Single Modal Motion Sensing for Low-Cost Data Collection in Sports

Submitted to
Graduate School of Information Science and Technology
Osaka University

January 2022

Ryosuke HASEGAWA

List of Publications

Related Journal Articles

(1) Ryosuke Hasegawa, Akira Uchiyama, Takuya Magome, Juri Tatsumi, Teruo Higashino, "Maneuver and Turn Classification in Wheelchair Basketball Using Inertial Sensors" *Journal of Information Processing*, vol. 29, pp. 70–80, January 2021.

(2) Ryosuke Hasegawa, Akira Uchiyama, Issei Ogasawara, Daigo Muramatsu, Fumio Okura, Hiromi Takahata, Ken Nakata, Teruo Higashino, "Close-Contact Detection Using a Single Camera for Sports Considering Occlusion" *IEEE Access*, (accepted)

Related Conference Papers

(1) Ryosuke Hasegawa, Akira Uchiyama, Teruo Higashino, "Maneuver Classification in Wheelchair Basketball Using Inertial Sensors", in *Proceedings of the 12th International Conference on Mobile Computing and Ubiquitous Networking 2019 (ICMU 2019)*, pp. 1–6, October 2019.

(2) Ryosuke Hasegawa, Akira Uchiyama, Issei Ogasawara, Daigo Muramatsu, Fumio Okura, Hiromi Takahata, Ken Nakata, Teruo Higashino, "Human Localization Using a Single Camera Towards Social Distance Monitoring During Sports", in *Proceedings of EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous 2021)*, November 2021.

Other Journal Articles

(1) Issei Ogasawara, Shigeto Hamaguchi, Ryosuke Hasegawa, Yukihiro Akeda, Naoki Ota, Gajanan S. Revankar, Shoji Konda, Takashi Taguchi, Toshiya Takanouchi, Kojiro Imoto, Nobukazu Okimoto, Katsuhiko Sakuma, Akira Uchiyama, Keita Yamasaki, Teruo Higashino, Kazunori Tomono, Ken Nakata, "Successful Reboot of High-Performance Sporting Activities by Japanese National Women' s Handball Team in Tokyo, 2020 during the COVID-19 Pandemic: An Initiative Using the Japan Sports–Cyber Physical System (JS–CPS) of the Sports Research Innovation

Project (SRIP)" *International Journal of Environmental Research and Public Health*, vol. 18, September 2021.

(2) Yusuke Hara, Ryosuke Hasegawa, Akira Uchiyama, Takaaki Umedu, Teruo Higashino, "FlowScan: Estimating People Flows on Sidewalks Using Dashboard Cameras Based on Deep Learning" *Journal of Information Processing*, vol. 28, pp. 55–64, January 2020.

Abstract

Recently, using data in sports has been attracting more and more attention for various purposes such as advanced strategic analysis and efficient player training. For example, if players' positions are obtained, we can know *distance covered* and *playing area* per game for each player. Also, by labeling actions such as passes and shoots in addition to the positional data, we can understand positions where attacks occur frequently. In fact, the German national football team acquired and analyzed various data in a game with high-resolution cameras, such as the speed of players and balls, the movement of the opponent team, etc. for creating a new training menu. As a result, they won the 2014 World Cup with the use of big data. The Japan national rugby team has also achieved great success by thoroughly analyzing positions of the players. Therefore, in many professional sports, data analysis experts participate in teams to collect and analyze practice and match data.

On the other hand, data utilization is not yet widespread in amateur sports such as universities' clubs. This is mainly due to the small number of staff. Even in a major sports data collection company, action data such as the numbers of passes and shoots are collected by manually labeling them. Many tools have been developed to collect data more easily. However, the actions are labeled by three experts in each game. The first person labels actions of the home team, the second one labels actions of the away team, and the third one checks the actions labeled by the other two. Therefore, in amateur sports, it is difficult to collect data and label them from daily practices and games. Additionally, it is hard to spend a lot of time for preparing for data collection because the reservation time for courts or fields is tight. Therefore, in order to secure enough practice time, it is necessary to mitigate the time and workload required to install sensor(s) for data collection. Single modal methods which use a single type of sensor(s) are useful to solve the problem. In these methods, we can start to collect data simply by attaching or deploying devices to the player's body, vehicle, or outside the court. For this reason, in this dissertation, we design single modal methods to collect players' positions and actions as the data necessary for the analysis.

Sports data collection methods are divided into two approaches: sensor-based methods that attach devices to players such as an inertial sensor and a heart rate sensor, and device-free methods that use non-contact devices such as cameras and radars. The movement and state of the player can be directly measured by attaching an inertial sensor to the player. However, there is concern that players' performance may deteriorate due to discomfort when attaching the device. Because of the risk of injury

in sports with contact, wearable devices such as smartwatches are sometimes not preferred as well. On the other hand, when using a camera, we can obtain images without any effect on players' performance. However, we need to detect and track the players in the video. Furthermore, action recognition from the video is required from movements of players. Therefore, we propose data collection methods that support various environments by two types of approaches: inertial sensors and a single camera.

Our goal is to design methods for data collection that support various environments in sports. We rely on a single type of devices for collecting position and action data with low cost in terms of workload. We design methods based on inertial sensors and a single camera. In this dissertation, four primary contributions will be made toward automatically collecting sports data for data analysis in various environments.

First, we design a localization method using inertial sensors. Focusing on wheelchair sports, we perform dead reckoning using the displacement and orientation of a player obtained from three inertial sensors attached to wheelchair wheels axles and under the chair. The relative position from the start position is estimated by accumulating the displacement for a short unit time in the orientation of the wheelchair. However, an accumulation of errors by sharp movement and/or peculiar to the sensor which is called drift becomes a problem. Therefore, in this study, we propose a method for position correction for dead reckoning using inertial sensors. We propose three types of correction methods: correcting by beacon attached to a goal, manually correcting from video, and correcting using collisions between wheelchairs. For evaluation, we collected data from actual wheelchair basketball games. From the result, we confirmed how localization errors accumulate in wheelchair sports and the effect of position correction frequency on position estimation accuracy.

Secondly, we design localization using video so that data can be collected even in a circumstance where an inertial sensor cannot be attached to a player. We detect the skeleton of a person moving in the image and propose a localization method that is robust to the pose during movement. When performing localization using video, it is necessary to first detect people in a video and get the coordinates in the image where they are. However, existing methods are limited in that all skeletons are visible and/or that both feet are not floating in the air. Therefore, in this study, to localize people using the coordinates of their waist, which is likely visible and little height fluctuation during movement. Furthermore, we propose a correction method using the skeleton of the lower body for various poses when the target does not move. For evaluation, we collected images including 4 orientations and 5 poses of people and compared our method with existing methods. As a result, we confirmed that our method mitigated the increase of the error in various environments and this result shows the versatility of our method. Furthermore, a distance between people has become important in sports where masks can be avoided due to the recent spread of coronavirus infection. Therefore, we developed a close-contact detection system using this localization method and we could lead the behavior modification to avoid the close-contact by using it in an actual sports competition.

Thirdly, we propose an action recognition method using inertial sensors. Attaching sensors directly

to the human body is sometimes avoided in sports where contact occurs. Therefore, we attach the sensor to sports equipment and estimate the movement of the arm. In our method, we focus on wheelchair sports, estimate the timing when a force is applied to the wheel from the change in rotation of the wheel, and divide sensor data in time where one maneuver action is likely performed. Next, one maneuver action is classified for each of the divided data sequences. Furthermore, by comparing the movements of the left and right wheels, wheelchair movement recognition such as sprints and turns is performed. By using our method in actual games, we could quantify the tendency and efficiency of maneuver actions for each player. This makes our method support understanding the maneuver behavior with high-level players and making an efficient training menu.

Fourthly, we propose an action recognition method using only video data. In existing action recognition methods, a person is first detected, and then the movement of the arm or foot is recognized for the person in the rectangle which is output from human detection. However, they focus on movements of the limbs without considering positions in the field. Therefore, in the existing methods, it is difficult to classify different actions with similar movements such as passes and shoots. In this study, we created a dataset including actions that are different even if they are similar movements and confirmed the classification problem of the existing methods. Furthermore, we assume that a player's position and movement on the court affect the action decision, and propose an action recognition method considering positional information. As an evaluation result, it can be seen that the actions in sports are related to the position where actions are performed and we can classify with higher accuracy by considering it.

Through these contributions, we have shown that it is possible to automatically collect sports data easily and with a low workload in various environments. This dissertation has established the foundation of a data collection and analysis system, which can be used even in environments with limits to the number of people in the team or time to use the facilities for sports.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recently, using data in sports has been attracting more and more attention, and it is used for various purposes such as advanced strategic analysis and efficient player training. The sports data required for analysis is a wide variety. For example, if players' positions are obtained, we can know *distance covered* and *playing area* per game for each player. By labeling actions such as passes and shoots in addition to the positional data, we can understand a team's tactics positions where attacks occur frequently. Furthermore, we focus on one action such as pushing an object, we may be able to understand how to apply force to the object more quickly and effectively by analyzing the movement in the action. For example, the German national football team acquired and analyzed various data in one game with high-resolution cameras, such as the speed of players and balls, the movement of the opponent team, etc. They created and executed a new training menu based on the data. As a result, it is said that the major factor of the victory in the 2014 World Cup is the use of big data [1]. The Japan national rugby team has also achieved great success by thoroughly analyzing positions of the players [2]. Therefore, in many professional sports, data analysis experts participate in teams to collect and analyze practice and match data. Additionally, such sports data utilization has attracted attentions of many people even in some minor and amateur sports.

