How AI-based Computer Vision Algorithms Are Impacting The Sports Industry: A Survey

Discover Artificial Intelligence

REVIEW

Open Access

How AI-based computer vision algorithms are impacting the sports industry: a survey

Luca Francesco Rossi[1,2*], Andrea Sanna[1†], Federico Manuri[1†] and Mattia Donna Bianco[2†]

[†]Andrea Sanna, Federico Manuri and Mattia Donna Bianco have contributed equally to this work.

*Correspondence:
Luca Francesco Rossi
lucafrancesco.rossi@polito.it
[1]Department of Control and Computer Engineering (DAUIN), Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, TO, Italy
[2]netventure R&D S.r.l., Via della Consolata 1/bis, 10122 Torino, TO, Italy

Abstract

The sports industry, a multibillion-dollar global enterprise, engages audiences and inspires athletes with its dynamic, fast-paced, and action-packed nature. However, this very dynamism poses challenges for analysis and performance optimization. In recent years, computer vision has emerged as a transformative technology with the potential to reshape the landscape of sports. This field, which empowers computers to interpret and process visual data from the real world, offers numerous opportunities for innovation across the industry. Indeed, by leveraging computer vision, researchers and developers have created tools and systems that enhance the analysis of the complex dynamics in sports, with possible applications including tracking player movements with precision, analyzing game footage to identify patterns and strategies, and providing real-time insights that enhance decision-making for coaches and athletes. Beyond performance improvement, computer vision technologies also elevate the fan experience by offering immersive and interactive features, such as augmented replays and detailed statistical visualizations. As the adoption of computer vision continues to grow, its impact extends across all levels of sports, from grassroots programs to elite competitions. This technology not only provides a competitive edge but also redefines how sports are analyzed, experienced, and understood, marking a significant evolution in the industry.

Keywords Artificial intelligence, Computer vision, Pattern recognition, Scene understanding, Sports

1 Introduction

Computer vision (CV) is revolutionizing the world of sports, transforming the way games are analyzed, played, and broadcast [245]. By leveraging advanced techniques to track players, balls, and camera movements, CV algorithms provide valuable insights that enhance team performance, referee accuracy, and fan engagement. CV plays a pivotal role across the sports pipeline, from training and coaching to live broadcasting and post-event analysis. By addressing the complexity of high-speed athletic activity, it enables novel opportunities for innovation and performance enhancement. Yet, its declination in such a field is inherently challenging: a primary challenge in applying CV to sports lies in the dynamic nature of athletic activities, which introduce complications

such as identity switching among players, posing variations, and severe occlusions. Additionally, the accuracy and reliability of single- or multi-player tracking in real-time sports video can be affected by factors such as insufficient resolutions, variable illuminations, or lighting effects. Furthermore, the recognition of player actions can be complicated by similarities among players, blurry video segments, and partial or full occlusions. The detection and tracking of the ball in various sports also poses a significant challenge due to its size, speed, and unstructured motion compared to the players and playfield. Similarly, camera calibration and viewpoints, such as close-up views, far views, and wide views, can also impact the accuracy of computer vision applications in sports at an higher level [178]. Nevertheless, given the economic scale of the sports industry, there is a growing demand for innovative solutions that can enhance the analysis, training, and broadcasting of sports events. This growing interest is expected to lead to the development of new commercial products and services that can transform the sports industry and, as a result, researchers and developers are increasingly exploring the use of computer vision in sports, driving advancements in areas such as player tracking, event detection, and video analysis.

Numerous surveys have addressed the field of computer vision in sports, covering various and diverse applications due to the inherently heterogeneous nature of its possible declinations in the sports industry. Bonidia et al. [26] conducted a systematic literature review (SLR) on data mining in sports, discussing current research topics, datasets, algorithms and opportunities. Rahmad et al. [205] proposed instead a survey on video-based sports intelligence systems used to recognize sports actions. van der Kruk and Reijne [126] presented a review summarizing multiple human motion capture systems, with the goal of helping researchers to select the most suitable setup for their experiments. Manafifard et al. [158] illustrated a survey on state-of-the-art player tracking algorithms in soccer videos, evaluating pros and cons of each one of them. Cust et al. [55] delivered a SLR on specific methods for sports-related movement recognition via inertial measurements and computer vision data. Kamble et al. [119] presented an exhaustive survey concerning ball-tracking techniques, reviewing their advantages and limitations. Shih [233] addressed the content analysis fundamental (e.g. sports genre classification) by analyzing related state-of-the-art techniques in literature. Beal et al. [18] focused on team sports, providing a review of the inherent challenges arising in match outcome prediction, player investments or tactical decision making. Apostolidis et al. [6] suggested a taxonomy concerning the existing video summarization algorithms, while providing as well a SLR of existing methods. Adesida et al. [2] surveyed the usage of wearable technology in sports to avoid injuries and improve performances. Sarlis and Tjortjis [221] presented a comprehensive study on the application of data mining and machine learning techniques to basketball analytics for players and team performance. Rana and Mittal [206] thoroughly reviewed the literature on the use of wereable devices for performance monitoring in multiple sports. Host and Ivašić-Kos [105] reviewed current human action recognition techniques in sports primarily based on computer vision, along with listing popular publicly available datasets. Naik et al. [178] proposed an extended review of computer vision in sports, surveying different aspects of sports video analysis, ranging from detection to tracking and event classification. Mendes-Neves et al. [167] proposed a thorough review of multiple advanced algorithms for sports analysis, with special emphasis on detection and 2D/3D pose estimation. Papageorgiou et al.

[193] focused instead on forecasting, presenting a comparative study evaluating multiple machine learning models for predicting future basketball player performance.

The rapid advancement of scientific knowledge necessitates continuous updating of state-of-the-art research to ensure relevance and accuracy. In AI-based fields, breakthroughs and new methodologies emerge at an accelerated pace, making previously established reviews quickly outdated. Moreover, a more fine-grained and revised categorization addresses the limitations of broad, generalized summaries that may overlook critical details or emerging subfields. Therefore, despite the growing body of literature addressing computer vision applications in sports, existing surveys often focus narrowly on specific subtopics, lacking a comprehensive integration of a complete understanding within a unified framework. This has resulted in a fragmented understanding of current capabilities, limitations, and emerging opportunities. Our survey aims to bridge this gap by presenting a cohesive and up-to-date synthesis of recent advancements in AI-based computer vision techniques, organized around two complementary dimensions: *physical* understanding (i.e., who is where) and *logical* understanding (i.e., what is happening). These terms are introduced as high-level abstractions to group related technical subdomains in a coherent and interpretable manner. The term "*physical*" highlights the geometric and spatial nature of the information extracted, grounded in pixel coordinates, field dimensions, and physical presence within the scene. It lays the foundational structure upon which further semantic interpretation can be built. Instead, the term "*logical*" underscores the move from raw spatial data to higher-order semantics, where temporal sequences and contextual reasoning are required to interpret complex interactions and outcomes. Such a hierarchical structure is qualitatively depicted in Figs. 1 and 2. This structure provides clarity to researchers and practitioners navigating a complex and evolving landscape, while also highlighting unresolved challenges and identifying promising directions for future research. This survey aims to address these limitations by offering a synthesized overview of recent advancements in AI-driven computer vision algorithms applied to the sports domain, equipping researchers with an updated and structured understanding of the field. Therefore, the research questions selected for the proposed survey are the following:

RQ1. What are the main fields of research?

RQ2. What are the limitations of current approaches?

RQ3. What are the most promising trends?

The rest of the paper is organized as follows. Section 2 tackles the *physical* understanding of the scene, that is, the comprehension of who is where, thus providing a structural representation to it. Section 3 deals instead with the *logical* understanding of the scene, that is, the perception and awareness of what is happening, bridging the gap between entities and actions. Lastly, Sects. 4 and 5 conclude by describing the final remarks and observations.

2 Where are the actors?

The *physical* understanding provides a first-level comprehension of a scene, shedding light on which actors are present and where they are located. Section 2.1 deals with camera calibration and camera motion understanding, surveying methods that try to provide a spatial dimension via a 2D-to-3D mapping between image coordinates and real-world ones. Indeed, camera calibration—a foundational process in both machine vision and

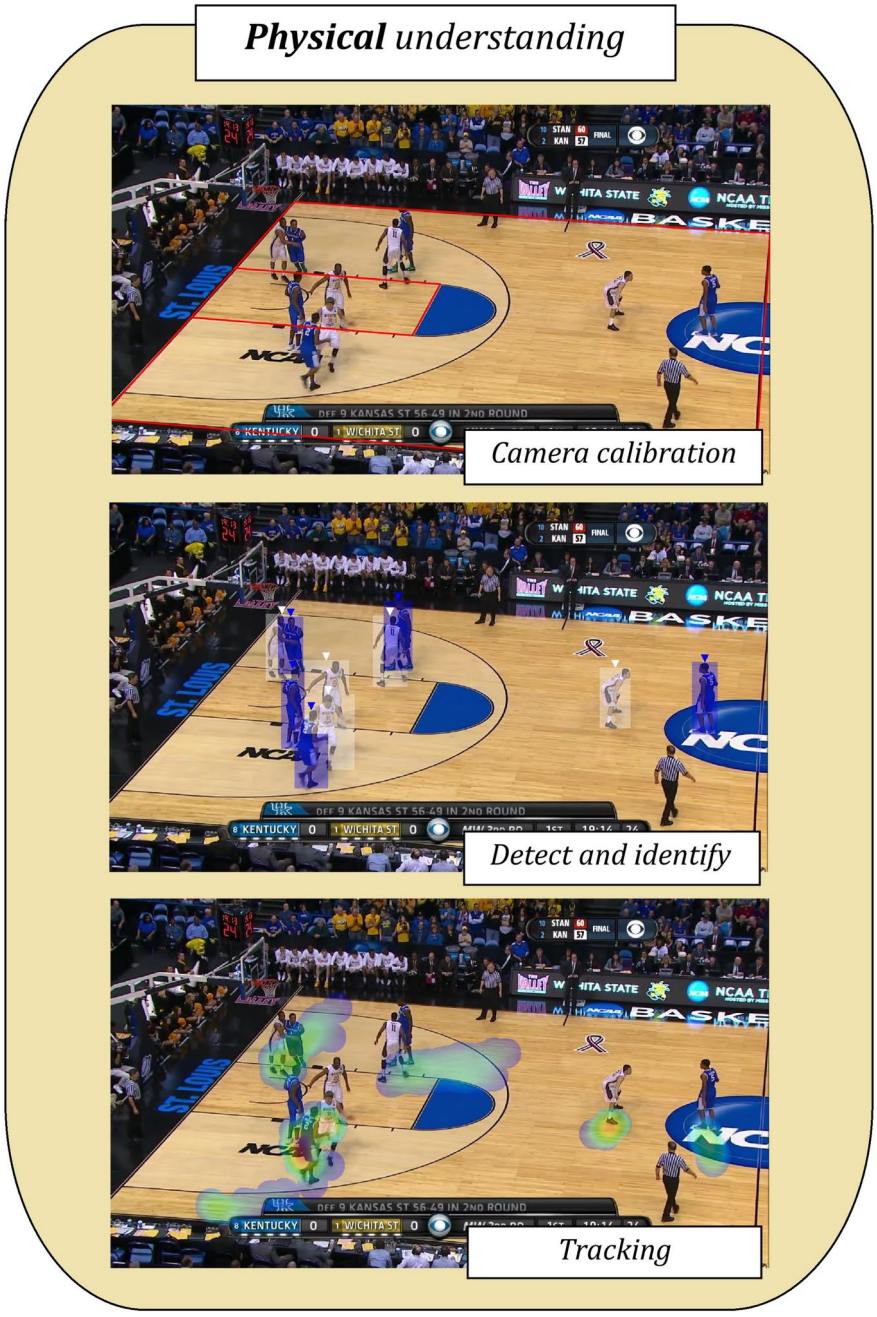

Fig. 1 Proposed physical understanding summary visualization. Image frame taken from the SportsHHI dataset by Wu et al. [273], licensed under Creative Commons Attribution-NonCommercial 4.0 International License

photogrammetry—is essential for quantifying the relationship between real-world scenes and their corresponding image representations, thereby allowing the extraction of geometric information from visual data, supporting accurate spatial interpretation.

Object detection and tracking are instead tasks that involve identifying and following objects in a video sequence, where the former aims to locate and classify objects within a single frame, while the latter aims to link detections across frames to form trajectories of objects over time. Sections 2.2 and 2.3 deal therefore with detecting, identifying and tracking desired entities such as players, teams, balls, and structural elements

Fig. 2 Proposed logical understanding summary visualization. Image frame taken from the SportsHHI dataset by Wu et al. [273], licensed under Creative Commons Attribution-NonCommercial 4.0 International License

within the sports domain, reviewing the most-recently proposed computer vision-based techniques.

2.1 Camera calibration

Camera calibration methods in sports can be meaningfully clustered based on their underlying strategy for estimating spatial transformations between image frames and real-world coordinates. By examining the methodological foundations, we identify two primary categories: (i) template-matching techniques, which leverage pre-generated

Table 1 Summary overview of the considered articles concerning camera calibration

Article	Camera	Dataset	Proposal	Advantage	Drawback
Chen and Little [41]	Single	Homayounfar et al. [101] Ibrahim et al. [112]	Template matching	Simple and interpretable	Requires human annotations
Sha et al. [228]	Single	Homayounfar et al. [101] Sha et al. [228]*	Template matching	Robust to noise and occlusions	Pose initialization is not differentiable
Cioppa et al. [49]	Single	Homayounfar et al. [101]	Template matching	First large-scale benchmark	Memory footprint and unexplainability
Zhang and Izquierdo [284]	Single	Homayounfar et al. [101]	Template matching	Robustness under shadow condition	More computationally expensive
Takahashi et al. [241]	Multi	Takahashi et al. [241]*	Geometric refinement	Cameras sync. not required	Performances dependent on early-stage detection
Citraro et al. [53]	Single	Homayounfar et al. [101] Citraro et al. [53]*	Geometric refinement	Near real-time approach	Lack of temporal consistency
Chu et al. [44]	Single	Chu et al. [44]*	Geometric refinement	Less ambiguity among keypoints	More computationally expensive
Baumgartner and Klatt [15]	Single	Baumgartner and Klatt [15]*	Geometric refinement	Readily adaptable to other sports	Advocates for domain-specificity
Zhang and Izquierdo [285]	Single	Homayounfar et al. [101] Zhang and Izquierdo [285]*	Geometric refinement	Near real-time approach	Performances dependent on early-stage segmentation
Theiner and Ewerth [244]	Single	Homayounfar et al. [101] Cioppa et al. [50]	Geometric refinement	One-step procedure	Rare local minima instability
Gutiérrez-Pérez and Agudo [91]	Single	Chu et al. [44] Cioppa et al. [50]	Geometric refinement	Extendable to multiple-view	Minimum number of geometries required

*Dataset privately collected by the authors themselves

Table 2 Emoji look-up legend for sports

	Athletics[1]		Cricket		Hockey		Swim/diving
	Badminton		Curling		Ice skating		Soccer
	Baseball		Darts		Karate		Table tennis
	Basketball		Fencing		Rehabilitation[2]		Tennis
	Boxing		Football		Rugby		Volleyball
	Canoeing		Golf		Skiing		Yoga/pilates

[1]Group of sporting events that involves competitive running, jumping and throwing

[2]Not a sport per se, it comprises rehabilitation purposes in sport scenarios

templates or feature libraries for matching and registration; and (ii) geometry-based methods, which rely on explicit detection of field markings or spatial correspondences as a proxy for geometric calibration. A summary of the considered works is presented in Table 1. Table 2 provides the correspondence of the emoji used to identify each sport in all the paper's tables.

2.1.1 Template matching

Chen and Little [41] proposed a highly automatic method for calibrating sport cameras from a single image with the help of synthetic data. The method first detects field markings from the input image, then retrieves an initial camera pose from a database of virtually-sampled configurations, and finally refines the camera pose using distance images, requiring minimal human intervention. It uses a feature-pose database built

using a novel camera pose engine, where the features are learned from a trained siamese network [93]. In order to extract field markings, the method uses a double Generative Adversarial Network (GAN) model, which consists of a segmentation GAN and a detection GAN: the segmentation GAN separates the foreground (e.g. grassland) from the background, and the detection GAN detects such field markings from the foreground image. Experimental results on both the 2014 Soccer World Cup dataset collected by Homayounfar et al. [101] and a revisited version of the volleyball action recognition dataset by Ibrahim et al. [112], showed that the proposed method achieves high accuracy in camera pose estimation, outperforming previous state-of-the-art methods [230].

Sha et al. [228] proposed an end-to-end camera calibration method for sport images by finding a homography that can register a target ground-plane surface of any frame from a broadcast video with a top-view field model. The method leverages three major techniques: semantic segmentation, camera pose initialization, and homography refinement. It uses a single network architecture that can be trained end-to-end, taking a single input image and calculating the homography. The network first performs semantic segmentation to extract key features via a UNet [212] style autoencoder, then uses the semantic map to select an appropriate template [288] and predict the relative homography, which is further refined via a transformer architecture. The network has four distinct loss functions and can be trained end-to-end since all losses are fully differentiable with respect to the network parameters. On a custom dataset of both basketball and soccer broadcast videos, the proposed framework achieves better results than the work by Chen and Little [41], considered as baseline, demonstrating the effectiveness of the end-to-end training approach and the template-based search space, allowing for faster and more accurate camera calibration (Table 2).

Cioppa et al. [49] discussed a method for camera calibration based on deep representations with the goal of exploiting them in a later step for the action spotting task of SoccerNet-v2 [86]. The proposed calibration method is based on the Camera Calibration for Broadcast Videos (CCBV) algorithm of Sha et al. [228], with some modifications. The method uses a dictionary of pairs of artificial field zone segmentations (templates) and homographies, which is built in a pre-processing step using a clustering algorithm based on Gaussian Mixture Models [64]. The dictionary associates an image projection of a synthetic reference field model to a homography used to produce such a projection. Each input image is first transformed to resemble a projection of the synthetic field, typically by a semantic segmentation of the field lines or areas defined by the field lines, which is then associated with its closest synthetic view in the dictionary, giving a rough estimate of the camera parameters, eventually refined to produce the final prediction. The model is trained as student via knowledge distillation considering a commercial tool as teacher, with the calibration information being finally used to improve the action spotting performance by concatenating the calibration features with the frame feature vectors extracted from the calibration representations extracted through a Context-Aware Loss Function (CALF) architecture [48]. The results show that the proposed method achieves a novel state-of-the-art performance on the SoccerNet-v2 action spotting leaderboard, outperforming other methods by a comfortable margin, thus confirming the validity of the proposed calibration procedure.

Zhang and Izquierdo [284] continued along a similar path, proposing an improved method aimed at real-time capabilities for sport videos. Their approach consists of three

main parts: semantic segmentation and feature extraction, feature database and search, and homography refinement. The approach uses a conditional GAN [113] for semantic segmentation and a feature extraction network inspired by VGG-16 and VGG-19. The feature extraction network is trained using a dataset of segmented images generated by warping a template with ground truth homographies. During testing, a video frame is input into the trained cGAN to obtain its semantic segmented image, and then the feature extraction network is used to extract features. The features are then searched in a feature database to find the best matching homography, which is further refined through image alignment.

2.1.2 Geometry-based methods

Takahashi et al. [241] decided to follow a different procedure for camera calibration, that is, by using human poses as calibration pattern. The proposed method for estimating 3D human pose uses unsynchronized and uncalibrated multiple cameras with wide baselines. It first detects 2D joint positions from multi-view videos and estimates initial values of each parameter in a Structure from Motion (SfM) manner [249] with assumed time shift. Then, it optimizes the initial values in terms of relaxed reprojection error and geometric constraints based on the a priori knowledge that the reference points are actually human joints. The method avoids the problem of detection errors by relaxing the reprojection errors, which is done by using a confidence map to weaken the influence of strict reprojections, selecting the output that yields the smallest values in terms of error. The authors showcased their approach on two custom-built datasets – a synthesized dataset on soccer players and a real-world application with recorded baseball players – illustrating also through the visualization of estimated camera positions and 3D joint positions in 3D space that the proposed method is more robust to noise-degraded data.

Citraro et al. [53] addressed the high cost and latency of current state-of-the-art calibration approaches, looking for a fast, robust, and generic framework for detecting keypoints in images of sports fields. Their idea leverages the fact that keypoints on a plane do not overlap to reduce inference time and detects a high number of interest points. The method uses a U-Net [212] architecture to detect semantic keypoints, such as lines and corners on the ground, and player keypoints, which are the projections of the players' center of mass on the ground. It then uses homography estimation, refinement, and intrinsics estimation [95] to compute the focal length and extrinsic camera parameters for each image in the sequence. The method also exploits the position of the players to increase robustness and accuracy in images lacking visible features. Additionally, it uses temporal consistency and spatial dropout [246] to improve the results. On a set of privately-built multi-sports datasets, this strategy achieves 20–30 frames per second and outperforms state-of-the-art methods on multi-sport custom datasets.

Chu et al. [44] tried instead to change perspective regarding sports field registration, drifting away from typical procedures, which estimate the homography via field-specific features extraction (e.g. corners, lines, etc.) from image. The proposed method consists of a standard encoder-decoder architecture and a keypoints-aware label condition module. The encoder takes a field image as input and extracts feature maps, which are then fed to the decoder. A keypoints-aware label condition module via dynamic filter learning [115] generates parameters for a dynamic head using the extracted features

and keypoints encoding vectors, providing then heatmaps, which are merged using soft aggregation [188] to produce the final output. The predicted homography is at last estimated from the merged heatmap using both DLT [1], i.e. Direct Linear Transformation, and RANSAC [95]. The results show that the proposed instance-based keypoints detection architecture achieves better accuracy and recall rate than other baseline methods [41, 49, 228, 230] on the TS-WorldCup dataset—an extended version of the original World Cup dataset [101]—with or without fine-tuning, and achieves the best performance in estimating soccer field registration.

