An Overview of Human Action and Activity Recognition in Sports

Abstract: Human action recognition (HAR) is a challenging task in Artificial Intelligence that can be used for recognizing players' actions and teams' activities in different sports such as volleyball, basketball, soccer, hockey, tennis, etc. Because there are no detailed overviews of HAR that include application in sports, such overview is presented in this paper as the main contribution. Furthermore, different methods for HAR implementation are mentioned, and popular public datasets for that purpose are presented. Additionally, an experiment on implementing HAR in handball is presented.

Key words: Computer Vision, Machine Learning, Deep Learning, Human Action Recognition, Sport, Action, Dataset, Sport Dataset, Human Action Recognition in Sport

Author: Kristina Host, Department of Informatics, University of Rijeka, Croatia

Mentor: Marina Ivašić-Kos, Department of Informatics, University of Rijeka, Croatia

Rijeka, November 2020

Contents

1. Introduction

Artificial intelligence (AI) is one of the biggest branches of computer science, whose main concept is to build intelligent systems that can perform tasks which require human intelligence, or in other words, to build machines which can behave like a human, think like a human, and be able to make decisions on their own [1]. One aspect of AI is to process visual data such as images and videos in the way humans do, to see an image or a video, extract useful information from it, and to understand its content [2]. This compelling field, called Computer vision (CV), is growing continuously due to big amounts of visual data uploaded online on a daily basis. For example, more than 3 billion images are uploaded every day on social networks like Instagram and Facebook, and hundreds of hours of videos are uploaded on YouTube, perhaps the largest search engine with videos. What also contributes to the exponential growth of CV are better hardware and easily accessible open source machine learning works of big companies doing AI research like Google [3], Facebook [4] , and Microsoft [5]. In other words, the growth of CV is based on greater computer power and no need to start everything from scratch, which in turn leads to less time spent on experiments (experiments that used to take weeks, now take only hours, or even minutes). Some of the main tasks of CV are to recognize whether there is an object or an action in the image or video (validation), which category it belongs to (classification), where it is situated (detection), and which pixels belong to the object (segmentation) [6]. The task of human action recognition (HAR), which is the focus of this paper, is to recognize which human actions are in a video, at what time they occur, and where they are located. In order to accomplish that, the field of CV closely interacts with the fields of Image processing and Machine Learning (ML). Image processing is used for normalizing photometric properties of visual data, removing digital noise, data augmentation, etc., and is not concerned with understanding the content of visual data. However, when it comes to understanding the content and automatization of CV tasks, the most important fields are ML and its subfield Deep Learning (DL) [6]. Before ML, computer vison tasks required a lot of manual coding and effort by developers (manually annotating a large amount of key points and measurements that define the unique characteristics of an object, and then comparing the results with little automation involved), but introducing the ML and later DL methods, that changed. Nowadays, due to innovations with neural networks that extract features automatically, the CV tasks are not only easier and faster to implement, but they also achieve better results (recent leap of the accuracy value, from 50% to nearly 99% [7]). Figure 1 shows the interaction of the mentioned fields in order to accomplish the HAR task.

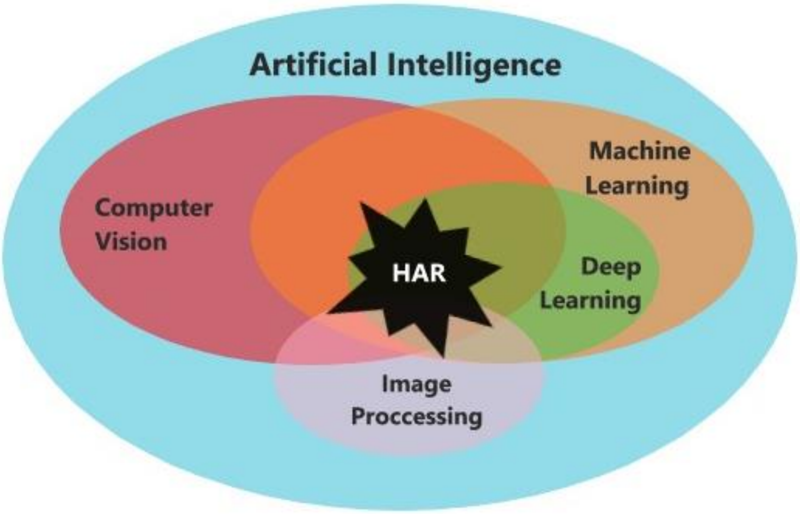

Figure 1.Human action recognition is part of different fields in Artificial Intelligence

In the next section, there is an introduction to human action recognition, what data can be used, which types of action exists, where HAR can be applied, what are the challenges, and how the process can be described. Then, different methods of HAR implementation are enumerated, and different public datasets available for the purpose of HAR are presented. The main contribution of this research is in section "HAR in sport domain", where is explained in which way HAR can be used in sport domain, and where an overview of HAR implementation in different sports with two players, such as tennis and badminton, or team sports, such as volleyball and soccer, is presented. In section "HAR experiment" is presented an HAR implementation on a handball dataset. In the last section, there is a conclusion and further work for HAR implementation in handball.

2. Human action recognition (HAR)

The goal of human action recognition is to understand, like a human, human actions taking place in a video. Some simple actions, like standing, could be recognized using just a single frame (image), but human actions are mostly much more complex and take place over a period of time; therefore, they need to be studied through consecutive frames (video). Due to successful results with image classification and object detection focusing on images, many researchers seek to expand the knowledge they gained on two-dimensional space to three-dimensional that include the temporal dimension.

According to the complexity of a human action, actions can be categorized as (1) primitive, actions that indicate basic movements like standing, lifting a hand, or any kind of gestures; (2) single person, more complex actions that consist of movements of the person, like walking, running, or jumping; (3) interaction, actions that involve objects, for example throwing a ball, dribbling, or picking a box; (4) group activity, actions that involve at least two persons performing a primitive action, single action, or interaction with objects, or a combination of them, for example, passing a ball or meeting.

The problems of action recognition can vary widely considering different applications and different data selection, so there is no single approach that suits all the challenges one may encounter. The studied applications cover many domains. For example, in the healthcare domain we can find applications in hospitals to detect heart attacks of patients, or in elderly houses to detect some abnormal activities or falling; in education, to detect presence of students; in video surveillance, to detect abnormal activity, in gaming and sports, to make players' behavior analysis, and in other domains such as content-based video summarization, entertainment, and human-computer interaction. Most of these applications are focused on RGB visual data, but there are also two other data types used for HAR, depth and skeleton data. HAR using RGB visual data is a demanding task due to limited understanding of biological vision, limitations with data, a lot of things happening simultaneously on the scene, cluttered background, different points of view for the same action, complexity of actions, occlusion, scaling, illumination, camera motion, etc., but by introducing other data like depth, some of these challenges may be overcome or at least seen from a different perspective.

2.1. HAR implementation

The process of HAR, presented in Figure 2, begins with data preparation, or in other words, data collection, data labeling, and data preprocessing such as removing digital noise, followed by feature extraction, mandatory if a ML-based technique is later implemented, or optional if a DL-based one is implemented. To implement an ML or DL technique, the dataset first needs same preparations, dimensionality reduction if it is required, and then splitting the data into training and testing sets. After implementing it, the obtained results need to be compared with the ground-truth data to evaluate the model. The performance evaluation is most often represented with confusion matrix and accuracy. Based on the descripted process, the HAR implementation can be divided into two categories: the traditional hand-crafted feature based, and the deep learning based.

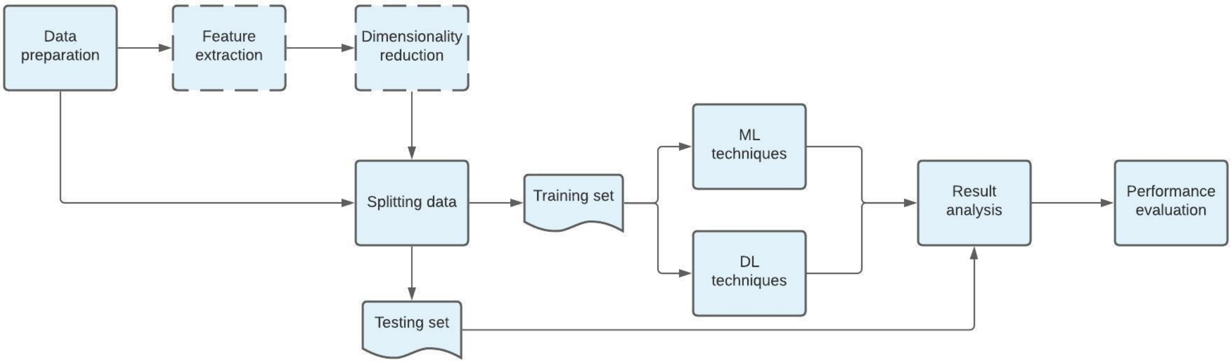

Figure 2. The HAR process

2.1.1. Implementation based on traditional hand-crafted features

Methods based on hand-crafted features involve two main steps. A feature extraction along with dimensionality reduction if needed, and classification. Feature extraction can be described as a pre-processing part to remove 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, while for high level features is important to include domain-knowledge to get structured information related to the action being taken [8]. Various features that can be extracted for the purpose of HAR are shown on Figure 3, and described in the survey [7].

