AI

Article

Robotics Perception: Intention Recognition to Determine the Handball Occurrence during a Football or Soccer Match

Mohammad Mehedi Hassan *,†, **Stephen Karungaru** † and **Kenji Terada** †

Computer Science and Mathematical Science Program, Tokushima University, Tokushima-shi 770-0814, Japan; karungaru@tokushima-u.ac.jp (S.K.); terada@is.tokushima-u.ac.jp (K.T.)
* Correspondence: metsys19@gmail.com
† Current address: Minamijosanjimacho 2-Chōme, Tokushima-shi 770-0814, Japan.

Abstract: In football or soccer, a referee controls the game based on the set rules. The decisions made by the referee are final and can't be appealed. Some of the decisions, especially after a handball event, whether to award a penalty kick or a yellow/red card can greatly affect the final results of a game. It is therefore necessary that the referee does not make an error. The objective is therefore to create a system that can accurately recognize such events and make the correct decision. This study chose handball, an event that occurs in a football game (Not to be confused with the game of Handball). We define a handball event using object detection and robotic perception and decide whether it is intentional or not. Intention recognition is a robotic perception of emotion recognition. To define handball, we trained a model to detect the hand and ball which are primary objects. We then determined the intention using gaze recognition and finally combined the results to recognize a handball event. On our dataset, the results of the hand and the ball object detection were 96% and 100% respectively. With the gaze recognition at 100%, if all objects were recognized, then the intention and handball event recognition were at 100%.

Keywords: robotics perception; emotion recognition; computer vision; machine learning; AI application

Citation: Hassan, M.M.; Karungaru, S.; Terada, K. Robotics Perception: Intention Recognition to Determine the Handball Occurrence during a Football or Soccer Match. *AI* **2024**, *5*, 602–617. https://doi.org/10.3390/ai5020032

Academic Editor: Andrea Calimera

Received: 14 March 2024
Revised: 26 April 2024
Accepted: 30 April 2024
Published: 8 May 2024

1. Introduction

Football or soccer is one of the most played and watched games in the world [1]. This all-year-played game involves 22 players in two teams, some complicated rules [2], and one main referee to control the play by making decisions. Sometimes, some controversial decisions can lead to brawls among the players and/or fans [3]. There are lots of events and actions going on in a football match. Of these, the handball event is one of the important events. This event results in penalty kicks or free kicks being awarded, which may lead to a goal-scoring opportunity [4] or a player receiving a yellow or red card [5]. A handball event occurs when a player, excluding the goalkeeper, intentionally makes contact with the ball using their hand or a part of their arm, and for the goalkeeper, if he or she touches the ball outside of the penalty area. According to the Fédération Internationale de Football Association (FIFA), there are a few things to consider before deciding what kind of handball event it is; the hand or arm touches the ball deliberately, the arm makes the body unnaturally bigger, or a goal is scored after a handball [4]. However, oftentimes, this decision is very difficult to make by the referees. For a single referee, it is a challenging task to monitor the whole game, make decisions, and also determine the player's intentions after an event. The player's intention influences the level of action taken by the referee. In the general case, during gameplay, the main referee monitors the game. If a player touches the ball, which is prohibited in the gameplay rules, the referee decides the occurrence of the handball event if the incident is visible to the referee. Currently, the referee can get help from VAR technology by pausing the game if the other team complains about the decision. Generally, the referee is in charge of the game. In our previous work [6], we initially defined a handball event solely based on player contact with the ball, without considering the intention behind the action. However, in this study, we have expanded our research scope to incorporate player intention into our analysis of handball events. In our study, an AI system or robot makes the decision based on the robot's perception. Therefore, In this study, we aim to decide the intention of a player during a handball event using computer vision. Intention, within the context of our study, serves as a facet of perception, representing the capacity to discern or become conscious of actions through sensory input. Specifically, in our system, the concept of intention revolves around the robot's ability to perceive actions, particularly the act of touching the ball with the hand, through visual input from a camera. This concept underpins the notion of robotic perception elucidated in the title, highlighting the system's capability to discern and determine the deliberateness of actions observed, thereby facilitating a deeper understanding of human-robot interactions in scenarios such as handball incidents during a football match.

2. Related Works

Several studies have already been conducted related to soccer or football games. Most of the work focused on various aspects of the game to analyze the it. Especially the tactics of the game or the overall data in the game for future usage. The article [7] introduces a method aimed at automatically creating summaries and highlights of soccer videos. This approach integrates audio and video features alongside an improved algorithm for extracting dominant colors. A key aspect of this method involves identifying the predominant color present on the football field.

In a separate investigation [8], researchers discovered all players on the field and pinpointed the player most frequently employing a specific handball technique. The 2001 study [9] introduced a novel methodology and effective methods for segmenting football videos, which organize content based on whether the ball is in play. Employing a specialized characteristic called grass-area-ratio, the initial stage categorizes each frame into three distinct perspectives. After that, skeleton extraction and target recognition are used to detect handball fouls in real-time while playing.

Another study [10] utilized artificial intelligence in Video-assisted refereeing technology (VAR), combining target recognition and pose extraction technology. Real-time detection is enabled through the migratory learning of the YOLO network. This study [11] outlines the application of Human Action Recognition (HAR) in sports, firstly utilizing computer vision as a principal component. Moreover, it discusses the utilization of openly available datasets extensively acknowledged in the domain. The research focuses on action recognition, recognizing 11 separate acts possible through a match of handball. A comparison is made between a standard CNN model categorizing every frame into groups of activities and LSTM and MLP-based methods incorporating temporal data from the video. While many studies have explored different facets of games, others have also looked into VAR technology.

Most of the work discussed above focused on specific areas of the game. Some focus on only player detection; some on summaries and highlights of the game; some on players' location by tracking the players to understand the tactics of the games and others check whether the game is in play. Some also utilize Video Assistant Referee (VAR) technology to define events. The Video Assistant Referee (VAR) technology in football has brought advancements in decision-making accuracy but also faces several limitations. Despite its potential to reduce errors, VAR decisions can still be subjective and lead to inconsistencies. Delays caused by VAR reviews disrupt the flow of the game, with the lack of transparency in decision-making processes. Additionally, VAR's scope is limited to specific types of decisions, leaving other aspects of the game unaffected. Our study takes a distinct strategy by focusing on the real-time identification of handball events throughout football or soccer matches. In this work, we also focused on intention recognition, which is one of the human emotions. There are very few work focused on this topic, especially on sports. Most researchers focused on human action in various sports like volleyball, basketball, soccer, and tennis for detecting players and understanding their actions during various activities. HAR [10] involves identifying the person in a video sequence, determining action duration, and categorizing action types. Another work [12] suggested the improvement of sports applications' perception, comprehension, and decision by surveying various deep learning methods in sports. There are few works focused solely on intention. This work [13] recognized intention among multiple agents by analyzing behaviors over time and refining a prescriptive behavioral model. The authors identify common intentions. They then transform the model into an unsupervised learning problem, employing a clustering algorithm to group intentions based on similarities in behavioral models. Another research [14] employed Bayesian Networks to enhance this process by integrating statistical evidence with contextual information. The authors advocate for an integrated approach, combining Logic Programming and Bayesian Networks, to effectively infer agents' intentions from their actions. Whereas our study took human-like approaches. Where we focused on the facial expression of a human and defined the intention. We did not attempt to study agents' behavior or humans' different kinds of actions. We only focused on behavior or the agent's awareness before the action. We aim to enhance referee decision-making procedures and foster equitable surroundings in the gameplay. Furthermore, the revelations gleaned from our study have the potential to promote the creation of tools or actions meant to diffuse disputes and improve the general fairness of the game.