Similarly, Baumgartner and Klatt [15] investigated the effects of camera calibration and partial sports field registration concerning monocular 3D Human Pose Estimation (HPE) for sports broadcasts. While current method exceed in providing 2D joints locations, 3D kinematics extraction requires knowledge about sequences of locations and displacements. The authors decided therefore to address field registration (and camera calibration) in order to learn about the geometry in a scene, with the goal of leveraging such information for 2D-to-3D projections. The proposed calibration method involves computing a collection of camera calibrations with 1-degree difference in azimuth, allowing for continuous selection of camera calibration [16]. For each frame, a virtual camera is picked that is consistent with the visible image. The method uses a lookup-grid of vanishing points to determine the visible track lanes and their vanishing point, and then creates a virtual camera for each azimuth in the lookup-grid. The elevation and roll of the camera are adjusted to result in a vanishing points in the correct direction, and the field-of-view is adapted to alter the distance of the vanishing point to the scene. The camera is then placed in a simulated scene and shifted to match the virtual track markings with the visible lanes, resulting in a translation that describes the camera's position. Finally, the overall projection matrix is computed by combining both extrinsic and intrinsic camera parameters. The framework is experimentally compared to state-of-the-art monocular HPE methods [218] on a proprietary virtually-built dataset and demonstrates improved performance in terms of re-projection error, 3D error, and knee angle error, showcasing the robustness of the calibration procedure as intermediate step.

Zhang and Izquierdo [285] further extended their previous work by refining it in a four-step strategy: firstly, a cGAN is adopted to generate semantically segmented video frames of the field. Then, a regression network is leveraged to estimate four points' coordinates from a single segmented frame. Furthermore, the DLT algorithm is used to estimated the homography matching the retrieved four points and the ones selected as control points on the template, thus imposing the 8 DoFs encoded by the 3×3 transformation matrix. Finally, an Enhanced Correlation Coefficient (ECC) technique [74] is performed with the goal of refining the predicted homography by correcting distortions. Experimental results on the 2014 Soccer World Cup dataset [101], the National Basketball dataset and the 2018 Soccer World Cup dataset highlight that the authors' approach is superior both in terms of accuracy and computational efficiency with respect to other competitive methods [41, 53, 117, 228, 230], achieving a satisfactory inference speed per image.

Theiner and Ewerth [244] focused instead on performances over latency, by proposing a state-of-the-art method for sports field registration. The proposed method, TVCalib, is a calibration approach that iteratively minimizes the segment reprojection loss to achieve accurate 2D sports field registration. It consists of several components, including

semantic segmentation, point selection, a calibration module, and result verification. The method takes into account different segment types, such as lines and point clouds, and uses a gradient-based iterative optimization approach to minimize the reprojection loss. The optimization process involves computing distances between reprojected segments and input pixels, and aggregating these distances to obtain the overall loss. The method also accounts for lens distortion correction and allows for parallelization and efficient computation using tensor operations. Results on both the SN-Calib dataset Cioppa et al. [50] and the World Cup dataset [101] show that TVCalib achieves superior performance in camera calibration and homography estimation tasks, with higher accuracy and completeness ratio compared to other methods [41, 44, 117, 228, 231]. The results also indicate that the homography decomposition introduces additional errors, particularly for goal segments, with homography estimation error being lower when ignoring goal posts and crossbars, suggesting that the decomposition is a major source of error. Instead, the use of multiple initializations and the selection of the best result is shown to improve the performance of the optimization process. Similarly, learning camera and radial lens distortion parameters jointly can improve results for samples with visible radial lens distortion.

Gutiérrez-Pérez and Agudo [91] pushed the task even further, looking for a minimalist pipeline solely based on 2D-3D correspondences, focusing on retrieving both extrinsic and intrinsic camera parameters from individual frames of sports TV broadcasts without any prior information about the camera position or orientation. The method consists of four processing components: soccer field modeling and keypoint generation, keypoint and line detection, DLT algorithm [1], and camera parameters retrieval. The approach is strongly based on pitch geometries, relying on the lines painted on the soccer field, their intersections, and the corners they define, due to their known position on the world coordinate system. A hierarchical structure is established for computing each set of keypoints, and an HRNetv2 encoder-decoder network [263] is used to produce heatmaps for keypoints and extremities of soccer field lines to extract their positions in the image space, with these keypoints being organized into subgroups based on the specific geometric features they represent, such as line-line intersections, extended line-line intersections, line-ellipse intersections, and ellipse tangent points. The authors experimentally validated their approach on several datasets—SN-Calib dataset [50], World Cup dataset [101], TS-World Cup dataset [44]—showing that the fine-tuned model outperforms other methods [41, 44, 53, 183, 244, 284, 285] in several metrics. The homography estimation evaluation reveals that the integration of keypoint sets derived from line-ellipse intersections and ellipse tangents enhances accuracy, particularly along the midfield line, and increases the completeness rate. The inclusion of ellipse tangents specifically augments the completeness rate by increasing the keypoint density within the field circles, yet decreasing its performance concerning accuracy metrics.

2.2 Detection and identification

To provide a more coherent understanding of recent advancements in detection and identification within sports settings, we organize the reviewed works into five thematic clusters based on their methodological focus and application scope: (i) multi-modal fusion, highlighting models performing inference by exploiting more than a single data modality; (ii) skeleton-driven modeling, leveraging keypoint-based annotations to

Table 3 Summary overview of the considered articles concerning detection and identification

Article	Target	Dataset	Proposal	Advantage	Drawback
Huda et al. [108]	Player|DET	Huda et al. [108]*	Multi-modal fusion	Readily adaptable to other sports	Custom set-up
Cioppa et al. [47]	Player|DET	Cioppa et al. [47]*	Multi-modal fusion	Near real-time	Custom set-up
Pandya et al. [192]	Player|DET	Pandya et al. [192]*	Multi-modal fusion	Does not require annotations	Custom st-up
Senocak et al. [227]	Player|ID	Senocak et al. [227]*	Pose-based modeling	Single player recognition	Ad hoc training required
Li and Bhanu [133]	Dribbling style|DET	Li and Bhanu [133]*	Pose-based modeling	Temporal evolution	More computationally expensive
Liu and Bhanu [138]	Player|ID	Liu and Bhanu [138]*	Pose-based modeling	Pose-guided supervision	Limitations under extreme poses
Bright et al. [30]	Player|DET	Bright et al. [31]	Pose-based modeling	Decoupling action from tracklets	More computationally expensive
Renò et al. [210]	Ball|DET	Renò et al. [210]*	Deep features	No a priori knowledge	Classifier required
Istasse et al. [114]	Team|ID	Istasse et al. [114]*	Deep features	No a priori knowledge	Less explainable
von Braun et al. [27]	Waterline|DET	von Braun et al. [27]*	Deep features	Kinematics derivation	Requires human annotations

*Dataset privately collected by the authors themselves

Table 4 Summary overview of the considered articles concerning detection and identification *(ctd.)*

Article	Target	Dataset	Proposal	Pros	Cons
Koshkina et al. [124]	Team|ID	Koshkina et al. [124]*	Deep features	No a priori knowledge	Less explainable
Koshkina and Elder [123]	Player|ID	Cioppa et al. [50] Koshkina and Elder [123]*	Deep features	General framework	More computationally expensive
Cioppa et al. [46]	Player|DET	Cioppa et al. [46]*	Data-efficient learning	Model and data agnostic	Vanilla dataset update strategy
Vandeghen et al. [253]	Player|DET Ball|DET	Cioppa et al. [50]	Data-efficient learning	Semi-supervised training	Heavily dependent on teacher model
Liu et al. [144]	Object pairing|DET	Liu et al. [144]*	Data-efficient learning	Detection and matching	Application specific
Vats et al. [259]	Player|ID	Vats et al. [257]	Data-efficient learning	Incorporates player shifts	Requires approximate labels

*Dataset privately collected by the authors themselves

improve performances; (iii) deep feature embeddings, emphasizing end-to-end learned visual descriptors and reID embeddings; and (iv) data-efficient learning, integrating tools such as knowledge distillation or domain adaptation to overcome the specificity requirements of sport-related labeled datasets. A summary of the considered works is presented in Tables 3 and 4.

2.2.1 Multi-modal fusion

Huda et al. [108] approached the task of detecting and estimating the number of soccer players for occupancy analysis by distinguishing between occluded and non-occluded blobs in images rendered via thermal cameras. The proposed method is a three-staged supervised player detection and counting system, consisting of candidate player

detection, occlusion detection, and estimation of the number of players in each occluded group. The system uses a detailed feature vector for each candidate region, formed by using the shape and geometry of the blobs (comprising connected point slope, distance, convex hull and bounding box length), and a bagged tree classifier [28] for detection of occlusions. To further classify the number of players in occlusions, a simulation-based method is suggested, following maximum likelihood density estimation with respect to learned blob sizes created in virtual environment. The proposed method, tested by the authors on broadcast non-commercial videos, showed that retrieving features via bagged tree classification models performed better than the state-of-the-art human detection histogram of oriented gradients (HOG) features [57].

Cioppa et al. [47] proposed a robust and cost-effective method for player detection and counting in a football field using a combination of data augmentation and motion detection algorithm. The method leverages detections made on a thermal image of a part of the field to detect all the players on the whole field on a fisheye image, similarly to what has been proposed by Huda et al. [109]. The data augmentation process creates artificial players with known bounding boxes in the region filmed by one camera, while the motion detection algorithm identifies areas that are guaranteed player-free, providing true negative areas where the network will be penalized for predicting player bounding boxes. The fast YOLOv3-tiny is trained using an online distillation process [143], allowing it to continuously adapt to changing weather and lighting conditions. The authors showcased the overall promising capabilities of their framework on some proprietary recordings, with the student model becoming more consistent as it trains, yet highlighting its tendency to slightly overestimate the actual number of players (that is, false positives).

Pandya et al. [192] investigated whether a field registration homography might prove beneficial for player detection in live sport broadcasts. The proposed method aims to automatically identify players on-screen in American football videos by leveraging Next Gen Stats (NGS) data, which provides key players' positions, speed, and acceleration information. The method involves computing a homography between the NGS plane and the video frame to register the field. Two approaches are proposed to propagate the homography across multiple frames: one uses player tracking information generated from RFID tags, and the other uses keypoints detected on the ground plane of the video frame. The method is robust in handling poor field landmark visibility but requires at least four players to be visible on the video frame to estimate the homography. The homography is computed for one key frame and then propagated across frames with low latency using player/key-point tracking techniques. In the actual experiments, keypoints are identified using SIFT [147] and matched between subsequent frames using FLANN [175] by looking for correspondences, showing better performances with respect to similar methods [183] in terms of homography recognition for field registration, and consequently in players identification.

2.2.2 Skeleton-driven modeling

Senocak et al. [227] tackled the challenge of automatically identifying players in broadcast sports videos captured from a single side-view medium distance camera. The proposed method for basketball player identification uses a combination of body and part representations to recognize players from the entire body in a broadcast camera view.

Taking inspiration from a previous work by Lu et al. [148] which adopted low-level hand-crafted features such as color histograms, MSER [160] and SIFT [147] to build visual representations, the authors' strategy leverages deep convolutional features at multi-scale to model the player representation, and exploits body parts to supplement coarse holistic body representation. The method first takes an input image at single scale, locates the body parts using Convolutional Pose Machines, extracts fixed-length activations from each part using a CNN, and aggregates all the activations into a single Fisher Vector. The final player representation is obtained by fusing the body and part representations in a late fusion manner, and is used for player identification. On a small proprietary dataset, the proposed system was able to identify players despite challenging conditions, such as varying lighting and pose. The results also show that the use of part representations in addition to body representations improved the accuracy of the system.

From an alternative perspective, Li and Bhanu [133] positioned themselves halfway with action recognition by deciding to investigate soccer dribbling techniques identification. The approach involves creating a single image representation, called a Dynamic Energy Image (DEI), from a video sequence of a dribbling action, which captures the spatio-temporal information of the action. More in detail, the framework consists of several modules, including dribbling player segmentation, pose detection, body parts, image registration, data augmentation, and dribbling styles classification. The method uses Mask R-CNN [97] to localize and segment players who are performing dribbling actions, and then extracts 2D pose information of each player using OpenPose [36]. The dribbling energy image (DEI) is then generated through an affine-transformation-based registration method, which encodes the spatial-temporal information of a video sequence into a single image [23]. Finally, the DEI is used as input to a CNN to classify the dribbling styles. The authors evaluated their framework on a proprietary custom dataset comprising tracklets of player performing three different styles of dribbling, showcasing how their approach outperforms other mainstream architectures, such as AlexNet [250], VGG-16 [234], and ResNet-18 [96], in terms of mean accuracy.

With the goal of tackling the identification task from a different perspective, Liu and Bhanu [138] discussed a framework for jersey number recognition in sports images consisting in two stages: the first stage uses a re-designed Region Proposal Network (RPN) that outputs candidate object bounding boxes across three classes: background, player, and digit. Person proposals and digit proposals are collected separately from a single RPN without adding many parameters. The second stage uses a modification of Faster R-CNN [209] that replaces ROIPool with RoIAlign and includes a human body key-point branch for predicting key-point masks. The classification and bounding-box regression are performed on pooled digit features concatenated with key-point masks. The framework improves localization performance of digits by associating person and digit Regions of Interest (RoI) and adding human pose supervision signal, allowing the model to target digits inside person proposals with the help of keypoint locations. The proposed pose-guided R-CNN model achieved better performance than the state-of-the-art models in jersey number recognition, outperforming Faster R-CNN in terms of number-level and digit-level accuracies, as well as mean average precision (mAP), AP50, and AP75, showcasing high robustness to pose variations.

Bright et al. [30] decided to focus specifically on baseball by introducing PitcherNet, an end-to-end automated system that analyzes pitcher kinematics from live broadcast videos capable of extracting valuable pitch statistics such as velocity, release point and extension, and pitch position. The framework consists of three components: pitcher tracking and identification, 3D human modeling, and pitch statistics. The system first detects and tracks players in a video, assigning unique labels to each tracklet, and then decouples the actions from the inferred poses of the players to facilitate player classification. The 3D human modeling component uses a 3D human model prior [146] to estimate the pose of the player, guided by masked modeling, distribution learning, and silhouette masks. Finally, the pitch statistics component leverages TCNs, i.e. Temporal Convolutional Networks [129], and kinematic-driven heuristics to capture various pitch metrics. The TCN architecture is designed to eliminate the dependence on specific player characteristics for classification, providing a more robust solution for identifying the target player in dynamic sports scenarios. The results show that the proposed model, D2A-HMR 2.0 (revisited version of the specific transformer architecture introduced by the same authors in Bright et al. [31], outperforms existing state-of-the-art Human Mesh Reconstruction (HMR) techniques in terms of 3D mesh alignment with the input image. The model demonstrates superior performance when incorporating both the player's vertices and 3D joints during the regression process. The use of additional pseudo-ground truth data from the transformer output tokens also improves the performance of recovering the 3D mesh. In terms of pitch statistics, the TCN model achieves perfect accuracy in classifying handedness and demonstrates impressive performance in classifying pitch position, release point, pitch velocity, and release extension.

2.2.3 Deep-features representation

Renò et al. [210] discussed instead the use of a convolutional neural network (CNN) for ball detection in tennis videos, specifically in a frame-by-frame basis without requiring any image preprocessing steps. Such a CNN-based classifier is used to decide whether an image patch can be labeled as "Ball" or "No Ball". The architecture is configured for RGB image patches and consists of four deep learning blocks of layers, followed by a classical neural network, a softmax, and a class output. The network is trained to preserve spatial relationships in the processed images by finding a large number of small filters that mimic the human vision system. The output of the CNN is used to compute a probability image, which gives a pixel-based evaluation of the ball probability, computed by taking the average probability value for the class "Ball" for each pixel, exploiting the output of the fully connected layer. This approach allows for a more detailed analysis of the image, including the identification of false positives and the visualization of probability values. On a dataset of privately acquired images, the retrieved true negative rate, accuracy, and balanced accuracy [235] indicated good performances, suggesting that the CNN can effectively classify images as "Ball" or "No Ball", even in cases where the ball is blurred or moving, thus resulting in a model that can correctly identify balls in new, unseen images.

Moving aside from human detection, von Braun et al. [27] decided to investigate waterline recognition in canoe sprint video, considering the waterline as a reference point for estimating kinematic parameters, with its exact position and orientation often being difficult to specify due to various factors such as turbulence, waves, and poor

image quality. The proposed method for determining a waterline in an RGB image from video sequences of canoe sprints involves a two-stage approach: the first stage uses image segmentation via Mask R-CNN [97] to separate the canoe hull from the water body, whereas the second stage employs an iterative procedure to derive a waterline from the binary segmentation mask obtained in the first stage. This procedure involves finding the contour of the canoe, identifying the bottom line of the contour, performing linear regression on the points in the bottom line, removing points above the regression line to account for small waves and splashes (or other disturbances), and finally performing another linear regression to predict the waterline. The results of the study showed a moderate and consistent segmentation quality for canoe segmentation on a custom-built dataset, highlighting promising results by assessing its accuracy against expert-based manual ground truths.

In order to tackle the task of team assignment, Istasse et al. [114] proposed a method that uses a variant of ICNet to both segment players and compute pixel-wise team discriminant embeddings. The ICNet variant is a type of Fully Convolutional Network (FCN) that outputs spatial feature maps, which are then used to separate the embeddings for distinct teams. The authors draw inspiration from the associative embedding framework outlined by Newell et al. [182], which distinguishes among object instances, yet diverging from the original work by deriving embeddings from a lightweight segmentation network and, more fundamentally, by assigning the same embedding to unconnected pixels representing different players from the same team. A clustering algorithm is then applied to group pixels into teams, using a simple and greedy method that relies on the assumption that a player pixel embedding, when surrounded by similar ones, is representative of its team embedding. ICNet learns this way a descriptor—a pixel-wise embedding vector—that is similar for pixels depicting players from the same team and dissimilar for pixels from different teams, eliminating this way the need for game-specific learning and enabling efficient team differentiation from the outset of the game. The proposed team discrimination method achieved solid results in correct team assignments on a custom dataset when testing on games and sport halls not seen during training, outperforming the baseline Bayesian color histogram clustering.

Similarly, Koshkina et al. [124] addressed the challenge of unsupervised classification of players in team sports based on their team affiliations, even in scenarios where jersey colors and designs are unknown beforehand. The proposed method involves a self-supervised training of an embedding network to learn feature representations of player images in a contrastive manner [93]. The process starts with clustering images from each game based on color features, generating initial input triplets, and training the embedding network using a triplet loss function. The network is then trained for several iterations, and the clustering is repeated based on the newly learned features, repeating this process until convergence. The trained network is then used to extract features from player images, which are then clustered into teams using k-means. The results show that the proposed unsupervised contrastive learning approach outperforms other unsupervised team labeling approaches, particularly when the number of frames available for learning cluster centers is small. The proposed approach also allows for visualization of team positioning over the course of a game or a portion of a game through the generation of heatmaps. These heatmaps are created by back-projecting [153] player positions onto a model of the playing surface using a learned homography and then applying

Gaussian kernel density estimation [196, 213], demostrating the effectiveness of the authors' approach in learning a more discriminative embedding space.

Following the idea of jersey number recognition in sports video, Koshkina and Elder [123] tried to suggest a general framework to improve tracklet-level classification. The proposed method involves a pipeline that detects, localizes, and recognizes jersey numbers in images by first identifying frames where the jersey number is visible and legible, and by using then a Scene Text Recognition (STR) system [17] to predict the number. Shifting from frame-level to tracklet-level recognition, the method filters out distractors and combines image predictions into a single tracklet-level prediction by also employing a Centroid-ReID [271] network to extract visual feature vectors, excluding outliers and combining predictions with the help of either heuristic or probabilistic consolidation from different frames, with options to bias towards two-digit numbers and threshold the sum of confidences. The ablation analysis performed on both a custom-made hockey dataset and the SoccerNet Jersey Number Recognition dataset demonstrates the effect of main subject filtering and different options for prediction consolidation, with the heuristic consolidation approach achieving the best results. The results are compared to previously published methods [13, 83, 131, 257, 259], showing the proposed pipeline's robustness and effectiveness.

2.2.4 Data-efficient learning

Semantic segmentation is a valuable tool for comprehending global scenes across various domains, including sports. However, it faces challenges like the dependency on pixel-level annotated training data and the absence of efficient real-time universal algorithms. To address these issues, Cioppa et al. [46] proposed a method, named ARTHuS, which consists in an adaptive real-time human segmentation network that evolves during a sports match without requiring manual annotation of a single frame. It achieves this through an online distillation process [143], where a slow but well-performing Mask R-CNN teacher [97] is used to train a fast TinyNet [45] student capable of real-time inference, making it match-specific. The student network is continuously updated and adjusted to the latest play conditions, allowing it to become more accurate and efficient over time. This approach tries to balance pros and cons of the two models by sacrificing some generalizability for performance and speed, resulting in a network that runs accurately and in real-time on the specific match it is analyzing. The results show that the adaptive network produced by ARTHuS outperforms pre-trained networks in human segmentation tasks, with experiments conducted on soccer and basketball videos highlighting that the networks produced by ARTHuS achieve satisfying results with respect to previous state-of-thae-art methods.

Similarly, Vandeghen et al. [253] suggested a teacher-student approach as a novel generic semi-supervised training method to improve both player and ball detection in soccer videos. More in detail, a teacher network is first trained on a labeled dataset in a fully supervised fashion, and then used to produce pseudo-labels on an unlabeled dataset. The pseudo-labeled dataset, along with the labeled dataset, is used to train a student network, whose training loss is parameterized based on the confidence score of the pseudo-labels. This allows the student to doubt unsure proposals by the teacher and achieve comparable performances on the test dataset. Subsequently, the student model is fine-tuned [134] on the labeled dataset and adopted as the new teacher in the

following training cycle. This process is repeated to improve the detection performance of the model. Experimental results on the SoccerNet-v3 dataset [50] show that the proposed method improves the detection performance compared to fully supervised methods, especially when considering few annotated data. Moreover, fine-tuning the student network on the labeled dataset at the end of the training process significantly improving the performance for all dataset sizes and parametrizations.