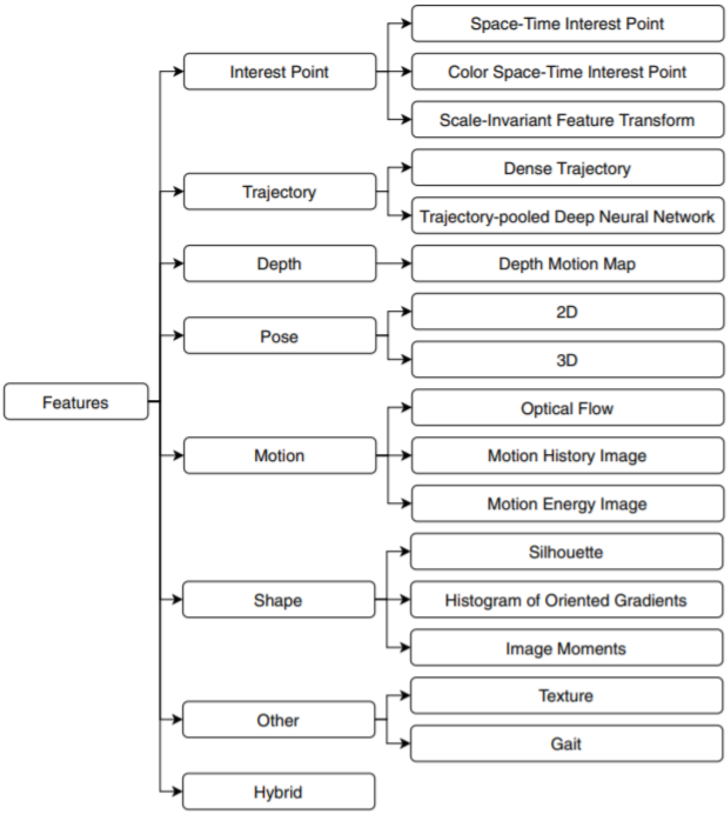

Figure 3. Feature representation [7]

The step of feature extraction in HAR may allow to identify the object movement in the scene, but these descriptors do not provide an understanding of the actions. Therefore, classification techniques must be used. In Figure 4, are shown different ML based techniques that can be used in HAR purpose, such as Support Vector Machine (SVM) and Hidden Markov Model (HMM). Description of ML based techniques and the overview of belonging HAR implementation can be found in [7].

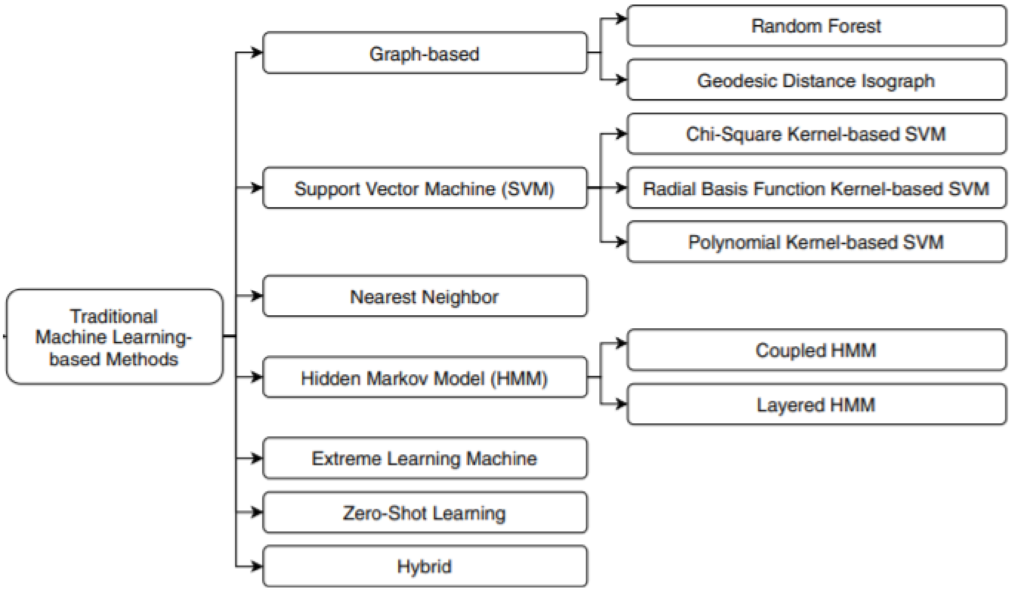
Traditional Machine Learning-based Methods
Graph-based
Random Forest
Geodesic Distance Isograph
Support Vector Machine (SVM)
Chi-Square Kernel-based SVM
Radial Basis Function Kernel-based SVM
Polynomial Kernel-based SVM
Nearest Neighbor
Hidden Markov Model (HMM)
Coupled HMM
Layered HMM
Extreme Learning Machine
Zero-Shot Learning
Hybrid

Figure 4. Machine Learning based methods [7]

Actions in HAR can also be classified, after feature extractions, using different DL-based techniques, although DL-based techniques are more often used on their own.

2.1.2. Implementation based on deep learning

Different from methods based on hand-crafted where features are engineered and pre-defined [9], DL-based methods are capable of automatically processing raw image and video data for feature extraction, description, and classification of HAR [10]. For example, a neural network that learns to classify labeled actions will contain in its hidden layers a representation of the input data that can be used as features to represent such data [9].

One of the most popular models being used in DL-based implementation of HAR is the Convolution Neural Network (CNN). Along with CNN, widely used networks for the HAR task Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Deep Belief Network (DBN), as well as Generative Adversarial Network (GAN). Description of DL-based techniques and the overview of belonging HAR implementation can be found in [7].

The main differences between the two implementations DL-based and hand-crafted based are: in the DL-based one, preprocessing phase is not required because deep learning eliminates the manual feature extraction phase because the network extract features directly from images during training, but DL requires larger dataset to compensate large size of hidden layer, it takes more time to train, and so demands the use of accelerators like Graphics Processing Unit (GPU) and Tensor Processing Unit (TPU).

2.2. HAR datasets

To evaluate the performance of HAR implementation there is a need of labeled data. Besides custom datasets that one can do on their own, there are some common public datasets available for the task of HAR. Table 1 presents popular datasets for HAR in chronological order, with the type of data used, number of samples, number of actions, and a brief description.

Table 1. Popular datasets for HAR

Dataset	Year	Type of data	No. of samples	No. of actions	Description
KTH [11]	2004	RGB	2391	6	Contains of six types of human actions (walking, jogging, running, boxing, hand waving, and hand clapping) performed several times by 25 people in four different scenarios. The sequences were down sampled to the spatial resolution of 160x120 pixels and have a length of four seconds in average.
Weizmann [12]	2005	RGB	4500	10	Contains sequences recorded with stationary camera, showing ten different people, each of which performing 10 different actions: walking, running, skipping, bending, jumping jack, jumping forward on two legs, jumping in place on two legs, galloping sideways, waving one hand and waving two hands.
IXMAS [13]	2006	RGB	1148	14	A multi view dataset (5 cameras) with 14 daily actions (nothing, crossing arms, checking watch, scratching head, getting up, turning around, throwing, sitting down, walking, waving, kicking, punching, pointing and picking up) performed by 11 actors.
HOLLYWOOD 2 [14]	2008	RGB	3.669	12	A dataset obtained from different movies containing eight actions (get out car, answer phone, handshake, hug person, sit down, sit up, stand up, kiss, run, driving car, eat, and fight.
UCF 11 YouTube Action [15]	2009	RGB	1.160	11	It contains 11 action categories: basketball shooting, biking/cycling, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. For each category, the videos are grouped into 25 groups with more than 4 action clips in it. The video clips in the same group share some common features, such as the same actor, similar background, similar viewpoint, and so on.
HMDB51 [16]	2011	RGB	6.766	51	Contains videos manually labeled and extracted from different sources, such as Google videos, YouTube and the Prelinger archive. It is divided into 51 action classes, grouped in five types: body motion for human interaction, general body motion, facial actions with object manipulation, general facial actions, and body motion with object interaction
UCF 50 [17]	2012	RGB	6.618	50	Extension of UFC 11 that contains realistic videos taken from YouTube and consist of 50 action categories (for example: Biking, Diving, Drumming, Playing Guitar, High Jump, , Horse Riding, Hula Hoop, Javelin Throw, Juggling Balls, Jump Rope, Pizza Tossing, Skate Boarding, Skiing, Skijet, etc.)
UCF 101 [18]	2012	RGB	13.320	101	Extension of UCF50 that contains 101 action classes. The action categories can be divided into five types: Human-Object Interaction, Body-Motion Only, Human-Human Interaction, Playing Musical Instruments, and Sports.
MSR DailyActivity 3D [19]	2012	RGB, depth, skeleton	320	16	Dataset captured by a Kinect device that contains 16 activity types: drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up, sit down. If possible. Each subject performs an activity in two different poses: “sitting on sofa” and “standing”.

Northwestern-UCLA [20]	2014	RGB, depth, skeleton	1.475	10	Data captured simultaneously by three Kinect cameras. This dataset includes 10 action categories: pick up with one hand, pick up with two hands, drop trash, walk around, sit down, stand up, donning, doffing, throw, carry. Each action is performed by 10 actors. Data is taken from a variety of viewpoints
MHAD [21]	2014	RGB, depth, skeleton	861	27	Contains a set of activities that have dynamic body movements. Some activities have dynamics in both upper and lower extremities. In this dataset, image resolution is 640 × 480
NTU-RGB+D [22]	2016	RGB, depth, skeleton	56.880	60	Contains 60 actions performed by 40 subjects, with 80 different views.
MultiTHUMOS [23]	2017	RGB	38.690	65	Contains dense, multilabel, frame-level action annotations for 30 hours across 400 videos. 38,690 annotations of 65 action classes, with an average of 1.5 labels per frame and 10.5 action classes per video.
HACS [24]	2019	RGB	1.5M	200	Large-scale dataset for recognition and temporal localization of human actions collected from Web videos. It has 504K videos retrieved from YouTube and then trimmed in a total of a total of 1.5M clips of 2-second duration.
Kinectics 700-2020 [25]	2020	RGB	650.000	700	Kinects is a collection of large-scale datasets (with 400/600/700 human action classes depending of version) taken from YouTube. The videos include human-object interactions such as playing instruments, as well as human-human interactions such as shaking hands and hugging. Each clip is human annotated with a single action class and it lasts around 10 seconds.

