training session, with sequences captured both indoors and outdoors by full HD cameras situated at different heights and locations on the sports field. In the dataset, there are more than 700 sequences of individual actions such as running (simple action) and shooting at the goal frame (interaction with an object), complex actions such as dribbling, and joint activities such as passing the ball between two players. The dataset is very challenging due to the large number of players, cluttered backgrounds, sunlight for outdoor sequences, artificial lighting for indoor sequences, etc.

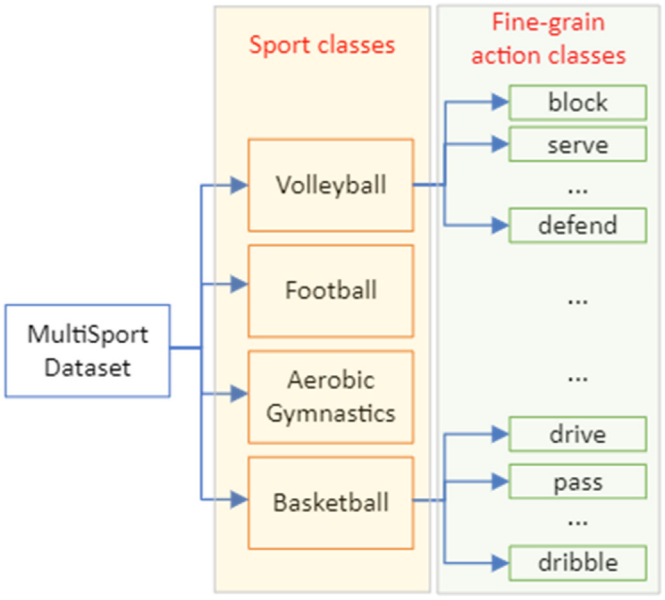

Figure 19. MultiSports has a two-level hierarchy of action vocabularies, were the actions of each sport are fine-grained.

It should be emphasized that for a practical assessment of HAR techniques in a sport, a dataset should be used that realistically describes the actions in that sport with all possible challenges and is large enough to learn the parameters of the method to build a reliable model. However, if there is no publicly available set, preparing such a set for learning is time-consuming, expensive, and tedious because it requires recording in natural conditions and later tagging and preparing the recordings.

5. HAR implementation in different sports

Some researchers focus on action recognition using benchmark datasets or custom datasets [70], which usually include significantly different actions of individual sports such as golf swinging, kicking, lifting, horseback riding, skiing, long jump, etc., or team sports focusing on one athlete, where the athlete performs an action or technique and possibly interacts with some sports equipment. However, recognizing similar actions is becoming more popular and challenging, especially if multiple players in the field are performing different actions simultaneously (Figure 20). HAR is applied in individual sports such as Taekwondo [71]. Still, here we analyze sports involving the ball (which moves fast and presents a challenge to find active players) and involving two or more players. In sports with multiple players, HAR usually encounters different viewpoints, changing lighting conditions, occlusions due to the presence of many players, and difficulties in localizing a player due to the complexity of the background.

Here is an overview of HAR implementations in sports with the ball divided into two players and team sports. The sports category with two players includes tennis, badminton, table tennis, and team sports, including soccer, basketball, volleyball, hockey, handball, baseball, cricket, rugby, and American football. Other sports fall into these two categories, but either no research papers were found or are not publicly available, so they cannot be analyzed.

5.1. Sports with two players

It should be noted that the following sports with two players can be played in pairs, but the principle of the game and the recognition of actions is the same whether there are 2 or 4 players on the field. The following sports are racket sports that mainly involve the movement of the arm.

5.1.1. Tennis

At the beginning of the 21st century, an automatic annotation method for tennis video content-based retrieval is proposed [72]. Players' actions are analyzed by 2D appearance-based matching using the transition of players' silhouettes and the Hidden Markov model. The researchers were focused on three actions: foreside-swing, backside-swing, and

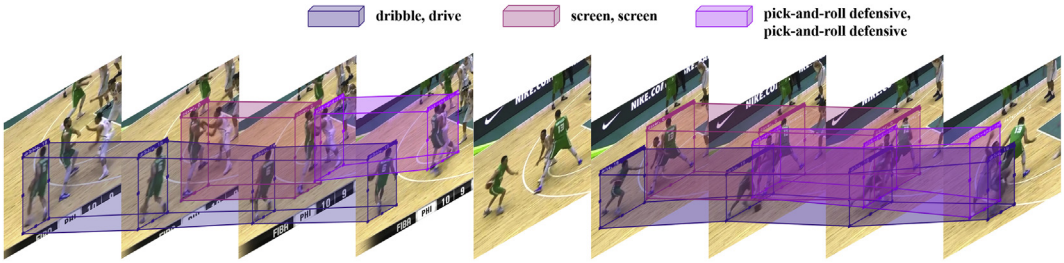

Figure 20. Example of multiple players performing different actions in basketball simultaneously (tubes with the same color represent the same person during time) [68].

over-shoulder-swing. Later in [73], their method was improved by combining player and ball position. Still, ball detection and tracking were difficult due to the poor quality of the videos.

Guided by their work, the authors of [74] focused on two primary actions: left-swing and right-swing, covering 90% of players' behavior in tennis. The main challenge is the far-view frames when a player figure might be only 30 pixels tall. They proposed a player action recognition model based on motion analysis and multimodal features for semantic and tactic analysis. They treat Optical Flow as spatial patterns of noisy measurements instead of precise pixel displacements.

In [31], the authors developed an approach to recognizing actions in the "medium" field (near-field - players approximately 300px tall, far-field - players approximately 3px tall). Their first step was to track and stabilize each human figure. Then, in order to obtain classification labels and other associated information like joint location and appearance, the authors used Optical Flow for each figure and an associated similarity measure. They prepared three different datasets to test their method, one of which concerns tennis actions. The dataset contains outdoor footages of two amateur tennis players (approximately 50px tall). The six considered actions are *stand, move left, move right, move left and swing,* and *move right and swing*. 4610 video frames of a female player were used to train a 5-nearest-neighbors classification, and 1805 frames of a male player were used to test the classification. In the results, it can be seen that the class *go left* is often confused with *go left and swing*. Moreover, some swing actions are misclassified when the Optical Flow misses the tennis racket.

[30] implemented transductive transfer learning methods for tennis action classification. Actions, in their work, are described using the Histogram of Oriented Gradient (HOG) 3D features, and for transfer, they used a method based on feature re-weighting and a method based on feature translation and scaling. They worked on a non-publicly available dataset and classified actions as "non-hit," "hit," and "serve" actions.

With the development of DL and the introduction of the THETIS dataset [75], more researchers focused on HAR in tennis. The mentioned dataset contains 1980 RGB videos of 12 tennis actions performed three times by 55 different players. Actions are performed using a tennis racket, but there is no tennis ball in the videos. The 12 actions are: backhand with two hands, backhand, backhand -slice, backhand volley, forehand flat, forehand with an open stance, forehand - slice, forehand volley, service flat, service kick, service slice, smash.

In [76], the authors proposed a model for recognizing tennis shot sequences. In their model, actions are regarded as a sequence of words (phrases) where an alphabet of possible motions is defined. They tested this approach with SVM and Conditional Random Fields on the RGB videos of the THETIS dataset.

Recent researchers [77] focused on fine-grained action recognition in tennis. In their model, videos are represented as feature sequences, extracted using the Inception neural network, trained on an independent dataset. Then a 3-layered LSTM network is trained for the classification. They test their results on the THETIS dataset.

In [78], the researchers explored the possibilities of using convolutional neural networks to recognize the type of tennis shots. They compared Inception-v3 [79] and MobileNet networks [80]. The focus was on action recognition of the following actions: the backhand preparation phase, backhand shot, forehand preparation phase, forehand shot, and non-shot. MobileNet network achieved better results.

The authors [81] proposed a weighted LSTM for 3D tennis shot recognition. From each video frame, the local spatial representations are extracted with a pre-trained Inception network. After that, a weighted LSTM decoder is introduced to model historical information, and the weighted LSTM is used to classify the action video content. The model was applied to the THETIS dataset.

In [82], the authors focused on recognizing two basic shots in tennis: forehand and backhand. They proposed using Spatial-Temporal Graph Neural Networks (ST-GCN) on images obtained from 3D tennis movements recorded by the Vicon motion capture system [83]. They compared two methods of putting data into the ST-GCN and achieved better results by fuzzing the data relative to results without fuzzing the data.

In [84], the authors used a knowledge-based approach where some intelligence is induced to the system to recognize tennis actions. They focused on masking the key functional object that acted as a base for sports scene detection and proposed an improvised mask RCNN technique.

In [85], the authors proposed a DL method introducing channel and spatial attention modules sequentially in the network. They use a CNN to extract the visual features passed through the attention module. The obtained transformed features maps are then given to the bi-directional LSTM network, whose output is then proceeded to a fully connected layer with a SoftMax classifier that assigns the probabilities to the actions performed. Finally, the network is trained and validated on six tennis players' actions from the THETIS dataset and obtained promising results.

5.1.2. Badminton

In addition to visual RGB data, authors sometimes use sensors to retrieve data and recognize actions and activities in a particular sport.

In [86], researchers recognized ten badminton actions from 300 depth map sequences acquired by the Microsoft Kinect sensor. Skeleton orientation details of badminton players were computed and extracted to form a bag of quaternions feature vectors. After conversion to a log-covariance matrix, the system is trained, and an SVM classifier classifies the badminton actions.

In [33], the authors proposed an approach for badminton stroke recognition using dense trajectories and trajectory-aligned HOG features, calculated inside local bounding boxes around players. A four-class SVM classifier is then used to classify badminton strokes from video footage to smash, forehand, backhand, etc.

In [17], the researchers inserted a sensor chip into the badminton racket for automatic action recognition to collect ten significant badminton actions. First, they achieved the best results with the proposed AFEB-AlexNet, where they proposed a specific block of adaptive feature extraction, then with AlexNet [43], and then with the LSTM networks. The inputs to the networks are original sensor data.

In [87], a model is proposed to recognize two badminton action classes on match images: hit and non-hit, using a pre-trained AlexNet for

features extraction and SVM for classification. Before using pre-trained AlexNet for automatic feature extraction, they have introduced a new local CNN extractor in the recognition pipeline.

Later in [88], they compared, for the same two actions, four different pre-trained models of deep CNN: AlexNet, GoogleNet [89], VGGNet-16, and VGGNet-19 [41]. The result shows that the GoogleNet model has the best performance.

In [90], the authors focused on an automated stroke detection and classification system for racket sports: badminton and table tennis. They extended existing sensor-based movement measuring methods with a multilayer hybrid clustering model using wristbands and smartphones. They presented a three-level K-means clustering to optimize feature extraction and segmentation with the density-based spatial clustering of applications with noise (DBSCAN) algorithm to determine the feature center of different movements. They analyzed four different actions considering badminton: Service, Drive, Smash, and Picking up, together with Walking. Their model shows good recognition for similar movements in badminton and table tennis.

In [91], the goal was to recognize badminton smash on broadcasted videos using pre-trained CNN methods. Smash and other badminton actions were studied, such as clear, drop, lift, and a net shot. They performed two experiments. The first experiment studies the performance of four existing pre-trained models: AlexNet, GoogleNet, VGGNet-16, and VGGNet-19, while recognizing five actions. The results show that the pre-trained AlexNet model has the highest performance accuracy. The second experiment is the study on the performance of two different pre-trained models: AlexNet and GoogleNet, to recognize smash and non-smash action only. The results show that the pre-trained GoogleNet model produces the best performance in recognizing smash action.

In [34], on the same badminton dataset, AlexNet and GoogLeNet neural networks are used for feature extraction to compare the performance between the two models. Then features were classified using SVM. The results showed that the feature extractor using the AlexNet CNN model has the best performance.

In [92], the author proposed a Hierarchical Multi-Classification framework for sensor-based badminton activity recognition with the help of the prior information on badminton activity categories and tested it on their dataset with video samples called the BSS-V2 dataset. Firstly, the samples are classified into primary classes. Then, they are classified into specific ones; prior human information of the badminton activity categories is manipulated to design the mapping from the main classes to the sub-classifiers.

In [93], the authors focused on recognizing badminton actions with wearable sensors. In their case, a single acceleration sensor is fixed on the end of the badminton racket handle, which is used to collect the data. To extract the hitting signals, a sliding window segmentation technique is used, and to identify ten different badminton strokes, the authors improved the hidden Markov model (HMM). The observed actions are: service, forehand chop, backhand chop, the goal, the forehand and backhand, forehand drive, backhand push the ball, forehand pick, pick the ball backhand, and forehand. They obtained satisfactory results in differentiation between actions and, with their method, improved the average recognition rate compared with traditional HMM.