Liu et al. [144] investigated the idea of detecting and matching related objects with the goal of retrieving more meaningful detections in sport videos, introducing a framework that uses a single proposal with multiple prediction heads to improve detection and matching performance. More in detail, the proposed method consists in a detection and matching framework that uses a single proposal box to output multiple predictions, which are then processed using a SetNMS procedure to suppress duplicated sets of detections, i.e. a revisited version of the standard Non-Maximum Suppression (NMS) algorithm [181]. The method is designed to jointly optimize the prediction of multiple associated objects from a single prediction box, and is particularly suited for applications with highly overlapped associations, such as in ice hockey and crowded scenes. The framework is trained and evaluated on a broadcast ice hockey dataset collected by the same authors, and is shown to perform better than baseline models in both detection and matching tasks.

Vats et al. [259] preferred instead to tackle the problem of player identification specifically targeting ice hockey. The proposed method uses a ResNet18 model [96] pre-trained on a jersey number dataset to identify player jersey numbers in tracklet frames as prepared by the authors in a previous work [257]. The model obtains approximate labels by multiplying the jersey number probability vector with binary shift vectors during inference, while being trained using a weakly-supervised training scheme that makes use of these approximate labels, which improves convergence. The method also handles class imbalance in the dataset by sampling null class tracklets with a lower probability and uses learned loss weights for the holistic jersey number and digit-wise losses. The overall network architecture is designed to effectively identify player jersey numbers in tracklet frames. The proposed network performs better than the current state-of-the-art on the dataset, demonstrating the effectiveness of the proposed approach, particularly showing higher accuracy due to its larger receptive field, data augmentation, and ability to handle class imbalance.

2.3 Tracking

To address the fragmentation noted in previous surveys, we introduce a thematic clusterization of recent tracking approaches in sports, organized by their core methodological strategies. This structure clarifies how different systems approach challenges such as occlusion, identity ambiguity, and temporal consistency. Specifically, we identify three main clusters: (i) optimization-based associations, which cast tracking as a global (graph) partitioning or simplification problem, optimizing over trajectories with spatio-temporal constraints; (ii) geometric reconstructions, which exploit camera geometry, multi-view cues or explicit motion models to reconstruct targets of interest from one or more views; and (iii) deep-features appearances, which focus on learning and leveraging rich appearance embeddings (often via deep networks) in order to detect, re-identify (ReID) or rescue a track. A summary of the considered works is presented in Table 5.

Table 5 Summary overview of the considered articles concerning tracking

Article	Target	Dataset	Proposal	Advantage	Drawback
Ullah and Cheikh [248]	Athlete	Ullah and Cheikh [248]*	Optimized association	Spatial constraints	More computationally expensive
Zecha et al. [283]	Human joints	Zecha et al. [283]*	Optimized association	Integer linear programming	Limitations under extreme poses
Liu and Wang [141]	Shuttlecock	Huang et al. [106] Liu andWang [141]*	Optimized association	Physics-based model	More computationally expensive
Bridgeman et al. [29]	Single	Bridgeman et al. [29]*	Geometric reconstruction	Near real-time	Requires multi-view videos
Dunnhofer et al. [71]	Skier	Dunnhofer et al. [71]*	Geometric reconstruction	Generalizes to other skiing disciplines	Application specific
Liu and Hafemann [142]	Player	Liu and Hafemann [142]*	Geometric reconstruction	Scale-invariant method	Limited in case of linear pattern
Maglo et al. [155]	Player	Maglo et al. [155]*	Deep features	Robust reID capabilities	Requires human annotations
Chen et al. [43]	Player	Kristan et al. [125]	Deep features	Neighbor tracklets	More computationally expensive
Majeed et al. [156]	Player	Cioppa et al. [51] Majeed et al. [156]*	Deep features	Near real-time	Prioritizes motion features

*Dataset privately collected by the authors themselves

2.3.1 Optimization-based associations

Aiming to solve the scalability burden of tracking algorithms, Ullah and Cheikh [248] proposed a Directed Sparse Graphical Model (DSGM) for multi-target tracking. The proposed method uses a block diagram where the input is a target hypothesis in each frame, which can be generated by a detector or manual annotation. A Hidden Markov Model (HMM) strategy [204] is adopted for spatial constraints, assuming linear Gaussian distribution for the nodes in the graph and a constant velocity model as the dynamic model. Each target is associated with an individual HMM, predicting highly probable positions and reducing the solution space, where the HMM helps to introduce sparsity in the graph, hence avoiding redundant connections. A fine-tuned CNN is then used to model the appearance of the targets and calculate probabilities [207]. The graphical model formulation and dynamic programming-based optimization strategy [201] are used to get the trajectories for the targets. The authors tested their approach on 3 sports games (sprint, football and basketball), with empirical results showing that the proposed Directed Sparse Graphical Model (DSGM) performs better than one existing method [63] and gives comparable performance to two others [171, 172], both on MOTA (Multiple Object Tracking Accuracy) and MOTP (Multiple Object Tracking Precision) metrics.

Zecha et al. [283] focused instead on improving human pose tracking via joint locations refinement and temporal consistency. A pipeline for joint tracking and joint regression is presented, which aims to improve pose estimation in sport applications. The framework involves modeling the problem as a graph, where edges represent possible connections among detection estimates produced by a Convolutional Pose Machine (CPM) foundation [269], and solving an optimization problem using the Embedded Conic Solver [66] and the convex optimization modeling language CVXPY [3]. The method also involves merging edges in the graph to reduce the number of weighted edges and associated constraints, leading to a more compact optimization problem that can be solved faster. On a custom swimming dataset, the results show that the rectification pipeline improves the

performance of the pose estimation systems, with CPM achieving the highest PCK values, that is Percentage of Correct Keypoints [217], for all swimming styles, when compared to a Mask R-CNN [97].

Liu and Wang [141] investigated instead the field of trajectory estimation, looking for and end-to-end system for the extraction and segmentation of 3D shuttle trajectories from monocular badminton videos. The system consists of a set of subsystems that analyze a given video to identify the court [77], player poses, and per-frame shuttlecock pixel positions [237]. A recurrent network is then trained to obtain the segmentation of shots, and a novel per-shot trajectory reconstruction method is proposed that leverages nonlinear optimization and domain knowledge as "particle-under-drag" physical approximation [54]. On a revisited version of the public TrackNetV2 dataset [106], empirical results show that the system provides accurate 3D shot trajectory estimation and shot segmentation, and requires no human intervention, significantly outperforming prior work, especially considering that neither manual annotation nor stereo camera cues are needed.

2.3.2 Geometric reconstructions

Bridgeman et al. [29] suggested instead a 3D pose tracking system from multi-view video. The proposed method is a greedy algorithm that seeks the best correspondences between 2D pose estimations [36] in multiple views to generate 3D skeletons. It first corrects errors in the output of the pose detector – such as part flipping, single-person splitting, or joint swapping – via per-frame heuristics, in contrast to other previous methods which adopted temporal filtering for tracking consistency [5, 166]. Then, it labels every 2D pose to ensure consistency among views. As a final step, the labelled 2D poses are used to produce a sequence of tracked 3D skeletons. Empirical results show that the proposed method is able to accurately track 3D skeletons of people in various sport datasets, including table-tennis [157], boxing, karate, and soccer [90]. The method achieves comparable results to state-of-the-art methods on both the Shelf [19] and Campus [20] datasets, despite not employing pose priors. The tracking stage of the method is able to maintain tracking for the duration of the sequence, even during close contact, with some exceptions due to poorly estimated distortion parameters or tracking failures.

On a different note, Dunnhofer et al. [71] decided to focus entirely on alpine skiing, aiming at tracking skiers and visualizing their trajectories in monocular videos. The authors' goal is to compute a trajectory that summarizes the position of the athlete in each frame, without requiring prior knowledge of the playing field, which is difficult to define for the chosen sport due to varying course conditions. The proposed algorithm, SkiTraVis, processes each frame of the video online, according to their temporal order. It first uses a visual object tracking algorithm [277] to model the motion of the skier and provide its position. Then, it estimates the camera's motion between consecutive frames by computing the homography transformation using static key-points [149, 232] and RANSAC. This transformation is used to map the points of the trajectory from the previous frame into the current frame's coordinate system, updating then the trajectory with the new position of the skier. Results on a dedicated custom-built evaluation dataset comprising real-world start-to-finish multi-camera broadcasting videos of professional alpine skiers show that SkiTraVis achieves a promising balance between trajectory error and processing speed, leading to a consistent perception of the trajectory, and

presenting therefore itself as a solid tool for situations where the real-time requirement is not so strict, such as broadcasting replays or training activities.

Liu and Hafemann [142] suggested instead a semi-automated data correction pipeline for trajectory simplification in sport scenarios. The method consists of a trajectory correction framework that utilizes tracking data as input and requires manual correction only for a small subset of this data to obtain the entire corrected sequence. The method categorizes observations into high-quality and low-quality subsets, simplifies trajectories by selecting a small set of keyframes via Directed Acyclic Graph (DAG) formulation and integral error minimization [35, 42], and then interpolates groundtruth trajectories from the cleaned simplified trajectories using linear interpolation, thus preserving information on the original sequence while reducing the number of frames that need to be manually reviewed or corrected. Extensive results demonstrate that the proposed scale-invariant trajectory simplification method outperforms existing methods [42, 168] in terms of mean IoU scores in all datasets considered, i,e. MOT20, SoccerNet-Tracking [51] and DanceTrack [238]. However, the method may not demonstrate an improvement over uniform sampling if the tracked motion predominantly follows a linear pattern, and may exhibit visible errors when the compression rate is over 10×. The study also evaluates the algorithm's performance under varied levels of data noise on a synthetic dataset, finding that it can filter noise while retaining necessary trajectory information.

2.3.3 Deep-features appearance

Maglo et al. [155] investigated the task of tracking players in absence of complete game-specific annotations. The proposed method is an offline approach to track and identify players in a team sport video with a single moving view. The method first generates non-ambiguous tracklets [22] by detecting and associating bounding boxes around persons across frames. A user then provides a few identity annotations to some of the generated tracklets, which are used to train a tracklet reID network. Once trained, the model generates classification scores and re-ID features for all tracklets, which are then matched to player identities using an association algorithm [98]. The method uses therefore an incremental learning approach, allowing the user to add more annotations to correct associations and improve the model's accuracy. The results of the proposed method on a custom-built tracking rugby sevens dataset show that the system is able to track players during a full match with few annotations (only 6 few seconds length tracklets per player).

Chen et al. [43] proposed instead NeighborTrack, a post-processor that leverages neighbor knowledge of the tracking target to improve and validate single-object tracking (SOT) results. Inspired by the principle of cycle consistency [267], the framework aims to improve the tracking results of a SOT network by incorporating past information for a period of time to reduce tracking inaccuracies resulting from occlusions or changes in appearance over time. The method uses a Kalman filter and hypothesis generation via bipartite matching [176] to provide candidates independent of appearance features, thus providing a greater range of options to the tracking system. NeighborTrack can be applied to any SOT network that meets two requirements: the ability to generate a tracking result given a target template patch and a sequence of frames, and the ability to generate a set of candidate bounding boxes and their confidence scores for a frame. The method has been shown to be effective in improving tracking results, especially in challenging scenarios such as sports-related videos. Indeed, improvements are observed

in terms of accuracy, precision, and success rate on the VOT dataset [125] by adding NeighborTrack to existing tracking methods [277, 281]. Moreover, the use of Kalman filter is also shown to be effective in preventing the tracker from being confused by neighbors similar to the target when the target is completely occluded, allowing the tracker to quickly return to the correct tracking position when the target reappears.

Majeed et al. [156], based on the assumption that soccer athletes moving on field will be outlined by motion vector discontinuities, proposed a novel method for real-time soccer player detection, instance segmentation, and tracking using motion vectors generated by a DenseNet-based motion estimation from absolute frame differences. The framework consists of three components: a backbone for feature extraction, a neck for video-frame resampling, and a head for motion compensation and tracking [25]. The backbone uses the Cross-Stage-Partial Network (CSPDarknet) to extract features from input frames, which are then concatenated and forwarded to the head stage for upsampling and image reconstruction, leveraging motion vectors to improve foreground/background separation and distinguish among players, especially in scenarios involving partial occlusion. Results on two public soccer datasets—DFL–Bundesliga Data Shootout [170], SoccerNet-Tracking [51]—and a private one (SoccerPro) show that the proposed framework outperforms existing methods that use only RGB images, despite the challenges imposed by noisy and low-resolution motion vectors, exhibiting exceptional tracking accuracy and outperforming contemporary SOTA trackers [4, 34, 69, 203, 286] in metrics such as HOTA, MOTA, and IDF1, which highlights its efficacy in accurately tracking soccer players' movements.

3 What is going on?

Driven by advances in image classification and object detection in both images and videos, researchers have shifted their focus to more complex action recognition tasks. Action recognition (AR) in sports refers to the use of computer vision methods to detect and recognize actions or activities happening in a sport-related environment. This can be done during warm-ups, fitness training, specific training for a sport, matches, or competitions. The goal of AR is to detect the person performing an action in an unknown video sequence, determine the duration of the action, and classify its type [105].

Actions in sports typically involve physical movements executed by a player to complete a task, either independently or through interaction with objects or other players. These actions can vary in complexity, depending on the nature of the sport and the task at hand.

Overall, AR has the potential to revolutionize the sports domain by providing valuable insights and improving performance, with the growing interest towards it leading to several possible applications, such as fitness applications that provide feedback on correct technique and suggest improvements, video analysis tools that help coaches guide players, applications that detect unusual situations on the field and alert medical personnel or real-time referees assistance. Section 3.1 deals with the general task of event spotting, that is, the recognition that a specific action in taking place on screen, let it be related to an athlete or an object. Section 3.2 deals instead with the specific scenario in which an action is taking place between two entities, hence gaining the status of relationship. The goal here shifts from solely spotting the event, to recognizing also the two entities having a role in the relationship under analysis. Section 3.3 concludes the overview by focusing

on the specific scope of performance assessment and rehabilitation, where action recognition and understanding becomes a key aspect of motion mentorship and therapeutic support.

3.1 Event spotting

To address the novelty of event spotting research in sports, we propose a thematic clusterization that organizes the literature into three key methodological categories, each representing a distinct axis of innovation and development. These clusters are: (i) skeleton-driven modeling, leveraging human joint detections or trajectories as the primary input for fine–grained event recognition; (ii) spatio-temporal modeling, reflecting on richer spatio-temporal context to detect and recognize the occurrence of an event; (iii) multi-modal fusion, fusing audio, vision–language embeddings or other side channels to support event detection. A summary of the considered works is presented in Tables 6, 7 and 8.

3.1.1 Skeleton-driven modeling

Einfalt and Lienhart [72] assessed the possibility of using human pose estimation as a compact description of an athlete's motion over time, looking to decouple the specific objective of motion event detection in particular sports from the low-level problem of inferring information from video data directly. In order to do so, the authors propose a modular and flexible approach that uses a modified variant of Mask R-CNN [97] to estimate the pose of the athlete of interest in every video frame as a continuous translation task [73]. They fine-tune the model on sampled and annotated video frames from their application domains, achieving high accuracy in keypoint estimates. Empirical results were reported on two privately-collected sport datasets, swimming and long/triple

Table 6 Summary overview of the considered articles concerning event spotting

Article	Event	Dataset	Proposal	Advantage	Drawback
Einfalt and Lienhart [72]	Game dynamics	Einfalt and Lienhart [72]*	Pose-based modeling	Readily applicable to other sports	Application specific
McNally et al. [163]	Automatic scorekeeping	McNally [162]	Pose-based modeling	Relies on single camera	Application specific
Kulkarni and Shenoy [127]	Stroke recognition	Kulkarni and Shenoy [127]*	Pose-based modeling	Spatio-temporal embeddings	Custom set-up
Zhu et al. [252]	Footwork recognition	Zhu et al. [252]*	Pose-based modeling	Spatio-temporal embeddings	Limitations under extreme poses
Deyzel and Theart [62]	Conditioning movement	Deyzel and Theart [62]*	Pose-based modeling	One-shot metric learning	Limitations under extreme poses
Ibh et al. [110]	Game dynamics	Ghosh et al. [84] Ibh et al. [110]*	Pose-based modeling	Spatio-temporal embeddings	Likely to overfit
Schlosser et al. [225]	Game dynamics	Jiang et al. [118]	Spatio-temporal inference	Optical flow inclusion	More computationally expensive
Kanojia et al. [121]	Diving classification	Kanojia et al. [121]*	Spatio-temporal inference	Representation learning	Application specific
Vats et al. [256]	Game dynamics	Deliège et al. [59] Vats et al. [256]*	Spatio-temporal inference	Robust against coarse annotations	Does not take contextual information into account
Sanford et al. [215]	Group activities	Sanford et al. [215]*	Spatio-temporal inference	Embeds physical location	Does not consider far distant context

*Dataset privately collected by the authors themselves

Table 7 Summary overview of the considered articles concerning event spotting *(ctd.)*

Article	Event	Dataset	Proposal	Advantage	Drawback
Voeikov et al. [261]	Game dynamics	Voeikov et al. [261]*	Spatio-temporal inference	Auto-referee system	Downscaled inputs
Giancola and Ghanem [85]	Game dynamics	Deliège et al. [59]	Spatio-temporal inference	Features pooling	Does not consider far distant context
Giancola et al. [87]	Game dynamics	Deliège et al. [59] Giancola et al. [87]*	Spatio-temporal inference	Reduced annotation effort	Limitations under extreme poses
Chappa et al. [40]	Group dynamics	Ibrahim et al. [112] Yan et al. [278]	Spatio-temporal inference	Long-range dependencies	More computationally expensive
Xarles et al. [274]	Game dynamics	Hong et al. [103] Xu et al. [275]	Spatio-temporal inference	Higher temporal resolution	More computationally expensive
Vanderplaetse and Dupont [254]	Game dynamics	Deliège et al. [59]	Multi-modal fusion	Multi-stage audio integration	Naïve coupling
Pidaparthy et al. [199]	Automatic video production	Ren et al. [209]	Multi-modal fusion	Trans-camera generalization	Application specific
Mkhallati et al. [174]	Automatic commentary production	Mkhallati et al. [174]*	Multi-modal fusion	Solid benchmarking	Complex temporal anchoring
Nonaka et al. [185]	Scene classification	Nonaka et al. [185] *	Multi-modal fusion	Explainability	Not comprehensive prompt tailoring
Gossard et al. [89]	Ball spin estimation	Gossard et al. [88]	Multi-modal fusion	Readily applicable to other sports	Custom set-up

*Dataset privately collected by the authors themselves

Table 8 Summary overview of the considered articles concerning event spotting *(ctd.)*

Article	Event	Dataset	Proposal	Advantage	Drawback
Nakabayashi et al. [179]	Ball spin estimation	Rebecq et al. [208]	Multi-modal fusion	2D-to-3D projection	Custom set-up
Decorte et al. [58]	Hit detection	Decorte et al. [58]*	Multi-modal fusion	Predictive capabilities	Custom set-up

*Dataset privately collected by the authors themselves

jump, indicating that the model produces reliable and precise keypoint estimates for the vast majority of test set keypoints, yet still struggling with high-precision estimation for swimming due to the aquatic environment and visual clutter.

McNally et al. [163] shifted instead their attention towards predicting dart scores from a single image of a dartboard taken from any camera angle. The proposed system, DeepDarts, consists of two stages: keypoint detection and score prediction. The keypoint detection stage models keypoints as objects, rather than using heatmaps, which can be problematic when multiple darts are close together. The system uses a deep learning-based object detector [262] to locate the exact coordinates where the dart tips strike the dartboard. The predicted calibration points are then used to map the predicted dart locations to a circular dartboard and calibrate the scoring area. Finally, the dart scores are then classified based on their relative position to the center of the dartboard. Empirical results on a manually-collected and annotated dataset [162] show promising results

in predicting dart scores, with the most common error occurring when darts were on the edge of a section.

Kulkarni and Shenoy [127] introduced a method to collect, analyze and perform stroke detection and classification in table tennis video data. The proposed method involves pose estimation [263] from a camera positioned in front of the player to extract high-level features, which are then fed into a stroke recognition models. Two approaches are discussed by the authors to deal with such a task: a machine learning approach, which flattens the data and uses different algorithms, and a temporal convolutional network (TCN) approach, which uses a single temporal convolutional layer followed by layer normalization. Experimental results on privately-collected videos show that the TCN-based approach outperforms other state-of-the-art LSTM models.

Zhu et al. [252] focused specifically their attention towards fencing by proposing FenceNet, a skeleton-based action recognition approach that incorporates temporal convolutional networks to extract relevant temporally-evolving information. More in detail, FenceNet uses a stacked temporal convolutional network (TCN) architecture and a multilayer perceptron for classification, employing also a feature selection algorithm to reduce dimensionality and a decision-level fusion scheme to combine the outputs of separate support vector machines trained on each feature set. By deploying FenceNet on the Fencing Footwork Dataset, the authors showcased its promising performances, highlighting how such a framework outperforms current top-scoring methods [122], and how including bidirectional temporal evolutions allows the model to achieve better performances.

Deyzel and Theart [62] addressed the problem of one-shot skeleton-based action recognition, specifically on strength and conditioning exercises scenarios. The proposed method involves training a state-of-the-art graph convolutional architecture—ST-GCN, a spatial-temporal graph convolutional neural network [279]—as an encoder model on the large-scale NTU RGB+D skeleton data [229]. The ST-GCN model learns to extract spatial-temporal features directly from skeleton sequence data and project them into an embedding space that clusters similar actions. The method uses a metric learning paradigm, where the feature vector is used directly as a learned embedding, and a k-nearest neighbor (k-NN) classifier is used for classification. The approach is therefore tailored for one-shot learning, where the model is trained on a set of classes and then tested on a novel set of classes with only one example per class provided as a reference. The authors proposed also a novel dataset, called Stellenbosch University Exercise Motion Dataset (SU-EMD), comprising multi-view RGB video clips with 3D pose estimation and trajectory data. After pretraining a GCN on the NTU RGB+D dataset, quantitative and qualitative evaluations on the unseen classes of the SU-EMD dataset demonstrate the effectiveness of the proposed approach for one-shot learning, able to achieve competitive performances with respect to state-of-the-art methodologies [140].