On the other hand, data utilization is not yet widespread in amateur sports such as universities' clubs. This is mainly due to the small number of staff. Even in a major sports data collection company, action data such as the numbers of passes and shoots are collected by manually labeling them. Many tools have been developed to collect data more easily. However, the actions are labeled by three experts in each game. The first person labels actions of the home team, the second one labels actions of the away team, and the third one checks the actions labeled by the other two. Therefore, in amateur sports, it is difficult to collect data and label them from daily practices and games. Additionally, it is hard to spend a lot of time for preparing for data collection because the reservation time for courts or fields is tight. Therefore, in order to secure enough practice time, it is necessary to mitigate the time and workload required to install sensor(s) for data collection. Single modal methods which use a single type of sensor(s) are useful to solve the problem. In these methods, we can start to collect data

simply by attaching or deploying devices to the player's body, vehicle, or outside the court. For this reason, in this dissertation, we design single modal methods to collect players' positions and actions as the data necessary for the analysis.

Sports data collection methods are divided into two approaches: sensor-based methods that attach devices to players such as an inertial sensor and a heart rate sensor, and device-free methods that use non-contact devices such as cameras and radars. The movement and state of the player can be directly measured by attaching an inertial sensor to the player. However, there is concern that players' performance may deteriorate due to discomfort when attaching the device. Because of the risk of injury in sports with contact, wearable devices such as smartwatches are sometimes not preferred as well. On the other hand, when using a camera, we can obtain images without any effect on players' performance. However, we need to detect and track the players in the video. Furthermore, action recognition from the video is required from movements of players. Therefore, we propose data collection methods that support various environments by two types of approaches: inertial sensors and a single camera.

To collect these data, many systems and studies have been proposed. For indoor localization with inertial sensors, dead reckoning is used. In dead reckoning, the relative movement trajectory from the initial position is estimated by accumulating the moving direction and displacement for each unit time, which are estimated by geomagnetism, gyroscope, and acceleration. At that time, if an absolute position at any time is given, the other absolute positions in the entire time can be estimated. Using this technique, several localization methods are proposed for smartphones [3] and mobile robots [4]. However, the moving objects in these studies are supposed to move forward and/or along the wall of the building. When using videos for localization, it is necessary to detect and track people and transform those people from the pixel coordinates in the image to the actual coordinates. Therefore, localization is realized by combining the technologies in each task. Several studies have proposed methods using detectors that output bounding boxes or skeletons [5–7]. In these methods, they regard the bottom edge of the bounding box or the midpoint of both ankles as the position of the person in the image. However, because the bounding box does not recognize the parts of the human body, if the lower body is invisible due to occlusion, the other parts such as the waist may become the bottom edge of the bounding box. In addition, even if the parts of the human body are recognized, when the midpoint between both ankles is regarded as the position of the person, the pose in which one leg floats in the air is not considered. It is a big problem because such a pose frequently occurs during exercise. Also, regarding action recognition, many methods specialized for each sport have been proposed because the movements of the arms and legs can be directly collected when the inertial sensor is used [8–10]. However, in sports with hard contacts, the installation on-body locations may be limited. Action recognition using videos is one of the main topics in computer vision, where many methods have been proposed [11, 12]. However, these studies focus on the movement of players itself without positions in the field. Some actions in sports, such as passes and shoots, need to be counted as different actions even if the movements are very similar. Even with the same action "throwing the ball", it is necessary

to recognize where the ball was thrown. However, it is difficult to classify them based on human centric movements. Therefore, these existing methods of localization and action recognition focus on the use in daily life without consideration of actions and pose changes that are unique in sports. It is also sometimes necessary to change the method of collecting data depending on the environment.

Our goal is to design methods for data collection that support various environments in sports. We rely on a single type of devices for collecting position and action data with low cost in terms of workload. We design methods based on inertial sensors and a single camera. In this dissertation, four primary contributions will be made toward automatically collecting sports data for data analysis in various environments.

First, we design a localization method using inertial sensors. Focusing on wheelchair sports, we perform dead reckoning using the displacement and orientation of a player obtained from three inertial sensors attached to wheelchair wheels axles and under the chair. The relative position from the start position is estimated by accumulating the displacement for a short unit time in the orientation of the wheelchair. However, an accumulation of errors by sharp movement and/or peculiar to the sensor which is called drift becomes a problem. Therefore, in this study, we propose a method for position correction for dead reckoning using inertial sensors. We propose three types of correction methods: correcting by beacon attached to a goal, manually correcting from video, and correcting using collisions between wheelchairs. For evaluation, we collected data from actual wheelchair basketball games. From the result, we confirmed how localization errors accumulate in wheelchair sports and the effect of position correction frequency on position estimation accuracy.

Secondly, we design localization using video so that data can be collected even in a circumstance where an inertial sensor cannot be attached to a player. We detect the skeleton of a person moving in the image and propose a localization method that is robust to the pose during movement. When performing localization using video, it is necessary to first detect people in a video and get the coordinates in the image where they are. However, existing methods are limited in that all skeletons are visible and/or that both feet are not floating in the air. Therefore, in this study, to localize people using the coordinates of their waist, which is likely visible and little height fluctuation during movement. Furthermore, we propose a correction method using the skeleton of the lower body for various poses when the target does not move. For evaluation, we collected images including 4 orientations and 5 poses of people and compared our method with existing methods. As a result, we confirmed that our method mitigated the increase of the error in various environments and this result shows the versatility of our method. Furthermore, a distance between people has become important in sports where masks can be avoided due to the recent spread of coronavirus infection. Therefore, we developed a close-contact detection system using this localization method and we could lead the behavior modification to avoid the close-contact by using it in an actual sports competition.

Thirdly, we propose an action recognition method using inertial sensors. Attaching sensors directly to the human body is sometimes avoided in sports where contact occurs. Therefore, we attach the

sensor to sports equipment and estimate the movement of the arm. Because small sensors are developed underway, we believe that the built-in sensors in sports equipment will develop in the future and it does not increase the risk of injury even in sports with many contacts. In our method, we focus on wheelchair sports, estimate the timing when a force is applied to the wheel from the change in rotation of the wheel, and divide sensor data in time where one maneuver action is likely performed. Next, one maneuver action is classified for each of the divided data sequences. Furthermore, by comparing the movements of the left and right wheels, wheelchair movement recognition such as sprints and turns is performed. By using our method in actual games, we could quantify the tendency and efficiency of maneuver actions for each player. This makes our method support understanding the maneuver behavior with high-level players and making an efficient training menu.

Fourthly, we propose an action recognition method using only video data. In existing action recognition methods, a person is first detected, and then the movement of the arm or foot is recognized for the person in the rectangle which is output from human detection. However, they focus on movements of the limbs without considering positions in the field. Therefore, in the existing methods, it is difficult to classify different actions with similar movements such as passes and shoots. In this study, we created a dataset including actions that are different even if they are similar movements and confirmed the classification problem of the existing methods. Furthermore, we assume that a player's position and movement on the court affect the action decision, and propose an action recognition method considering positional information. As an evaluation result, it can be seen that the actions in sports are related to the position where actions are performed and we can classify with higher accuracy by considering it.

In this dissertation, we propose sports data collection methods by two measures, in addition to a method of directly attaching inertial sensors to players, and a not invasive method of using only video taken from the outside, which does not need to consider problems of sensor attachment. As a result, we have realized the construction of automatic data collection platforms in the field of sports. And we focus on the collection of two types of data in each collection measure. One is positional data which is the basis of data analysis and another is action data which is important for understanding the team tactics and/or a player's performance. In the indoor localization task, we propose a method that can be applied even in an environment with sharp movement during exercise, using the characteristics of pose changes, vehicles, and actions specific to sports. Additionally, when using inertial sensors in an action recognition task, by using the device attached to the sports equipment, we recognize fine arm movements without directly attaching them to players. When using video in action recognition, we propose a model specialized for sports considering positional data.

Through these contributions, we have shown that it is possible to automatically collect sports data easily and with a low workload in various environments. This dissertation has established the foundation of a data collection and analysis system, which can be used even in environments with limits to the number of people in the team or time to use the facilities for sports.

The rest of this dissertation is organized as follows. Chapter 2 reviews related work on localization and action recognition. Chapter 3 explains the dead reckoning for localization, which is using inertial sensors. Chapter 4 proposes the localization method using a single camera and position correction method by considering human poses. Chapter 5 describes maneuver action recognition method using inertial sensors. Chapter 6 proposes action classification model considering positional data using only video. Finally, Chapter 7 summarizes and concludes this dissertation.

Chapter 2

Related Work

2.1 Indoor Localization by Attaching Device to Targets

Localization method by the sensor device alone is called dead reckoning [13]. In dead reckoning, the relative position from the start position is estimated by accumulating the displacement for a short unit time in the orientation of the targets. To deal with this problem, there is a method that uses deep neural networks to dynamically adapt the noise parameters of the filter. It achieve competitive performance with top-ranked methods which use LiDAR or stereo vision. However the targets of the method is only wheeled vehicle and the method does not assume sharp movements occurred in sports. Therefore, to reset the accumulated error, dead reckoning is generally used in combination with other localization methods such as using Wi-Fi [14, 15], beacon [3, 16], camera [17, 18], laser [4], and etc. For example, in Reference [3], Using a smartphone in a pedestrian's pocket, the number of steps is estimated from the accelerometer, autonomous navigation is performed in combination with the direction estimated from the angular velocity sensor, and the position is corrected when a signal is obtained from the beacon. Reference [4] proposes a method that combines dead reckoning based on the speed of the left and right wheels of a wheeled mobile robot and laser positioning.