5.1.3. Table tennis

In [35], the authors used Wi-Fi signals to recognize nine various table tennis actions. They used a discrete wavelet decomposition to decompose the Wi-Fi signal combined with SVM and KNN to classify the actions later. Wi-Fi devices provide channel status information for the Wi-Fi installation. By recording information on the carrier's status between the transmitter and the receiver, such installations can reflect the changes in the wireless signal and then obtain fine-grained wireless signal measurements.

In [94], the authors used body sensors placed on the lower arm, upper arm, and back of a player to collect information about angular velocity and acceleration in order to differentiate between players' stroke movements. Feature dimensions are reduced with Principal component analysis (PCA), and SVM, Decision trees, Naive Bayes, and KNN are applied to recognize strokes. Their dataset consists of 270 sample data from nine players, where the table tennis strokes are: *smash, block shot, forehand chop, backhand chop,* and *forehand drive*. The SVM achieved the best recognition accuracy of strokes.

In [95], the authors recognized hitting and missing actions in table tennis. Six participants wore wristband sensors, implemented using micro-motion sensors, on the wrist holding the bat. Various methods, such as KNN, SVM, decision tree, linear discriminant analysis, and Naive Bayes, were used to identify not only the hitting and missing of the ball but also the identities of the volunteers. By analyzing the feature selection, the authors concluded that the speed and agility of the hitting motion as well the hitting angle have a great influence on hitting the ball.

In [29], the author applies the Siamese Spatio-temporal Convolutional neural network to recognize 20 table tennis stroke actions with low inter-class variability. Their model takes as input an RGB image sequence from their TTStroke-21 dataset (129 videos representing 94 h of a table tennis game, introduced in MediaEval Challenge 2020), on which is computed the Optical Flow. After three Spatio-temporal convolutions, data are fused in a fully connected layer of the proposed Siamese network architecture. Later in [96] based on previous research, a temporal segmentation of table tennis strokes in videos is performed based on Optical Flow, temporal sliding windows, and a classifier performing detection and classification simultaneously. In [97], a region of interest (ROI) is inferred from the Optical Flow to give a probabilistic classification of all the table tennis strokes. In [45], a Twin Spatio-temporal Convolutional Neural Network, which takes as inputs an RGB image sequence and its computed Optical Flow, is proposed. It is a two-stream network comprising three Spatio-temporal convolutional layers, followed by a fully connected layer where data are fused. In [98], Twin convolutional neural networks are used with 3D convolutions both on RGB data and Optical Flow, then, they introduced 3D attention modules and examined their impact on classification efficiency. The use of attention blocks in the network speeds up the training step and improves the classification scores up to 5% with the Twin model. All the above works by the authors are united and presented in the thesis [99]. Later, in [100], they developed a two-stream Convolutional Neural Network processing in parallel the RGB Stream and its computed Optical Flow as part of the MediaEval 2021 benchmark for the Sport task.

In [28], the authors used action recognition to annotate sports videos with different table tennis strokes. First, they used K-means to classify the Optical Flow singularities into six clusters combined with spatial information and later with HOG features. They then used cross-validated linear SVM to classify the actions. The dataset obtained is also based on TTStroke-21 [29].

In [90], the authors focused on an automated stroke detection and classification system for racket sports, as mentioned in the Badminton section. They analyzed four different actions considering table tennis: Service, Stroke, Spin, Pick up, together with Walking action. Compared to vision-based methods, wristbands have great advantages in privacy and tolerance to external influences. Therefore, they can provide a reference solution for commercial action recognition in racket sports.

In [101], the authors proposed a multistage deep neural network pipeline for recognizing stroke types of table tennis using Spatio-temporal features, which predicts the final class with different aspects at each stage where RGB image-based, Optical Flow-based, pose-based, and region-of-interest-based methods are used. Outcomes of each stage are then fused to obtain the final prediction on the TTStroke-21 dataset. Finally, the best results are obtained using RGB images and Optical Flow-based methods together.

In [102], the authors proposed a well-established hybrid second-order Markov chain model to characterize and simulate the competition process in table tennis. Their method is the first to support the effective simulation of tactics. Also, they introduced Tac-Simur, a visual analytics system called based on the proposed model. Later in

[103], they proposed Tac-Valuer, an automatic stroke evaluation framework for analysts in table tennis teams. Finally, they integrated analyst's knowledge with Abductive Learning (integrates machine learning and logical reasoning).

5.2. Team sports

5.2.1. Soccer

At the begging of the 21st century, in [31], authors developed an approach to recognizing actions in the "medium" field described in the Tennis section. They prepared three different datasets to test their method, one of which concerns soccer actions. The dataset consists of 4500 frames extracted from 72 tracked sequences from the World Cup game and considers eight actions: *walk left, walk right, walk-in/out, run left, run right, run in/out, run left 45 ◦,* and *run right 45 ◦*. For classification, a 1-nearest-neighbor method was applied to the entire soccer dataset using the leave-one-sequence-out scheme for testing. Despite the small size of players on the field, the results are very satisfactory.

In [104], the authors proposed a multi-view framework for action recognition in soccer that extracts human silhouettes. Their focus was on scene dynamics analysis, that is, 3D motion analysis of a ball and player. Also, they merged the results of this approach with the Contourlet transform neural recognition tool to recognize actions.

In [105], for the task of recognition of group activities in videos of soccer, the authors proposed a local motion-based approach. They analyzed the local motion of individuals using SIFT key-point matches on two consecutive frames and proposed to group them into the background point set (to estimate camera motion) and the foreground point set (to represent group activity). They used the bag-of-words method to estimate the relative motion and then classified the results with SVM.

Later in the 21st century, a dataset called SoccerNet [106] was introduced for action spotting in soccer videos collected from online sources. The annotated action can be categorized into goals, cards, and substitutions. The dataset includes 6637 samples of actions. The feature extraction was performed with a 3D CNN [44], I3D CNN [42], and ResNet [107]. They used different pooling methods (mean pooling and max pooling), custom CNN, SoftDBOW, NetFV, NetVLAD, and NetRVLAD [108].

In [109], the pose-projected action recognition hourglass network (PARHN) is introduced for performing player-level action recognition in soccer. It includes an embedded pose projection component that regularizes the player's pose vector's range and incorporates the temporal information. A parallel structure is obtained for extracting projected pose vectors from all frames of an input sequence and using Long short-term memory (LSTM) layers to integrate the pose vectors across the input frames. They introduced a dataset SAR4 that includes 1292 video sequences for goalkeeper diving, player shooting, receiving a pass, and giving pass actions.

In [110], the authors used action recognition for summarizing long soccer videos. The recognition of five actions (centerline, corner-kick, free-kick, goal action, and throw-in), defined in their dataset Soccer5, was implemented by training an LSTM network on extracted soccer features and a ResNet based on 3D-CNN.

In [111], a deep learning, fine-grained action recognition method was proposed to analyze 132 soccer training videos to estimate whether a player has stopped a soccer ball successfully or not (2543 ball-stopping actions annotated). Because for these two actions, the motions and the scenes have no noticeable difference, it is important to consider the difference in human-object interaction motions. A cascaded scheme of deep networks based on the object-level trajectories is proposed, constructed by concatenating a YOLOv3 [112] network for detection with a classifier LSTM based network.

In [113], a framework for recognizing player actions in a live soccer game was proposed. The aim was to help text query-based video search, extract statistics during a soccer game, and generate textual commentary. They tested their framework on several action clips from the Soccer-8k dataset.

In [114], actions and group activities are considered in soccer videos. A proposed method has inferred eight individual actions and eleven group activities simultaneously from soccer videos. Player-centric snippets were used as inputs of the model. Player snippets are obtained using an Aggregated Channel Features person detector [115], and a virtual camera zooms in on each detected player, creating a standardized video frame cut-out. For feature extraction, I3D CNN based on RGB video and Optical Flow were used, and feature suppression and zero-padding in graph attention networks for the classification of actions.

In [116], the authors proposed different self-attention models that can consider relevant information from video data and trajectories of a group of soccer players to obtain activity recognition. They performed experiments on a large-scale dataset that includes three group activities: passing the ball, shooting the ball, and the reception of the ball. For the HAR task, the authors compared three different trajectory-based models, such as the Transformer, and four vision-based ones that used the I3D CNN as a backbone. In their work, the I3D-based models considering the entire frame outperformed other models, even the Transformer model. The authors concluded that there are some limitations, such as the fixed number of input frames, and therefore samples that necessitate a longer temporal context are not detected.

In [118], the author proposed a framework to classify soccer videos of approximately 5 s–9 s into six various actions: Goal, Head goal, Penalty save, Penalty goal, Red card, and Substitute. They preprocessed the data, extracted the HOG features, and classified each video with the multi-class SVM. The best results they obtained for the Red card action.

In [119], the authors dealt with temporally sparse actions within a complete soccer game, such as goals, player substitutions, and card scenes. They used a Transformer model, which allowed capturing important features before and after the action scenes and analyzed instances the model focuses on when predicting an action by observing the internal weights of the transformer. They tested their method on the SoccerNet dataset and improved significantly over existing methods. Furthermore, they analyzed the attention weights concluding that the model focuses on different temporal neighborhoods for different actions.

In [120], the authors proposed a lightweight and modular RMS-Net for soccer action spotting for the same problem, integrated with any existing backbone. The network can simultaneously predict the event label and its temporal offset using the same underlying features. The authors used two training strategies to balance the data and uniform the samples, mask ambiguous frames, and keep the most discriminative visual cues. They evaluated their method with ResNet as the backbone of the SoccerNet dataset and outperformed the existing method itself.

Also, for the same problem, authors in [121] proposed the dilated recurrent neural network (DilatedRNN) with LSTM units, grounded on Two-stream CNN features to model long-range and mid-range dependencies. The Two-stream CNN extracts local Spatio-temporal features, and the DilatedRNN makes the information obtained from distant frames available for the classifier and spotting algorithms. They also evaluated their work on SoccerNet. Similarly, authors from [122] presented an algorithm for automatically detecting events in soccer videos using 3D CNN. Their results show that the method can recognize events with high recall, low latency, and accurate time estimation, but there is a slightly lower precision than the current state-of-the-art.

In [123], the authors fine-tuned multiple action recognition models on soccer data to extract high-level semantic features and design a transformer-based temporal detection module to locate the target events. Their idea is to detect which action is happening when. They evaluated their method on a new version of a dataset, SoccerNet-v2 [124], which contains broadcast videos of 550 soccer games and 77 different games in Spain Laliga 2019–2021 season. The authors believe that the presented method can be extended to detect and locate events in other sports domains.

5.2.2. Basketball

In [125] is proposed a trajectory-based approach for automatic recognition of complex multi-player activities in basketball. A probabilistic model based on trajectory information is used to segment the game into activities (offense, defense, time out). Key elements are detected (starting formation, screen, and move), and their temporal orders are used to produce a semantic description of the observed activity.

In [126], a feature-representation method for recognizing actions in broadcast basketball videos focuses on the relationship between human actions and camera motions. Key-point trajectories are extracted as motion features in Spatio-temporal sub-regions called Spatio-temporal multiscale bags (STMBs). Global representations and local representations from one sub-region in the STMBs are combined to create a global pairwise representation (GPR). The GPR considers the co-occurrence of camera motions and human actions. The classification of actions is performed with two-stage SVM classifiers trained with STMB-based GPRs.

In [127], the authors researched action recognition in basketball to detect events and key actors in multi-person videos. The proposed model learns to detect eleven action/event classes on their custom dataset while automatically giving "attention" to the people responsible for the event. They tracked people in videos, used a recurrent neural network (RNN) to represent the track features, learned time-varying attention weights to combine these features at each time-instant, and finally used another RNN, Inception-v7 [79] network, for action detection and classification. They used a subset of the NCAA games available on YouTube as a dataset.

In [128], the authors focused on real-time detection and tracking of basketball players using deep neural networks and action recognition. For that purpose, they used a subset of broadcast basketball videos from the NCAA Basketball Dataset [127]. They used YOLOv2 [129] and SORT [130] for detection and track and LSTM for action classification.

In [51], is released a dataset with fine-grained actions in basketball game videos. They propose an approach by integrating the NTS-Net [131] into a two-stream network to locate the most informative regions and extract more discriminative features for fine-grained action recognition.