Ibh et al. [110] investigated the adoption of skeleton data for fine-grained action recognition in badminton scenarios. By highlighting the potential of transformer models for such a task, the authors proposed TemPose, a transformer-based architecture to capture temporal body dynamics and sequential interactions among players. The framework first extracts the skeleton data, player position, and shuttlecock data from RGB video input. The skeleton data is then mapped through a linear projection to a sequence of temporal tokens, and a learnable temporal embedding is added to capture the underlying temporal

structure. These tokens are then passed through transformer layers, where each layer is composed of a multi-head self-attention, layer normalization, and a multi-layer perceptron. Finally, the representation of the class token at the final transformer layer is used by the MLP head to make predictions. Empirical results show that the model outperforms previous state-of-the-art approaches on badminton action recognition tasks [84], but shows signs of overfitting, suggesting that taking additional steps to combat it, such as generating synthetic data, may further improve its performance. The authors also emphasize the importance of skeleton data in providing a detailed representation of body movement, enabling the extraction of features crucial for recognizing specific actions and movements.

3.1.2 Spatio-temporal modeling

Schlosser et al. [225] explored the need for temporal action proposals in the task of temporal localization of actions in long, untrimmed videos and the classification of said actions. Since traditional approaches [32] use sliding window methods, which are time-consuming and require classifying many time segments, the authors introduce temporal action proposals to reduce the number of time segments that need to be classified. Their idea extends traditional methods to a two-stream model architecture by introducing a second stream working on the corresponding images of optical flow, which allows efficiently making use of the dynamics of motion by applying 3D convolutions on it. Experiments on the THU-MOS'14 dataset [118] showed that all of the four two-stream models lead to improvements compared to the single-stream networks, highlighting as well that the additional usage of optical flow leads to improvements for major parts of all metrics considered.

Kanojia et al. [121] proposed an attention-guided neural network, based on LSTM, for the task of diving classification. The proposed method consists of a network architecture with four components: feature extractor, encoder, attention network, and decoder. The feature extractor uses a convolutional neural network, specifically ResNet with 18 layers [96], to obtain representations for each video frame. The encoder, attention network, and decoder are multi-layer LSTM networks. The encoder takes the feature vectors as input in temporal order, while the attention network takes the feature vectors in reverse order and generates attention vectors. The decoder takes the dot-product of the attention vectors with their corresponding feature vectors as input and outputs representations for each video frame. The authors collected also the Diving48 dataset, consisting in an action recognition collection of competitive diving clips. Experimental results show that the attention network is capable of identifying and localize the diver in the clip frames without being purposely trained to do so, significantly outperforming overall other state-of-the-art methods in both 2D and 3D frameworks [21, 247] concerning the dive classification task on the Diving48 dataset.

Vats et al. [256] discussed a new approach for event detection in sports videos, specifically in soccer and ice hockey, without the need for frame-level annotations. The proposed method for event detection in sports videos utilizes a multi-tower architecture with temporal 1D convolutional neural networks (CNNs) of varying kernel sizes and receptive fields to account for events occurring at different temporal scales. The input to the network is a sequence of frames sampled uniformly from an untrimmed video, which are first passed through a pre-trained 2D MobileNetV2 [214] to extract features

to fed into the 1D CNN towers, each of them processing the input on different temporal scales. Experimental results on both the NHL and SoccerNet dataset [86] show that their framework achieves competitive results compared to state-of-the-art methods, highlighting the advantage of using towers with different receptive fields.

Sanford et al. [215] similarly discussed a method for detecting activities in soccer game by including a model that uses transformers and self-attention to capture the context of the scene and predict the occurrence of activities. Specifically, their idea involves using vision-based approaches to predict events in soccer games via an Inflated 3D CNN [94] as the backbone, which takes short clips of size 3xTxHxW as input and outputs a multi-dimensional feature vector. The method then aggregates the results using either a simple aggregation (max over N), a Graph Convolutional Network (GCN), or a Transformer model. On a private dataset of multiple English Premier League soccer games, experiments showed that the framework performs interestingly in detecting events such as passes, receptions, and shots, demonstrating its effectiveness in detecting events in soccer games. The authors [258] declined their research also specifically towards multi-task event recognition in broadcast hockey videos. The proposed method starts by using a CNN to predict the probability of puck location on an ice rink from video clips of hockey games by including both spatial and temporal context. By leveraging attention components that learn spatial relationships between player locations and video features, the model is capable to produce the probability of puck location, which is then used for event recognition and team classification. On a custom dataset of broadcast NHL videos, each one annotated with puck location and event label, the framework achieves interesting results, yet still having lower accuracy in the bottom halves of the defensive and offensive zones due to the puck being occluded by the rink boards.

Voeikov et al. [261] focused their interest towards table tennis by proposing TTNet, a real-time multi-task network that, by using both temporal and spatial data, can detect in-game events such as bounces and net hits, perform semantic segmentation to identify the table, humans, and scoreboard, all while tracking the ball's position. The network takes as input a sequence of frames from a video and outputs the probabilities of in-game events, semantic masks, and the ball's coordinates. The event spotting branch acts on concatenated feature maps from global and local detectors, allowing gradients from event spotting loss to flow through both feature extractors, with adaptive balancing of loss components [52]. Furthermore, OpenTTGames, a dataset of table tennis videos was also collected by the authors and labeled to assess the proposed architecture, highlighting with empirical experiments that the proposed approach achieves high accuracy and efficiency in all the tasks considered.

Giancola and Ghanem [85] proposed a novel feature pooling method called NetVLAD++ for action spotting in soccer broadcasts. The method is an improvement over the original NetVLAD [7] and is designed to be temporally aware, allowing it to better capture the temporal relationships between frames in a video. More in detail, the method involves a novel temporally-aware NetVLAD pooling module that learns the past and future temporal context independently. The module is implemented in a comprehensive pipeline for action spotting, which includes a learnable projection to reduce the feature dimensionality. By considering window chunks of time along the video and by using a temporally-aware pooling approach that is aware of the temporal order of the frame features—therefore approaching past and future context independently—NetVLAD++

is capable to improve upon the original temporal pooling mechanism proposed in SoccerNet [86]. Empirical results on the SoccerNetV2 action spotting dataset [59] highlight that the proposed method outperforms the state-of-the-art in terms of Average-mAP, with ablation studies showing that each component of NetVLAD++ contributes to the overall improvement, concerning the temporally-aware feature pooling and learnable linear layer as the most significant contributors.

On a different note, Giancola et al. [87] suggested an active learning framework [79] for action spotting, which aims to train an accurate action spotting model using a minimal amount of labeled data. The framework consists of three key steps: training an action spotting model on a labeled dataset that grows at each active learning step, selecting the most informative data from an unlabeled pool using an active learning algorithm, and labeling the selected clips by an oracle and including the new data and annotations in the labeled set. The process is repeated iteratively until the desired performance is reached or the unlabeled dataset is empty, with its final goal being to minimize the number of times the oracle is queried by proposing an efficient active selection algorithm. Among the considered techniques [130], empirical results show that the Entropy Measure (EM) outperforms Random Sampling (RS) on all active learning setups, with adapting the active learning scheduler and continuing the training leading to an order of magnitude acceleration in running the experiments.

Chappa et al. [40] proposed SPARTAN, a self-supervised spatio-temporal transformer-based approach that aims to recognize group activities in videos without using person-bounding boxes or detectors via a self-supervised training approach within a teacher-student framework. More in detail, it uses a vision transformer (ViT, Caron et al. [37] to process video clips and applies individual attention to both temporal and spatial dimensions. The method processes two clips from the same video by changing their spatial-temporal characteristics and matches the features of the two dissimilar clips to impose consistency in motion and spatial changes. The ViT backbone has separate space-time attention and uses a multi-layer perceptron (MLP) to predict target features from online features. Experimental results on both basketball [278] and volleyball [112] datasets highlighted promising performances with respect to previous state-of-the-art frameworks, with such analysis also showing that the proposed inference method yields greater improvements on datasets that contain classes that can be more easily distinguished using motion information.

Xarles et al. [274] highlights the importance of processing videos in multiple temporal scales and enhancing token discriminability for precise predictions by proposing T-DEED, a model designed to address action spotting across various sports datasets. The authors' idea consists of three key components: a feature extractor, a temporally discriminant encoder-decoder, and a prediction head. The feature extractor produces per-frame representations using a compact 2D backbone and incorporates local temporal information through the use of GSF modules [236]. The temporally discriminant encoder-decoder captures local and global temporal information while enhancing token discriminability. Finally, the prediction head generates per-frame classifications and displacements for refinement. The method adopts an end-to-end approach, which has demonstrated benefits in learning more meaningful features, particularly in scenarios where precise predictions are essential. The authors validated their framework on the FigureSkating [103] and FineDiving [275] datasets, with empirical results showing that

T-DEED outperforms current state-of-the-art methods regarding precise event spotting in sport videos.

3.1.3 Multi-modal fusion

Vanderplaetse and Dupont [254] followed the intuition of multi-sensory analysis by finding a way to improve soccer action spotting using both audio and video streams. The main idea consists of a baseline processing pipeline with the audio and video stream representation vectors being concatenated at the merge points, leading to seven different methodologies. Empirical results on the SoccerNet dataset [86] show that using a combination of video and audio streams with a multi-clusters NetVLAD [7] as pooling layer provides the best performance, and that combining the two streams provides significantly improved performance with respect to the scenario in which only one modality is considered.

Pidaparthy et al. [199] proposed a novel system for automatic play-break detection in sports videography, specifically for amateur ice hockey games. The system uses visual cues from the video to identify stoppages in the game, such as breaks between plays, warm-ups, and intermissions, in order to abbreviate the video and provide a more condensed viewing experience. The proposed method uses visual cues from a single camera to automatically identify stoppages in amateur ice hockey games [209], allowing for the abbreviation of the video. A quantitative and qualitative assessment of the model shows that the deep visual cue outperforms the optic flow cue for all cameras. Additionally, while the optic flow cue benefits from integration with the audio cue, the deep visual cue is strong enough on its own and does not show any sensory integration benefit.

Mkhallati et al. [174] decided to tackle a completely different aspect, that is, the one of generating automated commentaries for soccer games, hoping to provide a similar level of engagement as a live game. The authors' goal is to enhance the accessibility and understanding of soccer content for a wider audience, bringing the excitement of the game to more people. The proposed method is a two-stage approach for dense video captioning, consisting of a spotting model and a captioning model, both of which use a shared frozen feature extractor [251] to generate a compact per-frame representation of the video. The spotting model uses an aggregator module to combine the frame features into a single clip feature representation, which is then passed to the spotting head to generate proposal timestamps. The timestamps are then used to trim clips of a fixed size, which are subsequently processed by the captioning model through the feature extractor, aggregator, and a captioning head to generate the anchored comment. Empirical results on a labeled subset of the SoccerNet dataset [86] show that the captioning model is able to generate meaningful comments but struggles with accurately generating scores and temporal anchoring of commentaries, highlighting the difficulty of the task, but still showing the potential of the proposed method to enhance the accessibility and understanding of soccer content for a wider audience.

Nonaka et al. [185] discussed a study on using Visual Language Models (VLMs) for analyzing and understanding rugby scenarios. The proposed method consists of three main components: a pre-trained VLM, i.e. a Visual Language Model [139], an image encoder, and a head module. The VLM takes natural language prompts and image data as input and outputs vector representations, but its parameters are not updated during training. While the encoder extracts image features, the head module takes the

output vectors from the VLM and image encoder and outputs vectors corresponding to the number of classes for the task. The authors empirically explored the use of different prompts to optimize performance for various tasks, with classification performances generally improving by leveraging the VLM output.

Gossard et al. [89] focused on estimating the spinning action of a table tennis ball with the help of event cameras, since they have oral resolution and can capture the movement of objects without motion blur. The authors adopt EROS, the Exponential Reduced Ordinal Surface [82] event representation, to process the events generated by the camera. EROS allows for continuous and asynchronous updates from the event stream, enabling sharp edges and accurate detection of objects regardless of their velocity. By comparing EROS with other accumulated event frames, the authors discuss the advantages of using event cameras, including low latency and reduced bandwidth requirements, mentioning also the potential of fusing data from frame-based and event-based cameras to compensate for each other's weaknesses. The proposed method for estimating the spin of flying balls using event cameras involves a three-phase pipeline. The method first uses a ball tracker to estimate the ball's position, velocity, and radius. This information is then used to extract events generated by the logo on the ball. Finally, the ball's spin is estimated from the extracted events [198]. On a dataset of privately-collected images generated via SpinDOE [88], empirical results show that the event-based approach has a success rate slightly above the frame-based approach in estimating the spin of flying balls, yet being still a bit behind in terms of spin axis and spin magnitude mean absolute error.

Nakabayashi et al. [179] further investigated the task of spin estimation, looking for an agnostic method regarding specific sport balls. Similarly to the previous work, the proposed approach uses an event camera to estimate the spin axis, angular velocity, and translational velocity of a ball. The method first projects the events into a three-dimensional space by calculating the depth, then estimates the parameters through iterative optimization using the Contrast Maximization framework [80]. Such a framework optimizes the parameter to maximize the variance of the events, allowing the method to accurately estimate the spin of the ball. Experimental results on a synthetic dataset [208] provided by the same authors highlight that the method consistently exhibits small errors regardless of the magnitude of the angular velocity, and is able to estimate spin independently of the ball's texture, making it applicable across various sports, outperforming other baseline methods [242], especially at high angular velocities.

Decorte et al. [58] discussed a multi-modal method for analyzing padel matches, specifically for detecting ball hits using audio analysis instead of video, while performing ball detection only on video frames close to the time intervals that the ball was hit. Differently from other approaches, the framework first analyzes the audio signal of the video to detect the exact moments when a player hits the ball [169], which produces a distinct sound. The timing windows of these hits are then propagated to the video pipeline for further analysis, which involves tracking the ball and detecting the player positions by observing all players on the field. By collecting a dataset comprising several hours of video summaries of padel matches, the authors validated their method through a standardized evaluation framework, providing valuable insights into player movements and interactions.

Table 9 Summary overview of the considered articles concerning visual relationship detection and prediction

Article	Target	Dataset	Proposal	Advantage	Drawback
Arbués-Sangë-sa et al. [8]	Pass prediction	Arbués-Sangësa et al. [8]*	Pose-based modeling	Solid mathematical modelization	Lack of whole-field passing feasability
Askari et al. [10]	Penalty recognition	Askari et al. [10] *	Pose-based modeling	Contextual informa-tion embedding	Application specific
Nonaka et al. [184]	Injury assessment	Nonaka et al. [184]*	Pose-based modeling	Injury risk classification	Presence of false positives
Askari et al. [11]	Penalty recognition	Askari et al. [11]	Pose-based modeling	Leverages unlabeled data	Specifically designed for downstream tasks
Askari et al. [12]	Penalty recognition	Askari et al. [12]	Pose-based modeling	Light and easy to train	Less sound tem-poral aggregation
Honda et al. [102]	Pass prediction	Honda et al. [102]*	Simulation forecasting	Body motion embedding	Lack of whole-field passing feasability
McNally et al. [164]	Ball flight prediction	McNally et al. [164]*	Simulation forecasting	Solid mathematical modelization	Application specific
Kaneko et al. [120]	Pass prediction	Kaneko et al. [120]*	Simulation forecasting	Easily scalable to multi-person scenario	Limited to specific simulation data
Ibh et al. [111]	Stroke prediction	Wang et al. [265] Ban et al. [14]	Simulation forecasting	Contextual informa-tion embedding	Application specific
Martin et al. [159]	Injury assessment	Martin et al. [159]*	Explainable assisting	Injury risk classification	Weak against strong occlusions

*Dataset privately collected by the authors themselves

Table 10 Summary overview of the considered articles concerning visual relationship detection and prediction *(ctd.)*

Article	Target	Dataset	Proposal	Advantage	Drawback
Sarkar et al. [220]	Pass prediction	Sarkar et al. [220]*	Explainable assisting	Near real-time	Lack of contextual passing information
Held et al. [99]	Interaction detection	Held et al. [99]*	Explainable assisting	Near real-time	Requires multi-view videos
Held et al. [100]	Interaction detection	Held et al. [99] Held et al. [100]*	Explainable assisting	Explainable assessment	Requires multi-view videos

*Dataset privately collected by the authors themselves

3.2 Visual relationship detection and prediction

This task frames the problem of recognizing interactions among entities as occurring actions, hence presenting itself as a sort of evolution from the previous goal of event recognition. Therefore, the proposed clusterization partly intersects with the former one, including: (i) skeleton-driven modeling, leveraging keypoint-based annotations to refine relationship occurrence; (ii) simulation forecasting, highlighting approaches aimed at prediction instead of detection; and (iii) explainable assisting systems, grouping those works which perform relationship detection as decision-support tools in downstream tasks. A summary of the considered works is presented in Tables 9 and 10.

3.2.1 Skeleton-driven modeling

Arbués-Sangësa et al. [8] presented a computational model to estimate the most feasible pass at any given time, given a monocular video of a soccer match. The proposed method involves a computational model that assigns a feasibility value to each potential receiver in a soccer game by taking into account the position and orientation [9] of the players

on a 2D field template and using a geometrical approach to assign discretized pass probability/EPV (Expected Possession Value, Cervone et al. [39] field values to each receiver. Moreover, the method integrates the probability/EPV values on a meaningful area that extends from the passer to the receiver, and the final individual value for each receiver is obtained by calculating the area of the region and integrating the probability/EPV values within that region, combining also the output of the pass probability and EPV models to enhance potentially good receivers in particular regions. Empirical results on a dataset comprising 11 whole games provided by F.C. Barcelona show that the effect of orientation is vital in the half-court, with a notable difference between successful and non-successful passes in the progression phase. Furthermore, both the Top-1 and Top-3 accuracy metrics are improved when taking orientation into account, leading to a more accurate model which can help in better understanding the decision-making process.

Askari et al. [10] investigated the problem of interaction recognition from multi-person videos. The proposed framework consists in a CNN-RNN based model that receives video frames and poses as input and classifies each single video. It uses a ResNet152 network [96] to extract scene features from video frames, which are then embedded into a 512-dimensional vector. A bidirectional LSTM (BiLSTM) extracts global information from the scene features, and these features, along with players' poses, are input to an attention model. The attention mechanism outputs key-points, and the final representation is the weighted sum of players' key point features. This representation is then concatenated with the output of the BiLSTM and input to an interaction classification LSTM, which acts as the final classifier. On a custom dataset comprising ice hockey penalty clips, the study compares the proposed method with other popular action/interaction recognition approaches [67, 70, 279] and notes that the proposed model not only improves their performances, but also offers both interaction classification and key actor detection, whereas the other methods only classify interactions.

Aligning with similar proof-of-concepts, Nonaka et al. [184] addressed the task of high-risk tackle detection by taking advantage of more recent progresses in deep learning methodologies. Differently from the former work—where tackle scenes were manually identified in advance—the authors aim here instead to develop a system which does not require human intervention. More in detail, the proposed method consists in a system to detect high-risk tackles from rugby match videos, which consists of four models: tackle frame selection model, tackle detection model, pose estimation model, and tackle risk classification model. The system takes a video as input, classifies frames by tackle frame selection model [96], combines tackle detection model and pose estimation model [289] to detect tackles, and finally classifies the risk of given tackles. On a custom-built dataset, this hierarchical procedure turns out to be able to detect actual head concussions half of the time, showcasing promising capabilities considering its fully-automated nature. Yet, the requirement of multiple networks poses strong limitations, especially regarding processing speed and false positive classification.

Askari et al. [11] further investigated the effectiveness of applying self-supervised learning in order to build useful representations from human skeleton pose data and gain a better understanding of visual relationship. Their suggested method adapts an image-based self-supervised learning (SSL) approach, specifically the Relational-SSL method [197], for sequences of human poses. The method first transforms the pose data into a format that can be processed by a convolutional neural network (CNN). The

representations retrieved by the CNN backbone are aggregated to form pairs, which are used to train a relation reasoning head through a simple classification task using binary cross-entropy loss, where the goal is to minimize the distance between the representations of positive pairs and maximize it for negative pairs [216]. Encouraging results on the same ice hockey dataset presented above [10] highlight how using SSL methods could be beneficial for the field of sport analytics, where abundant unlabeled data are often available and the task of creating a labeled sport-specific dataset is expensive and time-consuming.

The authors explored in Askari et al. [12] also an extension of relational networks (RN) via scaled dot product self-attention for video interaction recognition. The method defines a novel set of RN objects [216] that are simple, elegant, and effective in solving the task. These objects are generated from the skeletons of actors within individual frames and then concatenated to form object pairs. The object pairs serve as input to a relational module, which is followed by a self-attention module [255] responsible for aggregating the relationships, with such a final relational representation being then fed to a classifier to output class membership. On the same hockey dataset, the experimental results of the study demonstrate that using relative skeleton features consistently improves the model's performance. The inclusion of these features is a key contributor to the superiority of the approach, enabling it to outperform other top-performing models [10] and capture temporal dynamics without requiring temporal-specific architectures. Additionally, the study evaluates the effect of object definition on the interaction classification accuracy and finds that a noticeable decline in performance is observed when using a localized and joint-centric input definition.

3.2.2 Simulation forecasting

Honda et al. [102] approached the task of automatic prediction of the pass receiver by combining visual information with trajectories data of both players and ball. The proposed method for pass receiver prediction involves several steps. First, players' positions are detected in frames using YOLOv5, and then transformed player positions are aligned with those detected points to use tracked and labeled information in the original tracking data. To address the issue of undetected players, a rigid point registration algorithm called iterative closest point [287] is used to estimate undetected positions in the video, referred to as pseudo-detected points. Unnecessary points are removed using Hungarian matching. The method then extracts 20 features for each player, which are summed and used as input to learn the interaction between players. The relationship between players is modeled as a complete graph, and a transformer encoder with residual connections is used to update the players' features while aggregating their interaction. Finally, the features corresponding to the potential receivers are passed to fully connected layers and converted into a pass-receive probability using the softmax function. The authors collected a dataset comprising wide-angle videos in which all 20 players (goalkeepers excluded) appear at all times during the whole match. When compared with other CNN-based methods [107], experimental results showed that the proposed method achieved the highest prediction accuracy for several top-k metrics, confirming that including video information (alongside with trajectory data) is desirable and beneficial.