Two well-known datasets, that are used as benchmark for HAR implementation, are KTH and Weizmann. They were introduced in the early 20th century, and, compared to today's datasets, such as HACS and Kinectics 700-2020, they contain a small number of action classes and number of samples. The KTH dataset consists of 6 classes, and todays Kinectics dataset has 700 different classes, that is, 116 times more classes. Comparing samples of KTH (2.391) and HACS (1.5M), in HACS we have 6.273 times more samples.

It should be emphasized that for effective evaluation of HAR techniques there must be used a dataset that describes actions in a realistic way with all the challenges possible, but many of them are recorded in controlled environments. Although, there are more and more realistic ones taken from web videos and movies, there is not a unique dataset that present all scenarios possible, so to evaluate the performances of HAR researchers must use different public datasets or create some on their own for the domain they are working on.

2.2.1. Sport datasets

For HAR in sport domain there are three popular public action datasets: UCF Sports Action Data Set [8], Sports-1M Dataset [26], and Olympic Sports Dataset [27].

The UCF Sports Action Data Set, shown in Figure 5, consists of a set of actions collected from various sports which are typically featured on broadcast television channels. The dataset includes a total of 150 sequences with the resolution of 720 x 480 and frame rate 10fps. The dataset includes 10 actions: Diving, Golf Swing, Kicking, Lifting, Riding Horse, Running, SkateBoarding, Swing-Bench, Swing-Side, and Walking.

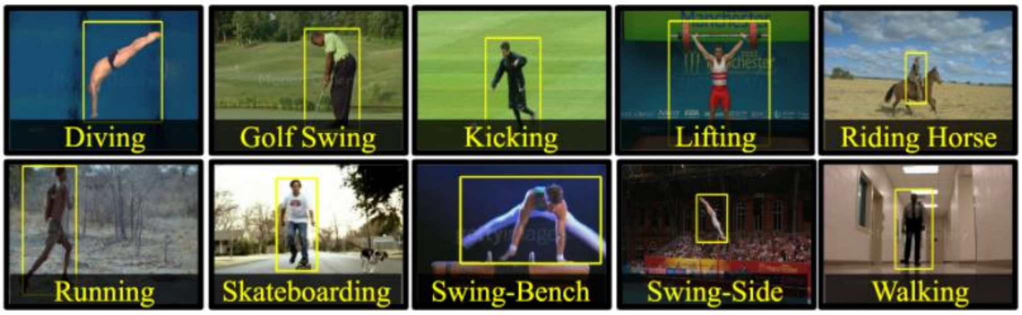

Figure 5. UCF Sports Action Data Set [8]

The Sports-1M dataset contains 1,133,158 video URLs which have been annotated automatically with 487 sports labels using the YouTube Topics API. Some examples of sport labels are shown in Figure 6.

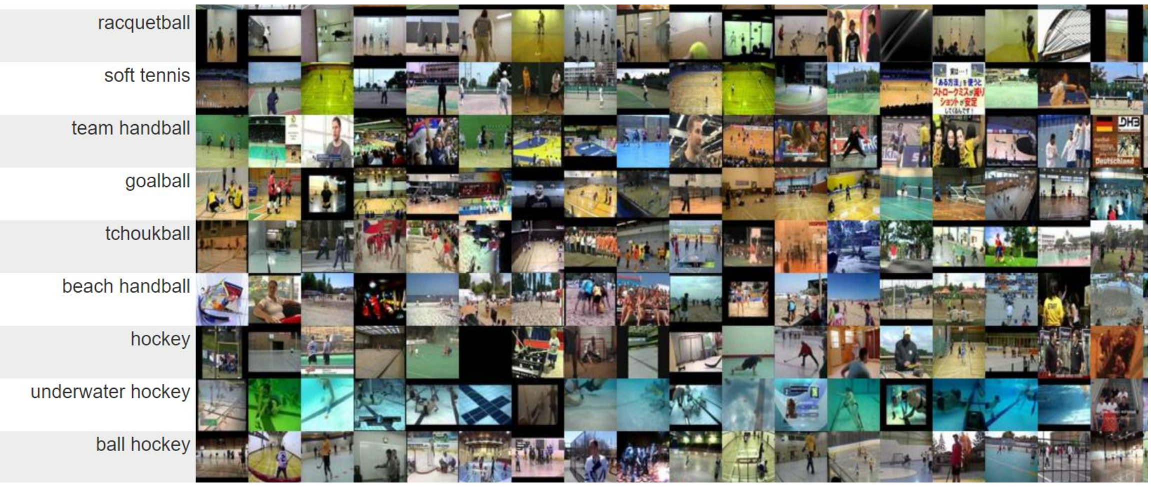

Figure 6. Sports-1M Dataset [26]

The Olympic Sports Dataset contains video sequences of athletes practicing different sports. The videos are taken from YouTube and annotated with the help of Amazon Mechanical Turk. The current release contains 16 sports shown on Figure 7.

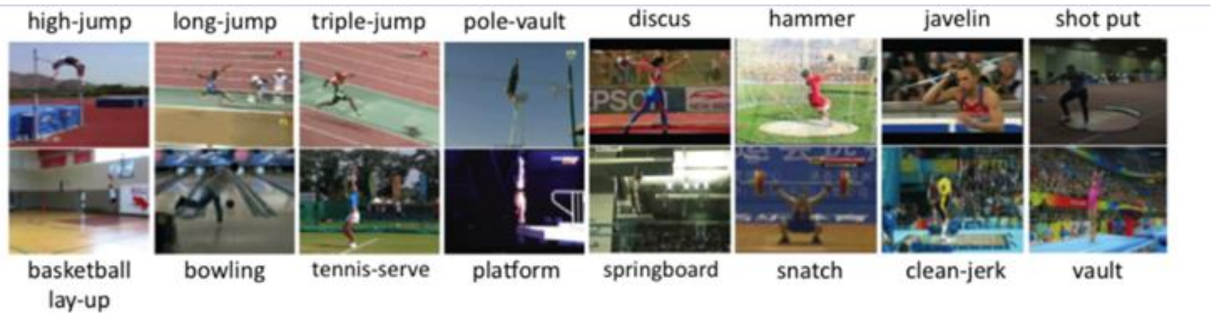

Figure 7. Olympic Sports Dataset [27]

These datasets are used for recognizing different sport actions such as bowling, swimming, or running, that have specific background that can help in recognizing the action. More challenging is to recognize action regarding one sport domain, but in that case, there is a problem of not having public dataset disponible. Therefore, when performing HAR on a certain sport domain, a lot of researchers creates appropriate dataset for that task on their own, as can be seen in the next chapter where different HAR implementation are described.

3. HAR in sport domain

Human action recognition in sports is commonly used to make automated statistical analysis about a sport match, such as the number of occurrences of a given action (e.g. number of shoots on a goal frame). But more importantly, is used to analyze tactics, key-stages in the game, or follow the activity (e.g. running distance) and performances of a player. This analysis can help the players and their trainers to find key elements of technique that should be adopted to achieve better results. The trainer, when it comes to team sports, should keep track of all the players on the field and all the action they perform, especially in which way they perform it. Therefore, with an automatic system that can track all of this, players could be even more successful. It should be emphasized that this is a challenging task for a computer because a lot of players is being on the sports field at the same time performing different actions simultaneously, which causes problems such as occlusion and cluttered scenes.

In this research, that is focused on sport domain, actions will be distinguished from activities, as shown in Figure 8. When there is a primitive action, single person action, or interaction with an object it is classified as action, but when more players are involved, and they perform some combination of the aforementioned actions is classified as activity.

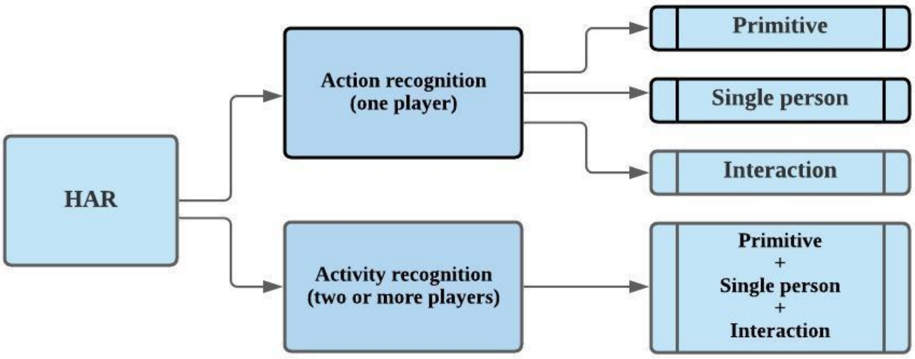
HAR
Action recognition (one player)
Activity recognition (two or more players)
Primitive
Single person
Interaction
Primitive + Single person + Interaction

Figure 8. HAR in sport domain, action vs activity recognition

There are different sports on which researchers are focused today, the one played with two players, such as tennis, and the one played in teams, such as basketball, soccer, baseball, hockey, volleyball, handball etc.

In sports with two players, there are only action recognition with interaction. In sports like tennis, badminton, and table tennis, there is a player interacting with a racket and a ball.