In [132], the focus is on recognizing group activities and the outcome (score or not score) in basketball. It proposed a scheme for global and local motion patterns and key visual information for recognition in basketball videos. A two-stream 3D CNN framework is utilized for group activity recognition over the separated global and local motion patterns.

In [133], the author used a combination of the Alpha pose [134] framework for pose estimation and the real-time multiple object tracker SORT [130] to obtain key point vectors as input to the Bidirectional Long Short-Term Memory (BiLSTM) [135] used for basketball pose-based action recognition for multiple players. He created a custom dataset with 37 annotated basketball videos for six different actions to be recognized: *run, walk, dribble, throw, receive_ball,* and *no_action.* He compared the result of BiLSTM classification with LSTM and MLP architecture and obtained higher accuracy. However, the *walk* and *no_action* classes were often misclassified, as well as the classes *walk* and *run*, because they look similar.

In [136], the authors proposed a recognition method for four basketball shooting categories: Shoot, Pass, Catch and Dribble. They used the Gaussian latent variable method of stochastic gradient descent combined with random forest, SVM, SOM neural network, and Bayesian network. They achieved the best results with the Bayesian network.

In [137], the authors used motion features, posture information, and skeleton sequence combined with KNN to create a basketball sports recognition model by leveraging motion block estimation. First, they collected a dataset that contains massive-scale basketball sports video clips from YouTube and manually labeled it.

In [138], the authors propose a human action recognition system using triple Kinect sensors for virtual reality applications in basketball. They designed a mark detection method to determine the front of the user and fusion skeleton data in real-time. They extract features like joint velocities, angles, and angular velocities from sequences, and they used the part-aware LSTM network for classification.

In [139], the authors created a new feature-enhanced skeleton-based method called LSTM-DGCN for basketball player action recognition based on LSTM and the deep graph convolutional network (DGCN) methods. They extracted the spatial features of the distances and angles between the joint points of basketball players, and built a large-scale dataset of 12 complex actions (32 kinds of atomic actions) for basketball players with RGB image data and Depth data captured at Northwestern Polytechnical University, named NPU RGB + D. The dataset consists of 2169 videos, i.e., 75 thousand frames, including RGB frame sequences, depth maps, and skeleton coordinates. The experimental results show that their method outperforms the state-of-the-art action recognition methods and that their dataset is very competitive.

In [140], the authors proposed a lightweight fine-grained action recognition model for basketball foul detection. Their network consists of multiple streams to extract temporal and spatial features with 3DCNN blocks. Their dataset consists of 182 video clips from NBA games (95 everyday actions, including players being touched but not fouled by the referee and 87 fouls). They point out that foul action recognition is challenging because fouls in basketball games are instantaneous and very similar to normal actions but obtain acceptable results for this problem.

5.2.3. Volleyball

Research in [141] focused on ball detection and trajectory extraction in volleyball videos. This paper presents a physics-based scheme that utilizes motion characteristics to extract ball trajectories from many moving objects. The ball trajectory can be exploited to recognize set types for tactics inference and detect basic actions in the volleyball game for close-up presentation based on game-specific properties.

In [142], the focus is on group activity recognition for volleyball. An LSTM model is designed to represent individual people's action dynamics in a sequence, and another LSTM model is designed to aggregate person-level information for complete activity understanding. They learned a temporal representation of person-level actions through a two-stage process and combined individual people's representations to recognize the group activity. The authors evaluated their work on two datasets: The Collective Activity Dataset and a custom volleyball dataset based on publicly available YouTube volleyball videos. Based on their work in [143], the authors proposed a Spatio-temporal graph representation and explored a generic feature representation based on Bag of Visual Words. Additionally, they applied random forest trees to the temporal features.

In [144], the authors focus on activity recognition in beach volleyball to prevent injuries through a monitoring system. They presented an obtrusive automatic monitoring system for beach volleyball based on wearable sensors applying Deep Learning CNN.

In [145] is presented a unified framework for understanding human social behaviors in raw image sequences. The architecture does not rely on external detection algorithms but is trained end-to-end to generate dense proposal maps refined with the inference scheme. The temporal consistency is handled with a person-level matching RNN (as a baseline, they used Inception-v3 [145], HDTM [142], and other models). The framework is evaluated on the dataset from [142].

In [146], the focus is on Decomposition and recognition of playing volleyball action based on the SVM algorithm. Firstly, the principal component analysis (PCA) is used for dimension reduction. The SVM classifier is trained for the sample dana obtaining the optimum parameters of the reduced dimension data by the grid search method. Lastly, the SVM classifier is reset to obtain the optimum SVM classifier parameters of the original sample data and realize the decomposition and recognition of playing volleyball action.

In [147], a method for volleyball action recognition is proposed by combining multi-view local motion features together with 3D global trajectories. The recognition is performed on game videos from a custom dataset of men's volleyball competitions.

In [148], a recognition framework based on 3D global and multi-view local features is proposed for action recognition in volleyball games. It combines global team formation features extracted from the 3D trajectories of all team members, ball motion features extracted from the 3D ball trajectory, and pose features that consist of hit frame and pose variation. These two features distinguish each action by focusing on the motion standard and stability between various quality actions. The recognition is performed on the same data [147].

In [149], the authors evaluated balanced and imbalanced learning methods with their proposed "super-bagging" method for volleyball action modeling. All methods are evaluated using six classifiers and four sensors (i.e., accelerometer, magnetometer, gyroscope, and barometer). They demonstrate that imbalanced learning provides better results for the non-dominant hand using a naive Bayes classifier than balanced learning. In comparison, balanced learning provides better results for the dominant hand using a tree bagger classifier than imbalanced learning.

In [150], the focus is on group activity recognition. The authors presented an attention semantic RNN, called stagNet, for understanding group activities and individual actions in videos, by combining the Spatio-temporal attention mechanism and semantic graph modeling. stagNet can extract discriminative and informative Spatio-temporal representations and capture inter-person relationships. Furthermore, they adopted a Spatio-temporal attention model to focus on key persons/frames for improved recognition performance. They evaluated the performance of stagNet on a video volleyball dataset.

In [151], the authors proposed the Actor Spatio-temporal Relation Networks (ASRN) to model the Spatio-temporal relation between participants in videos (of each individual and the relational features between multiple individuals). Firstly, they proposed a Spatio-temporal Relation Module (SRM) to calculate the relational information between each feature node. Then, they designed a Personal Spatio-temporal Feature Module (PSFM) and a Multi-actors Relation Module (MRM) to extract actor-level Spatio-temporal semantic information and the relation features between actors. They obtained good results on the Volleyball dataset [142], containing actions: right set, right spike, right pass, right win point, left set, left spike, left pass, and left win a point.

In [152], the authors introduce a novel deep learning-based group activity recognition approach called the Pose Only Group Activity Recognition System (POGARS), designed to use only tracked poses of players to predict the performed group activity. They used a 1D CNN to learn individuals' Spatio-temporal dynamics using their pose keypoints estimations and position tracklets, together with a spatial and temporal attention mechanism to infer person-wise importance and multi-task learning for simultaneously performing group and individual action classification. POGARS is also tested on the Volleyball dataset and achieves competitive results compared to other state-of-the-art methods. In their case, pose as only input achieves better results than methods that use RGB as input.

In [58], the authors strive to recognize individual actions and group activities from volleyball videos, proposing an Actor-transformer model able to learn and selectively extract information relevant for group activity recognition. They feed the transformer with rich actor-specific static and dynamic representations of features from a 2D pose network and 3D CNN. As a result, actor-transformers achieve state-of-the-art results on the Volleyball dataset [142].

Similarly, in [153], the authors proposed a novel group activity recognition network termed GroupFormer, which captures spatial-temporal contextual information jointly to augment the individual and group representations effectively with a clustered spatial-temporal transformer. The network models the spatial and temporal dependencies integrally and utilizes decoders to build the bridge between this information and models the clustered attention mechanism utilized to dynamically divide individuals into multiple clusters for better learning activity-aware semantic representations. GroupFormer outperformed the state-of-the-art methods on the Volleyball dataset.

Later, in [154], the authors also analyze group activity recognition in a short video clip, modeling the video as a series of tokens representing the video's multi-scale semantic concepts. The authors proposed a Multiscale Transformer-based architecture termed COMPOSER that performs attention-based reasoning over tokens at each scale and learns group activity compositionally, using only the keypoint modality, which reduces scene biases and improves the model's generalization ability. COMPOSER achieves a new state-of-the-art 94.5% accuracy on the Volleyball dataset.

5.2.4. Hockey

In [155], the authors focused on hockey action in a medium field, where a typical figure has a resolution of dozens of pixels in each dimension. A self-initializing tracker tracks the figures of hockey players. Then, a new stabilization algorithm uses a mixture of templates to estimate a figure's position and scale during the stabilization process. For action classification, motion and pose features are used. Image gradients are decomposed into four non-negative components used to characterize poses. Better results are obtained with pose features in comparison with motion ones.

In [156], a template-based algorithm is presented to track and recognize hockey players' actions in an integrated system using only visual information. This algorithm couples tracking and action recognition into a single framework, where tracking and action recognition assist one another. For the hockey sequences, images of players performing six actions are collected and transformed to the PCA-HOG descriptor, computed by first transforming the athletes to the HOG descriptor's grids and then projecting it to a linear subspace Principal Component Analysis (PCA).

In [157], the authors created a dataset of hockey videos to find violence in sports videos using the bag-of-words approach. Their dataset consists of 1000 clips of 50 frames of the *fight* and *non-fight* actions that were manually labeled. They used STIP [158] and MoSIFT [159], two Spatio-temporal descriptors to study violence videos, along with Optical Flow and HOG. They found that MoSIFT has higher performance than STIP. They also concluded that detecting violence in hockey is easier than detecting fights in movies.

In [160], the authors designed a CNN called Action recognition Hourglass Network (ARHN) to interpret player actions in ice hockey videos. Pose features from hockey images and videos are extracted and added to this network to produce action recognition. The first component of the network is the latent pose estimator. The second latent feature is transformed into a frame of reference, and in the third, action recognition is performed. A dataset of annotated hockey images is generated because no benchmark dataset for pose estimation or action recognition is available for hockey players.

In [161], the authors proposed a deep architecture to classify puck possession events in ice hockey. The model has three distinct phases: feature extraction, feature aggregation, learning, and inference. CNN is used for feature extraction and aggregation, followed by a late fusion model to extract and aggregate various features, including handcrafted homography features, to encode the camera information. Next, RNN is used for temporal extension and classification of the events to which CNN's output is passed. The team pooling and pre-trained model incorporate individual attributes of the players and their interaction. Only the player positions on the image and the homography matrix are needed for the model, simplifying the system's input. The model is evaluated on a new dataset called Ice Hockey Dataset and a volleyball dataset.

In [162], the focus is on finding fight scenes in hockey sports videos. Fast Fourier and Radon Transform (computes projections of an image matrix along the specified direction. It is used to reconstruct the frequency data from 2D Fast Fourier Transform into a two-dimensional form on which further calculations are being done) are applied to the local motion after being extracted in the video frames using blur information. The authors used transfer learning with the pre-trained deep learning

model VGGNet to identify fight scenes in video frames and feed-forward neural networks to compare the methodology. Finally, the authors used the National Hockey League dataset videos to present the model's outcome.

In [163], the authors presented a deep learning-based solution for hockey game action recognition in multi-label learning settings having a class imbalance problem. 3D CNN-based multilabel deep HAR system was implemented for multi-label class-imbalanced action recognition. The system was tested for two scenarios: an ensemble of k binary networks vs. a single k-output network on a publicly available hockey videos dataset.

In [164], the author designed and implemented a CNN automated method to determine the pose of a hockey player with and without a hockey stick from broadcast game video and perform action recognition via pose. Deep learning computer vision architecture HyperStackNet has been designed and implemented for a joint player and stick pose estimation. The action recognition hourglass network, or ARHN, interprets player actions in ice hockey videos using estimated poses. The first component of ARHN is the latent pose estimator, the second transforms latent features into a frame of reference, and the third performs action recognition. The authors built a custom annotated dataset for this purpose.

In [50], a two-stream architecture is proposed for action recognition in hockey. The pose is estimated via the Part Affinity Fields model to extract meaningful cues from the player. Temporal features are extracted using Optical Flow. These are then fused and passed to fully connected layers to estimate the hockey players' actions. A publicly available dataset, HARPET (Hockey Action Recognition Pose Estimation, Temporal), was created, composed of sequences of annotated actions and poses of hockey players, including their hockey sticks as an extension of the human body pose.