McNally et al. [164] placed their research on a completely different level, by aiming to investigate the relationship between a golf ball and its surrounding in order to simulate

its flight and predicting the landing position. The authors try to achieve so by integrating a neural network into a physics model of a golf ball in flight [243], combining therefore the strengths of physics-based modeling and deep learning to simulate multiple golf ball trajectories in parallel and obtain also final landing positions. The model uses a launch monitor to record a shot hit by a golfer and measures the launch conditions, which, together with the initial golf ball position, form the initial state of the golf ball. The differential equations describing the ball flight are numerically integrated from 0 to a simulation time, and a neural network [177] is used to estimate the aerodynamic coefficients. A soft-argmax operation is then used to obtain the predicted landing position. During training, the target landing positions are used to compute the landing position loss and update the neural network weights. The authors collected a ball flight dataset by using radar-based launch monitors, testing their model against two other published ball flight models [78]. Compared to such methods, the proposed approach has minimal computational overhead, running much faster while still achieving promising results in both simulating the ball's flight and predicting its landing position.

Kaneko et al. [120] focused on additional synthetic data generation to overcome the limitation of pass analysis datasets requiring costly human-performed annotations. The authors' idea is to generate diverse synthetic data that resembles real data and to utilize it for training. The method comprises three steps: learning behaviors similar to real data, generating diverse synthetic data, and training with a mix of synthetic and real data. The method uses a 3D soccer simulator [128] to replicate the soccer match environment with high fidelity, allowing for realistic actions and considerations of players' body axes and orientations. For training, the simulator produces precise data output, including the x, y coordinates of the ball and 20 field players, along with match video from an overhead perspective. The proposed method then uses this synthetic data, along with real data, to train a deep learning-based predictive model, leveraging techniques such as behavior cloning and attention mapping to capture intricate relationships between players. On a privately-collected dataset comprising trajectories and videos of successful passes from the main Japanese league, experimental results showcased the performance improvement with respect to previous methods [102] when correctly-generated synthetic data are included in the training process.

Ibh et al. [111] shifted instead the relationship detection task towards the domain of badminton, specifically focusing in predicting future strokes based on previous rally actions. The proposed method, called RallyTemPose, involves an adaptive cross-attention decoder that incorporates contextual stroke descriptors from high-dimensional embeddings of a pre-trained language model. The approach uses a transformer-based model [110] comprising both an encoder and a decoder. The encoder takes in input data and applies various transformations, including multi-headed self-attention, group pooling, and fully connected layers. The decoder takes the output of the encoder and uses it to predict the probability of the next stroke in a sequence. The decoder employs self-attention, dual cross-attention, and an adaptive fusion mechanism to make its predictions. The approach also uses latent variables [61] to represent stroke and player information, which can be used for match and play-style analysis. Experimental results show that RallyTemPose outperforms other baseline models [161, 239, 255] in standard and top-3 accuracy on both the ShuttleSet [265] and BadminDB [14] datasets. The results also show that the model's prediction prowess reflects its ability to select

the most logical outcomes, and that incorporating skeleton-motion and player-specific information improves the prediction logic compared to the baselines, specifically concerning the inclusion of the player ground position, emerging as the most critical factor in the model's performance.

3.2.3 Explainable assisting systems

Martin et al. [159] analyzed instead tackle-collision relationships to both improve the understanding of injury aetiology for risk assessment in contact-based sports, and assist referees by automatically flagging high-risk tackles. Their methodology starts by detecting the ball and players in a video using two YOLOv4 networks. The boundary box center coordinates of the ball and players are then filtered using a Kalman Filter [264] to track the ball and determine the tackle frame, ball-carrier, and tackler. Then, both tackler and ball-carrier boundary box coordinates are passed through the OpenPose network to produce key head points. The head-centers of the ball-carrier and tackler are then compared and evaluated to determine the tackle type, with a high-risk tackle defined as the orientation where the head-centers align, and a low-risk tackle defined as the orientation where the head-centers do not align. Overall, experimental results on a privately-collected dataset of royalty free images regarding rugby suggest that the system performs well in detecting high-risk tackles, yet there seems to be room for improvement, particularly in terms of precision due to the imbalance between low- and high-risk injuries and the presence of strong occlusions.

Sarkar et al. [220] suggested a method for detecting valid passes in soccer videos using dual interacting reinforcement learning (RL) agents. The proposed method takes a broadcast soccer video as input and initializes a temporal window on it. The agents, one for localization and one for identification, communicate with each other through a modified Dueling DQN architecture [268]. The localization agent is responsible for identifying the location of the ball and the possessing player, while the identification agent determines whether a pass is valid or not. Experiments performed on online videos show that the proposed dual interacting agent model shows competitive results compared to the state-of-the-art method [219] in generating possession statistics. Although the proposed method has a marginally higher error, it is significantly faster, demonstrating the effectiveness of the proposed framework.

Held et al. [99] focused on proposing an automated soccer decision-making system able to provide real-time feedback and guide referees in making informed decision and positively impacting the outcome of a game, hence capable of detecting and understanding visual relationship between entities. The system, called Video Assistant Referee System (VARS), takes multiple video clips from different views as input and extracts spatio-temporal features from each clip using a video encoder [76] and is designed to perform two tasks: fine-grained foul classification and offense severity classification. The feature vectors are then aggregated using a max or mean aggregation function to obtain a single multi-view representation. This representation is then fed into a classification head, which consists of two dense layers with softmax activation, to output a probability vector for the classification tasks. The authors proposed SoccerNet-MVFouls, which gathers 3901 action extracted from 500 soccer games retrieved from the SoccerNet datasets [49, 86]. Experimental results on such dataset highlight that performance varies considerably across classes, struggling to distinguish between genuine and deceptive

actions during a soccer game due to their shared characteristics. Despite the promising performances, some limitations still arise regarding offence classification, with VARS being likely to make bad predictions in neighboring classes.

Held et al. [100] further extended their previous framework by proposing X-VARS, which aimed at introducing explainability via multi-modal large language models (LLMs). It uses Video-ChatGPT as its foundation model, which is capable of understanding and generating detailed conversations about videos. The architecture consists of several components, including a CLIP ViT-L/14 model for extracting frame feature vectors and hidden states from input video clips. The model is fine-tuned on the SoccerNet-MVFoul [99] dataset to learn prior knowledge about football. The training process involves two stages: fine-tuning the CLIP model and then training a linear projection layer and a LLM to map spatio-temporal features into the same dimensional space as word embeddings. The model also incorporates explainability techniques, such as LIME [211], SHAP [152], and Grad-CAM [226], to provide insights into its decision-making process. Experimental results show that overall X-VARS performs promisingly in designing explanations for fouls and severity recognition decisions, making it a more transparent and trustworthy system and demonstrating that the model is capable of generating high-quality explanations that are comparable to those provided by human referees.

3.3 Performance assessment

This group surveys how action recognition is used to evaluate athlete performance and support recovery. We organize methods by their primary method, comprising: (i) deep-representation analysis, covering learned spatio-temporal and 3D-pose models that extract kinematic metrics and latent embeddings for scoring; (ii) multi-modality fusion, combining multi-modal input data formats to improve robustness and capture complementary contextual signals; and (iii) rule-based and symbolic approaches, describing interpretable pipelines that either mine or encode rules for transparent understanding. A summary of the considered works is presented in Tables 11 and 12.

3.3.1 Deep-representation analysis

Piergiovanni and Ryoo [200] explored the idea of detecting and predicting injuries in baseball pitchers using video data. Based on the considerations that early injury prediction can reduce severity and recovery time, the proposed method uses a 3D spatio-temporal CNN [38], trained on optical flow frames cropped to focus on the pitcher, aimed at detecting injuries in pitchers from videos. Empirical results on a collected dataset comprising TV broadcast MLB videos demonstrate that, overall, the model is able to capture spatio-temporal pitching motions, allowing it to to generalize to unseen pitcher injuries, and behaves promisingly concerning the prediction of injuries, even without seeing a specific pitcher with an injury.

Ludwig et al. [150] addressed the issue of standard skeleton definition in common Human Pose Estimation (HPE) datasets, that is, the shortcomings of requiring time-consuming manual annotations if more (or different) keypoints are needed for performance assessment. The proposed method involves generating ground truth keypoints on human limbs using segmentation masks of upper arms, forearms, thighs, and lower legs [68, 135]. To detect arbitrarily selected keypoints, two approaches are proposed: the vectorized keypoint approach and the norm pose approach. The vectorized keypoint

Table 11 Summary overview of the considered articles concerning performance assessment

Article	Aim	Dataset	Proposal	Advantage	Drawback
Piergiovanni and Ryoo [200]	Motion capture	Piergiovanni and Ryoo [200]*	Deep representation	Captures temporal evolution	Less explainable
Ludwig et al. [150]	Motion capture	Ludwig et al. [150]*	Deep representation	Extending human pose estimation	Limitations under extreme poses
Ludwig et al. [151]	Motion capture	Ludwig et al. [150] Ludwig et al. [151]*	Deep representation	Extending human pose estimation	Limitations under extreme poses
Suzuki et al. [240]	Motion capture	Suzuki et al. [240]*	Deep representation	Readily adaptable to other sports	Requires multi-view estimations
Yeung et al. [282]	Motion capture	Yeung et al. [282]*	Deep representation	Seamless integration	More computationally expensive
Qazi and Iqbal [202]	Personalised rehabilitation	Miron et al. [173]	Deep representation	Biomechanical insights	Relies on specific sensor technology
Wu et al. [272]	Learning	Wu et al. [272]*	Multi-modal fusion	Visual mimicking	Custom set-up
Evans et al. [75]	Motion capture	Evans et al. [75]*	Multi-modal fusion	Non-invasive approach	Application specific
Hachmann and Rosenhahn [92]	Motion capture	Hachmann and Rosenhahn [92]*	Multi-modal fusion	Adaptable to all kinds of surface tracking	Custom set-up
Li et al. [132]	Personalized rehabilitation	Li et al. [132]*	Multi-modal fusion	Benchmark dataset	Focused on skeletal data

*Dataset privately collected by the authors themselves

Table 12 Summary overview of the considered articles concerning performance assessment *(ctd.)*

Article	Aim	Dataset	Proposal	Advantage	Drawback
Ogata et al. [187]	Quality assessment	Ogata et al. [187]*	Rule-based reasoning	Individual-agnostic information	Limitations under extreme poses
Nekoui et al. [180]	Quality assessment	Parmar and Morris [194]	Rule-based reasoning	Robust against contorted poses	Application specific
Sarlis et al. [223]	Injury burden	Sarlis et al. [223]*	Rule-based reasoning	Pattern recognition	Application specific
Dittakavi et al. [65]	Pose correction	Verma et al. [260] Dittakavi et al. [65]*	Rule-based reasoning	Explainable assessment	Limitations under extreme poses
Sarlis and Tjortjis [222]	Injury burden	Sarlis and Tjortjis [222]*	Rule-based reasoning	Pattern recognition	Application specific
Sarlis et al. [224]	Injury burden	Sarlis et al. [224]*	Rule-based reasoning	Pattern recognition	Application specific
Okamoto and Parmar [189]	Quality assessment	Xu et al. [275]	Rule-based reasoning	Explainable assessment	Domain expertise required

*Dataset privately collected by the authors themselves

approach represents each keypoint as a combination of the projection point encoded in a keypoint vector and the thickness encoded in a thickness vector. The norm pose approach encodes each keypoint in normalized Euclidean coordinates on a norm pose (considered as a t-shape template pose). Qualitative and quantitative results on a subset of the standard COCO dataset [137] show that both approaches allow for the detection of desired arbitrary points on the limbs without additional annotations or

post-processing steps. The authors further experimented their framework in Ludwig et al. [151], addressing the task of arbitrary keypoints detection on the body of high, long and triple jump athletes. For such a reason, they release the jump-broadcast dataset, consisting in a collection of videos of competitions from professional broadcast TV footage. Empirical results indicate that thighs, knees, and head are the body parts with the best scores over all experiments, which can be attributed to the larger and thicker segmentation masks making it easier for the network to learn and detect precise keypoints, with the model performing surprisingly well on elbow keypoints compared to upper arm and forearm, despite the similar thickness.

Suzuki et al. [240] proposed a system to create a cost-effective motion capture system for sports motion using unsupervised fine-tuning of a monocular 3D pose estimation model. The process involves three steps: first, multi-view videos are input into a 2D pose estimation model [276] to estimate 2D keypoints, which are then input into a monocular 3D pose estimation model [165] to estimate 3D keypoints. Next, pseudo-labels are generated by triangulating the 2D keypoints with extrinsic camera parameters obtained from automatic camera calibration. Finally, the monocular 3D pose estimation model is fine-tuned using the triangulated 3D keypoints as pseudo-labels. Overall, the study demonstrates the effectiveness of the proposed unsupervised fine-tuning method for improving the performance of monocular 3D pose estimation, making the system capable to provide accurate 3D pose estimation with a reduced number of cameras and without the need for explicit camera calibration or labeled data.

On a differet note, Yeung et al. [282] suggested a novel framework for automated 3D posture analysis of soccer shot movements. Their method, AutoSoccerPose, aims to extract 3D poses from professional broadcast videos in order to further analyze them. It consists of several stages, including broadcast video processing, tracking, 2D and 3D pose estimation, and posture analysis. The method uses a combination of existing computer vision tasks in soccer, including tracking-by-detection and 2D-3D lifting approaches. Specifically, it adopts the RTMPose model [116] for 2D pose estimation and the MotionAGFormer model [165] for 3D pose estimation. On top of that, the method also proposes a non-linear graph-based sequential model, called 3DSP-GRAE, for posture analysis and sequences embedding. The authors introduced the 3D Shot Posture dataset (3DSP), serving as a benchmark for pose estimation in professional soccer broadcast videos, assessing the overall efficacy of AutoSoccerPose across each stage of the framework. Worth of notice is finally the block structure of AutoSoccerPose, which makes it capable of seamlessly integrate with other (sub-)models without requiring any finetuning.

Qazi and Iqbal [202] introduced a novel framework to analyze joint movements for personalized sports therapy. The proposed method utilizes a two-stage process to analyze human motion and body mechanics. First, a Random Forest Classifier (RFC) predicts the top priority joints based on the specific exercise or movement. Then a VQ-VAE, i.e. a Vector Quantized Variational Autoencoder [190], takes key-point estimation and depth maps data of individuals in images as input, focusing on the top priority joints identified by the RFC. The VQ-V AE learns a latent representation of the dataset and identifies the minimal essential joints that differentiate healthy from pathological movement patterns. This approach enables the distillation of complex biomechanical data into actionable insights, resulting in rehabilitation strategies that are both personalized

and automated. Experimental results on a dataset derived from the IntelliRehabDS [173] show a comparative analysis of healthy and unhealthy subjects' movements across nine different exercises performed in standing and sitting positions, with the analysis revealing a clear spatial dichotomy between divergent movement qualities, where the largest centroid distances is exhibited by the quantized embeddings. The results also show a high correlation with ground truth displacement fields derived from kinematic data, highlighting the robustness of the model. Moreover, the precise delineation of critical joints through RFC and VQ-VAE analyses significantly enhances the understanding of human movement biomechanics, offering pivotal insights into the substratum of healthy versus compensatory motion patterns.

3.3.2 Multi-modal fusion

Wu et al. [272] focused on proposing Virtual Reality (VR) vision augmentations as a learning methodology from a recorded expert skier motion. That is, the proposed method aims to enable a vision-based skill transfer between professional skiers and learners [186]. The system uses an indoor ski simulator, a VR system, and tracking sensors to capture the motion of skis and allow users to control them on a virtual slope. The method involves replaying the recorded motion of a professional skier as a virtual leading skier, providing visual cues to support users in following the expert's trajectory correctly. Among these, visual cues list a pose breakdown, which shows the sequential poses of the professional, and an expert shadow, which continuously shows the target position of the user. Additionally, a trail visualization is introduced as a more lightweight alternative, showing the lateral movement and rotation of the expert's skis. The authors' goal is therefore to provide a more natural and less invasive visualization of the expert's temporal motion, allowing users to learn from the professional skier's movements. Ad hoc on-site experiments highlight interestingly enough that additional feedback may not always be helpful as it can draw attention away from the task, with single trails getting more appreciation with respect to full poses breakdown. This is further validated by qualitative observations concerning task-specific metrics such as ankle rotation or lateral movement.

Evans et al. [75] addressed the task of velocity measurement of skeleton training by looking for a non-invasive vision based approach. Their solution involves using a multi-camera setup to capture images of a skeleton athlete and their sled as they move along a track [136]. The method first detects the corners of the sled and then uses this information to track the sled's movement. The corner detections are back-projected to a plane, while the method checks for consistency among the corner points to determine the most likely configuration of the sled, with such configuration being used to initialize the position of the sled and label the corner points. The method is therefore finally able to capture representative velocity data of the athlete's mass center and sled, providing a viable and accurate alternative to marker-based motion capture. Empirical results on privately-collected push trials clips showed very good agreement between the proposed system and the ground truth data (retrieved via marker-based motion capture), with a low bias for both athlete and sled step velocities. Quantitatively, the standard deviations fell within the limits of agreement, hence supporting the validity of the proposed method.

Hachmann and Rosenhahn [92] presented a marker-based multi-view human spine tracking method for spinal kinematic assessment. The authors propose the use of a

marker-based tracker that utilizes perforated kinesiology tape with a unique pattern of alternating rows of dots. The method first detects the kinesiology tape in images using a Mask R-CNN [97] and then finds the dots within the tape using a blob detector. The dots from multiple camera views are then mutually triangulated to create a 3D point cloud, and edges of appropriate lengths are found within the cloud. A linear program is used to select the most likely edges, and a 3D-based alignment Markov Random Field (MRF) model is iteratively registered to the selected edges. Finally, a second MRF optimized by particle belief propagation refines the model to all camera views and outputs the estimated 3D marker coordinates. A series of experiments conducted on both artificial and studio sequences show that the proposed tracking pipeline achieves a high level of precision in tracking markers on the human body with respect to other state-of-the-art methods [24, 145], even in the presence of strong occlusions.

Li et al. [132] described a study on movement quality assessment for rehabilitation purposes. Such a study consists in a multi-task and multi-modality dataset to evaluate movement quality based on three dimensions: completeness, correction, and smoothness, based on a 0-4 scale to assess multiple specific rehabilitative actions based on joint-specific performance. The collected dataset aims to enhance action recognition technology by assimilating spatio-temporal skeletal features from diverse representation subspaces, leading to improved cross-dataset generalization capabilities. The authors undertake a comprehensive comparison between two distinct skeleton-based action recognition methods [279, 280] to provide insights into their efficacy in various application scenarios, aiming for action quality assessment by probing into the subtleties of movement efficacy and accuracy.

3.3.3 Rule-based and symbolic approaches

Ogata et al. [187] addressed the issue of correcting and guiding human poses while self-working out by utilizing video data only, with particular focus on squatting. The authors' idea is based on exploiting pose information to improve generalization in squat classification. It uses distance matrices to represent poses, which are independent of orientation and location. The method concatenates distance matrices for all frames in a video to obtain a matrix where each column represents a pose and the y-axis represents time, adopting then one-dimensional convolutions to perform classification. Being based on ResNet [96] with a recursive structure, it treats temporal features using residual networks. Experimental results on privately-collected datasets show that this approach allows for improved accuracy and generalization in squat classification with respect to existing methods [266], showing also stronger robustness against background changes.

Nekoui et al. [180] addressed the problem of human judging support for sport scenarios presenting highly contorted poses such as diving. The proposed method, called FALCONS, is an engine for grading Olympic diving athletes based on execution and difficulty assessors. The execution evaluation is based on both visual and pose features of the action, similar to what human judges do. The method introduces the ExPose dataset to handle the estimation of pose in extreme body configurations and uses a pose estimation network trained on this dataset [279]. The extracted pose sequences are used by a bridge connector module to increase the contribution of the splash scene among other appearance clues. A simple assessor is proposed to work on the basis of pose features to extract the difficulty of the action. Finally, the overall score is provided by the

multiplication of the execution and difficulty scores. Empirical results show state-of-the-art performance compared to previous studies [191, 195] and acceptable generalization to unseen scenes from other sports, demonstrating the effectiveness of the proposed engine regarding the grading process.

Sarlis et al. [223] applied data science and machine learning techniques to investigate how injuries influence individual and team performance in the NBA, discovering that musculoskeletal injuries were the most common and had a significant effect on performance. The study also compared several machine learning models for their predictive accuracy in forecasting player absences due to injuries. Their work is further extended in Sarlis and Tjortjis [222], with the authors presenting a comprehensive data-driven analysis of NBA players over 24 seasons, aiming to evaluate the influence of player demographics, injuries, and positions on both performance and financial outcomes, identifying musculoskeletal injuries as the most financially burdensome since they account for nearly half of injury-related costs. Similarly, the authors explored in Sarlis et al. [224] how data science, sports analytics, and association rule mining can uncover relationships between NBA player injuries, recovery times, salaries, and team financial performance. Using over two decades of NBA data, the study applies data mining techniques to identify patterns in recovery times and financial outcomes. Key findings indicate strong correlations between low-salary players and higher team losses, and show that specific physical characteristics and positions are linked to longer recoveries and greater financial burden.

Dittakavi et al. [65] combined vision and pose skeleton models to obtain an explainable AI-based pose detection and correction system. Their framework, Pose Tutor, is a system for yoga pose classification that uses a coarse-to-fine framework to predict the pose class of an input image. More in detail, the system first uses a pre-trained vision model to obtain a coarse image-level prediction, which is then refined using a pose skeleton model. The pose skeleton model uses a pre-trained pose estimator to obtain noisy pose keypoints from the input image, and then generates a pose vector that summarizes the pose predicted by the keypoint prediction model. A K-Nearest Neighbors classifier is then used to predict the pose class based on the pose vector. For explainability [270], the system also uses a likelihood-based rationale generation method to explain the pose class predictions made by Pose Tutor. The explanation mechanism uses the pose vectors obtained for each training image to generate an angle distribution for each pose class, which is then used to identify the angles that contributed most to the output classification. Architectural experiments on the Yoga-82 dataset [260] and other two privately-curated datasets (respectively, Pilates-32 and Kungfu-7) highlight the promising behavior of Pose Tutor in terms of overall accuracy and explainability by denoting the joints with higher and lower impact concerning pose class prediction.