3. Proposed Methods

We divided the definition of the handball event into three parts as shown in Figure 1. After the ball touches the hand, determine if,

1. It's a deliberate action
2. The hand makes the body bigger
3. After touching the ball
    - It Creates a Goal scoring opportunity
    - A goal is scored

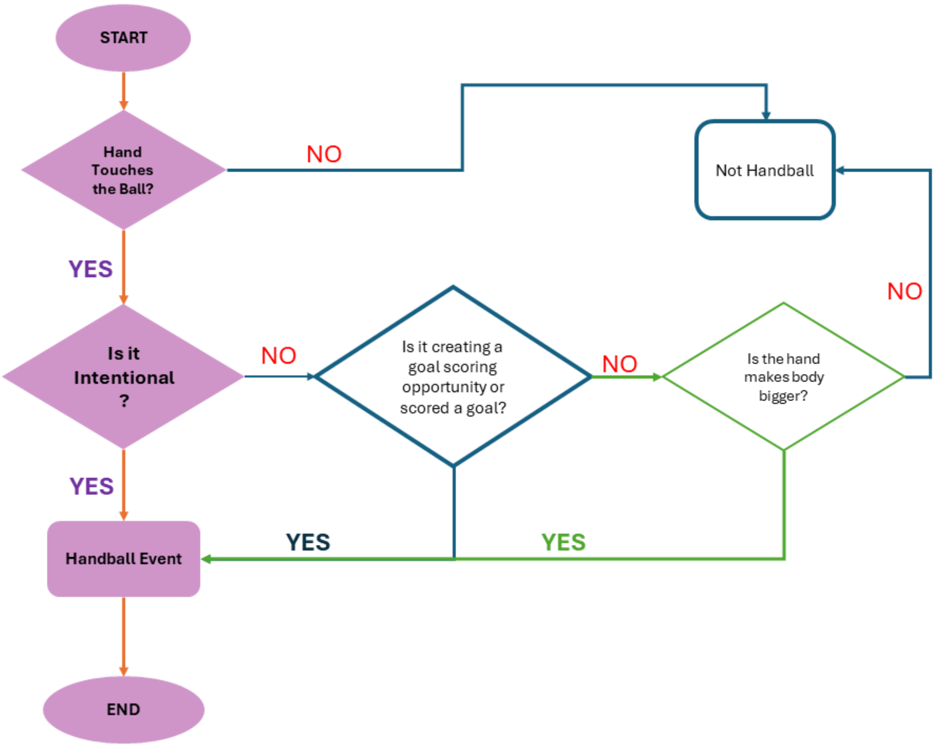

Figure 1. Flowchart of the proposed methods.

This research's main focus is to find the deliberate handball situation, i.e., determine whether the player touches the ball intentionally or not. To achieve the objective, we divided our work into three parts.

- Object Detection: Hand and Ball Recognition
- Intention Recognition: Perception Definition
- Event Definition: Recognizing Handball event

The details of these methods are discussed in the details below.

3.1. Object Detection: Hand and Ball Recognition

The important components of handball event recognition are hand and ball recognition. Accurate detection of the hand and ball is, therefore, vital. To determine hand and ball objects, we trained our system using Detectron2, a modular object detection library [15]. Detectron2 stands out as a versatile and powerful framework for object detection and segmentation tasks, offering a rich array of pre-trained models, efficient training pipelines, and robust inference capabilities. Its modular architecture facilitates easy customization and extension, allowing us to tailor models to specific domains and applications. We use instance segmentation [16] because there should be no gaps between bounding boxes and the objects. Simply put, when these two objects, the ball, and the hand bounding boxes, overlap, there is a high probability of a handball event, using Detectron2, an open-source library's straightforward and user-friendly API, enabled us to create and train our models. First, we import the necessary libraries and modules from Detectron2 and construct the configuration file for the instance segmentation model. The configuration file contains information about the model's architecture, training hyperparameters, and other configurations. As a training baseline, the previously trained model was used. Additionally, the number of classes was specified. Both the training and testing datasets were stored in the COCO [17] format, a popular format for object detection and instance segmentation tasks. To determine the accuracy of the training results for object detection, Fast R-CNN [18] and Mask R-CNN [19] were used. Fast R-CNN integrates a region proposal network with a deep convolutional neural network, allowing for end-to-end object detection. It proposes areas of interest and performs categorization and bounding box regression on these regions simultaneously, resulting in faster inference compared to previous methods. Additionally, it employs a multiple-task loss function to collaboratively optimize classification and bounding box regression tasks. By including a branch for segmentation mask prediction alongside the current classification and bounding box regression branches, Mask R-CNN expands Fast R-CNN. This enables instance segmentation, allowing the model to distinguish between different object instances in an image. By incorporating a mask branch, state-of-the-art performance in instance segmentation tasks are achieved with Mask R-CNN.

3.2. Intention Recognition: Perception Definition

3.2.1. Overview

Intention is an action or emotion where a person does something planned or deliberately [20]. In our daily lives, sometimes we do something habitually or follow a plan. In this work, we tried to visualize the intention based on the existing definition by dividing the definition into three parts, as shown in Figure 2. As Figure 2 illustrates, when we do something or take an action, first we observe and then make a decision based on the "Environment", "Objects", "Possible Actions", "Time", and finally we do the action based on our decision.

Our brain makes decisions to take action. Usually, it considers environmental factors, including location and object information. Possible actions and time are also important factors in making a decision. For the handball event, the environment is the football field, and the hand and ball information. Possible actions in this situation are whether a player touches the ball or not. Time is also important, because even if the objects are in sight, if a player does not have enough time to take actions, for example, avoiding moving objects, then that cannot be an intentional event.

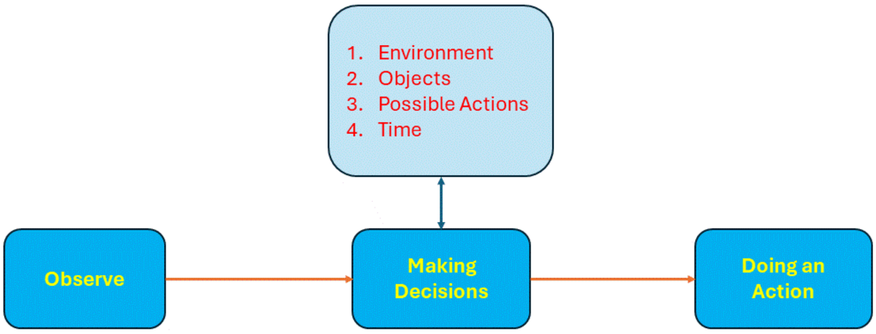

Figure 2. Summary of general procedures of taking action. (First, we observe the situation where we are in, make decisions based on different factors, and then take action.)