When using radio waves, the distance between transmitting and receiving devices can be measured from the time until the arrival of radio waves. Therefore, by placing multiple radio wave transmitters and receivers on the outside and having the tracking target also have a device, it is possible to localize. Many researchers have proposed various methods using different types of devices such as radio frequency (RF) [19–21]. Bluetooth and Wi-Fi are widely used for close-contact tracing owing to the wide availability of smartphones. However, the localization accuracy is typically up to a few meters [20, 21], which is not enough for distance-based close-contact detection. Recently, the millimeter wave has attracted the attention of researchers for localization because it has become available in IEEE 802.11ad and 5G cellular networks [22]. Although it provides centimeter-level localization accuracy [19], the deployment cost is still large. Furthermore, because RF signals are reflected, refracted, and attenuated by people and walls, there are concerns about vulnerability to dynamic environment.

This is a major problem of localization using radio frequency and there are many studies to handle the problem [23, 24].

2.2 Indoor Localization without Attaching Device to Targets

When localization from the outside without attaching a sensor to a player, methods using a camera or LiDAR (Light Detection And Ranging) has been proposed. Using LiDAR, we can measure the distance to objects and humans with centimeter accuracy by measuring the time of flight of laser pulses [25]. References [26,27] proposed target localization using LiDAR fixed in the target environment. However, we need to deploy LiDAR while incurring deployment cost although it can localize and track targets accurately.

When using a camera, a commercially available camera can be used, so we can realize localization lower cost than LiDAR. However, it is necessary to detect and track players in the video. For human detection in an image, methods for outputting a rectangle surrounding a person [28–31], methods for detecting the skeleton of the person body [32–34], and methods for distinguishing the person or not in pixel units of an image [35–37] have been proposed. Detecting a person on pixel-by-pixel is more accurate, but inference times are longer. When using detector which output bounding box of a person, it can work in real time even if not using graphics processing units with high cost. For tracking, many methods have also been proposed [38–41]. For example, in Reference [38],the authors proposed approach that consisted in predicting object motion using the Kalman filter [42] and then associating the detections together with the help of the Hungarian algorithm [43]. In addition, in order to analyze personal data, it is necessary to identify the person in the video. For this reason, a method of identifying by a player's uniform number [44], a method of identifying by a player's face [45], and a method combining a machine learning algorithm and textual information such as manually labeled actions [46] have been proposed. By combining the above techniques, the localization of a person in an image can be realized.

Next, in order to get the actual position in the court for the person, it is necessary to transform the position from coordinates in image to coordinates in actual world . When two or more cameras are used, the actual position of the person can be estimated from the parallax if the positional relationship between the cameras is given [47]. On the other hand, when using one camera, homography [48] is used. Homography is a transformation that projects a plane to another plane, given the four point correspondences between the two planes. Therefore, a homography transformation matrix can transform pixel coordinates in an image into the actual positions, given the distance between the four points in the real world.

Recently, with the spread of coronavirus, the distance between people has become important to prevent the infection, and along with this, several localization methods using a single camera have been proposed [5–7, 49, 50].. For example, References [5, 6] calculate the inter-person distance using

homography transformation with the bottom edge of the bounding box as the position of the person.

2.3 Action Recognition Using Inertial Sensors

The inertial sensor has many uses depending on the place to fix it and the environment in which it is used. Therefore, there are numerous studies on behavior recognition using sensors. When using the inertial sensor built into smartphones and smart watches, there are many studies on action recognition in daily life [51–54]. For example, Reference [54] is focusing on action recognition algorithm for similar gait actions using an inertial sensor in wearable and portable electronic devices such as smartphones, tablets, and smartwatches. In this paper, a method to classify similar gait action (such as walking on flat ground, up/down stairs, and up/down a slope) is proposed.

Because the smartwatch is fixed to the wrist and can be worn during exercise, it is also used for action recognition in sports [55–58]. For example, Reference [55] proposes a method that can recognize 18 types of shooting actions in basketball with high accuracy using a wristband with a built-in inertial sensor. In addition, Reference [56] proposes a method for recognizing the type of swimming and turn in swimming using a commercially available smartwatch.

Furthermore, there are also researches on action recognition by attaching inertial sensors to various parts of the body for the purpose of knowing more detailed movements of the body parts [10,59,60]. For example, Reference [59] proposes a method for recognizing six types of play such as pass and receive in field hockey from sensors attached to the chest, waist, and left and right wrists. Reference [10] propose a classification method of football kick types using ankle-mounted inertial sensors. Because important movements differ depending on each sport, it is necessary to carefully select the place to attach the device to recognize fine actions using inertial sensors.

In wheelchair basketball, there are no studies about action recognition as far as we know. However, there have been several studies on wheelchair basketball using inertial sensors. For example, References [61–63] study the relationship between the level of disabilities and the performance. This relationship must be determined in order to harmonize players with different level of disabilities, so a classification system is used to evaluate the functional abilities of players on a point scale of 1 to 4.5. Reference [61] reports the level of disability and the number of successful shots and passes are correlated for professional female wheelchair basketball players.These studies do not investigate the design of data analysis in wheelchair basketball because they focus on the medical aspect of wheelchair basketball rather than sports. Also, other studies from a medical perspective investigate the risk of heatstroke [64] or injury [65] during training and games.

Some research work on quantifying athletic performance is carried out by investigating the relationship between moving speeds and wheelchair configurations [66–68]. Such studies reveal the effectiveness of data analysis in wheelchair basketball although they rely on the measured raw data of acceleration and angular velocity in controlled environment. In wheelchair basketball, there are no studies about

action recognition as far as we know.

2.4 Action Recognition Using Cameras

Action recognition using cameras is one of the main topics in computer vision, and large datasets such as AVA [69] and Kinetics-600 [70] have been prepared and there are many studies [71–73]. For example, , Reference [73] propose an improved *Multi-scale Vision Transformer (MViT)* as a general hierarchical architecture for visual recognition. They use *Transformer* [74] that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data. It is used primarily in the field of natural language processing, however recently it began to be used in the field of computer vision. *MViT* is achieves state-ofthe-art accuracy on widely-used benchmarks across image classification, object detection, instance segmentation and video action recognition. In terms of action recognition, this model achieved 87.9 % for Kinetics-600.

To recognize actions in video, it is need to detect and track people as well as position estimation. However, in action recognition task we are given a video clip that one human performed one action, and then we classify it. Therefore, the human detection and tracking is out of scope of action recognition task. However, there are several studies that is trying to localize the action both spatially and temporally in video [11, 75]. For example, in Reference [11], *Slow Pathway*, which detects objects at a low frame rate and captures spatial features, and *Fast Pathway*, which detects moving objects at a high frame rate and captures temporal features. We propose a behavior recognition model *Slowfast* with. This model achieved behavior recognition with high accuracy of 81.8 % for data with 600 types of action labels called Kinetics-600. Recently, in Reference [12], the skeleton obtained by using a skeleton detector for a person is used as an input for *Slow Pathway* of *Slowfast* instead of a low frame rate image.

Chapter 3

Localization in Wheelchair Sports Using Inertial Sensors

3.1 Introduction

Recently, using big data in sports has been attracting more and more attention, and it is used for various purposes such as advanced strategic analysis and efficient player training [76,77]. For example, the German national football team acquired and analyzed about 40 million pieces of data in one game with high-resolution cameras, such as the speed of players and balls, the movement of the opponent team, etc. They then created a new training menu based on that data and executed it. As a result, they won the 2014 World Cup and it is said that using big data is a major factor [1]. The Japan national rugby team has also achieved great success by thoroughly analyzing the position of the players [2]. From such a background, data utilization is being promoted in minor and/or amateur sports such as universities' clubs. In wheelchair basketball, which is one of the sports for people with disabilities, it is also required to support skill improvement and strategy planning by using data, and several attempts have been made so far using inertial sensors [63, 78].

In order to collect such a huge amount of data, specialists or many staff for data analysis are required because it is needed to install cameras and sensors in right place for the data you need. Therefore, it is difficult to collect data for daily practice and games in minor sports. Therefore, it is desired to collect various data without complicated work as much as possible. Data collected in sports include bio-metric information such as heart rate in addition to positions and accelerations related to the movement of athletes and balls. In particular, players' position is the basic data with the widest range of applications. Therefore, in this chapter, we aim to localize each player in wheelchair basketball at the lowest possible workload.