In [165], the authors introduced a two-stream network utilizing player pose sequences and Optical Flow features for recognizing hockey actions. Players' pose sequences are compact representations of frame-by-frame human and stick joint locations and angles between joints. Two-layered LSTM network output is fused with Optical Flow features processed by a CNN. The authors demonstrate the efficacy of the method on the HARPET dataset.

In [166], the pre-trained VGGNet-16, a deep learning-based transfer learning model, has been proposed for activity recognition in hockey. Authors constructed their hockey dataset based on video samples collected from the International Hockey Federation and YouTube to recognize four main activities: free hit, goal, long corner, and penalty corner.

In [167], the authors introduced a network for localizing the hockey puck on the ice rink and recognizing actions in hockey. Their method estimates the puck location from video using the temporal context and leverages player location information with heatmaps using an attention mechanism. To create a dataset, they utilize 8,987 broadcast National Hockey League annotated videos by professionals of 2-s duration with a resolution of 1280 $\times$ 720 pixels and a framerate of 30 fps. In the videos, the pack location is annotated together with event labels: Faceoff, Advance (dump in/out), Play (player moving the puck with an intended recipient, e.g., pass, stickhandle), or Shot. Authors demonstrated that multitask learning with puck location improves event recognition accuracy.

In [168], the authors introduced a transformer network for recognizing players through their jersey numbers in broadcast National Hockey League videos that can be further analyzed and adapted for action recognition.

5.2.5. Handball

In [169], a handball action recognition method using Mask R–CNN [170] and space-time interest points (STIPS) [158] is proposed. The method combines the object detector's location information with a player activity measure based on Spatio-temporal interest points to track players performing relevant actions. In [171], the authors proposed a Mask R–CNN and Optical Flow-based method for detecting and marking handball actions called the MOF method. The method was evaluated on a dataset of handball practice videos recorded in the wild. In [172], the authors propose a leading player detection method MR-CNN + STIPs, that combines the Mask R–CNN object detector and Spatio-temporal interest points to recognize actions such as passing, shooting, jump shot, and dribbling.

In [173], the focus is on recognizing the throwing action in handball based on RGB-D data. The authors introduced an RGB-D dataset that compares and evaluates handball players' performance during throws. They examined the central angles responsible for throwing performance to analyze handball players' skills and adopted the dynamic time warping technique to compare the two athletes' throwing motions.

In [174], a method for temporal segmentation and recognition of team activities in sports is presented based on a new activity feature extraction. The focus is on the position of team players from a planned view of the playground. They constructed a position distribution along with each frame of the sequence. These methods extract activity features using the explicitly defined trajectories, where the players have specific positions. They classified six various team activities in European handball with SVM.

In [26], the authors differentiate between the performances of the CNN, MLP, and LSTM-based models while considering temporal information to recognize different actions in handball scenes on a custom dataset. The dataset consists of approximately 3000 annotated videos, considering classes such as Dribbling, Passing, Shot, Throw, etc. Given that the duration of different actions differs greatly, the authors examined the results considering combinations of the number of input frames with different strategies such as frame decimation. They obtained the highest accuracy with the MLP-based model and pointed out that in this case, more frames can positively influence the action recognition results.

5.2.6. Baseball

In [175], a dataset named MLB-YouTube is designed for fine-grained action recognition in baseball videos. It is used for segmented video classification and activity detection in continuous videos. The segmented video dataset consists of 4,290 video clips. Each clip is annotated with the various baseball activities, such as swing, hit, ball, strike, foul, etc. The video clips can contain multiple activities simultaneously. They compared different recognition approaches with temporal feature pooling for segmented and continuous videos, that is, I3D CNN or Inception-v3 combined with mean, max, pyramid pooling, LSTM, temporal convolution filters, and different learning of sub-events.

In [176], the authors introduced a large-scale dataset called the BBDB containing 400k samples of visually similar actions in baseball. It contains about 30 action classes such as *Two-base Hit, Infield Hit, Bunt Hit, Fly Out, Touch Out, Strike, Strike Out,* etc., which were semi-automatically annotated from 4200 h of broadcast videos. Apart from action recognition, the dataset can be used for other video understanding tasks such as text-video alignment and video highlight generation. The authors tested their dataset for action recognition using different algorithms such as the combination of HOG, OF, and Motion Boundary Histogram (MBH) along with SVM, then a VGG model, OF with a ResNet-50 model, then a Two-Stream combining the mentioned VGG with OF and ResNet-50. Furthermore, they combined CNN layers with RNN (CNN + GRU) and tested a 3DCNN called C3D and the I3D network. C3D shows slightly lower performance compared to the Two-stream network, but I3D outperforms Two-stream. CNN + GRU shows a large performance improvement compared to other methods.

In [177], the authors developed a high-quality custom dataset containing six pitch types classes. They focused on baseball pitch type recognition based on broadcast videos using a two-stream inflated 3D convolutional neural network (I3D).

In [178], the authors created a dataset with multimodal Kinect sensors and cameras to later analyze the signals and recognize baseball

player behaviors using the LSTM model with multimodal features. Ten different behaviors are proposed for the players' performance on the field using baseball pitch and baseball bat, and to understand the players' condition during warm-up or training, left and right stretch, left and right lunge, and deep squat are used. Preliminary results of the proposed model have shown that a potential baseball player's on-field and off-field behavior can be analyzed using multimodal sensor data to be later evaluated by a baseball coach.

5.2.7. Cricket

In [36], the authors tried to recognize Cricket strokes from Cricket telecast videos of match highlights. The predominant direction of motion is found by summing up the histograms of Optical Flow directions, taken for significant pixels, over the complete Cricket stroke clip. They used unsupervised K-Means clustering of the extracted clip feature vectors. They evaluated the results for 3-cluster K-Means by manually annotating the clusters as left, right, and Ambiguous strokes for 562 stroke instances.

In two consecutive works [179, 180] authors focused on the recognition of four different strokes in cricket (glance, block, drive, and cut) from images. In [179], the HOG features were combined with SVM, KNN, and AlexNet architecture. In [180], they used the OpenPose skeleton key points [181] as a set of defining features that are fed into an LSTM network. Later in [182], they proposed a scene recognition model to recognize two distinct classes: gameplay and stroke. They evaluated two different models, the combination of HOG and SVM and the AlexNet.

5.2.8. Rugby

In [183], the authors investigated action recognition to develop a tactics analysis system that could be implemented for various purposes, although they focused on rugby. The dataset used is a set of fixed images obtained by merging two fixed cameras into a left-right image to have a bird's-eye view. It includes seven actions (scrum, lineout, kick-off, kick counter, turnover, penalty, other play). They proposed a method that adds spatial information to time-series information as a new feature. Using the coordinates obtained by projectivity transforming the match video onto the bird's-eye view image, play classification was performed using the player, ball, and dense area positions as feature amounts.

In [184], the authors proposed a method that optimized the back-propagation neural network on a genetic algorithm and built a neural network classifier through extracting 10-dimension gray characteristics of samples and 12-dimension geometric characteristics of the body to recognize and count the tackle action in rugby. Their idea is to help trainers count the number of tackles that happened and analyze tactics to improve players' performance. Authors extracted and analyzed ten matched and extracted 6000 images containing different types of rugby tackle action (Side, Front, Back, Multiplayer). Experiment results show that their method compared with general methods, effectively avoids mistakes made by humans and has a higher rate of convergence for classification through the analysis of tackle action gray image.

5.2.9. American football

In [185] is presented a system for recognizing activities called plays (e.g., offense, defense, kickoff, punt, etc.) in amateur videos of American football games. Given a sequence of videos, where each shows a particular play in a football game, they run noisy play-level detectors on every video and then process the Hidden Markov Model's outputs, which encodes knowledge about the temporal structure.

In [186], a method for recognizing American football plays in video games is proposed. They used to shape and motion-based Spatio-temporal features and implemented Multiple Kernel Learning based on SVM to combine various features effectively. They evaluated their method on a dataset consisting of 78 video play instances collected from NCAA football games, including left-run, middle-run, right-run, short-pass, option-pass, rollout-pass, and deep-pass activities.

6. Discussion

Based on the reviewed papers that are found using Google Scholar, IEEE, Scopus, arXiv, and other sources, it can be seen that the most popular team sports around the world, soccer, volleyball, hockey, and basketball, are the most researched for HAR in sports, together with table tennis, tennis, and badminton where two players are involved. Out of 110 recently reviewed research on HAR in sports, there are at least ten different research for each sport mentioned above. The most common is research in soccer, which makes up to 15% of the selected papers, followed by table tennis and volleyball which are equally represented, and each makes up about 13%. Research related to other team sports such as handball, cricket, American football, baseball, and rugby is presented with a smaller number of papers in this review. Figure 21 shows the distribution (number of papers and percentage) of HAR investigation in different sports based on the reviewed papers.

There is at least one research on HAR in the early 21st century for most sports, followed by a gap without research until DL's development. As a result, many of the reviewed papers were written in the last couple of years because the HAR tasks were demanding, and before DL, no relevant results were achieved. In the last four years alone (2018–2021), there have been 82 out of 110 papers for HAR in sports. The cumulative sum showing the exponential increase in the number of research papers on HAR in sports in the last 21 years is shown in Figure 22.

As seen in Figure 23, which shows the number of implementations in the last four years, there is an increase in the interest among researchers in basketball, soccer, volleyball, and table tennis. Only for soccer and table tennis, there have been nine papers for each in the last two years. The number of recognized actions depends on the selected sport and the researchers' focus, whether individual actions, joint activities or team activities. Rarely focuses on all types of actions, but mostly on individual and joint or team activities. In sports with two players, the focus is on interactive actions (a player interacting with the racket and the ball), unlike team sports, where all actions can be recognized.

Some researchers are focused only on one action and whether a particular action has occurred or not. Others consider multiple actions and differentiate between them. The number of actions recognized in various sports is around 5–10 but can be much higher or lower.

It is also important to point out that most researchers have created a customized database to implement HAR methods. When they make the database publicly available, such as the THETIS dataset [56] and the Volleyball dataset [142], an increase in research for that domain can be noticed because it encourages others to compare their methods on the given dataset. These datasets on realistic conditions make HAR a demanding task due to simultaneous actions, cluttered background, various points of view for the same action, the complexity of actions, occlusion, scaling, illumination, camera motion, etc. However, by introducing multiple cameras and sensors, using additional data types such as depth or Optical Flow to existing RGB, 3D-based methods instead of 2D, more computational power, and more data, some of these challenges may be overcome, and better HAR results in sports can be expected.

Based on the collected data, i.e., the type, quality, and quantity of the data, researchers, over the years, have investigated various approaches for human action recognition in sports, trying to find the best combinations to achieve a high result HAR.

A distribution of largely used feature representations for HAR in sports is presented in Table 1., showing the percentage of the usage of a particular feature representation in relation to the total number of papers (e.g., the Optical Flow and other motion features are used in 28,18% of papers). Below, Table 2 shows the percentage of the usage of a particular ML and DL method in relation to the total number of papers (e.g., the SVM method is used in 15,45% of papers).

In the research, much attention has been paid to motion features to obtain the players' precise movements or the ball they interact with. The most commonly used features are Optical Flow (OF), used in 28,18% of

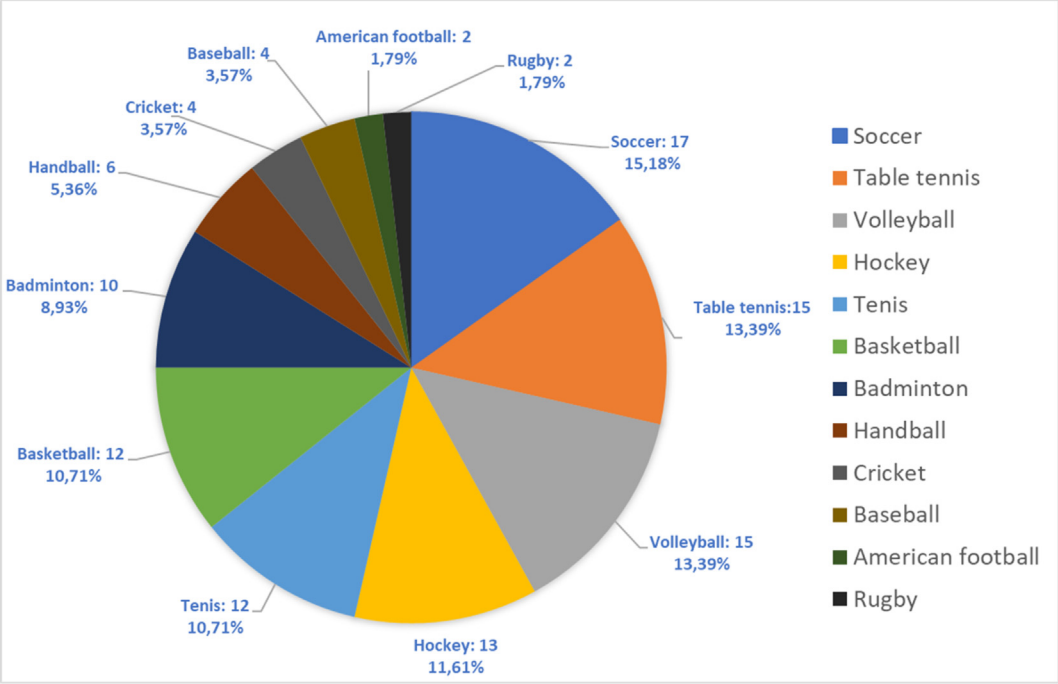

Figure 21. Distribution of HAR implementation in different sports (number of papers analyzed).