In team sports there can be all the combination of action and activity recognition. For example, for action recognition in handball as primitive is standing, for single person running, for interaction shot and jump-shot that includes the player interacting with a ball. For activity recognition the examples can be passing (a combination of actions, two players doing a single person action – running, and simultaneously performing interaction action with a ball - throwing/catching), crossing involving three players, or defense involving more team players.

3.1. Overview of HAR implementation in different sports

An overview of different HAR implementation in sports is presented, divided into two categories: sports with two players and team sports.

3.1.1. Sports with two players

Tennis

At the very beginning of the 20th century, a couple of authors tried to recognize player's actions in broadcast tennis videos. In [28], authors proposed an automatic annotation method of sports video for content-based retrieval. Players' actions are analyzed by 2D appearance-based matching using the transition of players' silhouettes and Hidden Markov model. The main actions of which the researchers focused are foreside-swing, backside-swing, and overshoulder-swing. Later in [29], improved their method by combining player and ball position, but ball detection and tracking were difficult due to poor quality of videos.

Guided by their work, authors of [30], had focused on two basic actions left-swing, and right-swing, which covers 90% player's behavior in tennis. The main challenge where 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 where they treat optical flow as spatial patterns of noisy measurements instead of precise pixel displacements, and they propose a framework combining players' action recognition with other multimodal features for semantic and tactic analysis. They accomplished better results than the appearance-based approaches for that time.

In [31], they implemented transductive transfer learning methods for action classification. Actions, in their work, are described using HOG3D features and for transfer they used a method based on feature re-weighting and a method based on feature translation and scaling. They applied it for action classification in tennis games for 'non-hit', 'hit' and 'serve' actions. They worked on non-publicly available data.

With the development of DL and with introduction of the THETIS dataset [32] 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 (open stance), forehand (slice), forehand (volley), service (flat), service (kick), service (slice), smash.

In [33], is proposed a model that considers the per-frame motion to be regarded as a word (within an alphabet of possible motions), and the sequence of frames as a phrase whose meaning is determined by the words given in a specific order. This feature extraction mechanism allows a semantic treatment of the classification stage using Conditional Random Fields. The system was applied on the RGB videos of the THETIS dataset.

Recent researchers, as [34], focused on the fine-grained action recognition in tennis. In their model, videos are represented as sequences of features, 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 THETIS dataset.

In [35], researchers explored the possibilities of using convolutional neural networks to recognize the type of tennis shots. They compared Inception-v3 [36] and MobileNet networks [37]. The focus was on action recognition of the following actions: backhand preparation phase, backhand shot, forehand preparation phase, forehand shot, and non-shot. MobileNet network achieved better results.

In [38], authors proposed weighted LSTM adopted with convolutional neural network representations for three-dimensional tennis shots recognition. First, the local two-dimensional convolutional neural network spatial representations are extracted from each video frame individually using a pre-trained Inception network. Then, a weighted LSTM decoder is introduced to take the output state at time t and the historical embedding feature at time t-1 to generate feature vector using a score weighting scheme. Finally, they use the adopted CNN and weighted LSTM to map the original visual features into a vector space to generate the spatial-temporal semantical description of visual sequences and classify the action video content. The model was applied on THETIS dataset.

Badminton

In [39], researchers tried to recognize 10 badminton actions from 300 depth map sequences acquired by Microsoft Kinect sensor. Bone orientation details of badminton players were computed and extracted in order to form a bag of quaternions feature vectors. After conversion to log-covariance matrix, the system is trained, and the badminton actions are classified by a SVM classifier.

In [40], authors proposed an approach for badminton stroke recognition using dense trajectories and trajectory aligned HOG features which are calculated inside local bounding boxes around players. A four-class SVM classifier is then used to classify badminton strokes from video footages to be either smash, forehand, backhand or other.

In [41], for automatic action recognition in badminton, researchers inserted a sensor chip into the badminton racket to collect the data of ten major badminton actions. The best results they achieved with the proposed AFEB-AlexNet, then with AlexNet [42], and then with the LSTM networks.

In [43], researchers developed a model for automated badminton action recognition from images of two classes, hit and non-hit action, using the deep learning pre-trained AlexNet for features extraction and linear Support-Vector Machine (SVM) to classify them. Before pre-trained AlexNet was directly extracting the features, they introduced a new local CNN extractor in recognition pipeline.

Later in [44], they compared, for the same two actions, four different pre-trained models of deep CNN: AlexNet, GoogleNet [45], VggNet-16 and VggNet-19. The models were evaluated by plotting its performance accuracy in form of confusion matrix. The result shows that the GoogleNet model has best performances.

In [46], the goal was to formulate an automated system for badminton smash recognition on broadcasted videos using pre-trained CNN methods. Smash and other badminton actions such as clear, drop, lift and net shot were studied. They performed two experiments. The first experiment is the study on the performance between four different existing pre-trained models which is AlexNet, GoogleNet, VggNet-16 and VggNet-19 in 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 which is 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 [47], on the same badminton dataset, they extracted features using Alexnet but introducing a new local feature extractor technique that extracts features at the fc8 layer. Features were being classified using SVM. The experiment was repeated using a normal global feature extractor technique. Lastly, both local and global feature extractor techniques were repeated using GoogleNet CNN model to compare the performance between AlexNet and GoogleNet model. The results showed that the new local feature extractor using AlexNet CNN model has the best performance.

Table tennis

In [48], authors used Wi-Fi signals to recognize nine different actions in table tennis. They used a discrete wavelet decomposition to decompose the Wi-Fi signal in combination with SVM, and later with KNN, used to classify the actions. Wi-Fi devices provide channel status information for the Wi-Fi installation. By recording information on the status of the carrier between the transmitter and the receiver, such installations can reflect very well the changes in the wireless signal, and then obtain fine-grained wireless signal measurements.

In [49], authors used action recognition to annotate sport videos with different strokes in table tennis. They used K-means to classify the Optical Flow singularities into six clusters in combination with spatial information, and later with HOG features, and then used cross-validated linear SVM to classify the actions. The dataset obtained is called MediaEval 2019.

In [50], author applied Siamese Spatio-Temporal Convolutional neural network for the purpose of recognition of 20 table tennis strokes, actions with low inter-class variability. Their model takes as inputs a RGB image sequence from their own TTStroke-21 dataset (129 videos representing 94 hours of table tennis game), on which is computed Optical Flow. After 3 spatio-temporal convolutions, data are fused in a fully connected layer of a proposed Siamese network architecture.

Later, in [51], based on previous research, a temporal segmentation of table tennis strokes in videos is performed, based on temporal sliding windows and their aforementioned classifier (performed both detection and classification simultaneously). In [52], from the optical flow, a region of interest-ROI-is inferred. Their classifier is then feed by RGB and optical flow ROIs stream to give a probabilistic classification over all the table tennis strokes

In [53], working on the same data, they present a new architecture, a Twin Spatio-Temporal Convolutional Neural Network for the same purpose, The proposed Twin architecture is a two stream network both comprising 3 spatiotemporal convolutional layers, followed by a fully connected layer where data are fused.

3.1.2. Team sports

Soccer

In the begging of the 20th century, in [54] is presented a multi-view action recognition framework able to extract human silhouette clues from different synchronized static cameras and then to validate them by analyzing scene dynamics. They implemented the neural recognition of the human body configuration by using a novel mathematical tool called Contourlet transform, and then performed 3D ball and player motion analysis. They then merged the outcomes to accomplish the final player action recognition task.

In [55], is proposed a local motion-based approach for recognizing group activities in soccer videos. Given the SIFT key-point matches on two successive frames, they proposed a method to group these key-points into the background point set (to estimate camera motion) and the foreground point set (to represent group activities). After camera motion compensation, they applied a local motion descriptor to characterize relative motion between corresponding key-points on two consecutive frames. The novel descriptor is effective in representing group activities since it focuses on local motion of individuals and excludes noise such as background motion caused by inaccurate compensation.

Later in the 20th century, in [56], is introduced a dataset called SoccerNet for action spotting in soccer videos collected from online sources. The action that are annotated can be categorized in goals, cards, and substitutions. The dataset includes 6637 samples of actions. The feature extraction was performed with an 3D CNN [57], I3D [58] , and ResNet [59]. They classified the action with different number of clusters for different pooling methods (mean pooling and max pooling, custom CNN, SoftDBOW, NetFV, NetVLAD and NetRVLAD [60]).

In [61], is introduced the pose-projected action recognition hourglass network (PARHN) for performing player-level action recognition in soccer. It includes an embedded pose projection component that regularizes the numerical range of the player's pose vector and incorporates the temporal information by having a parallel structure for extracting projected pose vectors from all frames of an input sequence and also by 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 pass, and giving pass actions.

In [62], 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 using a ResNet based 3D-CNN.

In [63], s proposed a deep learning based fine-grained action recognition method to analyze 132 soccer training videos for evaluating whether a player has stopped a soccer ball successfully or not (2543 ball-stopping actions annotated), that is, for an human-object (player-ball) interaction. Because for these two actions, the motions and the scenes have no obvious difference, it is important to consider difference in the human-object interaction motions. A cascaded scheme of deep networks based on the object-level trajectories is proposed, constructed by concatenating a YOLOv3 network for detection with a classifier LSTM based network

In [64], researchers proposed a framework that automatically recognize actions of players in live soccer game which will be helpful for text query based video search, for extracting stats in a soccer game and to generate textual commentary and the Soccer-8k dataset which consists of different action clips in the soccer play.

In [65], authors researched group activity recognition for video data in soccer. They proposed self-attention models to learn and extract relevant information from a group of soccer players for activity detection from both trajectory and video data. They explicitly modeled interactions between players and ball, and for visual data as backbone they used I3D CNN. I3D models trained on the whole frame performed better than directly modeling player interactions using transformers or Graph Convolutional Neural Networks [66]. They used a dataset of 74 soccer games from the 2018-2019 English Premier League and focused on detecting pass, shot and reception.