For human localization outdoors, GPS (Global Positioning System) is mainly used, and the position of a player can be tracked with high accuracy in a place where there are no tall buildings in the vicinity, such as a soccer field. It is expected that the accuracy of GPS will be further improved by the start of

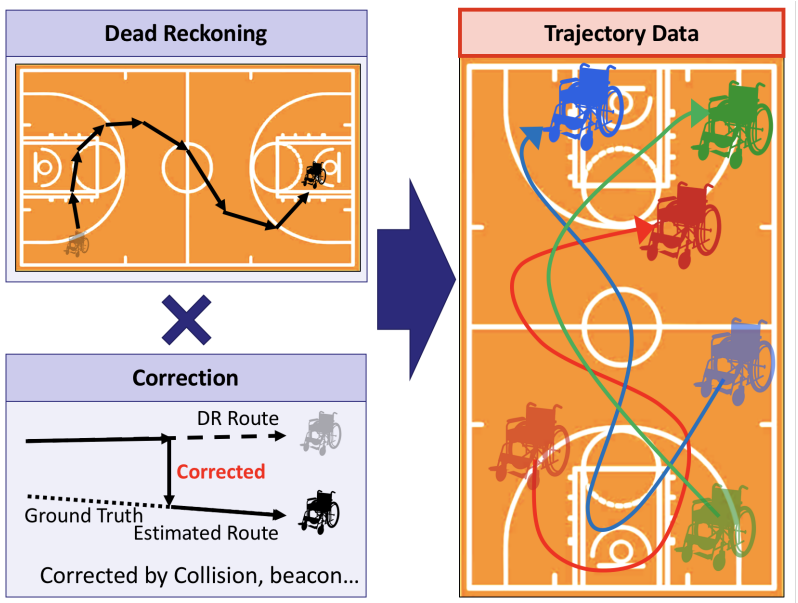

Figure 3.1: Method Overview

the service of Quasi-Zenith Satellite System [79]. On the other hand, because GPS satellite signals do not reach indoors, various localization methods such as Wi-Fi and cameras are proposed. Especially in sports, since there are few privacy issues, it is common to use a camera, and there are some products [80, 81]. However, in video-based localization, there are tracking failures due to occlusions. In order to reduce the false negatives in detection, it is necessary to install multiple cameras, which increases the time and effort for measurement. In addition, it is necessary to identify who the player is in the video, and under the present circumstances, manual correction is required each time tracking stops.

On the other hand, in the positioning method in which a sensor is attached to a player, it is known which player is wearing which sensor, so that it is not necessary to identify the person, which is a big merit. Because wheelchairs move by wheels, we can obtain players' trajectory by using the number of rotations of the wheels. In the case of robots, many methods that combine dead-reckoning by wheel rotations and distance estimation by laser have been proposed [4]. However, in the case of a non-electric wheelchair, unlike a robot, it is difficult to incorporate a sensor in the axle, so it is difficult to get the wheel rotation speed accurately. Therefore, in paper [82], the number of rotations of the wheel is measured by a magnet attached to the spoke of a wheelchair and they localize based on a floor map. However, because the number of spokes in a wheelchair is limited, the resolution of the measurable rotation speed becomes rough, and there is a concern that sufficient accuracy cannot be obtained with wheelchair basketball, which has a big and sharp movement.

Table 3.1: Sensor measurement range

Sensor	Unit	Range
Acceleration	G	[-16, +16]
Angular Velocity	dps	[-1500, +1500]
Magnetic Field	Gauss	[-10, +10]

In this chapter, we propose a localization system using a 9-axis sensor by dead reckoning in wheelchair basketball. We can use the system with even a low workload because it needs to attach only three sensors to each player. First, in order to estimate the displacement, the angular velocity is measured by sensors attached to the left and right axles, and we estimate the rotation speed of each wheel. Next, the geomagnetism is measured by a sensor attached under the chair. In wheelchair basketball, the quantity of change in orientation of wheelchairs may not be obtained accurately only from the rotation speed of the wheels because the wheels are sometimes floating in the air due to collisions between wheelchairs, so the orientation of wheelchairs are estimated using geomagnetism. In an environment for sports such as a gym, there are no large electronic devices that disturb the geomagnetism, so unlike general indoor localization, we can the orientation of wheelchairs from geomagnetism sensors. Furthermore, we examine several position correction methods for the errors accumulated by dead reckoning. Specifically, we considered three types: (1) collision between wheelchairs, (2) BLE (Bluetooth Low Energy) beacon installed at the goal, and (3) random manual correction. The third correction result helps you know how often you need to get the right position to achieve the desired accuracy. In wheelchair basketball, wheelchairs frequently collide with each other for the purpose of blocking the way to a goal. When a collision occurs, the wheelchairs that collided must be adjacent to each other, so it can be used for position correction. An accelerometer under a wheelchair is used to detect a collision.

For evaluation, we collected data on 6 players in a wheelchair basketball practice game. As a result, it can be seen that the average localization error is 5.3 m at the maximum for a game of about 5 minutes, although the error accumulates with time only by dead reckoning. In addition, we evaluated the effect of using position correction methods by simulation, it can be seen that the average error is up to 2.8 m when collision and beacon correction are used together. Although sufficient accuracy cannot be obtained for applications such as detailed tactical analysis, we can use the system for the purpose of collecting data on wheelchair movement such as movement patterns, sprints, and turns of each player.

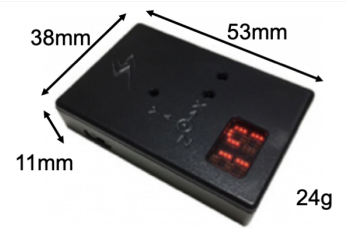

Figure 3.2: 9-axis motion sensor

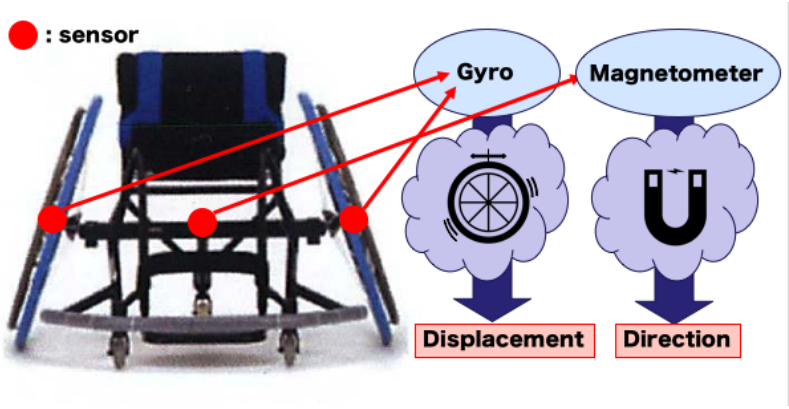

Figure 3.3: Sensor Equipment for Dead Reckoning

3.2 Method

3.2.1 Overview

Figure 3.1 shows the overview of the localization method examined in this chapter. In order to collect detailed data on wheelchair maneuver action, it is essential to attach an inertial sensor to a wheelchair. As shown in Fig. 3.3, inertial sensors are fixed to the axles of the left and right wheels. We use DSP wireless 9-axis motion sensors manufactured by SPORTS SENSING Co., LTD [1] (Figure 3.2). The sensor is capable of measuring 3-axis acceleration, 3-axis angular velocity, and 3-axis geomagnetic data at a sampling rate of 200 Hz. The measurement ranges of the sensor are shown in Table 3.1. Hereafter, we use $[radians/second]$ as the unit of angular velocity unless otherwise stated.

By combining these sensor data and wheelchair setting information (e.g. wheel size), the displacement and change in orientation of a wheelchair per unit time are estimated. By repeating this, the trajectory of each player can be estimated by dead reckoning. Because the position obtained by dead reckoning is a relative position, to get where a player is on the court it needs at least one absolute position. Therefore, the absolute position at the start of the match was given manually. Since the error

[1] https://www.sports-sensing.com/products/sensor/dspmotion/dspms.html

that accumulates over time is unavoidable with localization using only dead reckoning, we assumed methods to infrequently correct position. Specifically, we assumed two correction methods using absolute position which are obtained by BLE beacons or manually at random timing. Additionally, when wheelchairs collide they exist at the same position. Therefore we also assumed a correction method using relative position between two wheelchairs which collide.

3.2.2 Dead Reckoning

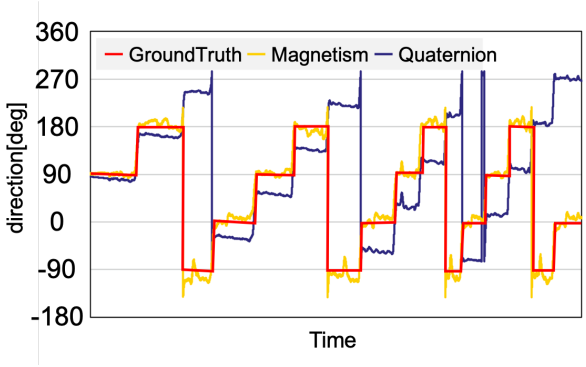

Figure 3.4: Geomagnetic Measurement Results

Localization

Position estimation by dead reckoning is performed by adding a movement vector in the current frame to the position before one frame. When the estimated position at time $t-1$ is (x^{t-1}, y^{t-1}) and the displacement from time $t-1$ to t is d, and the orientation at time $t-1$ is ψ^{t-1}, (x^t, y^t) which is the next estimated position at time t is shown as follows.

$$x^t = d^{t-1} \cos \psi^{t-1} + x^{t-1} \tag{3.1}$$

$$y^t = d^{t-1} \sin \psi^{t-1} + y^{t-1} \tag{3.2}$$

Displacement Estimation

Unlike pedestrians, wheelchairs have wheels, so when the angular velocity at time t is θ, the amount of rotation Θ at time T is shown as follows.

$$\Theta = \int^{T} \theta(t) dt \tag{3.3}$$

The angular velocity takes a positive value when the wheelchair moves forward. When the diameter of the wheel is R, the displacement of each of the left and right wheels can be expressed as follows.

$$d = R \times \frac{\Theta}{\pi} \tag{3.4}$$

The amount of movement d^t for the entire wheelchair is calculated using the displacement of left and right wheels as follows.