What we originally see is action. Other people's or players' decision-making process cannot be determined by a third person accurately. After seeing the action, we can determine whether it is intentional or not, but it is very difficult to determine with 100% accuracy the person's intention without a confession. Therefore, to determine the intention of a player, the eyes are very important. In our definition of intention, detecting a player's gaze is essential. Although Aristotle's De Anima (On the Soul) [21] is credited with originating the idea that humans have five primary senses, numerous philosophers and neuroscientists are currently discussing and investigating the possibility that humans may have more senses.

Using sight, it is possible to determine the person's observation area by determining the gaze angle.

According to [22] we should be able to see around 60 degrees nasally (toward your nose) and about 95 degrees temporally (toward your ear) from the center of an ocular. We should also be able to see 60 degrees above and 75 degrees below. This indicates that at any one moment, each eye provides a 155-degree horizontal field range and a 135-degree vertical field range. Figure 3 shows the field of view of an eye.

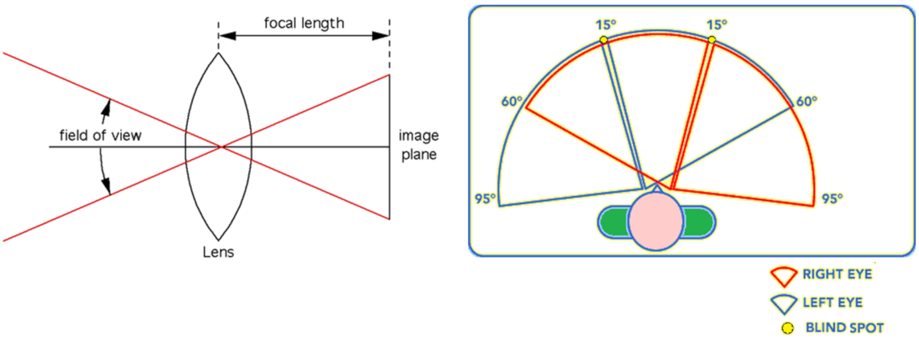

Figure 3. Field Of View of a person.

3.2.2. Making Decision

As Figure 4 shows, a player is standing and a ball is approaching the player from the right. To determine the action of the player, we need the gaze, hand, and ball objects.

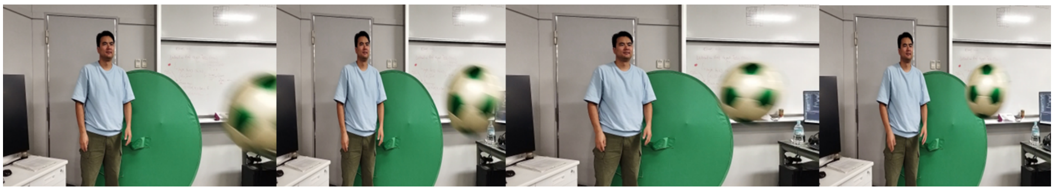

Figure 4. An event with multiple frames. Focusing on a football that is coming towards the player.

The Figure 5 shows the framework of the event recognition. First, we put the video as an input and then read the frames. After reading the frames we enhance the eyes of the player to determine the gaze directions and check the ball and hand positions. Then we calculate the decision-making time using Equation (1).

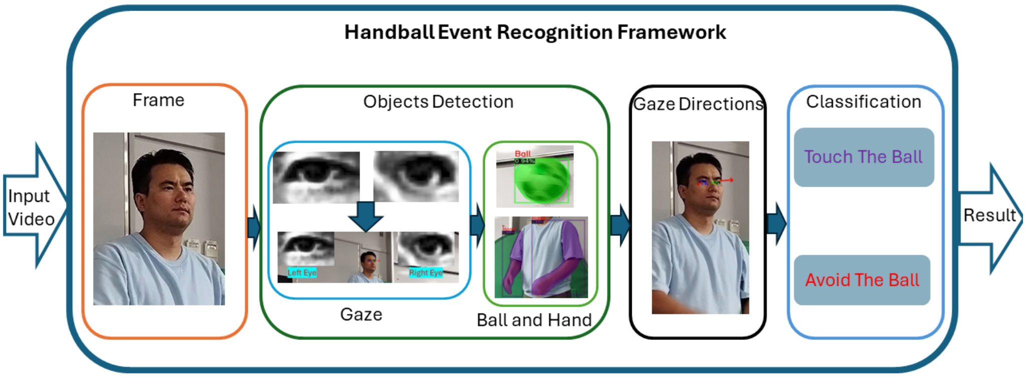

Figure 5. Handball event recognition procedures.

$$T_{\text{decision}} = \frac{D}{V - V_{\text{obj}} \cdot S} - R \cdot (1 - E) \tag{1}$$

where:

T_{decision}: Decision Making Time
D: Ideal distance between the player and the ball (meters)
S: Ideal speed of the ball (m/s)
V: Player's speed and agility (m/s)
V_{obj}: Visibility and size of the ball (Scaled from 0 to 1)
E: Environmental factors affecting avoidance (Scaled from 0 to 1)
R: Reaction time of the player (seconds)

Since our research mainly focused on intention recognition, we used the average value to determine the time the decision is made. The size and visibility of a soccer ball are generally good, especially when it is moving, so it can be considered to have high visibility and size. Its value is scaled between 0 for low and 1 for high visibility. If the environment is ideal then the Environment factors (E) value is 1 whereas if the condition is bad then the E will be 0. Distance between the player and the ball is calculated using their locations (p_1, b_1) and (p_2, b_2) respectively, in a two-dimensional coordinate system given by the formula:

$$\mathrm{D} = \sqrt{(p_2 - p_1)^2 + (b_2 - b_1)^2} \tag{2}$$

Equation (2) [23] calculates the distance between two objects. Speed is calculated using the equation $\mathrm{S} = \frac{\mathrm{d}}{\mathrm{t}}$ [24]. In our research case, the player's speed was zero because the player was not moving. However, in real situations, the average agility of a person is not zero, especially for the athletes [25]. In this case, we assumed that the player's agility or speed is 2 m/s. In general cases, an adult professional player can kick a ball between 29 m/s to 30 m/s [26]. For a simple task 0.2 to 0.3 s is the average reaction time for a healthy adult [27] person.

3.2.3. Taking Action

In this scenario, there are two actions available. Either the player touches, pushes, or grabs the ball by hand or avoids the ball. Equation (3) represents the probability of touching the ball by a player.

$$P(\text{Touch}) = 1 - P(\text{Avoid}) \tag{3}$$

In this equation:

- $P(\text{Touch})$ represents the probability of ball and person bounding boxes overlapping (Touching the ball).
- $P(\text{Avoid})$ represents the probability of the player avoiding the ball (ball and person bounding boxes not overlapping).

3.3. Event Definition: Recognizing Handball Event

3.3.1. Overview

The main objective of this study is to define the handball event in a football game using computer vision. We detect the hand and ball and determine the player's gaze. This section consists of two parts; the gaze direction and the hand/ball overlapping situation.