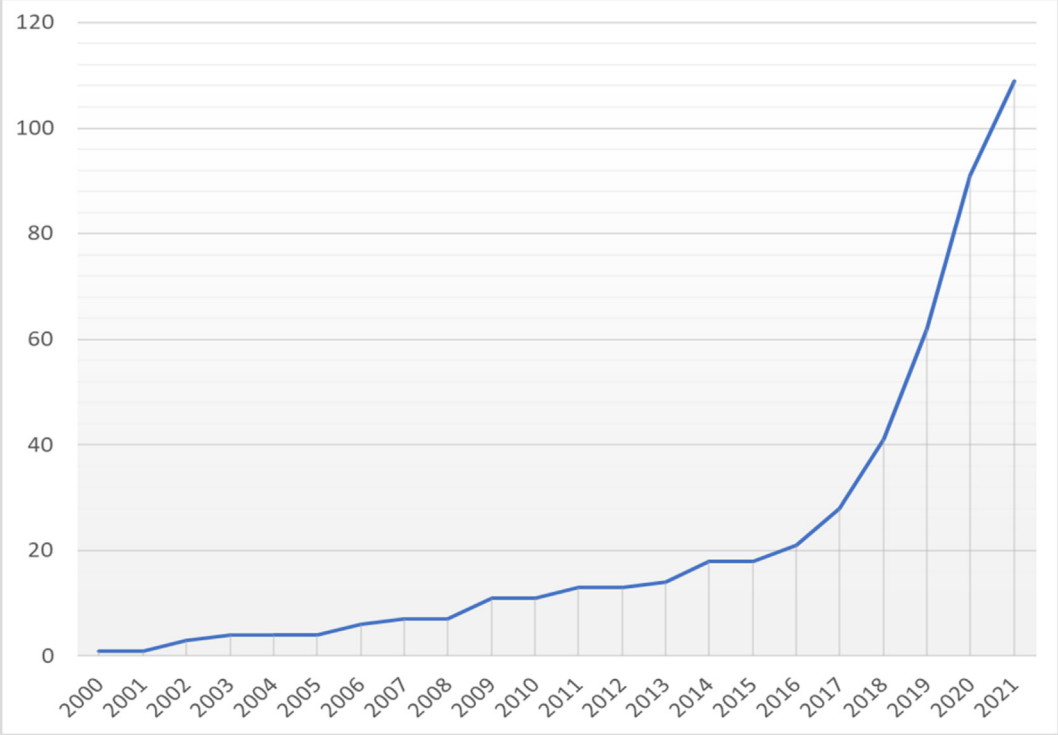

Figure 22. Number of research papers on HAR in sports over the years (2000–2021).

the total number of papers, and Histogram of oriented gradients (HOG), used in 6,37% of the total number of papers. However, some researchers combine DL networks to extract features automatically, and others use ML or DL methods to classify the actions and activities (e.g., [46, 63]).

The SVM classifier, the traditional ML method for classifying actions and activities, is often used (15,45% of the total number of papers) in various feature extraction combinations. Still, different DL methods have prevailed in the last couple of years thanks to higher performance results and automatically extracting features during the network training. The DL methods based on CNN (51,81% of the total number of papers) are implemented in most cases, but frame-by-frame processing when recognizing actions is their drawback, especially for actions spanning multiple frames. The LSTM-network, together with other RNNs, are used in 21,82% of the total number of papers because are better to implement

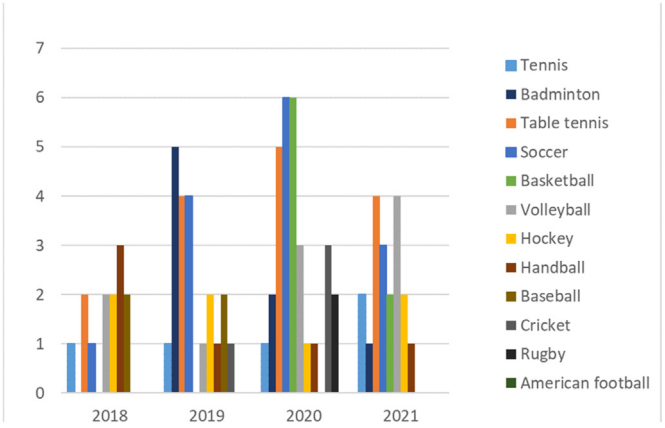

Figure 23. HAR implementation in sports in the last four years (2018–2021).

Table 1. The percentage of the usage of a particular feature representation in relation to the total number of selected representative papers.

OF and other motion features	28,18%
Multiple streams	12,72%
HOG	6,36%
Other features (BOW, Silhouette, SIFT, STIP,..)	16,36%

Table 2. The percentage of the usage of a particular ML and DL method in relation to the total number of selected representative papers.

SVM	17,27%
Other traditional methods (KNN, k-means, HMM, NB,..)	19,09%
CNN (AlexNet, GoogleNet, VGGNet, MobileNet,..)	51,81%
RNN/LSTM	21,82%
3D CNN (I3D, C3D,..)	20,00%
Attention/Transformer	9,09%
Other methods	16,36%

for video data since they can process sequences of frames and analyze all-action phases. Likewise, the 3D CNNs, which also consider the temporal dimension, are used in 20% of the total number of papers.

In addition, Table 3 shows the average value (percentage) of applying different feature representations by sports category (e.g., out of all features used in tennis, the HOG feature is used in 20% of cases).

Moreover, Table 4 shows the average value (percentage) of applying different ML and DL methods by sports category (e.g., out of all methods used in tennis, traditional methods, such as KNN, k-means, and others, are used in 21% of cases). According to the analyzed literature, it turned out that in badminton, many different CNNs, such as AlexNet and GoogleNet, are the most widely used (64%). At the same time, in table tennis, the interest in motion features and the use of Optical Flow (58%), along with the Spatio-temporal Convolutional Neural Networks, predominates. Considering all the articles reviewed, sensors are used in 10,91% of the articles. Still, the greatest interest in sensors is observed in these mentioned racket sports, where 21,62% of the total articles dealing with racket sports use sensors. In basketball, the interest in Recurrent Neural Networks, especially LSTM, and 3D CNN, is present, while in handball, the interest in Optical Flow and STIPs was noticed. On the other hand, two-stream networks, in some cases combining 3D CNNs, and, most recently, attention-based CNNs and transformers, have been proposed in Soccer, Hockey, and Volleyball to improve both the accuracy and efficiency of HAR from visual data. Researchers in other sports used different traditional ML and DL approaches in all combinations without noticeable consistency.

Table 3. The average value (percentage) of different feature representations used by sports category according to selected representative papers (only sports categories that used a feature representation are shown).

	OF and other motion features	HOG	Multiple streams	Other features (BOW, Silhouette, SIFT, STIP,..)
Tennis	40%	20%		40%
Badminton	75%	25%		
Table tennis	58%	5%	32%	5%
Soccer		20%	20%	60%
Basketball	43%		29%	29%
Volleyball	40%		10%	50%
Hockey	38%	13%	13%	38%
Handball	33%			67%
Baseball	50%	13%	38%	
Cricket		100%		

It can be concluded that each sport is different and contains different types of actions and activities with different complexity and duration, so running the same methods on various data may not give the same good results. It can be pointed out that in recent years, many researchers have focused on fine-grained action recognition, i.e., recognizing very similar actions, which is slowly becoming a new trend [187].

7. Conclusion

Human Action Recognition (HAR) can be applied in multiple areas such as education, content-based video summarization, video surveillance, human-computer interaction, entertainment, gaming, healthcare, and sports. For example, with the implementation of HAR in sports, researchers are trying to recognize individual actions of players, joint activities between multiple players, and teams' activities to provide statistical analysis of the game, follow players' behavior, and improve the performance of a player or the team. For this reason, a new systematization of human action recognition is proposed, taking into account the complexity of the actions, such as (1) individual actions, which are performed by one player and can be gestures, simple actions, interactions with objects, and complex actions, (2) joint activities, at least two players perform a combination of individual actions, and (3) team activities, where the majority of the team, strategically performs a combination of individual actions and joint activities.

Furthermore, the process of HAR implementation in sports is described, considering two different approaches, the traditional machine learning approach with hand-crafted feature extraction and the deep learning approach with automatic feature engineering.

The labeled data is the key for supervised learning so that some HAR datasets can be used for action recognition in the sport, such as UCF Sports Action Data Set, Sports-1M Dataset, and SVW. However, their focus is not on a particular sport but on actions from different ones. If a researcher wants to focus on a particular sport, the latest should create a dataset or search for existing ones for that domain. Significant efforts have been made to collect various types of action video data to advance research on action recognition in sports. However, due to the complexity of the actions, uneven distribution of the number of samples for a given action, multiple players in the scene moving quickly, occlusion, camera movement, etc., these datasets still pose a great challenge to the existing algorithms of ML and DL for the HAR task. Moreover, given the unpredictable number of video hours required to train a model successfully, the annotation process is extremely tedious. The introduction of semi-supervised and unsupervised learning algorithms for annotation in future research could at least slightly facilitate, speed up, and reduce the cost of the data preparation process.

An overview of implementations of HAR in sports is written for sports with two players (tennis, badminton, and table tennis) and for team sports (soccer, basketball, volleyball, hockey, handball, baseball, cricket,

Table 4. The average value (percentage) of using different ML and DL methods by sports category according to selected representative papers.

	SVM	Other traditional methods (KNN, k-means, HMM, NB,..)	CNN (AlexNet, GoogleNet, VGGNet,..)	3DCNN (I3D, C3D,..)	RNN (LSTM)	Attention/Transformer	Other methods
Tennis	7%	20%	27%		27%		20%
Badminton	18%	9%	64%		5%		5%
Table tennis	22%	39%	4%	26%		4%	4%
Soccer	5%	3%	35%	22%	11%	11%	14%
Basketball	13%	13%	25%	13%	31%		6%
Volleyball	6%	6%	33%	6%	17%	17%	17%
Hockey			43%	7%	14%	14%	21%
Handball	33%		33%		33%		
Baseball	8%		38%	31%	23%		
Cricket	17%	33%	33%		17%		
Rugby			50%				50%
Americanfootball	50%	50%					

rugby, and American football). In both cases, there are simple actions such as running, player interaction with a ball, or interaction with sports equipment, but in team sports, there are also interactions between players. In sports with two players, the players performing individual actions are opponents, while in team sports, multiple players collaborate in order to win. Accordingly, the number of actions that can be recognized in team sports is generally higher. Since an individual action in team sports can be recognized depending on the number of players involved, it can also be part of joint and team activities.

The most investigated sports are soccer, table tennis, volleyball, and hockey, where researchers mainly created their custom labeled datasets. A significant increase in papers has been visible in the last four years, although the increase can be seen even with the development of deep learning. Researchers mostly used DL methods based on CNN and LSTM, but there is also interest in SVM and motion features such as Optical Flow and HOG. Also, the authors have recently devoted themselves to both individual actions and group/team activities in sports, therefore have begun to implement CNNs with attention mechanisms and transformers, which is a new trend in HAR in general. It can be assumed that it will be applied more often in further works in the sports domain.

The future direction of the research is likely to include, to a more significant extent, team sports, given the achievements of DL methods and their accuracy in detecting persons and objects, improving the ability to track a larger number of objects. HAR research in sports is also influenced by appropriate databases prepared for machine learning models. It is expected that the number of image databases prepared for learning models for different types of actions and activities in specific sports will increase due to the increased interest in HAR with the development of data augmentation methods and the application of different transfer learning methods.