$$d^t = \frac{d_r + d_l}{2} \tag{3.5}$$

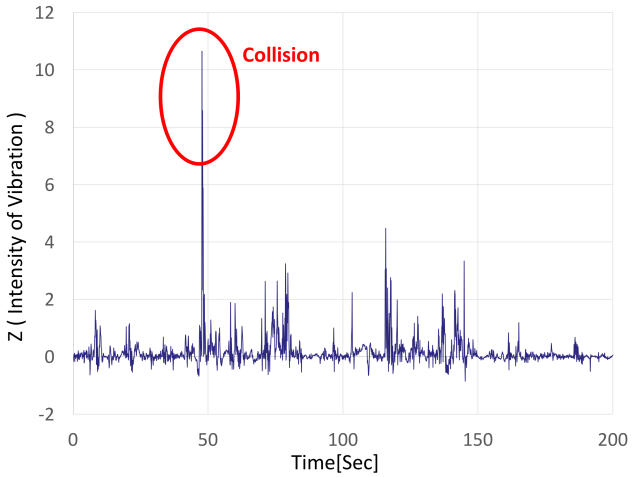

Figure 3.5: Difference in Intensity of Vibration between Collision and Normal

Orientation Estimation

The orientation of a wheelchair can be calculated from the displacement of both wheels and the distance between the wheels, but the errors accumulate by the errors of displacement of the wheels. However, wheels are frequently slipping or floating in the air during matches due to collision or hard braking. Therefore, it is hard to use the displacement of both wheels to get the orientation of a wheelchair. We thus use a geomagnetism sensor to get the orientation of the wheelchair. In the general indoor position estimation method, the human rotation is often estimated by a gyro sensor rather than a geomagnetism sensor, but in an environment such as a gymnasium, there are almost no electronic devices that cause disturbance of the geomagnetism, so even using geomagnetism we believe that the orientation can be obtained with high accuracy. In fact, we conducted a preliminary experiment in an environment where there are no electronic devices. In the preliminary experiment, a wheelchair orbited along the line of the outer circumference of the court. The result is shown in Figure 3.4. From this result, it can be seen that the orientation estimation by the geomagnetic sensor installed under the chair can be performed with higher accuracy than the quaternion combined with the accelerometer and gyro.

3.2.3 Localization Correction Method

Correction Using BLE Beacons

By using a BLE beacon, when the beacon signal is received we can know that the beacon is nearby. In wheelchair basketball, it is important to be under goals both when attacking and when defending, so players tend to gather under goals. Therefore, we consider correcting the position of the player under goals by attaching a beacon to goals. Position correction is performed after a beacon attached to a goal cannot receive the signal from a beacon attached to a wheelchair. Let $0, \ldots, F$ be a period from when a wheelchair moves to under the goal and starts receiving the beacon signal until it cannot be received, the estimated position P of the wheelchair during the beacon reception period is expressed as follows.

$$P = \{(x_0, y_0), (x_1, y_1), (x_2, y_2), ..., (x_F, y_F)\} \tag{3.6}$$

At that time, the position $P_{average} = (x_{average}, y_{average})$, which is the average of the estimated position, is calculated as follows.

$$P_{average} = (\frac{\sum_{i=0}^{F} x_i}{F}, \frac{\sum_{i=0}^{F} y_i}{F}) \tag{3.7}$$

Wheelchair position correction is performed by shifting the trajectory during the beacon reception period so that the $P_{average}$ is at the coordinates directly under the goal.

Correction Using Collisions between Wheelchairs

In wheelchair basketball, colliding a wheelchair is frequently occurred for stopping an opponent's wheelchair when defending. Therefore, we consider a method of detecting a collision from acceleration data and correcting the position of each wheelchair.

Collision Detection Intensity of vibration Z^t is obtained by subtracting the gravitational acceleration $1G$ from the resultant acceleration of the 3-axis accelerometer (a_x, a_y, a_z) as follows.

$$Z^t = |\sqrt{{a_x^t}^2 + {a_y^t}^2 + {a_z^t}^2} - 1| \tag{3.8}$$

Figure 3.5 shows the change in the intensity of vibration in a wheelchair basketball match. From the figure, it can be seen clearly that the intensity of vibration during collision is more pronounced than during normal movement. Therefore, collision is detected by threshold. If the time at which a collision is detected in both wheelchairs is close, it is considered that a collision has occurred between the two applicable wheelchairs. In the video of a match we collected, there were no more than two collisions per second. Therefore, we considered it is caused by the same collision if the time difference between the two big vibrations was within 0.5 seconds. If large vibrations are observed in three or more wheelchairs, we do not perform the correction. We have chosen the threshold empirically in which recall was high to prevent that the position would not be corrected by false positives of collision detection.

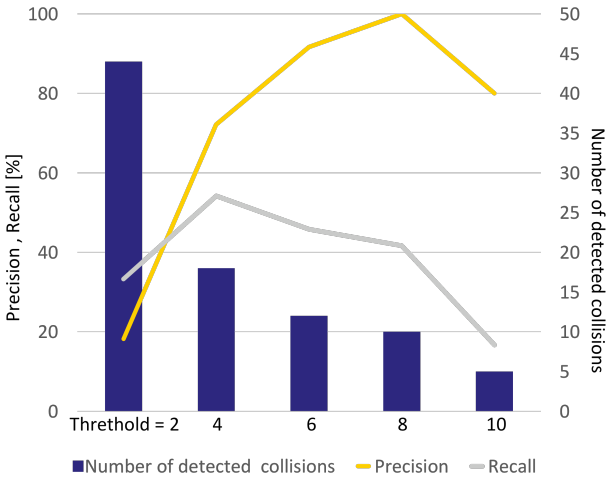

Figure 3.6: Result of Collision Detection

Correction Using Collision There are several possible methods for position correction when a collision is detected. In this chapter, we consider two types of correction methods. The first is a method of shifting both positions to the midpoint between the original position of two wheelchairs. The second method is to set reliability at each wheelchair and shift both positions to the weighted position according to the ratio of reliability score. The reliability score was set to 1 as the initial value and to be attenuated by 0.001 every frame. Please note that the lower limit of reliability score was set at 0.01. The reliability is reset to 1 when the position is corrected by beacons or manually. Assuming that the positions of both wheelchairs at the time of collision are $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$ and the reliability score is L_1 and L_2, respectively, the corrected positions $P_{collision}$ are as follows.

$$P_{collision} = P_1 + (P_2 - P_1) \times \frac{L_2}{L_1 + L_2} \tag{3.9}$$

Next, the reliability score $L_{collision}$ of both wheelchairs are updated according to the ratio of the original reliability score. The equation for the update is as follows.

$$L_{collision} = L_1 + (L_2 - L_1) \times \frac{L_2}{L_1 + L_2} \tag{3.10}$$

3.3 Evaluation

3.3.1 Evaluation Setting

We collected data in an actual wheelchair basketball match to evaluate the proposed method. The time of match was 5 minutes and 17 seconds long and the sampling rate of the sensor data and video were 200 and 30 frame per second, respectively. The ground truth of the players' position was obtained

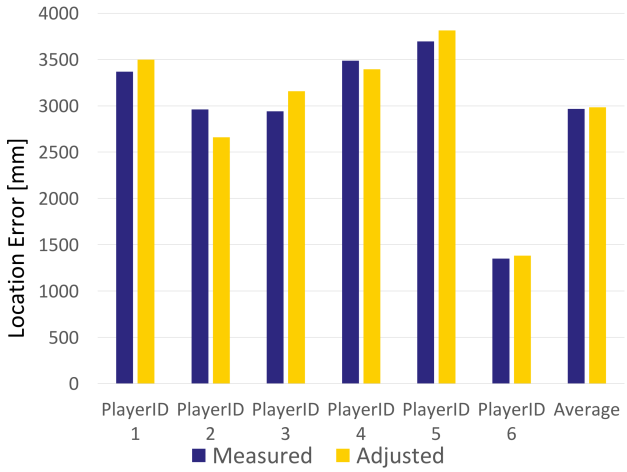

Figure 3.7: Effect of Adjustment of a Wheel Diameter

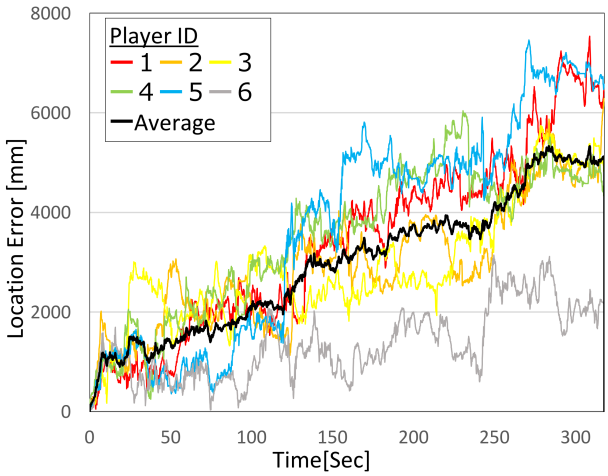

Figure 3.8: Changes in Location Error Using Dead Reckoning w/o Correction Over Time

from the video. Players were detected in the video and they were manually labeled player ID. For frames that were not detected due to occlusion, etc., the trajectory of the player was created by linear interpolation. To calculate the location error between the ground truth and the estimated position, it is necessary to compare the samples observed at the same time. Therefore, the sampling rate of each data has been unified to 10 frames per second because the greatest common divisor of each sampling rate is 10. Euclidean distance is used for location error There were 10 players who participated in the match, but due to a defect in the sensor or refusal to attach the sensor, only 6 players could be collected. The wheel diameter of wheelchairs used by players were 670 [mm] only for player ID 3 and

610 [mm] for the other players. The team of players who could collect data was divided into 4 and 2 players. Since the geomagnetism could not be recorded at the time of data collection, orientation of wheelchair extracted from the video was used in evaluation.