In [67] author researched individual action and group activity recognition in soccer video. He proposed a method that infers both eight individual actions and eleven group activities simultaneously from soccer videos. He used player-centric snippets as model input (the player snippets are obtained using an Aggregated Channel Features person detector [68] and a virtual camera that zooms in on each detected player, creating a standardized video frame cut-out). For feature extraction he used an I3D CNN based on RGB video and optical flow, and for classification of action and activities he used feature suppression and zero-padding in graph attention networks.

Basketball

In [69], 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 in order to segment the game into activities (offense, defense, time out). Key elements are detected (starting formation, screen, and move) and their temporal orders is used to produce a semantic description of the observed activity.

In [70] is presented a feature-representation method for recognizing actions in broadcast basketball videos which 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 then 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 [71], authors focused on real-time detection and tracking of basketball players using deep neural networks, but also on action recognition. For that purpose, they used a subset of broadcast basketball videos of the NCAA Basketball Dataset [72]. They used YOLOv2 [73] and SORT [74] for detection and tracking, and then LSTM for action classification.

In [75], authors researched action recognition in basketball, with the purpose of detecting events and key actors in multi-person videos. The proposed model learns to detect eleven action/event classes on their own dataset in such videos while automatically "attending" to the people responsible for the event. They tracked people in videos and 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, the Inception-7 [76] network, for action detection and classification. They used as dataset a subset of the NCAA games available on YouTube.

In [77], is released a dataset with fine-grained actions in basketball game videos. They propose an approach by integrating the NTS-Net [78] into two-stream network to locate the most informative regions and extract more discriminative features for fine-grained action recognition.

In [79], the focus is on recognition of group activities and the outcome (score or not score) in basketball. It is proposed a scheme for global and local motion patterns and key visual information for recognition in basketball videos. A two-stream a 3D CNN framework is utilized for group activity recognition over the separated global and local motion patterns.

Volleyball

In [80], researchers focused on ball detection and trajectory extraction in volleyball videos, this paper presents a physics-based scheme which utilizes the motion characteristics to extract ball trajectory from lots of moving objects. Based on game-specific properties, the ball trajectory can be exploited to recognize set types for tactics inference and to detect basic actions in the volleyball game for close-up presentation.

In [81], the focus is on group activity recognition for volleyball. A LSTM model is designed to represent action dynamics of individual people in a sequence and another LSTM model is designed to aggregate person-level information for whole activity understanding. Through a two-stage process, they learned a temporal representation of person-level actions and combined the representation of individual people to recognize the group activity. Authors evaluated their work on two datasets: The Collective Activity Dataset and a new volleyball dataset they created based on publicly available YouTube volleyball videos. Based on their work, in [82], 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 on the temporal features.

In [83] authors are focused on activity recognition in 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 [84], is presented a unified framework for understanding human social behaviors in raw image sequences. The architecture does not rely on external detection algorithms but rather is trained end-to-end to generate dense proposal maps that are refined with the inference scheme. The temporal consistency is handled with a person-level matching RNN (as baseline they used Inception-v3 [85], HDTM [86], and others models). The framework is evaluated on the dataset from [86]]

In [87], the focus is on Decomposition and recognition of playing volleyball action based on SVM algorithm. Firstly, is used the principal component analysis for dimension reduction, then 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 [88], authors proposed a 3D global trajectory and multi-view local motion combined volleyball player action recognition method. Global 3D feature and local motion feature mutually promote each other, and the actions are recognized well. The recognition is performed on game videos from the Semifinal and Final Game of 2014 Japan Inter High School Games of Men's Volleyball in Tokyo Metropolitan Gymnasium

In [89], is proposed a recognition framework based 3D global and multi-view local features combination with global team formation feature, ball state feature and abrupt pose features for action recognition in volleyball games. The team formation extracts the 3D trajectories from the whole team members rather than a single target player. The ball motion state feature extracts feature from the 3D ball trajectory. The pose feature extraction consists of two parts, hit frame pose and pose variation. These two features make difference of each action quality more distinguishable by focusing on the motion standard and stability between different quality actions. The recognition is performed on the same data as [88].

In [90], authors evaluated balanced and imbalanced learning methods with their proposed 'super-bagging' method for volleyball action modelling. 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, while balanced learning provides better results for the dominant hand using a tree bagger classifier than imbalanced learning.

In [91], the focus is on group activity recognition. The authors presented an attentive 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 capturing 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 an video volleyball dataset.

Hockey

In [92], the authors focused on hockey action in medium field, where typical figure has a resolution of dozens of pixels in each dimension. Self-initializing tracker tracks the figures of hockey players. Then, during the stabilization process a new stabilization algorithm uses a mixture of templates to estimate the position and a scale of a figure. Actions are classified by an action recognition system that uses motion and pose features for its classification. Image gradients are decomposed into four non-negative components which are used to characterize poses. Better results are obtained with pose features in comparison with motion ones.

In [93], is presented a template-based algorithm to track and recognize hockey player's actions in an integrated system using only visual information. This algorithm couples tracking and action recognition into single framework, where tracking and action recognition assists one another. For the hockey sequences, images of players performing 6 actions are collected and transformed to the PCA-HOG descriptor, which is computed by the first transforming the athletes to the grids of Histograms of Orianted Gradient (HOG) descriptor and then project it to a linear subspace by Principal Component Anaysis (PCA).

In [94], authors designed a CNN, called Action recognition Hourglass Network (ARHN) to interpret player actions in ice hockey video. 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, in the second latent features are transformed into a common 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 [95], authors proposed a deep architecture to classify puck possession events in ice hockey. Model has three distinct phases: feature extraction, feature aggregation and, learning and inference. CNN are used for feature extraction and aggregation, followed by a late fusion model on top to extract and aggregate different types of features that includes handcrafted homography features for encoding the camera information. RNN are used for temporal extension and classification of the events, to which output of CNN is passed. Team pooling and pre-trained model is used to incorporate individual attributes of the players and the interaction amongst them. Only the player positions on the image and the homography matrix are need for the model, which simplifies the input of the system. Model is evaluated on a new dataset called Ice Hockey Dataset and on a volleyball dataset.

In [96], is focused on finding fight scenes in hockey sport videos. Fast fourier and radon transform are applied on the local motion, after it is extracted in the video frames using blur information. Authors used transfer learning with pre-trained deep learning model VGG-Net to identify fight scenes in video frames, and feed forward neural networks to perform a comparison of the methodology. Author used videos from National Hockey League dataset to present the outcome of the model.

In [97], authors presented a deep learning-based solution for hockey game action recognition in multi-label learning settings having class imbalance problem. 3D CNN based multilabel deep HAR system was implemented for multi-label class-imbalanced action recognition. System was tested for two different scenarios: an ensemble of k binary networks vs. a single k-output network, on a publicly available hockey videos dataset.

In [98], the author designed and implemented an CNN automated method to determine the pose of a hockey player with and without a hockey stick from broadcast game video in addition to performing action recognition via pose. Deep learning computer vision architecture HyperStackNet has been designed and implemented for joint player and stick pose estimation. The action recognition hourglass network, or ARHN, is designed to interpret player actions in ice hockey video using estimated pose. The first component of ARHN is the latent pose estimator, the second transforms latent features to a common frame of reference, and the third performs action recognition. Authors generated an their own an annotated dataset for this purpose.

In [99], a two-stream architecture is proposed for action recognition in hockey. 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' action. A publicly available dataset HARPET (Hockey Action Recognition Pose Estimation, Temporal) was created, composed of sequences of annotated actions and pose of hockey players including their hockey sticks as an extension of human body pose.

In [100], authors introduced a two-stream network utilizing player pose sequences and optical flow features for recognizing hockey actions. Players pose sequences are compact representations consisting 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. Authors demonstrate the efficacy of the method on the aforementioned HARPET dataset.

In [101], a deep learning-based transfer learning model, VGG-16 has been proposed on activity recognition in field hockey. Pre-trained VGG-16 model identifies four main activities: free hit, goal, long corner, and penalty corner. Authors constructed their own hockey dataset consisting of four main activities.

Handball

In [102], authors presented a method for recognition of handball actions using mask R-CNN and STIPS focusing on minimizing the manual labor required to label the individual players performing the chosen action in a dataset. The method uses existing deep learning object recognition methods for player detection and combines the obtained location information with a player activity measure based on spatio-temporal interest points to track players that are performing the currently relevant action. In [103], authors proposed Mask R-CNN and Optical flow based method for detection and marking of handball actions called MOF method. The method was evaluated on a dataset of handball practice videos recorded in the wild.

In [104], the focus in on recognition of the throwing action in handball based on RGB-D data. Authors introduced a RGB-D dataset that can be used for an objective comparison and evaluation of handball player's performance during throws. They examined the main angles responsible for throwing performance in order to analyze individual skills of handball players and adopted the dynamic time warping technique to compare the throwing motion between two athletes.

In [105], a method for temporal segmentation and recognition of team activities in sports, based on a new activity feature extraction, is presented. The focus is on the position of team players from a plan view of the playground. They constructed a position distribution along each frame of the sequence. These methods extract activity features using the explicitly defined trajectories, where the players have specific positions. They classified six different team activities in European handball with SVM.

Baseball

In [106], is introduced a dataset named MLB-YouTube, designed for fine-grained action recognition in baseball videos. It is used for segmented video classification as well as activity detection in continuous videos. The segmented video dataset consists of 4,290 video clips. Each clip is annotated with the various baseball activities that occur, 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 both segmented and continuous videos, that is, I3D or InceptionV3 in combination with mean, max, and pyramid polling, LSTM, temporal convolution filters, and different learning of sub-events.