Okamoto and Parmar [189] focused on proposing a fully-automatic approach aimed at providing transparent, objective, and explainable scoring, addressing the limitations of traditional human judging and current fully neural model-based Action Quality Assessment (AQA) approaches [56]. Their idea consists in a neuro-symbolic approach [81] that combines the strengths of deep neural models and rules-based AI for fine-grained analysis of human movements and actions, specifically in the sport of diving. The approach first deconstructs the performance and its surrounding environment into interpretable "symbols" using neural models, and then uses a rules-based approach to construct a

hierarchy of representations, starting from identifying dive type and temporal segments to fine-grained action quality assessment and scoring. The system programmatically generates detailed performance analysis reports that check for all performance errors, compute their magnitudes, and provide an overall score obtained by aggregating the percentiles of all aspects of the dive using uniform weighted averaging. Experimental results showed that the NS-AQA system achieved high accuracy in fine-grained action recognition, outperforming other state-of-the-art methods [195] on the FineDiving dataset [275], with feedback from an Olympic diving coach and other experts showing that they agreed with the system's overall scores and individual error scores at least 90% of the time, highlighting this way the system's practical usefulness in diving competitions and training.

4 Discussion and remarks

By considering the analysis of the studies, it is clear that AI-based computer vision algorithms play a vital role in the area of sports video processing. Still, severe limitations are present when considering possible commercial applications and real-world requirements:

- Quantitatively determining the performance benchmark of algorithms is challenging due to the lack of standardized databases with ground truth data across different sports, which vary in numerous ways. Furthermore, factors such as the use of different video capture devices and variations in their settings add complexity, making it difficult to build an effective object tracking system for sports. Even if there are some proposals that are gaining more and more recognition, no benchmarking has imposed itself as standard yet. It is expected nonetheless that in the near future, considering the increasing interest towards the field, more and more all-encompassing and general proposals will rise, leading to a common standard able to provide solid comparisons among the deployed techniques;
- It is moreover undeniable that the "AI hype" trend has taken over the sports industry as well. Rising upon the encouraging results that fully-neural methods are capable of achieving, nearly all proposed techniques present some level of deep learning integration in them. A blind faith towards these models may turn out to be counter-productive in commercial scenarios, where the often noisy and unpredictable behavior of neural approaches could lead to unpleasant outcomes. It is expected that, sooner or later, the recent rise of explainable and trustworthy AI will impact this field as well, providing a more transparent decision-making process and a more fair integration with non-neural procedures;

4.1 Physical understanding

Current approaches to physical understanding in AI-based computer vision for sports are limited by several technical and practical challenges. The dynamic and unpredictable nature of sports scenes presents significant obstacles, as rapid player movements, frequent occlusions, and overlapping interactions complicate the accurate detection and tracking of entities. These issues often result in identity switching, fragmented trajectories, and reduced tracking continuity, particularly in fast-paced or crowded scenarios.

Camera calibration. Camera calibration is foundational for any spatial reasoning in sports video and therefore its limitations propagate to all downstream tasks.

The surveyed methods show two recurring practical implications: template-matching approaches, which either retrieve poses from prebuilt template libraries or learn a search-space, often enable faster initialization and homography prediction than geometry-based methods. Yet, they rely on ad-hoc prior knowledge about the views and limit out-of-the-box applicability to novel sports or unusual viewpoints. Moreover, there is a wide heterogeneity in evaluation procedures that complicates cross-method comparison. In this sense, particularly worth of notice is the work by Magera et al. [154], which suggested a universal protocol to benchmark camera calibration for sports, given the vast heterogeneity in evaluation procedures arose during the last years. The proposed method, named ProCC, provides an objective evaluation of calibration methods based on reprojection of arbitrary yet accurately known 3D objects. The protocol is designed to be agnostic to the camera model chosen for a camera calibration method but still relies on semantic point annotations of sports field markings as ground truth, which are valid for any type of camera.

Detection and identification. Detecting and identifying entities usually comprise the most pivotal aspects in the inference pipeline. This survey groups recent approaches into multi-modal fusion, skeleton-driven modeling, deep-feature representation and data-efficient learning. Each cluster brings concrete challenges: modality alignment and latency for fusion, severe occlusions and viewpoint or pose variability that undermine skeleton priors, the opacity and dataset-specific brittleness of end-to-end deep embeddings, and the risk that teacher-student or distillation schemes propagate teacher biases if unlabeled data or teacher quality are limited. Yet, practically, fusion and skeleton priors improve detection under ambiguity and motion-centric tasks, deep representations overcome the limitations of hand-crafted feature extractors, and data-efficient approaches such as online distillation or semi-supervision substantially reduce manual labeling needs for match-specific deployment. Sstem designers are therefore expected to balance accuracy, explainability, annotation budgets and latency when adopting any single cluster or hybrid pipeline.

Tracking. Tracking can be thought as maintaining consistency across time for frame-by-frame detection and identification. Current approaches are clustered into optimization-based associations, geometric reconstruction and deep-features appearance. Optimization methods impose global spatio-temporal constrained optimization that reduce identity switches and fragmented tracks but are often computationally heavy and sensitive to modeling choices. Geometric reconstruction approaches supply instead physically grounded 3D recovery and strong multi-view priors but require calibration or multiple views and degrade when those assumptions fail. Deep-feature methods excel at rescuing tracks in crowded or low-texture scenes and enable, on the contrary, (near-) real-time pipelines with efficient backbones, but demand substantial labeled data and are not as robust under domain shift while reducing interpretability.

4.1.1 Logical understanding

One of the primary challenges of logical understanding in AI-based computer vision for sports lies in the complexity and variability of sports activities, where interactions between players, objects, and the environment are dynamic and context-dependent. Many existing methods struggle to bridge the gap between detecting entities and understanding their roles and actions in the broader context of a game. Action recognition

models, while effective in controlled settings, often falter in real-world scenarios, limiting the generalizability of current approaches.

Event spotting. The thematic clusterization of event spotting approaches in skeleton-driven modeling, spatio-temporal modeling and multi-modal fusion neatly exposes complementary design choices: skeleton methods yield compact, interpretable priors for fine-grained spotting; spatio-temporal models capture the contextual dynamics required for multi-step plays; and multi-modal fusion improves robustness by incorporating audio/telemetry or language cues. While keypoint-driven methods are obviously sensitive to pose errors and occlusions, they do not suffer from the compute burden of spatio-temporal models and the reliance of multi-modal fusion on synchronized side channels that may be unavailable or noisy, making them more viable for live analysis, interactive applications or broadcasting requirements. In this sense, Cabado et al. [33] assessed the transfer learning capabilities concerning action spotting across diverse sport domains, analyzing the performance of five existing methods for action spotting tasks [48, 60, 85, 86, 104] and providing a detailed description of the implementation details and hyperparameters to ensure reproducibility and consistency across experiments.

Visual relationship detection and prediction. Visual relationship detection and prediction moves beyond event spotting to explicitly model interactions between entities and their temporal dependencies. The thematic division in skeleton-driven modeling, simulation forecasting and and explainable assistance both links such a task to the one of event spotting, but also project it towards forecasting. By focusing not only on identifying events but also on the predicates that connect them, these methods enable a richer interpretation of visual data, bridging the gap between perception and higher-level reasoning.

Performance assessment. While deep representations and multi-modal fusion show promise for automated assessment, their practical validity is limited by heterogeneous, often privately-collected datasets and the lack of unified benchmarking and ground-truth protocols, which hampers cross-study comparability and generalization. Moreover, deploying performance-assessment systems for injury prediction, rehabilitation or large-scale broadcast analysis raises concrete privacy and consent concerns, underscoring the need for explicit governance discussion. In this direction, we welcome the rise of rule-based and neuro-symbolic approaches, which complement deep perception modules by producing interpretable, structured outputs that map directly to human concepts used in scoring and coaching, underlying the need for transparency, interpretability and ethical safeguards.

5 Conclusion

The integration of AI-based CV algorithms in the sports industry represents a transformative leap forward, offering unprecedented opportunities to analyze, interpret and enhance various aspects of sports. Despite the significant progress made in both *physical* and *logical* understanding, challenges such as dynamic environments, occlusions and limited generalizability highlight the need for continued innovation. Promising trends are nonetheless emerging with the goal of mitigating such drawbacks.

Multi-modal approaches, combining visual data with additional inputs such as audio, textual annotations or player telemetry, are showing encouraging results in providing a richer and more comprehensive understanding of a scene. The development of

self-supervised and weakly-supervised learning techniques is another significant trend. These methods allow models to learn directly from raw video data without extensive manual labeling, making it easier to generalize across different sports and scenarios. Lastly, the emergence of explainable AI techniques is helping to make action recognition and event detection models more interpretable and reliable. By providing insights into how decisions are made, these methods increase trust in the systems and facilitate their integration into critical applications such as coaching, refereeing, and player performance evaluation.

Although many of the approaches reviewed cannot be considered real-time today, it is reasonable to expect steadily faster inference going forward. Advances in architectural design, model compression and smarter engineering will increasingly meet real-time or near-real-time needs in production. Yet, in this sense, broadcast and live-production companies will still need to strike a pragmatic balance between the richer capabilities that multi-modal systems provide and the operational simplicity required for rapid, reliable deployment in unpredictable live environments. That means prioritizing the modalities that deliver clear value, favoring lightweight yet robust models and standardized deployment stacks, while likely still keeping human-in-the-loop fallbacks. It is fundamental therefore to rally for publicly available, well-curated standard datasets and harmonized benchmarking protocols so researchers, vendors, and regulators can meaningfully compare methods and track progress. Open, diverse, and responsibly governed datasets, with clear annotation guidelines, standardized metrics and privacy-preserving sharing mechanisms, would improve reproducibility, surface biases, and speed safe deployment across real-world settings.

Concerning safe deployment, the on-going reduction in both memory footprint and inference time will also likely lead to a more widespread adoption of edge devices with incorporated AI capabilities. This sharply amplifies concerns about data privacy and ownership: on-device processing, strong encryption, transparent consent, clear ownership and portability rules will therefore be essential to help athletes and teams without putting their personal data at risk.

All these developments not only hold the potential to redefine player performance analysis, game strategy evaluation and fan engagement, but also open doors for interdisciplinary applications that extend beyond traditional sports. As research continues to address existing limitations and explore uncharted territories, the synergy between computer vision and sports will undoubtedly lead to profound advancements, shaping the future of how sports are played, experienced and understood.

Author contributions
LFR conceptualized and drafted this work. AS, FM and MDB equally revised it critically and approved the version to be published.

Funding
No funding was received to assist with the preparation of this manuscript.

Availability of data and materials
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors have no competing interests to declare that are relevant to the content of this article.

Received: 30 June 2025 / Accepted: 9 October 2025
Published online: 08 December 2025

References

1. Abdel-Aziz YI, Karara HM, Hauck M. Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. Photogramm Eng Remote Sens. 2015;81(2):103–7. https://doi.org/10.14358/PERS.81.2.103.
2. Adesida Y, Papi E, McGregor AH. Exploring the role of wearable technology in sport kinematics and kinetics: a systematic review. Sensors. 2019;19(7):1597. https://doi.org/10.3390/s19071597.
3. Agrawal A, Verschueren R, Diamond S, et al. A rewriting system for convex optimization problems. J Control Decis. 2018;5(1):42–60. https://doi.org/10.1080/23307706.2017.1397554.
4. Aharon N, Orfaig R, Bobrovsky B. Bot-sort: Robust associations multi-pedestrian tracking. CoRR. 2022. https://doi.org/10.48550/ARXIV.2206.14651, arXiv:2206.14651
5. Andriluka M, Roth S, Schiele B. People-tracking-by-detection and people-detection-by-tracking. 2008 IEEE Conference on Computer Vision and Pattern Recognition. 2008:1–8. https://doi.org/10.1109/CVPR.2008.4587583
6. Apostolidis E, Adamantidou E, Metsai AI, et al. Video summarization using deep neural networks: a survey. Proc IEEE. 2021;109(11):1838–63. https://doi.org/10.1109/JPROC.2021.3117472.
7. Arandjelovic R, Gronat P, Torii A, et al. Netvlad: CNN architecture for weakly supervised place recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:5297–5307. https://doi.org/10.1109/CVPR.2016.572
8. Arbués-Sangësa A, Martín A, Fernández J, et al. Using player's body-orientation to model pass feasibility in soccer. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020a:3875–3884. https://doi.org/10.1109/CVPRW50498.2020.00451
9. Arbués-Sangësa A, Martín A, Fernández J, et al. Always look on the bright side of the field: Merging pose and contextual data to estimate orientation of soccer players. 2020 IEEE International Conference on Image Processing (ICIP). 2020b:1506–1510. https://doi.org/10.1109/ICIP40778.2020.9190639
10. Askari F, Ramaprasad R, Clark JJ, et al (2022) Interaction classification with key actor detection in multi-person sports videos. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp 3579–3587 https://doi.org/10.1109/CVPRW56347.2022.00402
11. Askari F, Jiang R, Li Z, et al (2023) Self-supervised video interaction classification using image representation of skeleton data. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp 5229–5238 https://doi.org/10.1109/CVPRW59228.2023.00551
12. Askari F, Yared C, Ramaprasad R, et al (2024) Video interaction recognition using an attention augmented relational network and skeleton data. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp 3225–3234. https://doi.org/10.1109/CVPRW63382.2024.00328
13. Balaji B, Bright J, Prakash H, et al. Jersey number recognition using keyframe identification from low-resolution broadcast videos. Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports. 2023:123–130. https://doi.org/10.1145/3606038.3616162
14. Ban KW, See J, Abdullah J, et al. Badmintondb: A badminton dataset for player-specific match analysis and prediction. Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports. 2022:47–54. https://doi.org/10.1145/3552437.3555696
15. Baumgartner T, Paassen B, Klatt S. Extracting spatial knowledge from track and field broadcasts for monocular 3d human pose estimation. Sci Rep. 2023;13(1):14031. https://doi.org/10.1038/s41598-023-41142-0.
16. Baumgartner T, Klatt S. Monocular 3d human pose estimation for sports broadcasts using partial sports field registration. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2023:5109–5118. https://doi.org/10.1109/CVPRW59228.2023.00539
17. Bautista D, Atienza R. Scene text recognition with permuted autoregressive sequence models. Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII. 2022:178–196. https://doi.org/10.1007/978-3-031-19815-1_11
18. Beal R, Norman TJ, Ramchurn SD. Artificial intelligence for team sports: a survey. Knowl Eng Rev. 2019;34:e28. https://doi.org/10.1017/S0269888919000225.
19. Belagiannis V, Amin S, Andriluka M, et al. 3d pictorial structures for multiple human pose estimation. 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014:1669–1676. https://doi.org/10.1109/CVPR.2014.216
20. Berclaz J, Fleuret F, Turetken E, et al. Multiple object tracking using k-shortest paths optimization. IEEE Trans Pattern Anal Mach Intell. 2011;33(9):1806–19. https://doi.org/10.1109/TPAMI.2011.21.
21. Bertasius G, Feichtenhofer C, Tran D, et al. Learning discriminative motion features through detection. CoRR abs/1812.04172. 2018. arXiv:1812.04172
22. Bewley A, Ge Z, Ott L, et al. Simple online and realtime tracking. 2016 IEEE International Conference on Image Processing (ICIP). 2016:3464–3468. https://doi.org/10.1109/ICIP.2016.7533003
23. Bilen H, Fernando B, Gavves E, et al. Dynamic image networks for action recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:3034–3042. https://doi.org/10.1109/CVPR.2016.331
24. Bogo F, Romero J, Pons-Moll G, et al. Dynamic faust: Registering human bodies in motion. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:5573–5582. https://doi.org/10.1109/CVPR.2017.591
25. Bommes L, Lin X, Zhou J. Mvmed: Fast multi-object tracking in the compressed domain. 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA). 2020:1419–1424. https://doi.org/10.1109/ICIEA48937.2020.9248145
26. Bonidia RP, Rodrigues LAL, Avila-Santos AP, et al. Computational intelligence in sports: a systematic literature review. Adv Human-Comput Interact. 2018;1:3426178. https://doi.org/10.1155/2018/3426178.
27. Breiman L. Bagging predictors. Mach Learn. 1996;24:123–40. https://doi.org/10.1007/BF00058655.
28. Bridgeman L, Volino M, Guillemaut JY, et al. Multi-person 3d pose estimation and tracking in sports. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019:2487–2496. https://doi.org/10.1109/CVPRW.2019.00304

29. Bright J, Balaji B, Chen Y, et al. Pitchernet: Powering the moneyball evolution in baseball video analytics. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024a:3420–3429. https://doi.org/10.1109/CVPRW63382.2024.00346
30. Bright J, Balaji B, Prakash H, et al. Distribution and depth-aware transformers for 3d human mesh recovery. 2024b. arXiv:2403.09063
31. Buch S, Escorcia V, Shen C, et al. Sst: Single-stream temporal action proposals. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:6373–6382. https://doi.org/10.1109/CVPR.2017.675
32. Cabado B, Cioppa A, Giancola S, et al. Beyond the premier: Assessing action spotting transfer capability across diverse domains. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024:3386–3398. https://doi.org/10.1109/CVPRW63382.2024.00343
33. Cao W, Li Y. Dots: an online and near-optimal trajectory simplification algorithm. J Syst Softw. 2017;126:34–44. https://doi.org/10.1016/j.jss.2017.01.003.
34. Cao J, Pang J, Weng X, et al. Observation-centric sort: Rethinking sort for robust multi-object tracking. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023:9686–9696. https://doi.org/10.1109/CVPR52729.2023.00934
35. Cao Z, Simon T, Wei SE, et al. Realtime multi-person 2d pose estimation using part affinity fields. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:1302–1310. https://doi.org/10.1109/CVPR.2017.143
36. Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021:9630–9640. https://doi.org/10.1109/ICCV48922.2021.00951
37. Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:4724–4733. https://doi.org/10.1109/CVPR.2017.502
38. Cervone D, D'Amour A, Bornn L, et al. Pointwise: Predicting points and valuing decisions in real time with NBA optical tracking data. MIT Sloan Sports Analytics Conference 2014. 2014. https://www.lukebornn.com/papers/cervone_ssac_2014.pdf
39. Chappa NVR, Nguyen P, Nelson AH, et al. Spartan: Self-supervised spatiotemporal transformers approach to group activity recognition. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2023:5158–5168. https://doi.org/10.1109/CVPRW59228.2023.00544
40. Chen M, Xu M, Franti P. A fast $o(n)$ multiresolution polygonal approximation algorithm for GPS trajectory simplification. IEEE Trans Image Process. 2012;21(5):2770–85. https://doi.org/10.1109/TIP.2012.2186146.
41. Chen J, Little JJ. Sports camera calibration via synthetic data. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019:2497–2504. https://doi.org/10.1109/CVPRW.2019.00305
42. Chen YH, Wang CY, Yang CY, et al. Neighbortrack: Single object tracking by bipartite matching with neighbor tracklets and its applications to sports. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2023:5139–5148. https://doi.org/10.1109/CVPRW59228.2023.00542
43. Chu YJ, Su JW, Hsiao KW, et al. Sports field registration via keypoints-aware label condition. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022:3522–3529. https://doi.org/10.1109/CVPRW56347.2022.00396
44. Cioppa A, Deliège A, Giancola S, et al. Scaling up soccernet with multi-view spatial localization and re-identification. Sci Data. 2022;9(1):355. https://doi.org/10.1038/s41597-022-01469-1.
45. Cioppa A, Deliège A, Van Droogenbroeck M. A bottom-up approach based on semantics for the interpretation of the main camera stream in soccer games. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2018:1846–184609. https://doi.org/10.1109/CVPRW.2018.00229
46. Cioppa A, Deliège A, Istasse M, et al. Arthus: Adaptive real-time human segmentation in sports through online distillation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019:2505–2514. https://doi.org/10.1109/CVPRW.2019.00306
47. Cioppa A, Deliege A, Ul Huda N, et al. Multimodal and multiview distillation for real-time player detection on a football field. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020a:3846–3855. https://doi.org/10.1109/CVPRW50498.2020.00448
48. Cioppa A, Deliège A, Giancola S, et al. A context-aware loss function for action spotting in soccer videos. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020b:13123–13133. https://doi.org/10.1109/CVPR42600.2020.01314
49. Cioppa A, Deliège A, Magera F, et al. Camera calibration and player localization in soccernet-v2 and investigation of their representations for action spotting. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2021. https://doi.org/10.1109/CVPRW53098.2021.00511
50. Cioppa A, Giancola S, Deliège A, et al. Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022b:3490–3501. https://doi.org/10.1109/CVPRW56347.2022.00393
51. Cipolla R, Gal Y, Kendall A. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018:7482–7491. https://doi.org/10.1109/CVPR.2018.00781
52. Citraro L, Márquez-Neila P, Savarè S, et al. Real-time camera pose estimation for sports fields. Mach Vis Appl. 2020;31(3):16. https://doi.org/10.1007/s00138-020-01064-7.
53. Cohen C, Texier BD, Quéré D, et al. The physics of badminton. New J Phys. 2015;17(6):063001. https://doi.org/10.1088/1367-2630/17/6/063001.
54. Cust EE, Sweeting AJ, Ball K, et al. Machine and deep learning for sport-specific movement recognition: a systematic review of model development and performance. J Sports Sci. 2019;37(5):568–600. https://doi.org/10.1080/02640414.2018.1521769. (**pMID: 30307362**).
55. Dadashzadeh A, Duan S, Whone A, et al. Pecop: Parameter efficient continual pretraining for action quality assessment. 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2024: 42–52. https://doi.org/10.1109/WACV57701.2024.00012
56. Dalal N, Triggs B. Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). 2005;1:886–893. https://doi.org/10.1109/CVPR.2005.177