3.3.2 Results

Collision Detection

Figure 3.6 shows the evaluation results of collision detection. From this result, as the threshold becomes smaller, the more false positives increases because movements other than collisions are included. Even if there is an actual collision, if another large vibration is detected at the same time, the collision do not be detected due to matching issue. Therefore, even if threshold set lower the recall decreases in some cases. On the other hand it can be seen that the larger the threshold, the more precision increases. When the threshold was 8, the precision achieved 100Considering the use for position correction, it is desirable that the precision rate is high because correcting positions between the wrong players will greatly increase the location error. Based on this evaluation result, we decided to set the threshold to 8 in subsequent evaluations.

Localization without Correction

In dead reckoning, wheel diameter measurements error can have a significant impact on results. Therefore, we first evaluated the location error when the wheel diameter was set so that the actual length of the trajectory distance would be the same as the estimated length of the trajectory distance. The results are shown in Figure 3.7. From this result, it can be seen that adjusting the wheel diameter does not necessarily improve the performance, so factors other than the wheel diameter measurement error are large. Therefore, in the subsequent evaluations, the wheel diameter measured in advance is used.

Figure 3.8 shows the change in location error of dead reckoning over time.From this result, it can be seen that the error gradually increases with time. This is an unavoidable issue due to the characteristics of dead reckoning, and it is necessary to handle the problem of error accumulation by correcting the position using BLE beacons and the like. Nevertheless we used dead reckoning without correction that gave only the initial position, we achieved the average error 5.3 m at the maximum in the game of about 5 minutes long. It can also be seen that the increase in the average location error in 2 minutes was suppressed to about 2.2 m.

Effect of Correction

Correction by Beacon Because the BLE beacon could not be installed at the time of data collection, the correction effect of the BLE beacon was evaluated by simulation. We assumed a beacon attached a goal can receive a signal of a beacon attached a wheelchair in an position of a circle with a radius of 1 m centered under the goal. The result is shown in Figure 3.9, and it can be seen that

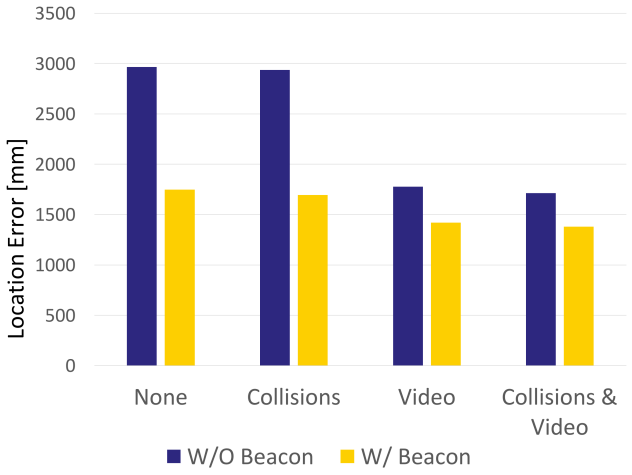

Figure 3.9: Location Error of Correction by Beacon

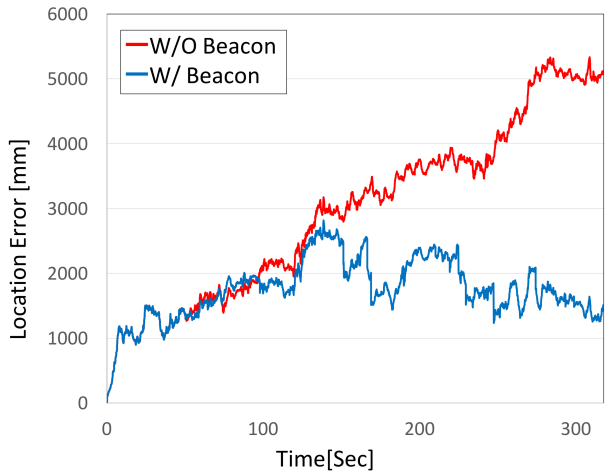

Figure 3.10: Changes in Location Error of Correction by Beacon Over Time

the average error can be suppressed to a maximum of about 2.8 m by correcting with a beacon in the game of about 5 minutes long. This reduces the location error by 41.1 % compared to dead reckoning without correction. It can also be seen that the error can be more suppressed by combining with other correction methods. Figure 3.10 shows the change in the average error over time when corrected using the beacon. From this result, it can be seen that the error continues to be accumulated without correction, but an average error does not exceed 2.8 m by correction using a beacon. From these results, although it is not sufficient accuracy for tactical analysis where the positional relationship between players is important, we believe it can use for the purpose of collecting data on wheelchair movements

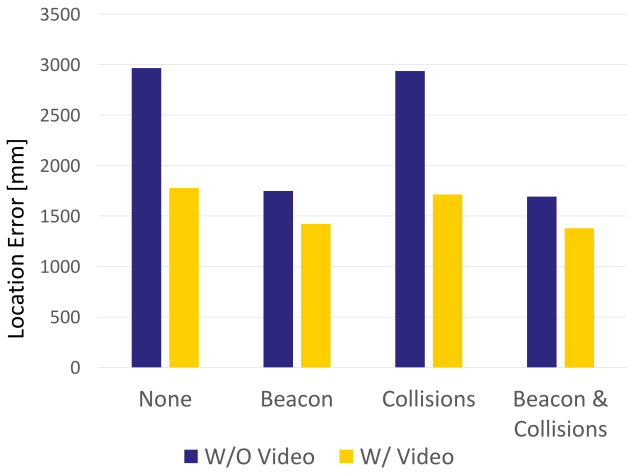

Figure 3.11: Location Error of Manual Correction

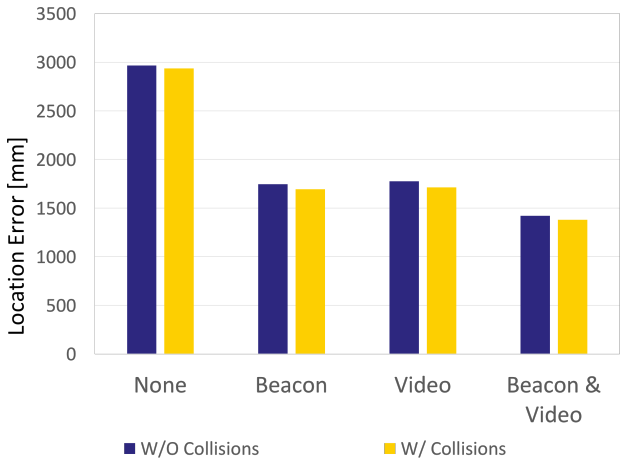

Figure 3.12: Location Error of Correction by Collisions

such as maneuver patterns, sprints and turns of each player.

Manual Correction For comparison with correction by beacon, we evaluated the location error when the estimated position was corrected to the ground truth position based on the recorded video. In this evaluation, the position is corrected with a probability of 1/200 for each frame. Because the sampling rate is 10 frame per second correction is performed about once every 20 seconds. When correcting, one targeted wheelchair was randomly chosen. Some wheelchairs are corrected many times, while others are never corrected, which is depends on the circumstances. Therefore, we conducted

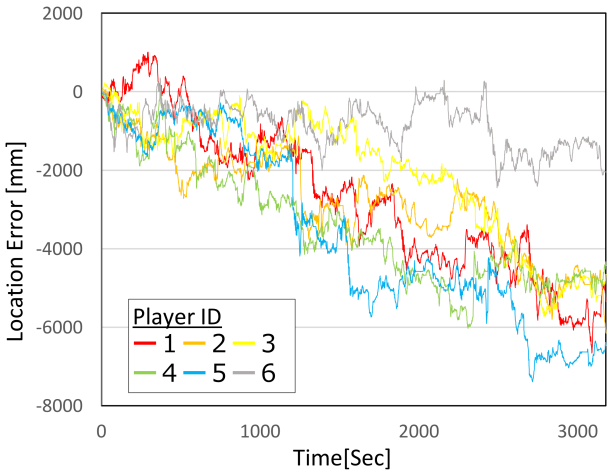

Figure 3.13: Location Error of Correction by Collisions in the Center Line Direction Over Time

the experiment 10 times and average error of them was used for evaluation. Figure 3.11 shows the result. From this result, it is possible to keep the average error within 2 m by manually correction, but it is high workload. Comparing with the result of correction by beacon (Figure 3.9), it can be seen that the position error is almost the same. Although the accuracy of the position that is used for manual correction is much higher than correction by beacon, it is because the manual correction is performed less frequently than correction by beacon. Actual correction by video requires additional camera setup and experts for image processing, which leads to an increase in workload. The workload and localization accuracy are trade-off relation.

Correction by Collisions Figure 3.12 shows the correction result using collision data. The error is slightly smaller overall. However, no significant improvement in accuracy was seen compared to correction using images (ground truth) and beacons. The reason is that the correction by collision is using relative position of each player unlike the above two correction methods, so it depends on both positions. In the data collected, there was a tendency for the estimated position to shift in the same direction as a whole due to reasons such as a bias in the actual movement. To confirm it, Figure 3.13 shows the location error in the center line direction over time. In this figure, the positive and negative values indicate the direction in which the estimated position is deviated. From this result, it can be seen that all estimated wheelchairs' positions are shifted in the same direction from ground truth. In such a case, even if the position is corrected by collision detection, one will move closer to the actual position and the other will move away from the actual position. It leaded that the average error could not be significantly reduced. This indicates that the error accumulation is not uniform, and it is necessary to find out the cause. All players in the data are right-handed, and it is possible that the

distance the wheels slipped differs due to braking power from right and left. In addition, the data of only 6 out of 10 players was available this time, and it is considered that the number of collisions was not enough. Therefore, it is necessary to further investigate the relationship between the number of collisions and the location error.