Although there have been visible improvements in HAR in sports, existing models are still not sufficient and stable to fully address the demanding challenges of this task. Therefore, much work will be needed to achieve competitive results and overcome current limitations. In improving the performance of recognizing an action to a level that is commercially interesting, all available information from the scene or video sequence can be helpful, and it is recommended to use them, such as the player's position on the field, the relationship between the players, the context of the game and rules and other similar data.

Furthermore, the performance can be improved by combining multiple architectures, using 3D-based methods instead of 2D, and improving existing algorithms with attention mechanisms. Also, better results can be obtained by multimodal fusion of RGB, skeleton, and depth data or by combining visual data with data from sensors, considering visual data for the same action from different angles, more computational power, etc. With the improvement of both accuracy and efficiency of HAR in sports, those algorithms could be integrated into applications that include more complex tasks that can help athletes, coaches, referees, commentators, and medical personnel. These applications with complex tasks could be, for example, a real-time fitness application that signals when an action is not performed correctly and makes suggestions to improve performance, a real-time application that replaces referees or helps them make easier decisions about events during a match (e.g., interruption, fouls, points scored, etc.), an application that recognizes interesting information from a game and summarizes it on an online website, a real-time video analysis application for a specific sport that helps coaches guide players to better execution of actions and activities, an application that detects unusual situations on the field and possible injuries of players and signals medical personnel to intervene, etc. Given the interest in HAR so far, shown in this review paper, it is expected that further research in the field will continue to rise exponentially.

Declarations

Author contribution statement

All authors listed have significantly contributed to the development and the writing of this article.

Funding statement

This work was supported by University of Rijeka (project number 18-222) and Hrvatska Zaklada za Znanost (IP-2016-2106-8345).

Data availability statement

Data included in article/supplementary material/referenced in article.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] M. Dhankar, N. Walia, An introduction to artificial intelligence, in: Emergin Trends in Big Data, IoT and Cyber Security, Maharaja Surajmal Institute, New Delhi, 2020, pp. 105–108.
[2] S.J. Prince, Computer Vision: Models, Learning, and Inference, Cambridge Univesity Press, 2012.

[3] IBM, What Is Computer Vision? IBM, 2021 (accessed March 11, 2021), https://www.ibm.com/topics/computer-vision.

[4] D.W. Stout, Social Media Statistics 2021: Top Networks by the Numbers • Dustin Stout, 2020 (accessed March 11, 2021), https://dustinstout.com/social-media-statistics/.

[5] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D.G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, TensorFlow: a system for large-scale machine learning, in: 12th USENIX Symposium on Operating Systems Design and Implementation, Savannah, 2016 (accessed March 10, 2021), https://research.google/pubs/pub45381/.

[6] Intel Corporation, OpenCV, 2021 (accessed March 11, 2021, https://opencv.org/.

[7] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, DeepFace: closing the gap to human-level performance in face verification, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2014, pp. 1701–1708.

[8] M.L. Microsoft, NET | Machine Learning Made for, NET, 2021 (accessed March 10, 2021), https://dotnet.microsoft.com/apps/machinelearning-ai/ml-dotnet.

[9] M.S. Nixon, A.S. Aguado, Introduction, in: Feature Extraction and Image Processing for Computer Vision, Elsevier, 2020, pp. 1–33.

[10] Y. Kong, Y. Fu, S. Member, Human Action Recognition and Prediction: A Survey, 2018 (accessed February 8, 2022), https://arxiv.org/abs/1806.11230v2.

[11] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, Machine recognition of human activities: a survey, IEEE Trans. Circ. Syst. Video Technol. 18 (2008) 1473–1488.

[12] Google, Google Trends, 2021 (accessed March 11, 2021), https://trends.google.com/trends/?geo=US.

[13] T.P. Moreira, D. Menotti, H. Pedrini, First-person action recognition through Visual Rhythm texture description, ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. (2017) 2627–2631.

[14] Digital Science & Research Solutions Inc, Dimensions, 2021 (accessed March 11, 2021), https://app.dimensions.ai/discover/publication.

[15] P. Pareek, A. Thakkar, A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications, Artif. Intell. Rev. (2020).

[16] M. Ivasic-Kos, K. Host, M. Pobar, Application of deep learning methods for detection and tracking of players, in: Artificial Neural Networks and Deep Learning - Applications and Perspective [Working Title], IntechOpen, 2021.

[17] Y. Wang, W. Fang, J. Ma, X. Li, A. Zhong, Automatic badminton action recognition using CNN with adaptive feature extraction on sensor data, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2019, pp. 131–143.

[18] A. Ahmadi, E. Mitchell, C. Richter, F. Destelle, M. Gowing, N.E. O'Connor, K. Moran, Toward automatic activity classification and movement assessment during a sports training session, IEEE Internet Things J. 2 (2015) 23–32.

[19] Y. Kong, Y. Fu, Human Action Recognition and Prediction: A Survey, 2018 (accessed February 25, 2022), https://arxiv.org/abs/1806.11230v3.

[20] M. Vrigkas, C. Nikou, I.A. Kakadiaris, A review of human activity recognition methods, Front. Robotics AI 2 (2015) 28.

[21] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, Machine recognition of human activities: a survey, IEEE Trans. Circ. Syst. Video Technol. 18 (2008) 1473–1488.

[22] I. Jegham, A. ben Khalifa, I. Alouani, M.A. Mahjoub, Vision-based human action recognition: an overview and real world challenges, Forensic Sci. Int.: Digit. Invest. 32 (2020), 200901.

[23] S. Herath, M. Harandi, F. Porikli, Going deeper into action recognition: a survey, Image Vision Comput. 60 (2016) 4–21.

[24] N.A. Rahmad, M.A. As'ari, N.F. Ghazali, N. Shahar, N.A.J. Sufri, A survey of video based action recognition in sports, Indones. J. Electr. Eng. Comput. Sci. 11 (2018) 987.

[25] K. Host, M. Ivašić-Kos, M. Pobar, Tracking handball players with the Deepsort algorithm, in: ICPRAM 2020 - Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods, SciTePress, 2020, pp. 593–599.

[26] K. Host, M. Ivasic-Kos, M. Pobar, Action recognition in handball scenes, Lecture Notes Netw. Syst. 283 (2022) 645–656.

[27] K. Soomro, A.R. Zamir, Action recognition in realistic sports videos, Adv. Compu. Vision Pattern Recognit. 71 (2014) 181–208.

[28] J. Calandre, R. Péteri, L. Mascarilla, Optical Flow Singularities for Sports Video Annotation: Detection of Strokes in Table Tennis, MediaEval, 2019.

[29] P.E. Martin, J. Benois-Pineau, R. Peteri, J. Morlier, Sport action recognition with siamese spatio-temporal CNNs: application to table tennis, in: Proceedings - International Workshop on Content-Based Multimedia Indexing, IEEE Computer Society, 2018.

[30] N. Farajidavar, T. de Campos, J. Kittler, F. Yan, Transductive transfer learning for action recognition in tennis games, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 1548–1553.

[31] A.A. Efros, A.C. Berg, G. Mori, J. Malik, Recognizing action at a distance, Proc. IEEE Int. Conf. Comput. Vision 2 (2003) 726–733.

[32] J.L. Barron, D.J. Fleet, S.S. Beauchemin, Performance of optical flow techniques, Int. J. Comput. Vis. 12 (1) (1994) 43–77.

[33] S. Ramasinghe, K.G.M. Chathuramali, R. Rodrigo, Recognition of badminton strokes using dense trajectories, in: 2014 7th International Conference on Information and Automation for Sustainability: "Sharpening the Future with Sustainable Technology", ICIAfS 2014, Institute of Electrical and Electronics Engineers Inc., 2014.

[34] N.A. Rahmad, M.A. As'ari, The new Convolutional Neural Network (CNN) local feature extractor for automated badminton action recognition on vision based data, J. Phys. Conf. 1529 (2020), 022021.

[35] C. Chen, Y. Shu, K.I. Shu, H. Zhang, WiTT:Modeling and the evaluation of table tennis actions based on WIFI signals, in: Proceedings - International Conference on Pattern Recognition, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 3100–3107.

[36] A. Gupta, A. Karel, M. Sakthi Balan, Discovering cricket stroke classes in trimmed telecast videos, in: Communications in Computer and Information Science, Springer, 2020, pp. 509–520.

[37] M. Al-Faris, J. Chiverton, D. Ndzi, A.I. Ahmed, A review on computer vision-based methods for human action recognition, J. Imaging 6 (2020) 46.

[38] P.H. Stakem, Graphics Processing Units, an Overview, Computer Architecture Series, 2017.

[39] Google Cloud, Cloud TPU Documentation | Google Cloud, 2021 (accessed March 11, 2021), https://cloud.google.com/tpu/docs.

[40] H. Sak, A. Senior, F. Beaufays, Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling, USA, 2014.

[41] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, International Conference on Learning Representations, ICLR, 2015 (accessed March 11, 2021), http://www.robots.ox.ac.uk.

[42] J. Carreira, A. Zisserman, Quo Vadis, action recognition? A new model and the kinetics dataset, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 4724–4733.

[43] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM 60 (2017) 84–90.

[44] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4489–4497.

[45] P.E. Martin, J. Benois-Pineau, R. Péteri, J. Morlier, Fine grained sport action recognition with Twin spatio-temporal convolutional neural networks: application to table tennis, Multimed. Tool. Appl. 79 (2020) 20429–20447.

[46] N. Crasto, P. Weinzaepfel, K. Alahari, C. Schmid, MARS: motion-augmented rgb stream for action recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2019-June, 2019, pp. 7874–7883.

[47] Y. Jiaxin, W. Fang, Y. Jieru, A review of action recognition based on Convolutional Neural Network, J. Phys. Conf. 1827 (2021), 012138.

[48] H. Fan, H.C. Ng, S. Liu, Z. Que, X. Niu, W. Luk, Reconfigurable acceleration of 3D-CNNs for human action recognition with block floating-point representation, in: Proceedings - 2018 International Conference on Field-Programmable Logic and Applications, FPL 2018, 2018, pp. 287–294.

[49] F. Malawski, Automatic Analysis of Techniques and Body Motion Patterns in Sport, AGH University of Science and Technology, 2019 (accessed February 24, 2022), https://www.researchgate.net/publication/332465274_Automatic_analysis_of_techniques_and_body_motion_patterns_in_sport.

[50] Z. Cai, H. Neher, K. Vats, D.A. Clausi, J. Zelek, Temporal hockey action recognition via pose and optical flows, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE Computer Society, 2019, pp. 2543–2552.

[51] X. Gu, X. Xue, F. Wang, Fine-grained action recognition on a novel basketball dataset, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2020, pp. 2563–2567.

[52] G. Varol, I. Laptev, C. Schmid, Long-term temporal convolutions for action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2018) 1510–1517.

[53] L.C. Jain, Recurrent Neural Networks : Design and Applications, 2000, p. 392.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. ukasz Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. v Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017, in: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[55] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2020 (accessed February 24, 2022), https://arxiv.org/abs/2010.11929v2.

[56] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, M. Chiaberge, Action Transformer: A Self-Attention Model for Short-Time Pose-Based Human Action Recognition, n.d. https://pic4ser.polito.it.

[57] R. Girdhar, J. Joao Carreira, C. Doersch, A. Zisserman, Video action transformer network, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2019-June, 2018, pp. 244–253.

[58] K. Gavrilyuk, R. Sanford, M. Javan, C.G.M. Snoek, Actor-transformers for group Activity recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020, pp. 836–845.

[59] J.K. Aggarwal, M.S. Ryoo, Human activity analysis: a review, ACM Comput. Surv. 43 (2011).

[60] C. Schüldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proceedings - International Conference on Pattern Recognition, 2004, pp. 32–36.