Cricket

In [107], 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 and they evaluated the results for 3-cluster K-Means by manually annotating the clusters as Left strokes, Right strokes and Ambiguous strokes for 562 stroke instances.

Rugby

In [108], authors researched action recognition in order to develop a tactics analysis system that could be implemented for different 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 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 projectively transforming the match video onto the bird's-eye view image, play classification was performed using the player position, the ball position, and the dense area position as feature amounts.

American football

In [109], 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, then processed the outputs by the Hidden Markov Model which encodes knowledge about the temporal structure.

In [110], a method for recognizing American football plays in video is propose games. They used shape and motion based spatio-temporal features and implemented Multiple Kernel Learning based on SVM to effectively combine different features. They evaluated their method on a dataset consisting of 78 videos of play instances collected from NCAA football games that includes left-run, middle-run, right-run, short-pass, option-pass, rollout-pass, deep-pass activities.

Based on the written review, it can be seen that the sports with two players are researched the most for the purpose of HAR in sport, thanks to the simplicity of actions (it includes only interactive actions). When it comes to team sports, soccer, then basketball, volleyball and hockey are mostly researched. For other sport there are just a few implementations. In Figure 9, is shown the distribution of HAR implementation in different sports based on the reviewed papers.

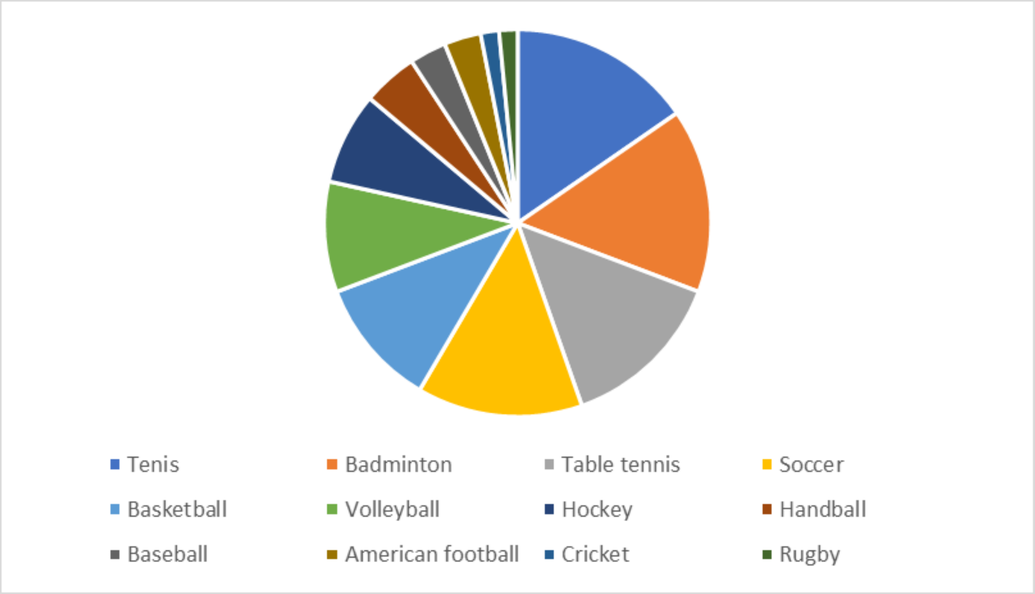

Figure 9. Distribution of HAR implementation in different sports

For most sports, there is at least one research on HAR in the early 20th century, followed by a gap without research until the development of DL. A large number of the reviewed papers is written in the last couple of years. Mostly, authors used different approaches and tried to get the best combination to obtain high result, so it cannot be concluded which method is the best overall, because every sport is different and contains different action and activities, and performing same methods on different data can't get same results. The SVM classifier is used a lot in various combination, but method based on CNN, along with LSTM yet prevail. It can be pointed out that a lot of papers in the last few years focused in fine-grained action recognition, that is, in recognizing very similar actions, which is slowly becoming a new trend.

4. HAR Experiment

In the experiment, the focus is on recognition of 11 players' actions and activities, shown in Table 2, that might occur during a handball match or practice. The dataset made for that purpose and split into training and testing sets, contains 2991 short high-quality video recordings of actions and activities in handball, recorded indoors during a handball school.

Table 2. Distribution of videos through classes and train/test sets

Class name	No. Videos (train)	No. frames (train)	No. Videos (test)	No. frames (test)
Throw	184	4907	32	913
Catch	202	4120	50	937
Shot	83	6097	22	1655
Jump-shot	270	18018	83	5676
Running	56	6049	14	1424
Dribbling	42	3549	11	753
Defense	97	6027	30	1668
Passing	509	30618	121	7252
Double-pass	35	2122	11	654
Crossing	238	18204	59	4482
Background	684	24929	158	8342
Total	2400	124640	591	33756

An example of consecutive frames in a sequence for the Catch action is shown in Figure 10.

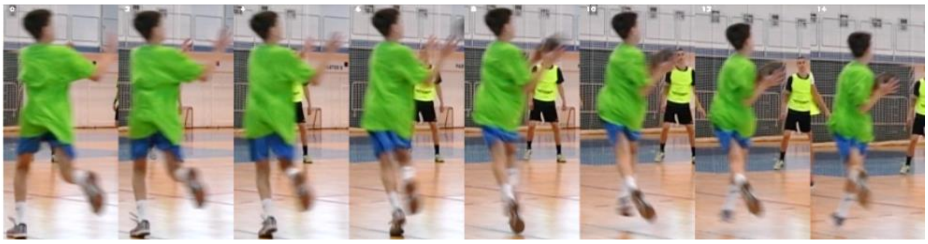

Figure 10.An example of consecutive frames for the Catch action

In the experiment, the performance of a baseline CNN-model that classifies each frame into an action class, the same CNN but with the majority voting scheme, LSTM, and MLP based models built on top of the baseline model, are compared. The models were trained and tested with different lengths of input sequences ranging from 20 to 80, since the action duration varies roughly in the same range. The results, presented as validation accuracy and obtained with a method of skipping frames (for bigger sequences) or adding ones (for smaller sequences) if needed, are shown on Figure 11.

The methods are implemented for all the 11 classes, but also for 9 of them, excluding the actions Throw and Catch because are fundamental part of other actions or activities such as passing and crossing.

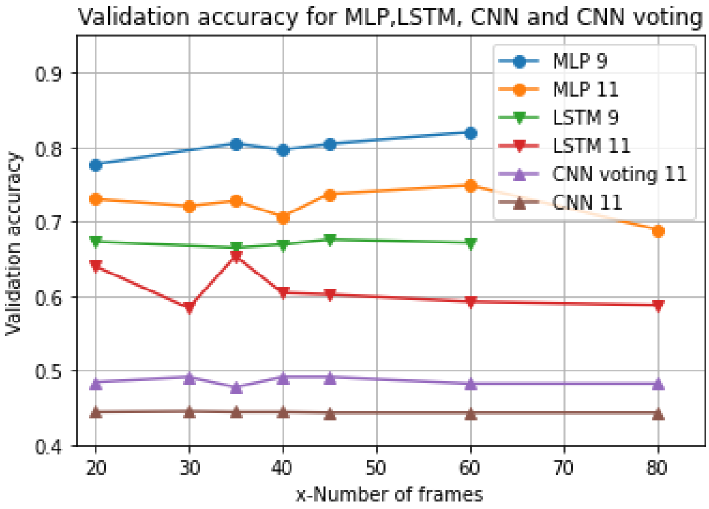

Figure 11. Validation accuracy values obtained by skipping frames for 9 and 11 classes

Out of all the models, the MLP model achieves the highest validation accuracy for both 9 and 11 classes. The best accuracy score of 81.95% the MLP model achieves with 60 input frames, but it achieves a similar score of 80.01% on average for all numbers of frames taken in the case of 9 classes. For 11 classes, MLP again has the highest accuracy score of 75% with the same number of input frames and the average accuracy for all tested number of input frames of approximately 72.25%. It can be seen that the validation accuracies for the model trained on 9 classes are higher than the ones trained on 11.

The LSTM model achieved significantly lower validation accuracy values compared to the MLP model. It achieved the best score for 9 classes of 67.58% with 45 input frames (67.05% on average), while for 11 classes the best accuracy was 65.31% for 35 input frames (60.88% on average).

The worst performance is achieved by the CNN model that has no temporal dimension with the best score of 44.5% for 30 input frames, but with minimal difference between scores for different input lengths (44.37% on average). With the majority voting scheme, the score improved for about 4% to 49.1% for 45 input frames, or 48.54% on average. From this result, it can be concluded that the temporal dimension plays a major role in action recognition.

5. Conclusion

Human action recognition can be used in different domains, such as healthcare, education, video surveillance, content-based video summarization, entertainment, human-computer interaction, gaming, and, finally, sports.

HAR is mostly focused on RGB visual data, so there are many public video datasets available for research that are presented in this one. When it comes to sport, there are three popular datasets, but their focus is not on a certain sport, but on actions from different ones, and if a researcher wants to focus on a certain sport, must create a dataset or search for existing ones for that domain.

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, rugby, and American football)). With the implementation of HAR in sport researchers tries to for recognizing players' actions and teams' activities in order to do statistical analysis for the game, to analyze tactics or to follow players behavior, all in order to improve the performances of the team or a player.