3.4 Conclusion

In this chapter, we investigated a localization method using three 9-axis sensors installed on the left and right wheels and under the chair in wheelchair basketball. For evaluation, we collected data in a practice match of about 5 minutes long. As a result, it can be seen that the average error was up to 5.3m only by dead reckoning. Furthermore, it was confirmed that the average error would be up to 2.8 m when correction by the BLE beacon installed at the goal was assumed. From these results, it is hard to use for applications such as detailed tactical analysis, however, we can use the system for the purpose of collecting data on wheelchair movements such as movement patterns, sprints, and turns of each player.

Chapter 4

Localization Focusing on Human Poses Using a Single Camera Towards Social Distance Monitoring During Sports

4.1 Introduction

The Coronavirus disease 2019 (COVID-19) is still prevalent in the world. Meanwhile, sports are important to maintain our health physically and mentally. Social distancing is more important during sports because we may not be able to wear masks to avoid breathing problems, heatstroke, etc [83]. Because vision-based human detection and tracking has been actively evaluated since before the pandemic, vision-based systems have been developed to support the management of social distancing [5–7, 49, 50]. These systems detect and track the skeletons or bounding boxes of humans to estimate interpersonal distance. However, the position error may increase during sports because the human pose changes frequently. Moreover, the tracking duration of close-contact is important in addition to distance among people because longer contact leads to higher risk [84]. For the supporting management of social distancing, the real-time warning of close-contacts is an effective way to avoid the risk of infection. It is also important to be able to analyze when and where the risk is high. This enables managers of sports facilities and teams to improve their behavior and rules.

To achieve the goal, we have developed a system designed for sports to detect and track close-contacts. Our system uses a single camera for low deployment costs and detects skeletons of people using *OpenPose* [32]. We select the waist position estimated by OpenPose to represent the position of the person for its stability in human detection. We then detect a close-contact when the distance between two persons becomes less than 2 m based on the definition of social distancing in Japan [85]. To improve the position error owing to the pose variation, we adjust the height of the waist according

to the pose of the legs. We proposed the basic concept of human localization with the waist height adjustment in Ref. [86]. In this work, we further propose the tracking of people and close-contacts based on the estimated positions of people. Specifically, for the tracking of close-contacts, the challenge is occlusion because we rely on a single camera. To solve the problem, we leverage the observation that most of the false negatives in human detection are caused by occlusion owing to other people. This is because there are few obstacles in sports facilities. Based on the above observation, we assume that a person still exists near the last detected position even when s/he disappeared in the proximity of other people.

To identify people at high risk of infection, it is necessary to identify the close-contact members. However, person identification [87–89] is another challenging topic which has been addressed by many researchers. Therefore, we exclude person identification (and inevitably human tracking) out of the scope of this paper. Instead, we focus on the tracking of close-contacts, i.e. measurement of close-contact duration regardless of involved persons. This is enough for the real time warning and the analysis of the time and locations with high risk to improve behavior and rules in sports facilities.

For evaluation, we recorded 834 videos that were 112 min in total including various scenarios with 2724 close-contacts. The results show that we achieve an F1-score of 83.6% for close-contact detection and an IDF1 [90] of 67.3% for close-contact tracking[1]. We also confirmed that the start and end times of more than 80% of the close-contacts are within 1 s, indicating that the close-contacts were correctly detected and tracked spatially and temporally. Additionally, we applied the system to an actual tennis tournament to support the management of social distancing. Through feedback on time and locations with frequent occurrences of close-contacts, we successfully suppressed the occurrence of close-contacts by changing the behavior of people.

Our contributions are summarized as below.

- We develop a close-contact detection and tracking system using a single camera for sports.
- To reduce the effect of pose variation on the position estimation, we adjust the position of the waist according to the pose of the legs.
- We design a close-contact tracking system, which is robust to occlusion based on the observation that occlusion in sport facilities is mostly caused by other people.
- To the best of our knowledge, this is the first study to evaluate the spatio-temporal correctness of close-contact detection and tracking.

4.2 System Overview

Figure 4.1 illustrates the overview of our system. Our target environment is sports activities in sports schools, gyms, etc. Our system consists of a single fixed camera and a computer for video processing.

[1] IDF1 is the ratio of correctly identified detections over the average number of ground-truth and computed detections.

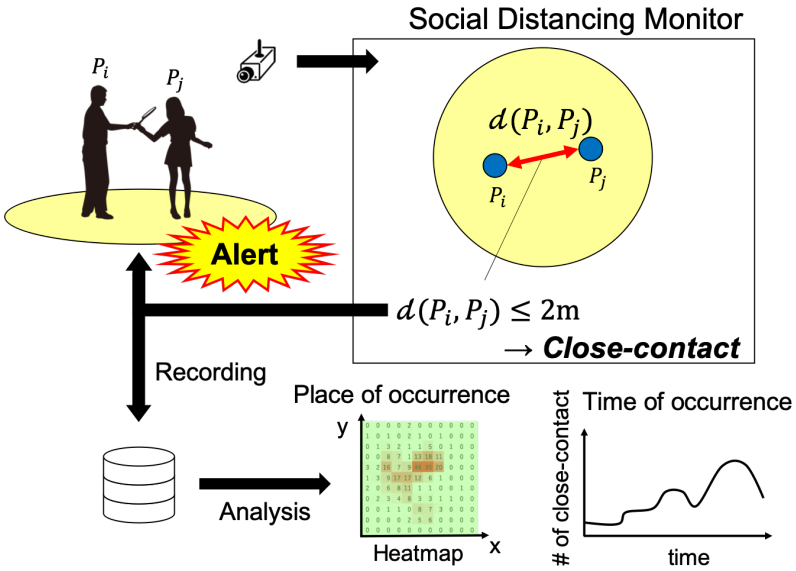

Figure 4.1: System Overview

The camera is installed at a high place such as a ceiling to capture the target area in the angle of view. We detect and estimate the positions of humans in the captured image. We then calculate the interpersonal distance based on the estimated positions to detect the close-contact in real time. The system notifies a close-contact if detected. The system also records the positions and time of the detected close-contacts. By analyzing the records, managers of sports facilities and teams can find the time and places when and where close-contacts occur frequently for the improvement of their behavior and rules.

4.3 Method

4.3.1 Overview

Figure 4.2 shows the flow of our method. In each frame, we first detect persons using a state-of-the-art skeleton detector called *OpenPose-STAF* [32]. *OpenPose-STAF* detects and tracks a skeleton of a person in a video. Next, we estimate the position of the detected person based on the skeleton and the coordinates of four points whose positions are known. The four points correspond to the scene in the real world and we can transform coordinates of skeletons in an image into the actual positions. Finally, we detect a close-contact by calculating the interpersonal distance based on their positions. To mitigate the effect of occlusion, we track the detected people using *OpenPose-STAF*. We then assume that a person still exists near the last detected position even if s/he disappears in the proximity of

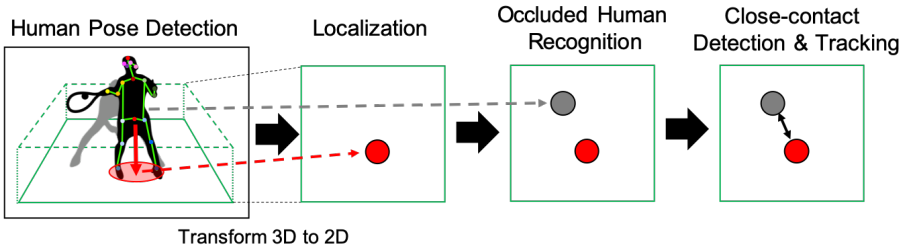

Figure 4.2: Method Overview

other people. In this way, we avoid false negatives in the detection and tracking of close-contacts.

4.3.2 Localization

Homography Transformation

For each frame, we estimate the position of the person whose skeleton is detected using *OpenPose-STAF*. For localization, we use the homography [48], which is a transformation that projects a plane to another plane, given the four point correspondences between the two planes. Therefore, a homography transformation matrix can transform pixel coordinates in an image into the actual positions, given the distance between the four points in the real world. This means that it is necessary to measure the distance between these four points in advance.

When a coordinate in an image is (u, v)[pixel], the corresponding coordinate (x, y)[m] in the real world is obtained by the following equation.

$$(x, y) = H(u, v) \tag{4.1}$$

H is the homography transformation matrix represented by the following equation.

$$H = \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & 1 \end{bmatrix} \tag{4.2}$$

For each point with a given coordinate, we obtain two equations. Because H has eight variables, we can solve H, given the actual positions of the four points in the image.

Our method uses the key point of the waist for the reference key point whose position is regarded as the position of the person. This is because the waist key point is stably detected even during movement compared with other key points such as the legs. We conducted a preliminary experiment to see how the height of each key point changes during movement. A subject moved across the front of a camera deployed at the height of 3m. We asked the subject to follow one of the three types of movements: walking, jogging, and running. The standard deviations of the key point heights for each movement type are shown in Table 4.1. From this result, we see that the waist height is more stable

than the other key points. The height of the head fluctuated slightly less than the waist in walking and jogging. However, we see that the fluctuation becomes larger with the increase of movement intensity (i.e. running). Therefore, the height of the four points for the homography transformation matrix is set to 0.9 m, which is the average waist height for adults.