3.3.2. Handball Event

To determine the hand and ball touching situation, first, a list of 2D points representing the vertices of the two objects' polygons is obtained and tracked by Kalman filter [28]. To track the detected objects (hand and ball) we used the Kalman filter. Its recursive estimation process enables real-time updates of system states based on new observations, making it well-suited for applications requiring continuous monitoring and prediction. It combines measurements from multiple sources to calculate the true condition of a system while minimizing the effects of noise and uncertainty. By iteratively updating its estimates based on new measurements, the Kalman filter provides optimal estimates of the system state over time. According to our research approach, we commenced by setting the initial state to zero. Subsequently, we examined a video dataset containing numerous frames, utilizing our tailored model crafted for detecting the hand and ball in every frame. This bespoke model facilitated precise identification and localization of the hand and ball within the video frames. Following the detection stage, we used the Kalman filter to fine-tune the location of the objects that were detected—the ball and the hand. Next, we check if there is any polygonal overlap between the two objects. To ascertain whether the two polygons belonging to distinct classes overlapped, we applied the Separating Axis Theorem [29]. Algorithm 1 shows the separating axis theorem algorithm. This theorem uses the direction of an object's or other property as a dividing axis in each zone. This enables the fast and efficient detection of collisions between two convex polygons.

Algorithm 1 Using Separating Axis Theorem to check if two polygons overlapped or not.

```
for $poly$ in $[poly1, poly2]$                      ▷ Loop through both polygon; do
    $poly\_arr \leftarrow$ NumPy array;
    ▷ Get the converted polygon vertices to a NumPy array;
    for $i \leftarrow 0$ to length of $poly\_arr$        ▷ Loop through each edge of the polygon. do
        1. $p1 \leftarrow i$-th vertex of $poly\_arr$
        2. $p2 \leftarrow (i+1)$-th vertex of $poly\_arr$ (mod length of $poly\_arr$)
        ▷ Get the two points ($p1$ and $p2$) that define the edge
        3. $normal \leftarrow$ cross-product of $p2 - p1$ and $[0, 0, 1]$
        4. $normal \leftarrow$ NumPy array of $normal$
        ▷ Calculate the normal vector of the edge using the cross-product
        5. $min1, max1 \leftarrow$ project($poly1$, $normal$)
        6. $min2, max2 \leftarrow$ project($poly2$, $normal$)
        ▷ Project both polygons onto the normal vector and get the min and max projections
        7. if $max1 < min2$ or $max2 < min1$ then
            return False
            ▷ If the projections of the two polygons don't overlap, the polygons don't collide
        end
    end
    return True
    ▷ If the loop completes without returning False, the polygons collide
end
```

In this case, a collection of 2D points denoting the vertices of the second polygon is referred to as $poly2$, while a set of 2D points representing the vertices of the first polygon is denoted by $poly1$. To find the normal vector of an edge in a 2D space between two points, p1 and p2, we employ the cross-product method: Initially, determine the vector v pointing from $p1$ to $p2$ by subtracting $p1$ from $p2$: $v = p2 - p1$. Following this, find the normal vector n by taking the cross-product of v with the 2D unit vector in the positive z direction: $n = \text{cross}(v, [0, 0, 1])[: 2]$, where $\text{cross}()$ expressed the 3D cross-product function, and $[: 2]$ is utilized to obtained the x and y variables of the generated 3D vector [6].

3.4. Gaze Directions

In this work, to predict the gaze direction, we used UnityEye [30] to generate synthesized computer-generated eyes to train a model. We used almost 60,000 images of different directions of the eye. We estimate the gaze direction and head pose of a person from a video stream using facial landmarks. It starts by initializing a video capture and defining a 3D face model. Gaze scores are calculated based on the movement of facial landmarks, representing the relative gaze direction. Camera calibration is performed and estimates the rotation and translation vectors for each eye. The head pose is then adjusted based on the gaze scores. The 3D axis of rotation for each eye is projected onto the image, and lines representing pitch, roll, and yaw are drawn.

The gaze direction is visualized in the image by drawn lines. Figure 6 shows the different directions of the eye.

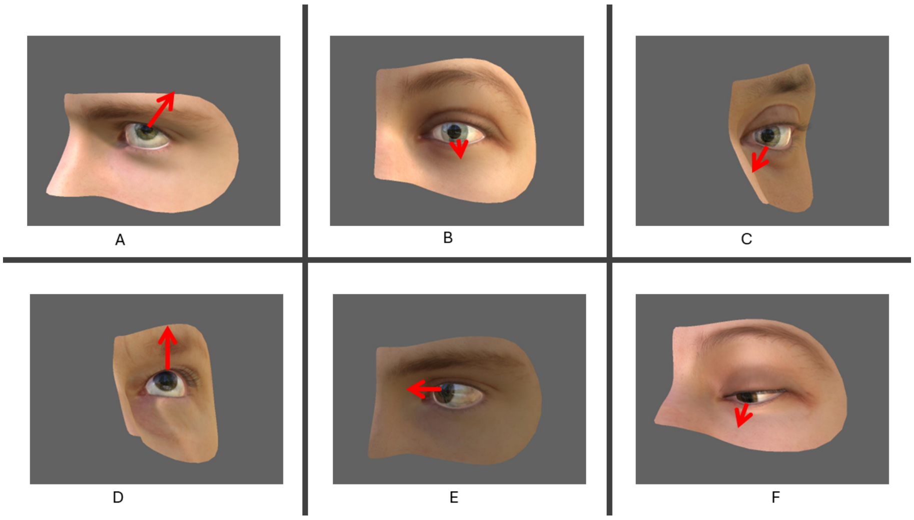

Figure 6. Eyes Directions. (**A**,**D**) Upper direction, (**B**) Straight direction, (**E**) Side direction, (**C**,**F**) Down direction.

To determine the focused view of the point, first, we calculate the left x gaze score. W used an array face_2d_head containing specific facial landmarks: Nose, Chin, Left Eye Left Corner, Right Eye Right Corner, Left Mouth Corner, and Right Mouth Corner. Then check if the horizontal distance between the Nose and the Left Eye Left Corner is not zero. If not zero, it calculates the ratio of the horizontal gap between the Right Gaze and Left Eye Left Corner to the gap between the Nose and Left Eye Left Corner and updates the previous ratio for the next iteration. Then we calculate the left y gaze score by checking if the vertical gap between the Left Eye Top and Left Eye Bottom is not zero. If not zero, it calculates the ratio of the vertical distance between the Nose Top and Left Eye Bottom to the distance between Left Eye Top and Left Eye Bottom and updates the previous ratio of the vertical distance. In a similar way, we calculate the right eye x and y gaze score. After calculating gaze scores for the left and right eyes in both the x and y directions based on the relative positions of specific facial landmarks, we perform smoothing to avoid abrupt changes in scores. Then we define the camera matrix using the focal length and image dimensions.

This matrix represents the intrinsic parameters of the camera. We also estimated the pose (rotation and translation vectors) of the left and right eyes in 3D space. Then we drew the lines by projecting the axis of rotation between the left and right eyes and its gaze direction.