After the overview of existing implementation of HAR in sports, an experiment performed on a handball dataset is presented to see how different methods for HAR perform on it. In future work the focus will be on the performances for every single action class, on enlarging the dataset for those actions for which there are less videos, on implementing different methods of selecting frames, and on combining different methods of HAR to improve the obtained results.

References

[1] M. Dhankar and N. Walia, "An Introduction to Artificial Intelligence," in *Emerging Trends in Big Data, IoT and Cyber Security, Maharaja Surajmal Institute*, New Delhi, 2020.

[2] S. J. D. Prince, Computer Vision: Models, Learning, and Inference, Cambridge Univesity Press, 2012.

[3] 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 and Vi, "TensorFlow: A System for Large-Scale Machine Learning," in *12th USENIX Symposium on Operating Systems Design and Implementation*, Savannah, 2016.

[4] Y. Taigman, M. Yang and M. R. L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 2014.

[5] M. .NET, "ML.NET," Microsoft, 2020. [Online]. Available: https://dotnet.microsoft.com/apps/machinelearning-ai/ml-dotnet.

[6] M. Nixon and A. Aguado, Feature Extraction and Image Processing for Computer Vision, Elsevier, 2020.

[7] P. Pareek and A. Thakkar, *A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications,* Springer Nature, Artificial Intelligence Review, 2020.

[8] K. Soomro and A. Zamir, "Action Recognition in Realistic Sports Videos," *Springer,* 2014.

[9] N. A. Rahmad, M. A. As'ari, N. F. Ghazali, N. Shahar and N. A. J. Sufri, "A Survey of Video Based Action Recognition in Sports," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 11, pp. 987-993, 2018.

[10] M. Al-Faris, J. Chiverton, D. Ndzi and A. Ahmed, "A Review on Computer Vision-Based Methods for Human Action Recognition.," *J. Imaging,* 2020.

[11] C. Schuldt, I. Laptev and B. Caputo, "Recognizing human actions: A local SVM approach.," in *17th International Conference on Pattern Recognition, ICPR 2004*, Cambridge, 2004.

[12] M. Blank, L. Gorelick, E. Shechtman, M. Irani and R. Basri, "Actions as space-time shapes," in *IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, Beijing, 2005.

[13] A. Kojima, T. Tamura and K. Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," *International Journal of Computer Vision,* pp. 171-184, 2020.

[14] M. Marszalek, I. Laptev and C. Schmid, "Actions in context.," in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, Miami Beach, 2009.

[15] J. Liu, J. Luo and M. Shah, "Recognizing Realistic Actions from Videos "in the Wild"," in *IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*, Miami, 2009.

[16] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio and T. Serre, "HMDB: A large video database for human motion recognition," in *IEEE International Conference on Computer Vision (ICCV)*, Barcelona, 2011.

[17] K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos.," in *Machine Vision and Applications Journal (MVAP)*, 2012.

[18] K. Soomro, A. Zamir and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in the Wild.," 3 December 2012. [Online]. Available: https://arxiv.org/abs/1212.0402.

[19] J. Wang, L. Zicheng, W. Ying and Y. Junsong, "Mining actionlet ensemble for action recognition with depth cameras.," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, 2012.

[20] J. Wang, X. Nie, Y. Xia, Y. Wu and S. Zhu, "Cross-view Action Modeling, Learning and Recognition," in *Computer Vision and Pattern Recognition*, Columbus, 2014.

[21] C. Chen, R. Jafari and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognitionutilizing a depth camera and a wearable inertial sensor," in *IEEE InternationalConference on Image Processing (ICIP)*, Quebec City, 2015.

[22] A. Shahroudy, J. Liu, T.-T. Ng and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human ActivityAnalysis," in *IEEE Conference on Computer Vision and Pattern Recognition 2016*, Las Vegas Valley, 2016.

[23] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori and L. Fei-Fei, "Every Moment Counts: Dense Detailed Labeling of Actions in Complex Videos," *International Journal of Computer Vision,* 2017.

[24] Z. Hang, T. Antonio, T. Lorenzo and Y. Zhicheng, "HACS: Human Action Clips and Segments Datasetfor Recognition and Temporal Localization," 2019. [Online]. Available: https://arxiv.org/pdf/1712.09374.pdf.

[25] L. Smaira, J. Carreira, E. Noland, E. Clancy, A. Wu and A. Zisserman, "A Short Note on the Kinetics-700-2020 Human Action Dataset," 2020. [Online]. Available: https://arxiv.org/abs/2010.10864.

[26] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 2014.

[27] J. C. Niebles, "Olympic Sports Dataset," 2010. [Online]. Available: http://vision.stanford.edu/Datasets/OlympicSports/. [Accessed 15 11 2020].

[28] H. Miyamori and S. Iisaku, "Video annotation for content-based retrieval using human behavior analysis and domain knowl-edge," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[29] H. Miyamori, "Improving accuracy in behavior identification for content-based retrieval by using audio and video information," in *IEEE International Conference on Pattern Recognition, vol. 2*, 2020.

[30] G. Zhu, C. Xu, Q. Huang, W. Gao and L. Xing, "Player action recognition in broadcast tennis video with applications to semantic analysis of sports game," in *14th ACM International Conference on Multimedia*, Santa Barbara, 2006.

[31] N. F. Davar, T. d. Campos, J. Kittler and F. Yan, "Transductive transfer learning for action recognition in tennis games," in *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011.

[32] S. Gourgari, G. Goudelis, K. Karpouzis and S. Kollias, "THETIS: Three dimensional tennis shots a human action dataset," in *CVPR Workshops*, 2013.

[33] J. V. Maguitman, J. F. Manera, P. Negri and C. Delrieux, "Modeling Video Activity with Dynamic Phrases and Its Application to Action Recognition in Tennis Videos," in *Iberoamerican Congress on Pattern Recognition*, 2014.

[34] S. V. Mora and W. J. Knottenbelt, "Deep Learning for Domain-Specific Action Recognition in Tennis," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[35] M. Skublewska-Paszkowska, E. Lukasik, B. Szydlowski, J. Smolka and P. Powroznik, "Recognition of Tennis Shots Using Convolutional Neural Networks Based on Three-Dimensional Data," in *International Conference on Man–Machine Interactions*, 2019.

[36] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.," 2015. [Online]. Available: https://arxiv.org/abs/1502.03167.

[37] M. Z. Andrew G. Howard, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," 2017. [Online]. Available: https://arxiv.org/abs/1704.04861.

[38] J. Cai and X. Tang, "RGB Video Based Tennis Action Recognition Using a Deep Historical Long Short-Term Memory," *Computer Science, Mathematics,* 2018.

[39] H. Ting, K. Sim and F. Abas, "Automatic Badminton Action Recognition Using RGB-D Sensor," *3rd International Conference on Key Engineering Materials and Computer Science, KEMCS,* 2014.

[40] S. Ramasinghe, K. G. M. Chathuramali and R. Rodrigo, "Recognition of badminton strokes using dense trajectories," in *7th International Conference on Information and Automation for Sustainability*, 2014.

[41] Y. Wang, W. Fang, J. Ma, X. Li and A. Zhong, "Automatic Badminton Action Recognition Using CNN with Adaptive Feature Extraction on Sensor Data," in *International Conference on Intelligent Computing*, 2019.

[42] A. Krizhevsky, I. Sutskever and a. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing*, Curran, 2012.

[43] N. A. Rahmad, M. A. As’ari, M. F. Ibrahim, N. A. J. Sufri and K. Rangasamy, "Vision Based Automated Badminton Action Recognition Using the New Local Convolutional Neural Network Extractor," in *International Conference on Movement, Health and Exercise*, 2019.

[44] N. A. Rahmad, N. A. J. Sufri, M. A. As'ari and A. Azaman, "Recognition of Badminton Action Using Convolutional Neural Network," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI).*

[45] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going Deeper with Convolutions," 2014. [Online]. Available: https://arxiv.org/abs/1409.4842.

[46] N. A. Rahmad, K. S. M. A. As’ari and I. Zulkapri, "Automated badminton smash recognition using convolutional neuralnetwork on the vision based data," in *Sustainable and Integrated Engineering International Conference*, 2019.

[47] N. A. Rahmad and M. A. As’ari, "The new Convolutional Neural Network (CNN) local feature extractor for automated badminton action recognitionon vision based data," in *JICETS*, 2019.

[48] C. Chen, Y. Shu, K.-I. Shu and H. Zhang, "WiTT:Modeling and the evaluation of table tennis actions based on WIFI signals," in *24th International Conference on Pattern Recognition (ICPR)*, 2018.

[49] J. Calandre, R. Péteri and L. Mascarilla, "Optical Flow Singularities forSports Video Annotation: Detection of Strokes in Table Tennis," *MediaEval,* 2019.

[50] P.-E. Martin, J. Benois-Pineau, R. Péteri and J. Morlier, "Sport Action Recognition with Siamese Spatio-Temporal CNNs: Application to Table Tennis," in *International Conference on Content-Based Multimedia Indexing (CBMI)*, 2018.

[51] P.-E. Martin, J. Benois-Pineau and R. Péteri, "Fine-Grained Action Detection and Classification in Table Tennis with Siamese Spatio-Temporal Convolutional Neural Network," in *IEEE International Conference on Image Processing (ICIP)*, 2019.

[52] P.-E. Martin, J. Benois-Pineau, B. Mansencal, R. Péteri and J. Morlier, "Siamese Spatio-temporal convolutional neural network for stroke classification in Table Tennis games," in *Medi-aEval 2019 Workshop*, 2019.

[53] P.-E. Martin, J. Benois-Pineau, R. Peteri and J. Morlier, "Fine grained sport action recognition with Twin spatio-temporal convolutional neural networks Application to table tennis," *Springer Nature 202,* 2020.