[61] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: Proceedings of the IEEE International Conference on Computer Vision, 2005, pp. 1395–1402.
[62] H. Zhao, A. Torralba, L. Torresani, Z. Yan, HACS: Human action clips and segments dataset for recognition and temporal localization, in: Proceedings of the IEEE International Conference on Computer Vision, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 8667–8677.
[63] L. Smaira, J. Carreira, E. Noland, E. Clancy, A. Wu, A. Zisserman, A short note on the kinetics-700-2020 human action dataset, ArXiv (2020) 10864 (accessed March 11, 2021), http://arxiv.org/abs/2010.
[64] J.C. Niebles, Olympic Sports Dataset, 2010 (accessed March 11, 2021), http://vision.stanford.edu/Datasets/OlympicSports/.
[65] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, F.F. Li, Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2014, pp. 1725–1732.
[66] D. Shao, Y. Zhao, B. Dai, D. Lin, FineGym: a hierarchical video dataset for fine-grained action understanding, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020, pp. 2613–2622.
[67] S.M. Safdarnejad, X. Liu, L. Udpa, B. Andrus, J. Wood, D. Craven, Sports Videos in the Wild (SVW): a video dataset for sports analysis, in: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015, 2015.
[68] Y. Li, L. Chen, R. He, Z. Wang, G. Wu, L. Wang, MultiSports: A Multi-Person Video Dataset of Spatio-Temporally Localized Sports Actions, 2021 (accessed February 25, 2022), https://arxiv.org/abs/2105.07404v2.
[69] M. Pobar, M. Ivasic-Kos, Active player detection in handball scenes based on activity measures, Sensors 20 (2020) 1475.
[70] M.D. Rodriguez, J. Ahmed, M. Shah, Action MACH: a spatio-temporal maximum average correlation height filter for action recognition, 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2008).
[71] J. Lee, H. Jung, TUHAD: taekwondo unit technique human action dataset with key frame-based CNN action recognition, Sensors 20 (2020) 4871.
[72] H. Miyamori, S.I. Iisaku, Video annotation for content-based retrieval using human behavior analysis and domain knowledge, in: Proceedings - 4th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2000, IEEE Computer Society, 2000, pp. 320–325.
[73] H. Miyamori, Improving accuracy in behaviour identification for content-based retrieval by using audio and video information, in: Proceedings - International Conference on Pattern Recognition, 2002, pp. 826–830.
[74] G. Zhu, C. Xu, Q. Huang, W. Gao, L. Xing, Player action recognition in broadcast tennis video with applications to semantic analysis of sports game, in: Proceedings of the 14th Annual ACM International Conference on Multimedia, MM 2006, 2006, pp. 431–440.
[75] S. Gourgari, G. Goudelis, K. Karpouzis, S. Kollias, THETIS: three dimensional tennis shots a human action dataset, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 676–681.
[76] J. Vainstein, J.F. Manera, P. Negri, C. Delrieux, A. Maguitman, Modeling video activity with dynamic phrases and its application to action recognition in tennis videos, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2014, pp. 909–916.
[77] S.V. Mora, W.J. Knottenbelt, Deep learning for domain-specific action recognition in tennis, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE Computer Society, 2017, pp. 170–178.
[78] M. Skublewska-Paszkowska, E. Lukasik, B. Szydlowski, J. Smolka, P. Powroznik, Recognition of tennis shots using convolutional neural networks based on three-dimensional data, in: Advances in Intelligent Systems and Computing, Springer, 2020, pp. 146–155.
[79] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: Proceedings of the 32nd International Conference on Machine Learning, PMLR, PMLR, 2015, pp. 448–456 (accessed March 11, 2021), http://proceedings.mlr.press/v37/ioffe15.html.
[80] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: efficient convolutional neural networks for mobile vision applications, ArXiv (2017) (accessed March 11, 2021), http://arxiv.org/abs/1704.04861.
[81] J. Cai, X. Tang, RGB Video Based Tennis Action Recognition Using a Deep Weighted Long Short-Term Memory, 2018.
[82] M. Skublewska-Paszkowska, P. Powroznik, E. Lukasik, Learning three dimensional tennis shots using graph convolutional networks, Sensors 20 (2020) 1–12.
[83] Vicon Motion Systems Ltd UK, Capture.U | Real Time Data with Blue Trident Sensors, Vicon, 2021 (accessed March 11, 2021), https://www.vicon.com/software/capture-u/.
[84] S. Kanimozhi, T. Mala, A. Kaviya, M. Pavithra, P. Vishali, Key object classification for action recognition in tennis using cognitive mask RCNN, Lecture Notes Netw. Syst. 287 (2022) 121–128.
[85] M. Ullah, M.M. Yamin, A. Mohammed, S.D. Khan, H. Ullah, F.A. Cheikh, Attention-based LSTM network for action recognition in sports, in: IS and T International Symposium on Electronic Imaging Science and Technology, 2021, p. 2021.
[86] H.Y. Ting, K.S. Sim, F.S. Abas, Automatic badminton action recognition using RGB-D sensor, in: Advanced Materials Research, Trans Tech Publications Ltd, 2014, pp. 89–93.
[87] N.A. Rahmad, M.A. As'ari, M.F. Ibrahim, N.A.J. Sufri, K. Rangasamy, Vision based automated badminton action recognition using the new local convolutional neural network extractor, in: Lecture Notes in Bioengineering, Springer, 2020, pp. 290–298.
[88] N.A. Rahmad, N.A.J. Sufri, M.A. As'Ari, A. Azaman, Recognition of badminton action using convolutional neural network, Indones. J. Electr. Eng. Informat. 7 (2019) 750–756.
[89] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2015, pp. 1–9.
[90] K. Xia, H. Wang, M. Xu, Z. Li, S. He, Y. Tang, Racquet sports recognition using a hybrid clustering model learned from integrated wearable sensor, Sensors 20 (2020) 1638.
[91] N.A. Rahmad, M.A. As'ari, K. Soeed, I. Zulkapri, Automated badminton smash recognition using convolutional neural network on the vision based data, IOP Conf. Ser. Mater. Sci. Eng. 884 (2020), 012009.
[92] Y. Wang, J. Ma, X. Li, A. Zhong, Hierarchical multi-classification for sensor-based badminton activity recognition, in: International Conference on Signal Processing Proceedings, ICSP, Institute of Electrical and Electronics Engineers Inc., 2020, pp. 371–375.
[93] C. Ma, D. Yu, H. Feng, Recognition of badminton shot action based on the improved hidden Markov model, J. Healthcare Eng. (2021) 2021.
[94] R. Liu, Z. Wang, X. Shi, H. Zhao, S. Qiu, J. Li, N. Yang, Table tennis stroke recognition based on body sensor network, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 11874 LNCS, 2019, pp. 1–10.
[95] X. Sha, G. Wei, X. Zhang, X. Ren, S. Wang, Z. He, Y. Zhao, Accurate recognition of player identity and stroke performance in table tennis using a smart wristband, IEEE Sensor. J. 21 (2021) 10923–10932.
[96] P.E. Martin, J. Benois-Pineau, R. Peteri, Fine-grained action detection and classification in table tennis with siamese spatio-temporal convolutional neural network, in: Proceedings - International Conference on Image Processing, ICIP, IEEE Computer Society, 2019, pp. 3027–3028.
[97] P.E. Martin, J. Benois-Pineau, B. Mansencal, R. Péteri, J. Morlier, Siamese Spatio-temporal convolutional neural network for stroke classification in Table Tennis games, in: MediaEval 2019 Workshop, 2019.
[98] P.E. Martin, J. Benois-Pineau, R. Péteri, J. Morlier, 3D attention mechanism for fine-grained classification of table tennis strokes using a Twin Spatio-Temporal Convolutional Neural Networks, in: Proceedings - International Conference on Pattern Recognition, 2020, pp. 6019–6026.
[99] P.-E. Martin, Fine-grained Action Detection and Classification from Videos with Spatio-Temporal Convolutional Neural Networks. Application to Table Tennis, Archive ouverte HAL, Université de Bordeaux, Université de la Rochelle, 2020 (accessed February 24, 2022), https://hal.archives-ouvertes.fr/tel-03099907.
[100] A. Zahra, P.-E. Martin, Two Stream Network for Stroke Detection in Table Tennis, 2021.
[101] K. Aktas, M. Demirel, M. Moor, J. Olesk, C. Ozcinar, G. Anbarjafari, Spatiotemporal based table tennis stroke-type assessment, Signal, Image Video Proc. 15 (2021) 1593–1600.
[102] J. Wang, K. Zhao, D. Deng, A. Cao, X. Xie, Z. Zhou, H. Zhang, Y. Wu, Tac-simur: tactic-based simulative visual analytics of table tennis, IEEE Trans. Visual. Comput. Graph. 26 (2020) 407–417.
[103] J. Wang, D. Deng, X. Xie, X. Shu, Y.-X. Huang, L.-W. Cai, H. Zhang, M.-L. Zhang, Z.-H. Zhou, Y. Wu, Tac-valuer: knowledge-based stroke evaluation in table tennis; tac-valuer: knowledge-based stroke evaluation in table tennis, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD `21), August 14â•fi18, 2021, Virtual Event, Singapore. 1, 2021.
[104] M. Leo, T. D'Orazio, P. Spagnolo, P.L. Mazzeo, A. Distante, Multi-view player action recognition in soccer games, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2009, pp. 46–57.
[105] Y. Kong, W. Hu, X. Zhang, H. Wang, Y. Jia, Learning group activity in soccer videos from local motion, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2010, pp. 103–112.
[106] S. Giancola, M. Amine, T. Dghaily, B. Ghanem, SoccerNet: a scalable dataset for action spotting in soccer videos, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE Computer Society, 2018, pp. 1792–1802.
[107] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2016, pp. 770–778.
[108] A. Miech, I. Laptev, J. Sivic, Learnable pooling with Context Gating for video classification, ArXiv (2017) (accessed March 11, 2021), http://arxiv.org/abs/1706.06905.
[109] M. Fani, K. Vats, C. Dulhanty, D.A. Clausi, J. Zelek, Pose-projected action recognition hourglass network (PARHN) in soccer, in: Proceedings - 2019 16th Conference on Computer and Robot Vision, CRV 2019, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 201–208.
[110] R. Agyeman, R. Muhammad, G.S. Choi, Soccer video summarization using deep learning, in: Proceedings - 2nd International Conference on Multimedia Information Processing and Retrieval, MIPR 2019, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 270–273.
[111] J. Xiong, L. Lu, H. Wang, J. Yang, G. Gui, Object-level trajectories based fine-grained action recognition in visual IoT applications, IEEE Access 7 (2019) 103629–103638.