3.5. Event Definition

After defining the intention, we extracted the polygons of the objects of two classes. To determine the hand and touches, we extracted the hand and ball objects polygons, then we examined each polygon through iteration, calculating the normal for each edge. Subsequently, this normal is employed to project the vertices of both polygons onto a common axis. If the projected intervals show no overlap, indicating no intersection between the polygons, the function returns False. Conversely, if there is an overlap on all axes, signifying an intersection, the function returns True. Which means the two objects overlapped each other. Figure 7 represents the skeletons of the players and the red circle is the ball objects polygon whereas green polygons represent the hands or arms of the players.

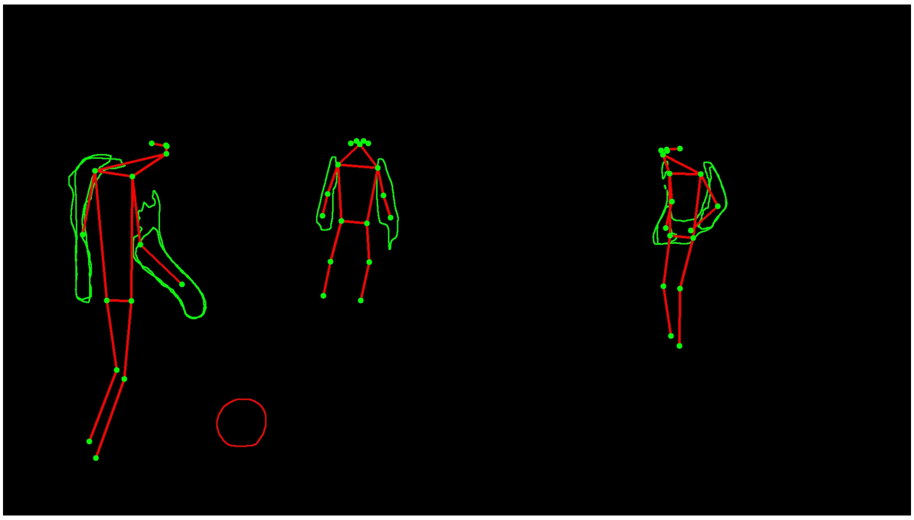

Figure 7. Extraction of the polygons and players' skeletons.

4. Experiment and Results

This section is divided into three parts based on the experiment's method,

- Results of Object Detection;
- Results of Perception Definition;
- Results of Event Definition;

4.1. Results of Object Detection

There are 676 instances in total in the dataset that were utilized for object detection, Table 1.

Table 1. Instances of the objects [6].

Category	Instances of the Objects
Hand	567
Ball	109

567 instances for the Hand category and 109 for the Ball category. Random sampling was used to make sure the dataset was representative of the population. The Hand category has far more instances than the Ball category due to player density. Figure 8 shows a sample of the results of instance segmentation training.

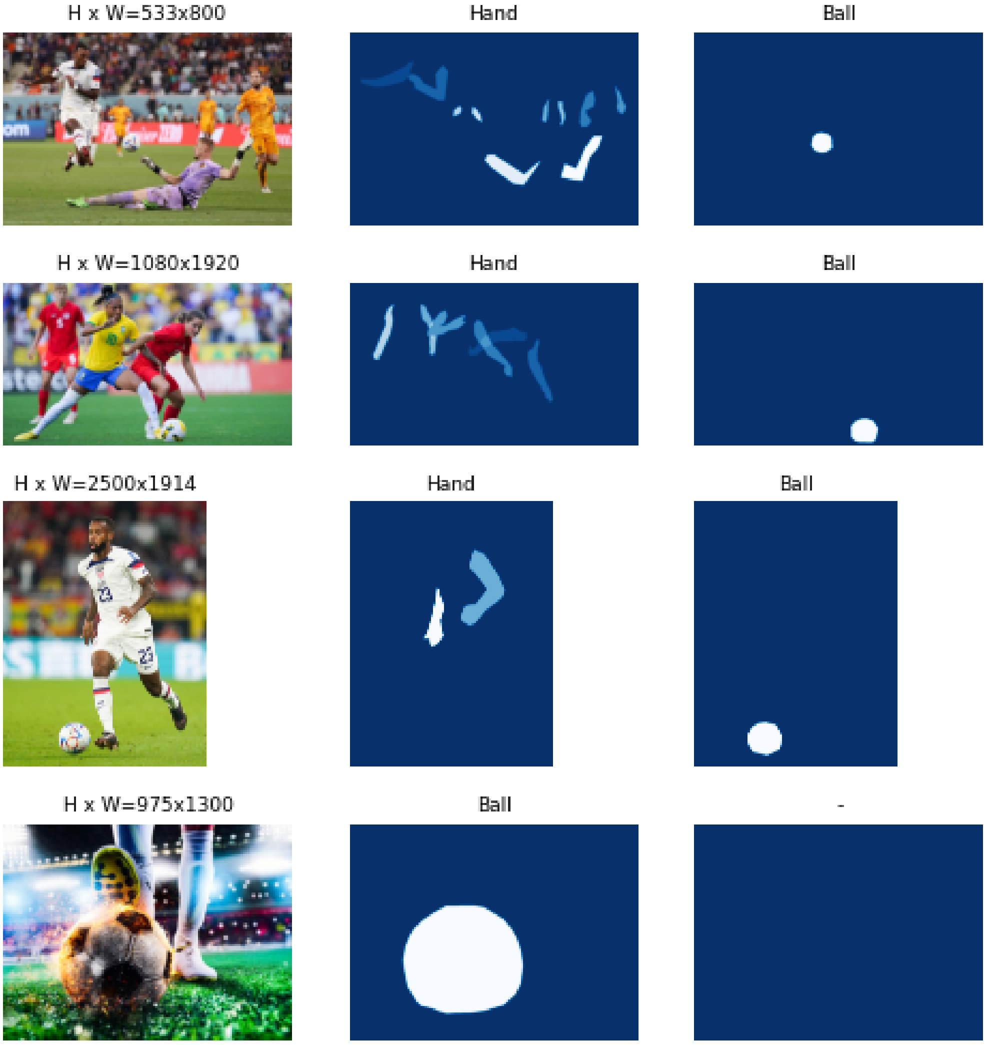

Figure 8. The process of learning to identify "Hand" and "Ball" objects [6].

Hand detection achieved an accuracy of approximately 97% and 96% in training and testing respectively. The ball detection achieved 100% accuracy in both situations. The test results' confusion matrix is displayed in Figure 9. Additionally, it was discovered that the instance segmentation outcomes during training were adequate. The accuracy of a Fast R-CNN object detection model is shown in Figure 10, demonstrating the beneficial effects of longer training times and more data on accuracy.

The decrease in false negatives over time and the increase in correctly classified foreground objects indicate improved performance. Nonetheless, as the model's performance is visually shown, there is still potential for improvement. In the Mask R-CNN model (Figure 11), accuracy and false negatives demonstrate positive outcomes, but fluctuations in false positives indicate areas for improvement, especially in cases where legs or other objects are occasionally misidentified as hands due to similar lengths.