[54] M. Leo, T. D’Orazio, P. Spagnolo, P. L. Mazzeo and A. Distante, "Multi-view Player Action Recognition in Soccer Games," in *International Conference on Computer Vision / Computer Graphics Collaboration Techniques and Applications*, 2009.

[55] Y. Kong, W. Hu, X. Zhang, H. Wang and Y. Jia, "Learning Group Activity in Soccer Videos from Local Motion," in *Asian Conference on Computer Vision*, 2009.

[56] S. Giancola, M. Amine, T. Dghaily and B. Ghanem, "SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018.

[57] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," *ICCV,* 2015.

[58] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," 2017. [Online]. Available: https://arxiv.org/abs/1705.07750.

[59] K. He, S. R. X. Zhang and J. Sun, "Deep residual learning for image recognition," *CVPR,* 2016.

[60] A. Miech, I. Laptev and J. Sivic, "Learnable pooling with context gating for video classification," 2017. [Online]. Available: https://arxiv.org/abs/1706.06905.

[61] M. Fani, K. Vats, C. Dulhanty, D. A. Clausi and J. Zelek, "Pose-Projected Action Recognition Hourglass Network (PARHN) in Soccer," in *16th Conference on Computer and Robot Vision (CRV)*, 2019.

[62] R. Agyeman, R. Muhammad and G. S. Choi, "Soccer Video Summarization Using Deep Learning," in *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019.

[63] J. Xiong, L. Lu, H. Wang, J. Yang and G. Gui, "Object-Level Trajectories Based Fine-Grained Action Recognition in Visual IoT Applications," in *IEEE Access (Volume: 7)*, 2019.

[64] Y. Ganesh, A. S. Teja, S. Munnangi and G. R. Murthy, "A Novel Framework for Fine Grained Action Recognition in Soccer," in *IWANN*, 2019.

[65] R. Sanford, S. Gorji, L. G. Hafemann, B. Pourbabaee and M. Javan, "Group Activity Detection from Trajectory and Video Data in Soccer," in *EEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[66] J. Wu, L. Wang, L. Wang, J. Guo and G. Wu, "Learning actor relation graphs for group activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[67] B. Gerats, "Master's Thesis: Individual action and group activity recognition in soccer videos," University of Twente, Student Theses, 2020.

[68] D. Piotr, A. Ron, B. Serge and P. Pietro, "Fast feature pyramids for object detection," in *IEEE transactions on pattern analysis and machine intelligence*, 2014.

[69] M. Peršea, M. Kristana, S. Kovačiča, G. Vučkovič and J. Perša, "A trajectory-based analysis of coordinated team activity in a basketball game," *Computer Vision and Image Understanding,* 2009.

[70] M. Takahashi, M. Naemura, M. Fujii and J. J. Little, "Recognition of Action in Broadcast Basketball Videos on the Basis of Global and Local Pairwise Representation," in *IEEE International Symposium on Multimedia*, 2013.

[71] 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.

[72] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. N. Gorban, K. Murphy and L. Fei-Fei, "Detecting events and key actors in multi-person videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016 .

[73] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017 .

[74] A. Bewley, Z. Ge, L. Ott, F. Ramos and B. Upcroft, "Simple online and realtime tracking," in *IEEE International Conference on Image Processing (ICIP)*, 2016.

[75] V. Ramanathan, J. Huang, S. Abu-El-Haija, K. M. A. Gorban and L. Fei-Fei, "Detecting Events and Key Actors in Multi-person Videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[76] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proceedings of The 32nd International Conference on Machine Learning*, 2015.

[77] X. Gu, X. Xue and F. Wang, "Fine-Grained Action Recognition on a Novel Basketball Dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[78] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao and L. Wang, "Learning to Navigate for Fine-grained Classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[79] L. Wu, Z. Yang, Q. Wang, M. Jian, B. Zhao, J. Yan and C. W. Chen, "Fusing motion patterns and key visual information for semantic event recognition in basketball videos," *Neurocomputing,* 2020.

[80] H. Chen, H. Chen and S. Lee, "Physics-Based Ball Tracking in Volleyball Videos with its Applications to Set Type Recognition and Action Detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP*, 2007.

[81] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat and G. Mori, "A Hierarchical Deep Temporal Model for Group Activity Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),*, 2016.

[82] M. S. Ibrahim, A. E. Abdelaal, M. Lu and H. Wu, "Improved Hierarchical Deep Temporal Model forGroup Activity Recognition," [Online]. Available: http://www.ece.ubc.ca/~aabdelaal/index.html/Improved%20Hierarchical%20Deep%20Temporal%20Model%20for%20Group%20Activity%20Recognition.pdf.

[83] T. Kautz, B. Groh, J. Hannink and e. al., "Activity recognition in beach volleyball using a Deep Convolutional Neural Network," *Data Mining and Knowledge Discovery,* 2017.

[84] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua and S. Savarese, "Social Scene Understanding: End-to-End Multi-person Action Localization and Collective Activity Recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, 2017.

[85] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna., "Rethinking the inception architecture for computer vision.," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[86] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[87] Y. Yang, "Decomposition and recognition of playing volleyball action based on SVM algorithm,," *Journal of Interdisciplinary Mathematics, 21:5, 1181-1186,* 2018.

[88] L. Y., H. S., C. X. and I. T., "3D Global Trajectory and Multi-view Local Motion Combined Player Action Recognition in Volleyball Analysis," *Advances in Multimedia Information Processing – PCM 2018, Lecture Notes in Computer Science, vol 11166. Springer,* 2018.

[89] "3D Global and Multi-View Local Features Combination Based Qualitative Action Recognition for Volleyball Game Analysis," [Online].

[90] F. Haider, F. Salim, D. Postma, R. Van Delden, D. Reidsma, B.-J. BEIJNUM and S. Luz, "A Super-Bagging Method for Volleyball Action Recognition Using Wearable Sensors," *Multimodal Technologies and Interaction,* 2020.

[91] M. Qi, Y. Wang, J. Qin, J. L. A. Li and L. V. Gool., "stagNet: An Attentive Semantic RNN for Group Activity and Individual Action Recognition," *IEEE Transactions on Circuits and Systems for Video Technology, vol. 30,* 2020.

[92] T.-b. a. r. :. c. h. p. movement, "Template-based action recognition : classifying hockey players' movement," Master's Thesis at The University of Calgary, 2002.

[93] W.-L. Lu and J. J. Little, "Simultaneous Tracking and Action Recognition using the PCA-HOG Descriptor," in *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*, 2006.

[94] M. Fani, H. Neher, D. A. Clausi, A. Wong and J. Zelek, "Hockey Action Recognition via Integrated Stacked Hourglass Network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

[95] M. R. Tora, J. Chen and J. J. Little, "Classification of Puck Possession Events in Ice Hockey," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[96] S. Mukherjee, R. Saini, P. Kumar, P. P. Roy, D. P. Dogra and B.-G. Kim, "Fight Detection in Hockey Videos using Deep Network," *Journal of Multimedia Information System,* 2017.

[97] K. Sozykin, S. Protasov, A. Khan, R. Hussai and J. Lee, "Multi-label Class-imbalanced Action Recognition in Hockey Videos via 3D Convolutional Neural Networks," in *IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence*, 2018.

[98] H. Neher, "Hockey Pose Estimation and ActionRecognition using ConvolutionalNeural Networks to Ice Hockey," Master's Thesis at the University of Waterloo, Waterloo, Ontario, Canada, 2018.

[99] H. N. Zixi Cai, K. Vats, D. A. Clausi and J. Zelek, "Temporal Hockey Action Recognition via Pose and Optical Flows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[100] K. Vats, H. Neher, D. A. Clausi and J. Zelek, "Two-Stream Action Recognition in Ice Hockey using Player Pose Sequences and Optical Flows," in *16th Conference on Computer and Robot Vision (CRV)*, 2019.

[101] K. Rangasamy, M. A. As'ari, N. A. Rahmad and N. F. Ghazali, "Hockey activity recognition using pre-trained deep learning model," in *ICT Express Volume 6, Issue 3*, 2020.

[102] M. Ivasic-Kos and M. Pobar, "Building a labeled dataset for recognition of handball actions using mask R-CNN and STIPS," in *7th European Workshop on Visual Information Processing (EUVIP)*, 2018.

[103] M. Pobar and M. Ivasic-Kos, "Mask R-CNN and Optical Flow Based Method for Detection and Marking of Handball Actions," in *11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2018.

[104] A. Elaoud, W. Barhoumi, E. Zagrouba and B. Agrebi, "Skeleton-based comparison of throwing motion for handball players," *Journal of Ambient Intelligence and Humanized Computing volume 11,* 2020.

[105] C. Direkoğlu and N. E. O'Connor, "Temporal segmentation and recognition of team activities in sports," *Machine Vision and Applications volume 29,* 2018.

[106] A. Piergiovanni and M. S. Ryoo, "Fine-grained Activity Recognition in Baseball Videos," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[107] G. A., K. A. and S. B. M., "Discovering Cricket Stroke Classes in Trimmed Telecast Videos," in *Computer Vision and Image Processing CVIP 2019, Communications in Computer and Information Science*, 2019.

[108] A. Y. R., "Action Recognition in Sports Video Considering Location Information," in *Frontiers of Computer Vision, IW-FCV 2020, Communications in Computer and Information*, 2020.

[109] S. C. e. al., "Play type recognition in real-world football video," in *IEEE Winter Conference on Applications of Computer Vision*, 2014.

[110] B. Siddiquie, Y. Yacoob and L. S. Davis, "Recognizing Plays in American Football Videos," 2009.