[112] J. Redmon, A. Farhadi, YOLOv3: an incremental improvement, ArXiv (2018) (accessed March 11, 2021), http://arxiv.org/abs/1804.02767.
[113] Y. Ganesh, A. Sri Teja, S.K. Munnangi, G. Rama Murthy, A novel framework for fine grained action recognition in soccer, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2019, pp. 137–150.
[114] B.G.A. Gerats, Individual Action and Group Activity Recognition in Soccer Videos, University of Twente, 2020 (accessed March 11, 2021), http://essay.utwente.nl/84037/.
[115] P. Dollar, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2014) 1532–1545.
[116] R. Sanford, S. Gorji, L.G. Hafemann, B. Pourbabaee, M. Javan, Group Activity detection from trajectory and video data in soccer, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2020-June, 2020, pp. 3932–3940.
[118] H. Ullah, M. Sajjad, Mustaqeem, Salient Event Detection in Soccer Videos Using Histogram of Oriented Gradient, 2020.
[119] H. Minoura, T. Hirakawa, T. Yamashita, H. Fujiyoshi, M. Nakazawa, Y. Chae, B. Stenger, Action spotting and temporal attention analysis in soccer videos, in: Proceedings of MVA 2021 - 17th International Conference on Machine Vision Applications, 2021.
[120] M. Tomei, L. Baraldi, S. Calderara, S. Bronzin, R. Cucchiara, RMS-Net, Regression and masking for soccer event spotting, in: Proceedings - International Conference on Pattern Recognition, 2020, pp. 7699–7706.
[121] B. Mahaseni, E.R.M. Faizal, R.G. Raj, Spotting football events using two-stream convolutional neural network and dilated recurrent neural network, IEEE Access 9 (2021) 61929–61942.
[122] O.A. Nergård Rongved, S.A. Hicks, V. Thambawita, H.K. Stensland, E. Zouganeli, D. Johansen, M.A. Riegler, P. Halvorsen, Real-time detection of events in soccer videos using 3D convolutional neural networks, in: Proceedings - 2020 IEEE International Symposium on Multimedia, ISM 2020, 2020, pp. 135–144.
[123] X. Zhou, L. Kang, Z. Cheng, B. He, J. Xin, Feature Combination Meets Attention: Baidu Soccer Embeddings and Transformer Based Temporal Detection, 2021 (accessed February 27, 2022), https://arxiv.org/abs/2106.14447v1.
[124] A. Deliege, A. Cioppa, S. Giancola, M.J. Seikavandi, J.v. Dueholm, K. Nasrollahi, B. Ghanem, T.B. Moeslund, M. van Droogenbroeck, SoccerNet-v2: a dataset and benchmarks for holistic understanding of broadcast soccer videos, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 4503–4514.
[125] M. Perše, M. Kristan, S. Kovačič, G. Vučkovič, J. Perš, A trajectory-based analysis of coordinated team activity in a basketball game, Comput. Vis. Image Understand. 113 (2009) 612–621.
[126] M. Takahashi, M. Naemura, M. Fujii, J.J. Little, Recognition of action in broadcast basketball videos on the basis of global and local pairwise representation, in: Proceedings - 2013 IEEE International Symposium on Multimedia, ISM 2013, 2013, pp. 147–154.
[127] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, L. Fei-Fei, Detecting events and key actors in multi-person videos, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2016, pp. 3043–3053.
[128] D. Acuna, Towards real-time detection and tracking of basketball players using deep neural networks, in: 31st Conference on Neural Information Processing Systems, Long Beach, 2017 (accessed March 11, 2021), http://goo.gl/ZPWStU.
[129] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 6517–6525.
[130] A. Bewley, Z. Ge, L. Ott, F. Ramos, B. Upcroft, Simple online and realtime tracking, in: Proceedings - International Conference on Image Processing, ICIP, IEEE Computer Society, 2016, pp. 3464–3468.
[131] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, L. Wang, Learning to navigate for fine-grained classification, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2018, pp. 438–454.
[132] L. Wu, Z. Yang, Q. Wang, M. Jian, B. Zhao, J. Yan, C.W. Chen, Fusing motion patterns and key visual information for semantic event recognition in basketball videos, Neurocomputing 413 (2020) 217–229.
[133] I. Zakharchenko, Basketball Pose-Based Action Recognition, Ukrainian Catholic University, Faculty of Applied Sciences, L'viv, 2020.
[134] H.S. Fang, S. Xie, Y.W. Tai, C. Lu, RMPE: regional multi-person pose estimation, in: Proceedings of the IEEE International Conference on Computer Vision, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 2353–2362.
[135] A. Graves, S. Fernández, J. Schmidhuber, Bidirectional LSTM networks for improved phoneme classification and recognition, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2005, pp. 799–804.
[136] R. Ji, Research on Basketball Shooting Action Based on Image Feature Extraction and Machine Learning, IEEE Access 8 (2020) 138743–138751.
[137] Z. Pan, C. Li, Robust basketball sports recognition by leveraging motion block estimation, Signal Process. Image Commun. 83 (2020).
[138] B. Yao, H. Gao, X. Su, Human motion recognition by three-view kinect sensors in virtual basketball training, in: IEEE Region 10 Annual International Conference, Proceedings/TENCON, Institute of Electrical and Electronics Engineers Inc., 2020, pp. 1260–1265.
[139] C. Ma, J. Fan, J. Yao, T. Zhang, S. Nowaczyk, M.-R. Bouguelia, H. Fanaee, S. Yeom, NPU RGBD dataset and a feature-enhanced LSTM-DGCN method for action recognition of basketball Players+, Appl. Sci. 11 (2021) 4426.
[140] C.H. Lin, M.Y. Tsai, P.Y. Chou, A lightweight fine-grained action recognition network for basketball foul detection, in: 2021 IEEE International Conference on Consumer Electronics-Taiwan, ICCE-TW 2021, 2021.
[141] H.T. Chen, B.S. Chen, S.Y. Lee, Physics-based ball tracking in volleyball videos with its applications to set type recognition and action detection, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2007.
[142] M. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, G. Mori, A hierarchical deep temporal model for group Activity recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2016, pp. 1971–1980.
[143] M. Ibrahim, A.E. Abdelaal, M. Lu, H. Wu, Improved Hierarchical Deep Temporal Model for Group Activity Recognition, 2016.
[144] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, S. Savarese, Social scene understanding: end-to-end multi-person action localization and collective activity recognition, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 3425–3434.
[145] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2016, pp. 2818–2826.
[146] Y. Yang, Decomposition and recognition of playing volleyball action based on SVM algorithm, J. Interdiscipl. Math. 21 (2018) 1181–1186.
[147] Y. Liu, S. Huang, X. Cheng, T. Ikenaga, 3D global trajectory and multi-view local motion combined player action recognition in volleyball analysis, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2018, pp. 134–144.
[148] X. Cheng, Y. Liu, T. Ikenaga, 3D global and multi-view local features combination based qualitative action recognition for volleyball game analysis, in: IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences. E102A, 2019, pp. 1891–1899.
[149] F. Haider, F.A. Salim, D.B.W. Postma, R. van Delden, D. Reidsma, B.J. van Beijnum, S. Luz, A super-bagging method for volleyball action recognition using wearable sensors, Multimodal Technol. Interact. 4 (2020) 1–14.
[150] M. Qi, Y. Wang, J. Qin, A. Li, J. Luo, L. van Gool, StagNet: an attentive semantic RNN for group Activity and individual action recognition, IEEE Trans. Circ. Syst. Video Technol. 30 (2020) 549–565.
[151] Y. Zhou, S. Tan, D. Wang, J. Mu, Actor spatiotemporal relation networks for group Activity recognition, in: 2021 11th International Conference on Information Science and Technology, ICIST 2021, 2021, pp. 505–510.
[152] H. Thilakarathne, A. Nibali, Z. He, S. Morgan, Pose Is All You Need: the Pose Only Group Activity Recognition System (POGARS), 2021 (accessed February 25, 2022), https://arxiv.org/abs/2108.04186v1.
[153] S. Li, Q. Cao, L. Liu, K. Yang, S. Liu, J. Hou, S. Yi, GroupFormer: Group Activity Recognition with Clustered Spatial-Temporal Transformer, 2021 (accessed February 25, 2022), https://arxiv.org/abs/2108.12630v1.
[154] H. Zhou, A. Kadav, A. Shamsian, S. Geng, F. Lai, L. Zhao, T. Liu, M. Kapadia, H.P. Graf, COMPOSER: Compositional Learning of Group Activity in Videos, 2021 (accessed February 27, 2022), https://arxiv.org/abs/2112.05892v1.
[155] X. Wu, B. Sc, Template-based Action Recognition: Classifying Hockey Players' Movement, Master, The University of Calgary, 2002.
[156] W.L. Lu, J.J. Little, Simultaneous tracking and action recognition using the PCA-HOG descriptor, in: Third Canadian Conference on Computer and Robot Vision, CRV 2006, IEEE Computer Society, 2006, p. 6.
[157] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, R. Sukthankar, Violence detection in video using computer vision techniques, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 6855 LNCS, 2011, pp. 332–339.
[158] I. Laptev, On space-time interest points, Int. J. Comput. Vis. 64 (2005) 107–123.
[159] M. Chen, A.G. Hauptmann, MoSIFT: Recognizing Human Actions in Surveillance Videos, 2009.
[160] M. Fani, H. Neher, D.A. Clausi, A. Wong, J. Zelek, Hockey action recognition via integrated stacked hourglass network, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE Computer Society, 2017, pp. 85–93.
[161] M.R. Tora, J. Chen, J.J. Little, Classification of puck possession events in ice hockey, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE Computer Society, 2017, pp. 147–154.
[162] S. Mukherjee, R. Saini, P. Kumar, P.P. Roy, D.P. Dogra, B.-G. Kim, Fight detection in hockey videos using deep network, J. Mult. Informat. Syst. 4 (2017) 225–232.
[163] K. Sozykin, S. Protasov, A. Khan, R. Hussain, J. Lee, Multi-label class-imbalanced action recognition in hockey videos via 3D convolutional neural networks, in: Proceedings - 2018 IEEE/ACIS 19th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2018, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 146–151.
[164] H. Neher, Hockey Pose Estimation and Action Recognition Using Convolutional Neural Networks to Ice Hockey, Master, The University of Waterloo, 2018 (accessed March 11, 2021), https://uwspace.uwaterloo.ca/handle/10012/13835.
[165] K. Vats, H. Neher, D.A. Clausi, J. Zelek, Two-stream action recognition in ice hockey using player pose sequences and optical flows, in: Proceedings - 2019 16th Conference on Computer and Robot Vision, CRV 2019, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 181–188.

[166] K. Rangasamy, M.A. As'ari, N.A. Rahmad, N.F. Ghazali, Hockey activity recognition using pre-trained deep learning model, ICT Express 6 (2020) 170–174.
[167] K. Vats, M. Fani, D.A. Clausi, J. Zelek, Puck localization and multi-task event recognition in broadcast hockey videos, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2021, pp. 4562–4570.
[168] K. Vats, W. McNally, P. Walters, D.A. Clausi, J.S. Zelek, Ice Hockey Player Identification via Transformers, 2021 (accessed February 27, 2022), https://arxiv.org/abs/2111.11535v1.
[169] M. Ivasic-Kos, M. Pobar, Building a labeled dataset for recognition of handball actions using mask R-CNN and STIPS, in: Proceedings - European Workshop on Visual Information Processing, EUVIP, Institute of Electrical and Electronics Engineers Inc., 2019.
[170] K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 2980–2988.
[171] M. Pobar, M. Ivasic-Kos, Mask R-CNN and optical flow based method for detection and marking of handball actions, in: Proceedings - 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, CISP-BMEI 2018, Institute of Electrical and Electronics Engineers Inc., 2019.
[172] M. Pobar, M. Ivasic-Kos, Detection of the leading player in handball scenes using Mask R-CNN and STIPS, in: Eleventh International Conference on Machine Vision, SPIE-Intl Soc Optical Eng, 2019, p. 3.
[173] A. Elaoud, W. Barhoumi, E. Zagrouba, B. Agrebi, Skeleton-based comparison of throwing motion for handball players, J. Ambient Intell. Hum. Comput. 11 (2020) 419–431.
[174] C. Direkoğlu, N.E. O'Connor, Temporal segmentation and recognition of team activities in sports, Mach. Vis. Appl. 29 (2018) 891–913.
[175] A.J. Piergiovanni, M.S. Ryoo, Fine-grained activity recognition in baseball videos, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE Computer Society, 2018, pp. 1821–1830.
[176] M. Shim, Y.H. Kim, K. Kim, S.J. Kim, Teaching machines to understand baseball games: large-scale baseball video database for multiple video understanding tasks, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 11219 LNCS, 2018, pp. 420–437.
[177] R. Chen, D. Siegler, M. Fasko, S. Yang, X. Luo, W. Zhao, Baseball pitch type recognition based on broadcast videos, in: Communications in Computer and Information Science. 1138 CCIS, 2019, pp. 328–344.
[178] S.W. Sun, T.C. Mou, C.C. Fang, P.C. Chang, K.L. Hua, H.C. Shih, Baseball player behavior classification system using long short-term memory with multimodal features, Sensors 19 (2019) 1425.
[179] T. Moodley, D. van der Haar, Cricket stroke recognition using computer vision methods, in: Lecture Notes in Electrical Engineering, Springer, 2020, pp. 171–181.
[180] T. Moodley, D. van der Haar, CASRM: cricket automation and stroke recognition model using OpenPose, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer, 2020, pp. 67–78.
[181] Z. Cao, G. Hidalgo, T. Simon, S.E. Wei, Y. Sheikh, OpenPose: realtime multi-person 2D pose estimation using Part Affinity fields, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 172–186.
[182] T. Moodley, D. van der Haar, Scene recognition using AlexNet to recognize significant events within cricket game footage, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Science and Business Media Deutschland GmbH, 2020, pp. 98–109.
[183] R. Ichige, Y. Aoki, Action recognition in sports video considering location information, in: Communications in Computer and Information Science, Springer, 2020, pp. 150–164.
[184] Z. Zhang, L. Gao, Y. Xiang, Application of optimized BP neural network based on genetic algorithm in rugby tackle action recognition, in: Proceedings of 2020 IEEE International Conference on Artificial Intelligence and Computer Applications, ICAICA 2020, 2020, pp. 95–99.
[185] S. Chen, Z. Feng, Q. Lu, B. Mahasseni, T. Fiez, A. Fern, S. Todorovic, Play type recognition in real-world football video, in: 2014 IEEE Winter Conference on Applications of Computer Vision, WACV 2014, IEEE Computer Society, 2014, pp. 652–659.
[186] B. Siddiquie, Y. Yacoob, L.S. Davis, Recognizing Plays in American Football Video, 2009.
[187] D. Shao, Y. Zhao, B. Dai, D. Lin, FineGym: A Hierarchical Video Dataset for Fine-grained Action Understanding, 2020 (accessed February 22, 2022), https://sdolivia.github.io/FineGym/.