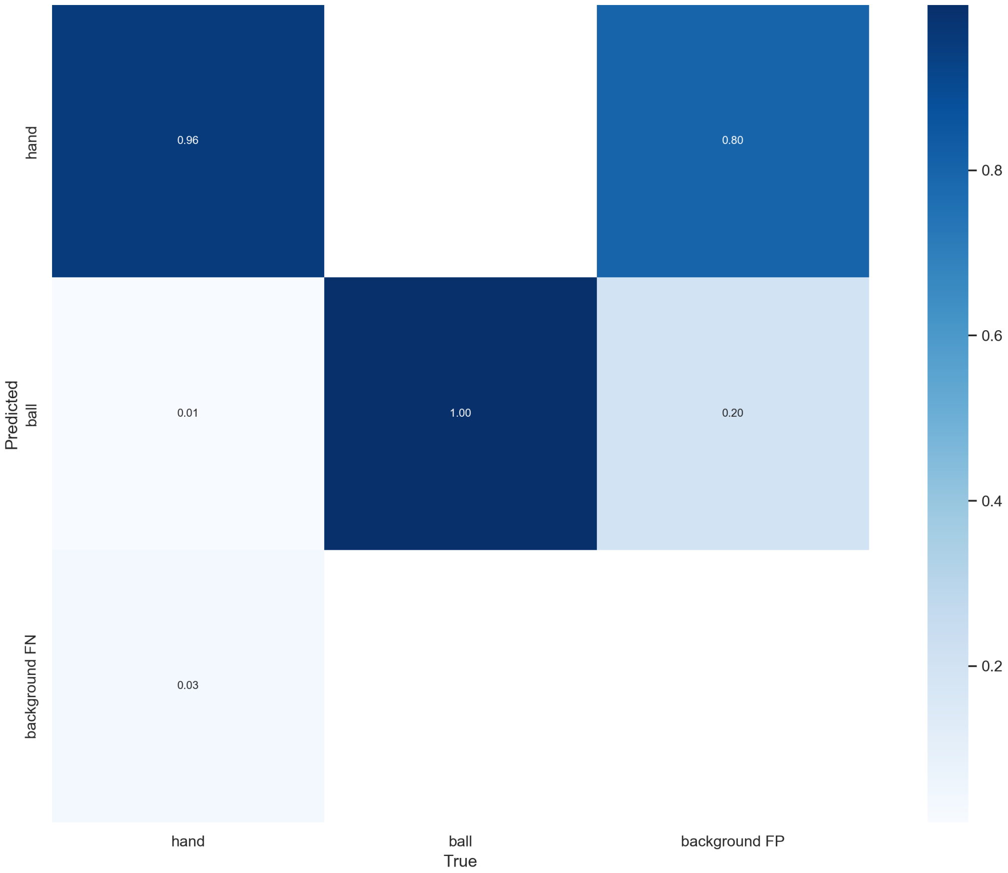

Figure 9. Confusion Matrix of Test results [6].

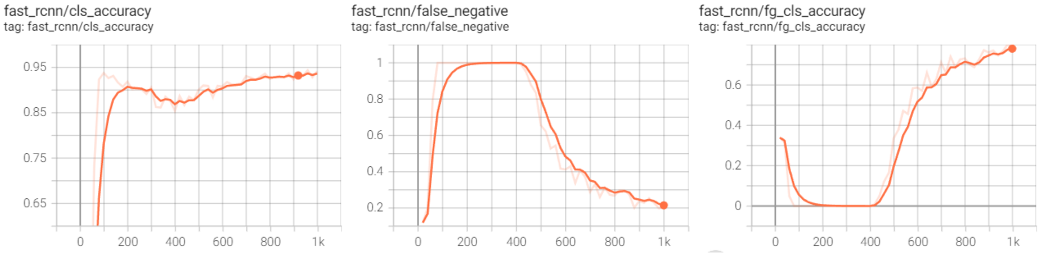

Figure 10. Fast R-CNN Object Identification Model Precision.

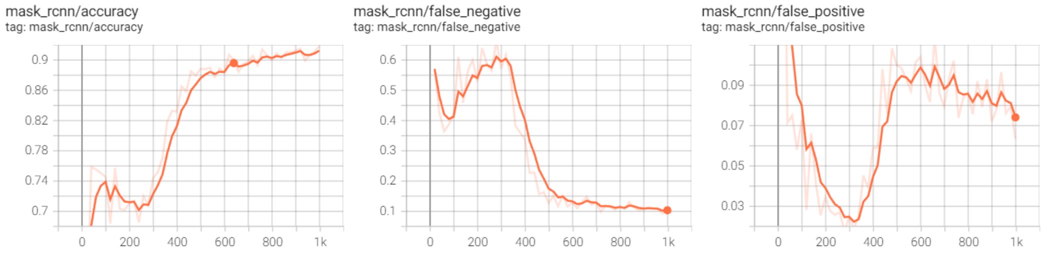

Figure 11. Accuracy Mask R-CNN Object Detection Model.

Figure 12 displays instances of low accuracy results, such as failure to recognize a player's hand (situation A) and false positives where a body part is misclassified as a hand (situation B) and a bag is identified as a ball (situation C) [6].

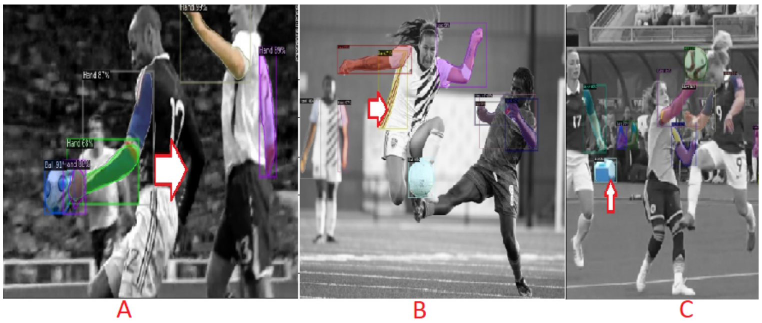

Figure 12. Low accuracy results [6].

4.2. Results of Perception Definition

Results of Gaze Detection

We only experimented with focused conditions in this research because unfocused conditions are more dangerous if the player cannot catch the ball. Figure 13 shows that a player focused on an incoming ball. We determine the focused vision or intended ball touching by object the ball is in the inside of the pitch, the roll, and yaw of the eyes. Distance is also calculated by the length of the lines, if the the ball is inside of this angle, then we determine that the next action should be intended.

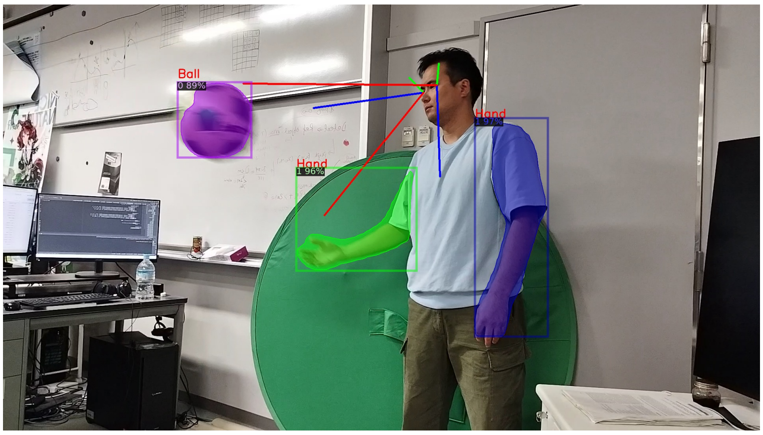

Figure 13. Focusing on an object.