57. Decorte R, Paré M, Vanhaeverbeke J, et al. Multi-modal hit detection and positional analysis in padel competitions. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024:3306–3314. https://doi.org/10.1109/CVPRW63382.2024.00335
58. Deliège A, Cioppa A, Giancola S, et al. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2021. https://doi.org/10.1109/CVPRW53098.2021.00508, kAUST Repository Item: Exported on 2021-09-03
59. Denize J, Liashuha M, Rabarisoa J, et al. Comedian: Self-supervised learning and knowledge distillation for action spotting using transformers. 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). 2024:518–528. https://doi.org/10.1109/WACVW60836.2024.00060
60. Devlin J, Chang MW, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019;1:4171–4186. https://doi.org/10.18653/v1/N19-1423
61. Deyzel M, Theart RP. One-shot skeleton-based action recognition on strength and conditioning exercises. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2023:5169–5178. https://doi.org/10.1109/CVPRW59228.2023.00545
62. Dicle C, Camps OI, Sznaier M. The way they move: Tracking multiple targets with similar appearance. 2013 IEEE International Conference on Computer Vision. 2013:2304–2311. https://doi.org/10.1109/ICCV.2013.286
63. Dinov ID. Expectation maximization and mixture modeling tutorial. UCLA: Statistics Online Computational Resource https://escholarship.org/uc/item/1rb70972. 2008.
64. Dittakavi B, Bavikadi D, Desai SV, et al. Pose tutor: An explainable system for pose correction in the wild. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022:3539–3548. https://doi.org/10.1109/CVPRW56347.2022.00398
65. Domahidi A, Chu E, Boyd S. Ecos: An SOCP solver for embedded systems. 2013 European Control Conference (ECC). 2013:3071–3076. https://doi.org/10.23919/ECC.2013.6669541
66. Donahue J, Hendricks LA, Rohrbach M, et al. Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans Pattern Anal Mach Intell. 2017;39(4):677–91. https://doi.org/10.1109/TPAMI.2016.2599174.
67. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. CoRR . arXiv:2010.11929. 2020.
68. Du Y, Zhao Z, Song Y, et al. Strongsort: make deepsort great again. IEEE Trans Multimedia. 2023;25:8725–37. https://doi.org/10.1109/TMM.2023.3240881.
69. Duan H, Zhao Y, Chen K, et al. Revisiting skeleton-based action recognition. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022:2959–2968. https://doi.org/10.1109/CVPR52688.2022.00298
70. Dunnhofer M, Sordi L, Micheloni C. Visualizing skiers' trajectories in monocular videos. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2023:5188–5198. https://doi.org/10.1109/CVPRW59228.2023.00547
71. Einfalt M, Lienhart R. Decoupling video and human motion: Towards practical event detection in athlete recordings. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020:3901–3910. https://doi.org/10.1109/CVPRW50498.2020.00454
72. Einfalt M, Dampeyrou C, Zecha D, et al. Frame-level event detection in athletics videos with pose-based convolutional sequence networks. Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports. 2019:42–50. https://doi.org/10.1145/3347318.3355525
73. Evangelidis GD, Psarakis EZ. Parametric image alignment using enhanced correlation coefficient maximization. IEEE Trans Pattern Anal Mach Intell. 2008;30(10):1858–65. https://doi.org/10.1109/TPAMI.2008.113.
74. Evans M, Needham L, Colyer SL, et al. A non-invasive vision based approach to velocity measurement of skeleton training. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020:3885–3892. https://doi.org/10.1109/CVPRW50498.2020.00452
75. Fan H, Xiong B, Mangalam K, et al. Multiscale vision transformers. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021:6804–6815. https://doi.org/10.1109/ICCV48922.2021.00675
76. Farin D, Krabbe S, de With PH, et al. Robust camera calibration for sport videos using court models. Storage and Retrieval Methods and Applications for Multimedia, San Jose (CA). 2004:80–91. https://doi.org/10.1117/12.526813
77. Ferguson S, McNally W, McPhee J. Predicting the flight of a golf ball: comparing a physic-based aerodynamic model to a neural network. Engineering of Sport: Proceedings of the 14th Conference of the Internatoinal Sports Engineering Association 14. 2022. https://doi.org/10.5703/1288284317493
78. Gal Y, Islam R, Ghahramani Z. Deep bayesian active learning with image data. Proceedings of the 34th International Conference on Machine Learning. 2017;70:1183–1192. https://doi.org/10.5555/3305381.3305504
79. Gallego G, Rebecq H, Scaramuzza D. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018:3867–3876. https://doi.org/10.1109/CVPR.2018.00407
80. Garcez AS, Gabbay DM, Broda KB. Neural-Symbolic Learning System: Foundations and Applications. Springer-Verlag Berlin Heidelberg. 2002. https://doi.org/10.1007/978-1-4471-0211-3.
81. Gava L, Monforte M, Iacono M, et al. Puck: Parallel surface and convolution-kernel tracking for event-based cameras. 2022. https://doi.org/10.48550/arXiv.2205.07657
82. Gerke S, Müller K, Schäfer R. Soccer jersey number recognition using convolutional neural networks. 2015 IEEE International Conference on Computer Vision Workshop (ICCVW). 2015:734–741. https://doi.org/10.1109/ICCVW.2015.100
83. Ghosh A, Singh S, Jawahar CV. Towards structured analysis of broadcast badminton videos. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). 2018:296–304. https://doi.org/10.1109/WACV.2018.00039
84. Giancola S, Ghanem B. Temporally-aware feature pooling for action spotting in soccer broadcasts. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2021:4485–4494. https://doi.org/10.1109/CVPRW53098.2021.00506
85. Giancola S, Amine M, Dghaily T, et al. Soccernet: A scalable dataset for action spotting in soccer videos. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2018:1792–179210. https://doi.org/10.1109/CVPRW.2018.00223

86. Giancola S, Cioppa A, Georgieva J, et al. Towards active learning for action spotting in association football videos. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2023:5098–5108. https://doi.org/10.1109/CVPRW59228.2023.00538
87. Gossard T, Tebbe J, Ziegler A, et al. Spindoe: A ball spin estimation method for table tennis robot. 2023. https://doi.org/10.1109/IROS55552.2023.10342178
88. Gossard T, Krismer J, Ziegler A, et al. Table tennis ball spin estimation with an event camera. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024:3347–3356. https://doi.org/10.1109/CVPRW63382.2024.00339
89. Guillemaut JY, Hilton A. Joint multi-layer segmentation and reconstruction for free-viewpoint video applications. Int J Comput Vis. 2011;93(1):73–100. https://doi.org/10.1007/s11263-010-0413-z.
90. Gutiérrez-Pérez M, Agudo A. No Bells, Just Whistles: Sports Field Registration by Leveraging Geometric Properties. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024:3325–3334. https://doi.org/10.1109/CVPRW63382.2024.00337
91. Hachmann H, Rosenhahn B. Human spine motion capture using perforated kinesiology tape. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2023:5149–5157. https://doi.org/10.48550/arXiv.2306.02930
92. Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). 2006:2:1735–1742. https://doi.org/10.1109/CVPR.2006.100
93. Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3d CNNs retrace the history of 2d CNNs and imagenet? 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018:6546–6555. https://doi.org/10.1109/CVPR.2018.00685
94. Hartley RI, Zisserman A. Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, ISBN: 0521540518. 2004. https://doi.org/10.1017/CBO9780511811685
95. He Y, Wei X, Hong X, et al. Multi-target multi-camera tracking by tracklet-to-target assignment. IEEE Trans Image Process. 2020;29:5191–205. https://doi.org/10.1109/TIP.2020.2980070.
96. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:770–778. https://doi.org/10.1109/CVPR.2016.90
97. He K, Gkioxari G, Dollár P, et al. Mask r-cnn. 2017 IEEE International Conference on Computer Vision (ICCV). 2017:2980–2988. https://doi.org/10.1109/ICCV.2017.322
98. Held J, Cioppa A, Giancola S, et al. Vars: Video assistant referee system for automated soccer decision making from multiple views. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2023:5086–5097. https://doi.org/10.1109/CVPRW59228.2023.00537
99. Held J, Itani H, Cioppa A, et al. X-vars: Introducing explainability in football refereeing with multi-modal large language models. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024:3267–3279. https://doi.org/10.1109/CVPRW63382.2024.00332
100. Homayounfar N, Fidler S, Urtasun R. Sports field localization via deep structured models. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:4012–4020. https://doi.org/10.1109/CVPR.2017.427
101. Honda Y, Kawakami R, Yoshihashi R, et al. Pass receiver prediction in soccer using video and players' trajectories. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022:3502–3511. https://doi.org/10.1109/CVPRW56347.2022.00394
102. Hong J, Fisher M, Gharbi M, et al. Video pose distillation for few-shot, fine-grained sports action recognition. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021:9234–9243. https://doi.org/10.1109/ICCV48922.2021.00912
103. Hong J, Zhang H, Gharbi M, et al. Spotting temporally precise, fine-grained events in video. Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV. 2022:33–51. https://doi.org/10.1007/978-3-031-19833-5_3
104. Host K, Ivašic-Kos M. An overview of human action recognition in sports based on computer vision. Heliyon. 2022;8(6):e09633. https://doi.org/10.1016/j.heliyon.2022.e09633.
105. Huang YC, Liao IN, Chen CH, et al. Tracknet: A deep learning network for tracking high-speed and tiny objects in sports applications. 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). 2019:1–8. https://doi.org/10.1109/AVSS.2019.8909871
106. Hubáček O, Šourek G, Železný F. Deep learning from spatial relations for soccer pass prediction. Machine Learning and Data Mining for Sports Analytics. 2019:159–166. https://doi.org/10.1007/978-3-030-17274-9_14
107. Huda NU, Hansen BD, Gade R, et al. The effect of a diverse dataset for transfer learning in thermal person detection. Sensors. 2020;20(7):1982. https://doi.org/10.3390/s20071982.
108. Huda NU, Jensen KH, Gade R, et al. Estimating the number of soccer players using simulation-based occlusion handling. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2018:1905–190509.https://doi.org/10.1109/CVPRW.2018.00236
109. Ibh M, Grasshof S, Witzner D, et al. Tempose: a new skeleton-based transformer model designed for fine-grained motion recognition in badminton. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2023:5199–5208. https://doi.org/10.1109/CVPRW59228.2023.00548
110. Ibh M, Grasshof S, Hansen DW. A stroke of genius: Predicting the next move in badminton. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024:3376–3385. https://doi.org/10.1109/CVPRW63382.2024.00342
111. Ibrahim MS, Muralidharan S, Deng Z, et al. A hierarchical deep temporal model for group activity recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:1971–1980. https://doi.org/10.1109/CVPR.2016.217
112. Isola P, Zhu JY, Zhou T, et al. Image-to-image translation with conditional adversarial networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:5967–5976. https://doi.org/10.1109/CVPR.2017.632
113. Istasse M, Moreau J, De Vleeschouwer C. Associative embedding for team discrimination. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019:2477–2486. https://doi.org/10.1109/CVPRW.2019.00303
114. Jia X, De Brabandere B, Tuytelaars T, et al. Dynamic filter networks. Advances in Neural Information Processing Systems 29. 2016. https://doi.org/10.5555/3157096.3157171
115. Jiang T, Lu P, Zhang L, et al. Rtmpose: Real-time multi-person pose estimation based on mmpose. 2023. arXiv:2303.07399

116. Jiang W, Gamboa Higuera JC, Angles B, et al. Optimizing through learned errors for accurate sports field registration. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). 2020:201–210. https://doi.org/10.1109/WACV45572.2020.9093581
117. Jiang YG, Liu J, Zamir AR, et al. THUMOS challenge: Action recognition with a large number of classes. 2014. http://crcv.ucf.edu/THUMOS14/
118. Kamble PR, Keskar AG, Bhurchandi KM. Ball tracking in sports: a survey. Artif Intell Rev. 2019;52(3):1655–705. https://doi.org/10.1007/s10462-017-9582-2.
119. Kaneko T, Kawakami R, Naemura T, et al. Augmenting pass prediction via imitation learning in soccer simulations. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024:3194–3203. https://doi.org/10.1109/CVPRW63382.2024.00325
120. Kanojia G, Kumawat S, Raman S. Attentive spatio-temporal representation learning for diving classification. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019:2467–2476. https://doi.org/10.1109/CVPRW.2019.00302
121. Ke Q, An S, Bennamoun M, et al. Skeletonnet: mining deep part features for 3-d action recognition. IEEE Signal Process Lett. 2017;24(6):731–5. https://doi.org/10.1109/LSP.2017.2690339.
122. Koshkina M, Elder JH. A general framework for jersey number recognition in sports video. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. 2024:3235–3244. https://arxiv.org/abs/2405.13896
123. Koshkina M, Pidaparthy H, Elder JH. Contrastive learning for sports video: Unsupervised player classification. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2021:4523–4531. https://doi.org/10.1109/CVPRW53098.2021.00510
124. Kristan M, Matas J, Leonardis A, et al. A novel performance evaluation methodology for single-target trackers. IEEE Trans Pattern Anal Mach Intell. 2016;38(11):2137–55. https://doi.org/10.1109/TPAMI.2016.2516982.
125. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90. https://doi.org/10.1145/3065386.
126. Kulkarni KM, Shenoy S. Table tennis stroke recognition using two-dimensional human pose estimation. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2021:4571–4579. https://doi.org/10.1109/CVPRW53098.2021.00515
127. Kurach K, Raichuk A, Stanczyk P, et al. Google research football: a novel reinforcement learning environment. Proceed AAAI Conf Artif Intell. 2020;34(04):4501–10. https://doi.org/10.1609/aaai.v34i04.5878 (https://ojs.aaai.org/index.php/AAAI/article/view/5878).
128. Lea C, Flynn MD, Vidal R, et al. Temporal convolutional networks for action segmentation and detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017:1003–1012. https://doi.org/10.1109/CVPR.2017.113
129. Lewis DD, Catlett J. Heterogeneous uncertainty sampling for supervised learning. Mach Learn Proceed. 1994;1994:148–56. https://doi.org/10.1016/B978-1-55860-335-6.50026-X.
130. Li G, Xu S, Liu X, et al. Jersey number recognition with semi-supervised spatial transformer network. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2018:1864–18647. https://doi.org/10.1109/CVPRW.2018.00231
131. Li J, Xue J, Cao R, et al. Finerehab: A multi-modality and multi-task dataset for rehabilitation analysis. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024:3184–3193. https://doi.org/10.1109/CVPRW63382.2024.00324
132. Li R, Bhanu B. Fine-grained visual dribbling style analysis for soccer videos with augmented dribble energy image. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019:2439–2447. https://doi.org/10.1109/CVPRW.2019.00299
133. Li Y, Huang D, Qin D, et al. Improving object detection with selective self-supervised self-training. Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX. 2020:589–607. https://doi.org/10.1007/978-3-030-58526-6_35
134. Li Y, Zhang S, Wang Z, et al. Tokenpose: Learning keypoint tokens for human pose estimation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021:11293–11302.https://doi.org/10.1109/ICCV48922.2021.01112
135. Lin TY, Maire M, Belongie S, et al. Microsoft coco: common objects in context. Comput Vis ECCV. 2014;2014:740–55. https://doi.org/10.1007/978-3-319-10602-1_48.
136. Lin K, Wang L, Luo K, et al. Cross-domain complementary learning using pose for multi-person part segmentation. IEEE Trans Circuits Syst Video Technol. 2021;31(3):1066–78. https://doi.org/10.1109/TCSVT.2020.2995122.
137. Liu J, Shahroudy A, Perez M, et al. Ntu rgb+d 120: a large-scale benchmark for 3d human activity understanding. IEEE Trans Pattern Anal Mach Intell. 2020;42(10):2684–701. https://doi.org/10.1109/TPAMI.2019.2916873.
138. Liu H, Bhanu B. Pose-guided r-CNN for jersey number recognition in sports. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019:2457–2466.https://doi.org/10.1109/CVPRW.2019.00301
139. Liu H, Li C, Wu Q, et al. Visual instruction tuning. Adv Neural Inf Process Syst. 2023;36:34892–34916. https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf
140. Liu P, Wang JH. Monotrack: Shuttle trajectory reconstruction from monocular badminton video. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022:3512–3521. https://doi.org/10.1109/CVPRW56347.2022.00395
141. Liu Y, Hafemann LG. A scale-invariant trajectory simplification method for efficient data collection in videos. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2023:5129–5138. https://doi.org/10.1109/CVPRW59228.2023.00541
142. Liu Y, Chen K, Liu C, et al. Structured knowledge distillation for semantic segmentation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019:2599–2608. https://doi.org/10.1109/CVPR.2019.00271
143. Liu Y, Hafemann LG, Jamieson M, et al. Detecting and matching related objects with one proposal multiple predictions. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2021:4515–4522. https://doi.org/10.1109/CVPRW53098.2021.00509
144. Loper M, Mahmood N, Black MJ. Mosh: motion and shape capture from sparse markers. ACM Trans Graph. 2014;33(6):220–1. https://doi.org/10.1145/2661229.2661273.

145. Loper M, Mahmood N, Romero J, et al. Smpl: a skinned multi-person linear model. ACM Trans Graph. 2015:34(6). https://doi.org/10.1145/2816795.2818013
146. Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis. 2004;60(2):91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94.
147. Lu WL, Ting JA, Little JJ, et al. Learning to track and identify players from broadcast sports videos. IEEE Trans Pattern Anal Mach Intell. 2013;35(7):1704–16. https://doi.org/10.1109/TPAMI.2012.242.
148. Lucas BD, Kanade T. An iterative image registration technique with an application to stereo vision. Proceedings of the 7th International Joint Conference on Artificial Intelligence. 1981;2:674–679. https://doi.org/10.5555/1623264.1623280
149. Ludwig K, Kienzle D, Lienhart R. Recognition of freely selected keypoints on human limbs. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022:3530–3538. https://doi.org/10.1109/CVPRW56347.2022.00397
150. Ludwig K, Lorenz J, Schön R, et al. All keypoints you need: detecting arbitrary keypoints on the body of triple, high, and long jump athletes. 2023 IEEE/CVF International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 17 2023 to June 24 2023, Vancouver, BC, Canada. 2023:5179 – 5187. https://doi.org/10.1109/CVPRW59228.2023.00546
151. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30. 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
152. Luong QT, Faugeras OD. The fundamental matrix: theory, algorithms, and stability analysis. Int J Comput Vis. 1996;17(1):43–75. https://doi.org/10.1007/BF00127818.
153. Magera F, Hoyoux T, Barnich O, et al. A Universal Protocol to Benchmark Camera Calibration for Sports . 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024:3335–3346. https://doi.org/10.1109/CVPRW63382.2024.00338
154. Maglo A, Orcesi A, Pham QC. Efficient tracking of team sport players with few game-specific annotations. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022:3460–3470. https://doi.org/10.1109/CVPRW56347.2022.00390
155. Majeed F, Gilal NU, Al-Thelaya K, et al. Mv-soccer: Motion-vector augmented instance segmentation for soccer player tracking. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024:3245–3255. https://doi.org/10.1109/CVPRW63382.2024.00330
156. Malleson C, Gilbert A, Trumble M, et al. Real-time full-body motion capture from video and imus. 2017 International Conference on 3D Vision (3DV). 2017:449–457. https://doi.org/10.1109/3DV.2017.00058
157. Manafifard M, Ebadi H, Moghaddam HA. A survey on player tracking in soccer videos. Comput Vis Image Underst. 2017;159:19–46. https://doi.org/10.1016/j.cviu.2017.02.002. (**computer Vision in Sports**).
158. Martin Z, Hendricks S, Patel A. Automated tackle injury risk assessment in contact-based sports—a rugby union example. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2021:4589–4598. https://doi.org/10.1109/CVPRW53098.2021.00517
159. Matas JG, Chum O, Urban M, et al. Robust wide-baseline stereo from maximally stable extremal regions. Image Vis Comput. 2004;22(10):761–7. https://doi.org/10.1016/j.imavis.2004.02.006. (**british Machine Vision Computing 2002**).
160. Mazzia V, Angarano S, Salvetti F, et al. Action transformer: A self-attention model for short-time pose-based human action recognition. Pattern Recogn. 2022:124(C). https://doi.org/10.1016/j.patcog.2021.108487.
161. McNally W. Deepdarts dataset. 2021. https://doi.org/10.21227/05e7-xs69
162. McNally W, Walters P, Vats K, et al. DeepDarts: Modeling Keypoints as Objects for Automatic Scorekeeping in Darts using a Single Camera . 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2021:4542–4551. https://doi.org/10.1109/CVPRW53098.2021.00512
163. McNally W, Lambeth J, Brekke D. Combining physics and deep learning models to simulate the flight of a golf ball. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2023:5119–5128. https://doi.org/10.1109/CVPRW59228.2023.00540
164. Mehraban S, Adeli V, Taati B. Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2024:6905–6915. https://doi.org/10.1109/WACV57701.2024.00677
165. Mehta D, Sridhar S, Sotnychenko O, et al. Vnect: real-time 3d human pose estimation with a single RGB camera. ACM Trans Graph. 2017;36(4):1–14. https://doi.org/10.1145/3072959.3073596.
166. Mendes-Neves T, Meireles L, Mendes-Moreira J. A survey of advanced computer vision techniques for sports. 2301.07583 2023.
167. Meratnia N, de By RA. Spatiotemporal compression techniques for moving point objects. Adv Database Technol EDBT. 2004;2004:765–82. https://doi.org/10.1007/978-3-540-24741-8_44.
168. Mesaros A, Heittola T, Dikmen O, et al. Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015:151–155.https://doi.org/10.1109/ICASSP.2015.7177950
169. Michalczyk MJ, Janetzke M, Mücke MM, et al. Dfl—bundesliga data shootout. 2022. https://kaggle.com/competitions/dfl-bundesliga-data-shootout
170. Milan A, Schindler K, Roth S. Detection- and trajectory-level exclusion in multiple object tracking. 2013 IEEE Conference on Computer Vision and Pattern Recognition. 2013:3682–3689. https://doi.org/10.1109/CVPR.2013.472
171. Milan A, Gade R, Dick A, et al. Improving global multi-target tracking with local updates. Computer Vision—ECCV 2014 Workshops. 2015:174–190. https://doi.org/10.1007/978-3-319-16199-0_13
172. Miron A, Sadawi N, Ismail W, et al. Intellirehabds (irds)—a dataset of physical rehabilitation movements. Data. 2021;6(5):46. https://doi.org/10.3390/data6050046.
173. Mkhallati H, Cioppa A, Giancola S, et al. Soccernet-caption: Dense video captioning for soccer broadcasts commentaries. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2023:5074–5085. https://doi.org/10.1109/CVPRW59228.2023.00536
174. Muja M, Lowe DG. Fast approximate nearest neighbors with automatic algorithm configuration. International Conference on Computer Vision Theory and Application VISSAPP'09). 2009:331–340. https://api.semanticscholar.org/CorpusID:7317448