4.3. Results of Event Definition

After defining intention if the player touches the ball with their hands then the framework defines the event as a handball event in a football game, as shown in Figure 14.

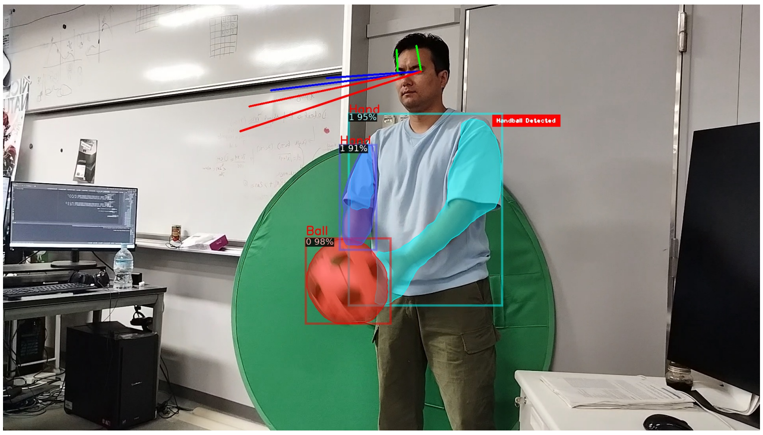

Figure 14. Recognizing the handball event in a football game.

Intention was recognized at 100% and handball event recognized at 100% in an ideal environment where player was not moving and the light of the environment was ideal with a single camera. The overall results are shown in Table 2.

Table 2. Summary of the Experiments Results.

Object Detection	Results of Object Detection	Collective Objective	Results of Collective Objective
Hand Detection	96%	Intention Recognition	100%
Ball Detection	100%	Event Recognition	100%
Eye Gaze Detection	100%		

5. Experiment Environment

This study utilized the Canon PowerShot SX740 HS Camera, featuring a 40× Optical Zoom and a 20.3 Megapixel CMOS Sensor, within an indoor setting. Video datasets were captured under ideal conditions, with the distance between the player and the camera ranging approximately from 1 to 2 m, predominantly from a front-facing angle. The proposed model was validated using a PC equipped with the following specifications: GPU—NVIDIA GeForce GT 730, CPU—12th Gen Intel(R) Core(TM) i7-12700 with a clock speed of 2.10 GHz.

6. Discussion

Object detection training for hand and ball was previously conducted [6] and, in this paper, we extended to train the eye gaze directions and intention definition. The object detection training was implemented with fewer images thus there were some mistakes and lower accuracy. Whereas in the intention definition, if the eyes were not visible or distanced from the camera, then the results could not be determined whether the player was focused or not. Thus, camera calibration and objects were very important to determine the overall results. We used synthesized data to avoid copyright issues. In addition, in this model, the synthesized data comes with direction levels and annotations. A novel approach for efficiently generating extensive sets of diverse eye region images for use as training data. Handball recognition speed is not calculated, but the system operates in real-time (30 fps). Gaze directions are one important factor in determining handball events in a game. However, if the player's eyes are obstructed from view due to factors such as other players, the ball, or body parts, the system may fail to recognize the event even if the player intentionally touches the ball. Moreover, environmental conditions like snow, darkness, or rain can further impede visibility, potentially leading to unexpected outcomes due to obscured views of the player's eyes. In addition, in this study, a single camera was employed. Therefore, the method's effectiveness may be compromised if objects are obstructed from view. Despite this limitation, our findings indicated that the method remains effective when all necessary objects are visible, yielding satisfactory results. Therefore, in future endeavors, we aim to explore the utilization of multiple cameras to enhance the detection of handball events during football matches.

Environmental factors play a pivotal role in decision-making processes. Our formulated equation (Equation (1)), denoted by (E), ranges from 0 to 1, with 1 representing an ideal environment and 0 indicating the worst environment possible. In real-world scenarios, such as on a football field, the efficacy of our method may be compromised if the camera fails to capture or detect all the essential elements required to identify a handball event accurately. However, with the advancement of camera technology, modern cameras boast robust capabilities to capture intricate player movements and details. By leveraging multiple cameras strategically positioned across the field, we can overcome these limitations and ensure comprehensive coverage for precise event definition. If all objects were detected 100% then the handball event recognition is at 100% but if one of the objects is missing, then the handball event can not be recognized. To conclude, the hand

and ball objects are needed to accurately identify the handball incident during a football game. In real-world scenarios, such as on a football field, the efficacy of our method may be compromised if the camera fails to capture or detect all the essential elements required to identify a handball event accurately. However, with the advancement of camera technology, modern cameras boast robust capabilities to capture intricate player movements and details. By leveraging multiple cameras strategically positioned across the field, we can overcome these limitations and ensure comprehensive coverage for precise event definition. If this system is implemented in a live game, then the system can avoid controversial decisions without pausing the game, like goal-line technology. This system can swiftly process events in real time, augmenting results efficiently within the timeframe of human perception. Consequently, game interruptions are obviated, offering a smoother and more efficient decision-making process. Implementing the system for a soccer match in a live setting requires a combination of technological infrastructure and referee training. First and foremost, the system would rely on high-quality cameras strategically positioned around the field to capture comprehensive footage. These cameras should provide clear views of the players and the ball to accurately identify handball events. The captured footage would then be processed in real-time using this system. To fully integrate this technology, referees would require specialized training to understand how the system operates and how to interpret its findings. This training would include familiarization with the system's capabilities and limitations, as well as guidelines on when to rely on the technology's decisions versus making their own judgments. Additionally, the infrastructure supporting the system would need to be robust and reliable, with redundant backups in place to ensure uninterrupted operation during matches. This includes backup power sources, redundant data storage, and failover mechanisms to prevent system downtime.

In addition, intention recognition, a process that infers the goals or intentions behind observed actions, can be adapted to various ambiguous game situations beyond handball occurrences in football. By leveraging machine learning techniques and advanced perception algorithms, intention recognition systems can analyze player behaviors and infer their intentions in real time. For instance, in basketball, intention recognition can help detect fouls, determine whether a player intends to shoot or pass the ball, or predict defensive strategies based on player movements. Similarly, in tennis, intention recognition can aid in identifying whether a player intends to serve, volley, or execute a baseline shot, enabling predictive analytics for opponents' strategies. Moreover, intention recognition can be applied in team sports like volleyball to anticipate players' positioning and actions during rallies, facilitating proactive decision-making and strategic planning. Overall, the adaptability of intention recognition lies in its ability to analyze contextual cues, infer underlying motivations, and predict future actions, thereby enhancing situational awareness and decision-making in various ambiguous game scenarios.

7. Conclusions

In this study, we chose a handball event that occurs in a football game (Not to be confused with the game of Handball) and technically defined it using object detection and robotic perception and decided whether it was intentional or not. In our experiments, we detected hand and ball objects at an accuracy of 96% and 100% respectively, then we determined the intention of a player by their gaze direction. Unity Eye has been used to generate eyes and direction to train a model to detect the gaze direction. Gaze direction was used to define the intention of a player. In the scenario of accurately detected objects, the results of the handball event definition are satisfactory using a single camera.

In future work, we want to train the object detection model with more datasets and include more cameras to collaborate and define the handball event. Intention recognition in other environments is also included in our future work.

Author Contributions: M.M.H.: Conceptualization, Methodology, Software, Writing—Original Draft Preparation. S.K.: Investigation, Data Curation, Visualization, Supervision, Project Administration, Writing—Review & Editing. K.T.: Resources, validation—Review & Editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. FIFA. FIFA Survey: Approximately 250 Million Footballers Worldwide. 15 September 2006. Available online: http://access.fifa.com/infoplus/IP-199_01E_big-count.pdf (accessed on 5 June 2023).
2. The International Football Association Board. Laws of the Game 21/22. Available online: https://downloads.theifab.com/downloads/laws-of-the-game-2021-22?l=en (accessed on 5 June 2023).
3. NPR. A Stampede at a Soccer Match Has Killed at Least 125 People in Indonesia. 2 October 2022. Available online: https://www.npr.org/2022/10/01/1126439213/indonesia-soccer-riot-fans-dead (accessed on 5 June 2023).
4. The International Football Association Board/Guardians of the Laws of the Game, "Law 12 Fouls and Misconduct". Available online: https://www.theifab.com/laws/latest/fouls-and-misconduct/#restart-of-play-after-fouls-and-misconduct (accessed on 5 June 2023).
5. The International Football Association Board/Guardians of the Laws of the Game, "Law 1 the Field of Play". Available online: https://www.theifab.com/laws/latest/the-field-of-play/#field-surface (accessed on 5 June 2023).
6. Hassan, M.M.; Karungaru, S.; Terada, K. Recognizing football game events: Handball based on Computer Vision. In Proceedings of the 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Busan, Republic of Korea, 28–31 August 2023; pp. 1158–1163. [CrossRef]
7. Tabii, Y.; Rachid, O. A Framework for Soccer Video Processing and Analysis Based on Enhanced Algorithm for Dominant Color Extraction. *Int. J. Image Process.* **2009**, *3*, 131.
8. Xu, P.; Xie, L.; Chang, F.; Divakaran, A.; Vetro, A.; Sun, H. Algorithms And System for Segmentation And Structure Analysis in Soccer Video. In Proceedings of the IEEE International Conference on Multimedia and Expo 2001, Tokyo, Japan, 22–25 August 2001. [CrossRef]
9. Xu, J.; Zhang, Y.; Ye, A.; Dai, F. Real-time detection of game handball foul based on target detection and skeleton extraction. In Proceedings of the 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), Fuzhou, China, 24–26 September 2021; pp. 41–46. [CrossRef]
10. Host, K.; Ivašić-Kos, M. An overview of Human Action Recognition in sports based on Computer Vision. *Heliyon* **2022**, *8*, e09633. [CrossRef] [PubMed]
11. Host, K.; Ivašić-Kos, M.; Pobar, M. Action Recognition in Handball Scenes. In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 645–656. [CrossRef]
12. Zhao, Z.; Chai, W.; Hao, S.; Hu, W.; Wang, G.; Cao, S.; Song, M.; Hwang, J.-N.; Wang, G. A Survey of Deep Learning in Sports Applications: Perception, Comprehension, and Decision. *arXiv* **2023**, arXiv:2307.03353.
13. Zhang, Z.; Zeng, Y.; Jiang, W.; Pan, Y.; Tang, J. Intention recognition for multiple agents. *Inf. Sci.* **2023**, *628*, 360–376. [CrossRef]
14. Han, T.A.; Pereira, L.M. State-of-the-art of intention recognition and its use in decision making. *AI Commun.* **2013**, *26*, 237–246. [CrossRef]
15. Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; Girshick, R. *Detectron2*. 2019. Available online: https://github.com/facebookresearch/detectron2 (accessed on 10 July 2023).
16. Kurtanović, J. Deep Learning—Instance Segmentation. 10 May 2021. Available online: https://serengetitech.com/tech/deep-learning-instance-segmentation/ (accessed on 9 July 2023).
17. Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context. *arXiv* **2014**, arXiv:1405.0312. https://doi.org/10.48550/arXiv.1405.0312.
18. Girshick, R. *Fast R-CNN*. CoRR, vol. abs/1504.08083, 2015. Available online: http://arxiv.org/abs/1504.08083 (accessed on 5 May 2023).
19. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017. [CrossRef]
20. Cambridge Dictionary. (n.d.) Intention. In Cambridge English Dictionary. Available online: https://dictionary.cambridge.org/us/dictionary/english/intention (accessed on 10 December 2023).

21. Aristotle, *On the Soul*, Provided by The Internet Classics Archive. See Bottom for Copyright. Available online: http://classics.mit.edu/Aristotle/soul.html (accessed on 15 January 2024).
22. Strasburger, H.; Rentschler, I.; Jüttner, M. Peripheral vision and pattern recognition: A review. *J. Vis.* **2011**, *11*, 13. [CrossRef] [PubMed]
23. Varsity Tutors. (n.d.). Distance Formula. Available online: https://www.varsitytutors.com/hotmath/hotmath_help/topics/distance-formula (accessed on 10 June 2023).
24. Speed, Distance, Time Calculator. Calculatorsoup.com. Available online: https://www.calculatorsoup.com/calculators/math/speed-distance-time-calculator.php (accessed on 12 August 2023).
25. Goral, K. Examination of Agility Performances of Soccer Players According to Their Playing Positions. *Sport J.* **2015**, *24*. Available online: https://thesportjournal.org/article/examination-of-agility-performances-of-soccer-players-according-to-their-playing-positions/ (accessed on 8 January 2024). [CrossRef] [PubMed]
26. Beau Bridges. "How Fast Can A Soccer Ball Be Kicked?" Soccer Novo, 22 March 2023. Updated 3 January 2024. Available online: https://soccernovo.com/how-fast-can-a-soccer-ball-be-kicked/ (accessed on 7 January 2024).
27. ARCCA. Human Reaction Time in Emergency Situations. 1 October 2021. Available online: https://arcca.com/blog/human-reaction-time-in-emergency-situations/#:~:text=Reaction%20time%20is%20defined%20simply,time%20is%20actually%20more%20complex (accessed on 13 December 2023).
28. Lacey, T. Tutorial: The Kalman Filter. Massachusetts Institute of Technology. Available online: https://web.mit.edu/kirtley/kirtley/binlustuff/literature/control/Kalman%20filter.pdf (accessed on 5 April 2023).
29. Bittle, W. SAT (Separating Axis Theorem). 1 January 2010. Available online: https://dyn4j.org/2010/01/sat/ (accessed on 23 May 2023).
30. Wood, E.; Baltrušaitis, T.; Morency, L.-P.; Robinson, P.; Bulling, A. Learning an Appearance-Based Gaze Estimator from One Million Synthesized Images. In Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications, Charleston, SC, USA, 14–17 March 2016; pp. 131–138.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.