175. Munkres J. Algorithms for the assignment and transportation problems. J Soc Ind Appl Math. 1957;5(1):32–8. https://doi.org/10.1137/0105003.
176. Murtagh F. Multilayer perceptrons for classification and regression. Neurocomputing. 1991;2(5–6):183–97. https://doi.org/10.1016/0925-2312(91)90023-5.
177. Naik BT, Hashmi MF, Bokde ND. A comprehensive review of computer vision in sports: open issues, future trends and research directions. Appl Sci. 2022;12(9):4429. https://doi.org/10.3390/app12094429.
178. Nakabayashi T, Higa K, Yamaguchi M, et al. Event-based ball spin estimation in sports. Proceedings—2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2024. 2024:3367–3375. https://doi.org/10.1109/CVPRW63382.2024.00341
179. Nekoui M, Tito Cruz FO, Cheng L. Falcons: Fast learner-grader for contorted poses in sports. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020:3941–3949. https://doi.org/10.1109/CVPRW50498.2020.00458
180. Neubeck A, Van Gool L. Efficient non-maximum suppression. 18th International Conference on Pattern Recognition (ICPR'06). 2006;3:850–855. https://doi.org/10.1109/ICPR.2006.479
181. Newell A, Huang Z, Deng J. Associative embedding: End-to-end learning for joint detection and grouping. Advances in Neural Information Processing Systems. 2017;30. https://dl.acm.org/doi/10.5555/3294771.3294988
182. Nie X, Chen S, Hamid R. A robust and efficient framework for sports-field registration. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). 2021:1935–1943. https://doi.org/10.1109/WACV48630.2021.00198
183. Nonaka N, Fujihira R, Nishio M, et al. End-to-end high-risk tackle detection system for rugby. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022:3549–3558. https://doi.org/10.1109/CVPRW56347.2022.00399
184. Nonaka N, Fujihira R, Koshiba T, et al. Rugby scene classification enhanced by vision language model. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024:3256–3266. https://doi.org/10.1109/CVPRW63382.2024.00331
185. Nozawa T, Wu E, Perteneder F, et al. Visualizing expert motion for guidance in a vr ski simulator. ACM SIGGRAPH 2019 Posters. 2019. https://doi.org/10.1145/3306214.3338561
186. Ogata R, Simo-Serra E, Iizuka S, et al. Temporal distance matrices for squat classification. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019:2533–2542. https://doi.org/10.1109/CVPRW.2019.00309
187. Oh SW, Lee JY, Xu N, et al. Video object segmentation using space-time memory networks. Proceedings—2019 International Conference on Computer Vision, ICCV 2019". 2019:9225–9234. https://doi.org/10.1109/ICCV.2019.00932
188. Okamoto L, Parmar P. Hierarchical neurosymbolic approach for comprehensive and explainable action quality assessment. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024;3204–3213. https://doi.org/10.1109/CVPRW63382.2024.00326
189. Pan JH, Gao J, Zheng WS. Action assessment by joint relation graphs. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019:6330–6339. https://doi.org/10.1109/ICCV.2019.00643
190. Pandya Y, Nandy K, Agarwal S. Homography based player identification in live sports. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2023:5209–5218. https://doi.org/10.1109/CVPRW59228.2023.00549
191. Papageorgiou G, Sarlis V, Tjortjis C. Evaluating the effectiveness of machine learning models for performance forecasting in basketball: a comparative study. Knowl Inf Syst. 2024;66(7):4333–75. https://doi.org/10.1007/s10115-024-02092-9.
192. Parmar P, Morris BT. Learning to score olympic events. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2017:76–84. https://doi.org/10.1109/CVPRW.2017.16
193. Parmar P, Morris BT. What and how well you performed? A multitask learning approach to action quality assessment. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019. https://doi.org/10.48550/arXiv.1904.04346
194. Parzen E. On estimation of a probability density function and mode. Ann Math Stat. 1962;33(3):1065–76. https://doi.org/10.1214/aoms/1177704472.
195. Patacchiola M, Storkey A. Self-supervised relational reasoning for representation learning. Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020. https://doi.org/10.5555/3495724.3496061
196. Pfrommer B. Frequency cam: Imaging periodic signals in real-time. 2022. https://doi.org/10.48550/arXiv.2211.00198
197. Pidaparthy H, Dowling MH, Elder JH. Automatic play segmentation of hockey videos. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2021:4580–4588. https://doi.org/10.1109/CVPRW53098.2021.00516
198. Piergiovanni AJ, Ryoo MS. Early detection of injuries in MLB pitchers from video. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019:2431–2438. https://doi.org/10.1109/CVPRW.2019.00298
199. Pirsiavash H, Ramanan D, Fowlkes CC. Globally-optimal greedy algorithms for tracking a variable number of objects. CVPR. 2011;2011:1201–8. https://doi.org/10.1109/CVPR.2011.5995604.
200. Qazi A, Iqbal A. ExerAIde: AI-assisted Multimodal Diagnosis for Enhanced Sports Performance and Personalised Rehabilitation . 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024:3430–3438. https://doi.org/10.1109/CVPRW63382.2024.00347
201. Qin Z, Zhou S, Wang L, et al. Motiontrack: Learning robust short-term and long-term motions for multi-object tracking. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023:17939–17948. https://doi.org/10.1109/CVPR52729.2023.01720
202. Rabiner LR, Juang BH. An introduction to hidden markov models. IEEE ASSP Mag. 1986;3(1):4–16. https://doi.org/10.1109/MASSP.1986.1165342.
203. Rahmad NA, As'ari MA, Ghazali NF, et al. A survey of video based action recognition in sports. Indones J Electr Eng Comput Sci. 2018;11(3):987–93. https://doi.org/10.11591/ijeecs.v11.i3.
204. Rana M, Mittal V. Wearable sensors for real-time kinematics analysis in sports: a review. IEEE Sens J. 2021;21(2):1187–207. https://doi.org/10.1109/JSEN.2020.3019016.

205. Razavian AS, Azizpour H, Sullivan J, et al. CNN features off-the-shelf: An astounding baseline for recognition. 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014:512–519. https://doi.org/10.1109/CVPRW.2014.131
206. Rebecq H, Gehrig D, Scaramuzza D. Esim: an open event camera simulator. Proceedings of The 2nd Conference on Robot Learning. 2018;87:969–982. https://rpg.ifi.uzh.ch/esim.html
207. Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis & Machine Intelligence. 2017;39(06):1137–49. https://doi.org/10.1109/TPAMI.2016.2577031.
208. Renò V, Mosca N, Marani R, et al. Convolutional neural networks based ball detection in tennis games. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2018:1839–18396. https://doi.org/10.1109/CVPRW.2018.00228
209. Ribeiro MT, Singh S, Guestrin C. "why should i trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:1135–1144. https://doi.org/10.1145/2939672.2939778
210. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention—MICCAI. 2015;2015:234–41. https://doi.org/10.1007/978-3-319-24574-4_28.
211. Rosenblatt M. Remarks on some nonparametric estimates of a density function. Ann Math Stat. 1956;27(3):832–7. https://doi.org/10.1214/aoms/1177728190.
212. Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018:4510–4520. https://doi.org/10.1109/CVPR.2018.00474
213. Sanford R, Gorji S, Hafemann LG, et al. Group activity detection from trajectory and video data in soccer. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020:3932–3940. https://doi.org/10.1109/CVPRW50498.2020.00457
214. Santoro A, Raposo D, Barrett DG, et al. A simple neural network module for relational reasoning. Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017:4974–4983. https://doi.org/10.5555/3295222.3295250
215. Sapp B, Taskar B. Modec: Multimodal decomposable models for human pose estimation. 2013 IEEE Conference on Computer Vision and Pattern Recognition. 2013:3674–3681. https://doi.org/10.1109/CVPR.2013.471
216. Sarkar S, Mukherjee DP, Chakrabarti A. From soccer video to ball possession statistics. Pattern Recogn. 2022;122:108338. https://doi.org/10.1016/j.patcog.2021.108338.
217. Sarkar S, Mukherjee DP, Chakrabarti A. Watch and act: Dual interacting agents for automatic generation of possession statistics in soccer. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022:3559–3567. https://doi.org/10.1109/CVPRW56347.2022.00400
218. Sarlis V, Chatziilias V, Tjortjis C, et al. A data science approach analysing the impact of injuries on basketball player and team performance. Inf Syst. 2021;99:101750. https://doi.org/10.1016/j.is.2021.101750.
219. Sarlis V, Tjortjis C. Sports analytics—evaluation of basketball players and team performance. Inf Syst. 2020;93:101562. https://doi.org/10.1016/j.is.2020.101562.
220. Sarlis V, Tjortjis C. Sports analytics: Data mining to uncover nba player position, age, and injury impact on performance and economics. Information. 2024;15(4):242. https://doi.org/10.3390/info15040242.
221. Sarlis V, Papageorgiou G, Tjortjis C. Leveraging sports analytics and association rule mining to uncover recovery and economic impacts in NBA basketball. Data. 2024:9(7). https://doi.org/10.3390/data9070083
222. Schlosser P, Münch D, Arens M. Investigation on combining 3d convolution of image data and optical flow to generate temporal action proposals. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019:2448–2456. https://doi.org/10.1109/CVPRW.2019.00300
223. Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision (ICCV). 2017:618–626. https://doi.org/10.1109/ICCV.2017.74
224. Senocak A, Oh TH, Kim J, et al. Part-based player identification using deep convolutional representation and multi-scale pooling. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2018:1813–18137. https://doi.org/10.1109/CVPRW.2018.00225
225. Sha L, Hobbs J, Felsen P, et al. End-to-end camera calibration for broadcast videos. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020:13624–13633. https://doi.org/10.1109/CVPR42600.2020.01364
226. Shahroudy A, Liu J, Ng TT, et al. Ntu rgb+d: A large scale dataset for 3d human activity analysis. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:1010–1019. https://doi.org/10.1109/CVPR.2016.115
227. Sharma RA, Bhat B, Gandhi V, et al. Automated top view registration of broadcast football videos. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). 2018:305–313. https://doi.org/10.1109/WACV.2018.00040
228. Shi F, Marchwica P, Gamboa Higuera JC, et al. Self-supervised shape alignment for sports field registration. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2022:3768–3777. https://doi.org/10.1109/WACV51458.2022.00382
229. Shi J, Tomasi C. Good features to track. 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 1994:593–600. https://doi.org/10.1109/CVPR.1994.323794
230. Shih HC. A survey of content-aware video analysis for sports. IEEE Trans Circuits Syst Video Technol. 2018;28(5):1212–31. https://doi.org/10.1109/TCSVT.2017.2655624.
231. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556. 2014. https://doi.org/10.48550/arXiv.1409.1556
232. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45(4):427–37. https://doi.org/10.1016/j.ipm.2009.03.002.
233. Sudhakaran S, Escalera S, Lanz O. Gate-shift-fuse for video action recognition. IEEE Trans Pattern Anal Mach Intell. 2023;45(9):10913–28. https://doi.org/10.1109/TPAMI.2023.3268134.
234. Sun NE, Lin YC, Chuang SP, et al. Tracknetv2: Efficient shuttlecock tracking network. 2020 International Conference on Pervasive Artificial Intelligence (ICPAI). 2020:86–91. https://doi.org/10.1109/ICPAI51961.2020.00023
235. Sun P, Cao J, Jiang Y, et al. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022:20961–20970. https://doi.org/10.1109/CVPR52688.2022.02032

236. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Proceedings of the 27th International Conference on Neural Information Processing Systems. 2014;2:3104–3112. https://doi.org/10.5555/2969033.2969173
237. Suzuki T, Tanaka R, Takeda K, et al. Pseudo-label based unsupervised fine-tuning of a monocular 3d pose estimation model for sports motions. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024:3315–3324. https://doi.org/10.1109/CVPRW63382.2024.00336
238. Sárándi I, Linder T, Arras KO, et al. Metrabs: Metric-scale truncation-robust heatmaps for absolute 3d human pose estimation. IEEE Transactions on Biometrics Behavior and Identity Science. 2021;3(1):16–30. https://doi.org/10.1109/TBIOM.2020.3037257.
239. Takahashi K, Mikami D, Isogawa M, et al. Human pose as calibration pattern: 3d human pose estimation with multiple unsynchronized and uncalibrated cameras. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2018:1856–18567. https://doi.org/10.1109/CVPRW.2018.00230
240. Tamaki T, Wang H, Raytchev B, et al. Estimating the spin of a table tennis ball using inverse compositional image alignment. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2012:1457–1460. https://doi.org/10.1109/ICASSP.2012.6288166
241. Thain E. Science and Golf IV. Routledge. 2002. https://doi.org/10.4324/9780203715000
242. Theiner J, Ewerth R. Tvcalib: Camera calibration for sports field registration in soccer. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2023:1166–1175. https://doi.org/10.1109/WACV56688.2023.00122
243. Thomas G, Gade R, Moeslund T, et al. Computer vision for sports: current applications and research topics. Comput Vis Image Underst. 2017;159:3–18. https://doi.org/10.1016/j.cviu.2017.04.011.
244. Tompson J, Goroshin R, Jain A, et al. Efficient object localization using convolutional networks. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015:648–656. https://doi.org/10.1109/CVPR.2015.7298664
245. Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018:6450–6459. https://doi.org/10.1109/CVPR.2018.00675
246. Ullah M, Cheikh FA. A directed sparse graphical model for multi-target tracking. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2018:1897–18977. https://doi.org/10.1109/CVPRW.2018.00235
247. Ullman S, Brenner S. The interpretation of structure from motion. Proc R Soc Lond B. 1979;203(1153):405–26. https://doi.org/10.1098/rspb.1979.0006.
248. Vandeghen R, Cioppa A, Van Droogenbroeck M. Semi-supervised training to improve player and ball detection in soccer. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022:3480–3489. https://doi.org/10.1109/CVPRW56347.2022.00392
249. Vanderplaetse B, Dupont S. Improved soccer action spotting using both audio and video streams. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020:3921–3931. https://doi.org/10.1109/CVPRW50498.2020.00456
250. van der Kruk E, Reijne MM. Accuracy of human motion capture systems for sport applications; state-of-the-art review. Eur J Sport Sci. 2018;18(6):806–19. https://doi.org/10.1080/17461391.2018.1463397.
251. van den Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning. Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017:6309–6318. https://doi.org/10.5555/3295222.3295378
252. von Braun MS, Frenzel P, Käding C, et al. Utilizing mask r-CNN for waterline detection in canoe sprint video analysis. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020:3826–3835. https://doi.org/10.1109/CVPRW50498.2020.00446
253. Vaswani A, Shazeer N, Parmar N, et al. the most critical factor in the model's performance. all you need. Advances in Neural Information Processing Systems 30. 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
254. Vats K, Fani M, Walters P, et al. Event detection in coarsely annotated sports videos via parallel multi receptive field 1d convolutions. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020:3856–3865. https://doi.org/10.1109/CVPRW50498.2020.00449
255. Vats K, Fani M, Clausi DA, et al. Multi-task learning for jersey number recognition in ice hockey. Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports. 2021a:11–15. https://doi.org/10.1145/3475722.3482794
256. Vats K, Fani M, Clausi DA, et al. Puck localization and multi-task event recognition in broadcast hockey videos. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2021b: 4562–4570. https://doi.org/10.1109/CVPRW53098.2021.00514
257. Vats K, McNally W, Walters P, et al. Ice hockey player identification via transformers and weakly supervised learning. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022:3450–3459. https://doi.org/10.1109/CVPRW56347.2022.00389
258. Verma M, Kumawat S, Nakashima Y, et al. Yoga-82: A new dataset for fine-grained classification of human poses. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020:4472–4479. https://doi.org/10.1109/CVPRW50498.2020.00527
259. Voeikov R, Falaleev N, Baikulov R. Ttnet: Real-time temporal and spatial video analysis of table tennis. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020:3866–3874. https://doi.org/10.1109/CVPRW50498.2020.00450
260. Wang W, He N, Yao K, et al. Improved kalman filter and its application in initial alignment. Optik. 2021;226:165747. https://doi.org/10.1016/j.ijleo.2020.165747.
261. Wang J, Sun K, Cheng T, et al. Deep high-resolution representation learning for visual recognition. IEEE Trans Pattern Anal Mach Intell. 2021;43(10):3349–64. https://doi.org/10.1109/TPAMI.2020.2983686.
262. Wang CY, Bochkovskiy A, Liao HYM. Scaled-YOLOv4: Scaling Cross Stage Partial Network . 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021a:13024–13033. https://doi.org/10.1109/CVPR46437.2021.01283
263. Wang WY, Huang YC, Ik TU, et al. Shuttleset: A human-annotated stroke-level singles dataset for badminton tactical analysis. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023:5126–5136. https://doi.org/10.1145/3580305.3599906
264. Wang X, Girshick R, Gupta A, et al. Non-local neural networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018:7794–7803. https://doi.org/10.1109/CVPR.2018.00813

265. Wang X, Jabri A, Efros AA. Learning correspondence from the cycle-consistency of time. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019:2561–2571. https://doi.org/10.1109/CVPR.2019.00267
266. Wang Z, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning. Proceedings of The 33rd International Conference on Machine Learning. 2016;48:1995–2003. https://doi.org/10.5555/3045390.3045601
267. Wei SE, Ramakrishna V, Kanade T, et al. Convolutional pose machines. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:4724–4732. https://doi.org/10.1109/CVPR.2016.511
268. Weld DS, Bansal G. The challenge of crafting intelligible intelligence. Commun ACM. 2019;62(6):70–9. https://doi.org/10.1145/3282486.
269. Wieczorek M, Rychalska B, Dabrowski J.On the unreasonable effectiveness of centroids in image retrieval. Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part IV 2021):212–223. https://doi.org/10.1007/978-3-030-92273-3_18
270. Wu E, Nozawa T, Perteneder F, et al. Vr alpine ski training augmentation using visual cues of leading skier. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020:3836–3845.https://doi.org/10.1109/CVPRW50498.2020.00447
271. Wu T, He R, Wu G, et al. Sportshhi: A dataset for human-human interaction detection in sports videos. 2404.04565. 2024.
272. Xarles A, Escalera S, Moeslund TB, et al. T-deed: Temporal-discriminability enhancer encoder-decoder for precise event spotting in sports videos. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024:3410–3419. https://doi.org/10.1109/CVPRW63382.2024.00345
273. Xu J, Rao Y, Yu X, et al. Finediving: A fine-grained dataset for procedure-aware action quality assessment. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022:2939–2948. https://doi.org/10.1109/CVPR52688.2022.00296
274. Xu Y, Zhang J, Zhang Q, et al. Vitpose: simple vision transformer baselines for human pose estimation. Proceedings of the 36th International Conference on Neural Information Processing Systems. 2024. https://doi.org/10.5555/3600270.3603065
275. Yan B, Peng H, Fu J, et al. Learning Spatio-Temporal Transformer for Visual Tracking. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021:10428–10437. https://doi.org/10.1109/ICCV48922.2021.01028
276. Yan R, Xie L, Tang J, et al. Social adaptive module for weakly-supervised group activity recognition. Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII. 2020:208–224. https://doi.org/10.1007/978-3-030-58598-3_13
277. Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. 2018. https://doi.org/10.5555/3504035.3504947
278. Yang D, Wang Y, Dantcheva A, et al. UNIK: A unified framework for real-world skeleton-based action recognition. CoRR abs/2107.08580. 2021. arXiv:2107.08580
279. Ye B, Chang H, Ma B, et al. Joint feature learning and relation modeling for tracking: A one-stream framework. Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII. 2022:341–357. https://doi.org/10.1007/978-3-031-20047-2_20
280. Yeung C, Ide K, Fujii K. Autosoccerpose: Automated 3d posture analysis of soccer shot movements. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2024:3214–3224.https://doi.org/10.1109/CVPRW63382.2024.00327
281. Zecha D, Einfalt M, Lienhart R. Refining joint locations for human pose tracking in sports videos. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2019:2524–2532. https://doi.org/10.1109/CVPRW.2019.00308
282. Zhang Z. Iterative point matching for registration of free-form curves and surfaces. Int J Comput Vis. 1994;13:119–52. https://doi.org/10.1007/BF01427149.
283. Zhang Z. A flexible new technique for camera calibration. IEEE Trans Pattern Anal Mach Intell. 2000;22(11):1330–4. https://doi.org/10.1109/34.888718.
284. Zhang N, Izquierdo E. A four-point camera calibration method for sport videos. IEEE Trans Circuits Syst Video Technol. 2023;33(8):3811–21. https://doi.org/10.1109/TCSVT.2023.3243126.
285. Zhang N, Izquierdo E. A high accuracy camera calibration method for sport videos. 2021 International Conference on Visual Communications and Image Processing (VCIP). 2021:1–5. https://doi.org/10.1109/VCIP53242.2021.9675379
286. Zhang Y, Sun P, Jiang Y, et al. Bytetrack: Multi-object tracking by associating every detection box. Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27 2022, Proceedings, Part XXII. 2022:1–21. https://doi.org/10.1007/978-3-031-20047-2_1
287. Zhou X, Koltun V, Krähenbühl P. Tracking objects as points. Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV. 2020:474–490. https://doi.org/10.1007/978-3-030-58548-8_28
288. Zhou X, Kang L, Cheng Z, et al. Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection. CoRR abs/2106.14447. 2021. arXiv:2106.14447
289. Zhu K, Wong A, McPhee J. Fencenet: Fine-grained footwork recognition in fencing. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2022:3588–3597.https://doi.org/10.1109/CVPRW56347.2022.00403

Publisher's Note