Table 4.1: Standard Deviation of Key Points Height During Movement [cm]

Type \ Key Point	Waist	Head	Right Ankle	Left Ankle
Walking	3.32	2.94	11.89	5.88
Jogging	4.53	3.78	14.89	7.96
Running	4.51	8.55	37.54	24.26
Average	4.12	5.27	19.86	12.30

Waist Height Correction

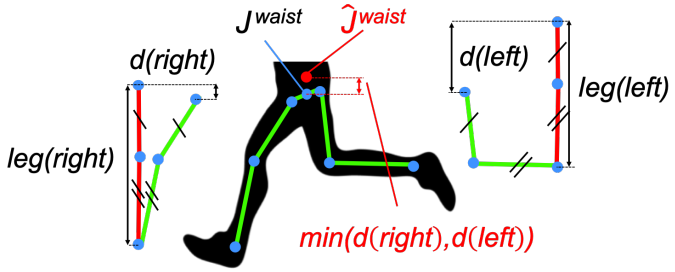

Figure 4.3: Waist Height Correction

While walking and running, the height of the waist does not change significantly. However, it can change significantly depending on poses such as sitting on a chair or the ground. Because the height error leads to a position error after the transformation, we mitigate the effect by mapping the position of the waist onto the plane with the height of 0.9 m. The correction is performed before the homography transformation.

The overview of the correction is shown in Figure 4.3. We let a coordinate of key point k be $J^k = [u^k, v^k]$. The length $l(p, q)$ between key points p and q is defined as below.

$$l(p, q) = \sqrt{(u^p - u^q)^2 + (v^p - v^q)^2} \tag{4.3}$$

For each leg, *OpenPose-STAF* outputs three key points, which are the hip, knee, and ankle. The length $|leg|$ of the leg is obtained by combining the lengths between these joints as follows.

$$|leg| = l(hip, knee) + l(knee, ankle) \tag{4.4}$$

We refer to the difference between the ankle-to-hip height and $|leg|$ as the *correction distance* d. The correction distance is defined as below.

$$d = |leg| - (v^{hip} - v^{ankle}) \tag{4.5}$$

If the leg angle against the ground decreases, d increases. This means that the height of the reference key point (i.e. the waist) in the image is less than the assumed average waist height (i.e. 0.9 m). Therefore, we correct the hip height by adding d to the original hip height. However, there are some cases where the leg is not on the ground because of jumping, balancing, etc. For the waist height correction, we need to use d of the grounded leg because d is calculated assuming that the pose of the grounded leg lowers the waist height. If both legs are not on the ground, its duration is usually short. Therefore, we simply ignore such cases. However, when only one of the left and right legs is not on the ground, the vertical ankle-to-hip distance of the ungrounded leg becomes shorter than that of the grounded leg. In other words, d of the ungrounded leg is larger than the other because the lengths of the left and right legs should be almost the same. Therefore, we use either the left or right leg with the smaller correction distance. The coordinate of the waist $\hat{J}^{waist}$ after correction $\hat{v}^{waist}$ is given below.

$$\hat{v}^{waist} = v^{waist} + \min(d(left), d(right)) \tag{4.6}$$

We note that, if either of the legs is not detected, we do not perform the correction because we cannot determine whether the detected leg is on the ground.

4.3.3 Human Tracking

We use *Openpose-STAF* for human tracking. As mentioned earlier, occlusion is a major challenge in a single camera setting. We leverage the observation that most of the false negatives in human detection are caused by occlusion owing to other people in sports facilities. Therefore, we assume that a person still exists near the last detected position even when s/he disappeared in the proximity of other people.

Specifically, suppose ID^k is the set of IDs of humans detected in frame k. The IDs are given by *OpenPose-STAF*. For person i satisfying the following equation, the coordinate (x_i^{k-1}, y_i^{k-1}) in frame $k-1$ is defined as a *missing point*.

$$i \in ID^{k-1} \wedge i \notin ID^k \tag{4.7}$$

When a person is temporarily not detected because of occlusion, the corresponding missing point is defined. If we detect a person with a new ID within θ_d[m] from the missing point, we consider that the occlusion is resolved and delete the missing point. Meanwhile, a person may move during the occlusion, e.g. when multiple people are walking in a line. To deal with such cases, we delete a missing point if we do not detect any person with a new ID within θ_d[m] from the missing point for more than $\theta_t s$. In this chapter, we empirically set $\theta_d = 2.0$[m] and $\theta_t = 1/3$[sec].

Algorithm 1 Tracking Close-contacts

Require: $C^p = \{c_1^p, c_2^p, ..., c_k^p\}(k \geq 0)$
$C^c = \{c_1^c, c_2^c, ..., c_n^c\}(n \geq 0)$
1: $sort(C^p)$ // sort descending order of duration
2: **for each** $c_i^p \in C^p$ **do**
3: $NearestID \leftarrow 0$
4: $NearestDistance \leftarrow \theta_d$
5: **for each** $c_j^c \in C^c$ **do**
6: **if** $l(c_i^p, c_j^c) < NearestDistance$ **then**
7: $NearestID \leftarrow j$
8: $NearestDistance \leftarrow l(c_i^p, c_j^c)$
9: **end if**
10: **end for**
11: **if** $NearestID \neq 0$ **then**
12: $Associate(c_i^p, c_{NearestID}^c)$
13: sub $(c_{NearestID}^c)$ from C^c
14: **end if**
15: **end for**

4.3.4 Close-contact Detection and Tracking

We perform the close-contact detection and tracking based on the result of human tracking and estimated skeletons with the waist height correction. First, we calculate the distance between each pair of persons including the missing points. We denote the position of a person with ID i as P_i. We then calculate the distance $d(P_i, P_j)$ between persons i and j. If the following condition is satisfied, we detect a close-contact and define the midpoint between P_i and P_j as the point of the close-contact occurrence.

$$d(P_i, P_j) \leqq 2.0[\mathrm{m}] \tag{4.8}$$

To obtain the duration and trajectory of the close-contact, we also track the close-contacts. For this purpose, it is necessary to associate the close-contacts detected in the previous frame and those detected in the current frame. Because the close-contacts with a longer duration have a higher risk, we associate the close-contact with the longest duration in the previous frame with the nearest close-contact in the current frame within a distance θ_d. Our association algorithm is shown in Algorithm 1.

4.4 Evaluation

4.4.1 Evaluation Setting

We conducted four types of experiments to evaluate the performance of our system in terms of 1) the effect of waist height correction in localization, 2) the effect of human orientation in localization, 3) localization comparison with other methods, and 4) the accuracy of close-contact detection and

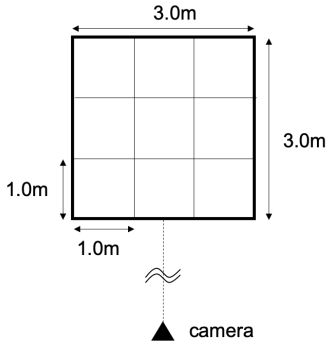

Figure 4.4: Evaluation Area (Effect of Waist Height Correction in Localization)

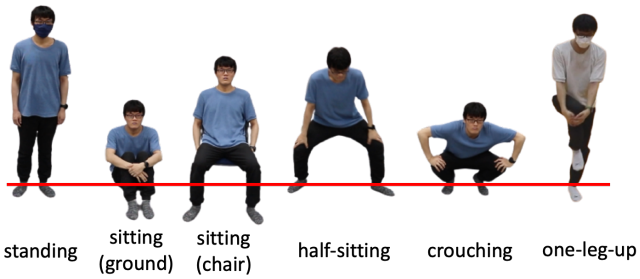

Figure 4.5: Poses Used in Evaluation

tracking. The details of each experiment are as below.

Effect of Waist Height Correction in Localization

For the evaluation of the effect of the waist height correction in the human localization, we collected images from one participant. We regard the position on the surface of the floor which is straight down from the waist as the actual position of the subject. We note that we have not obtained the ground truth of the waist positions directly due to its difficulty. Instead, the subject located at one of the lattice points in Figure 4.4 with his waist straightly above the lattice points. He took five types of poses except one-leg-up as shown in Figure 4.5. The poses are standing, sitting (ground), sitting (chair), half-sitting, and crouching. In this way, we collected data with the ground truth of the positions. Please note that we have not obtained the ground truth of the height of the waist. For each pose and position, we obtained images in which all key points of the lower body were detected. For this purpose, we recorded videos of the subject facing the camera with the height of 2.5m. Finally, we collected 80 images.

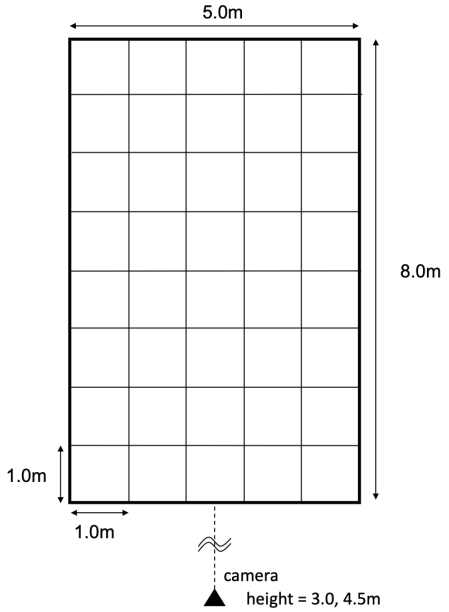

Figure 4.6: Evaluation Area (Effect of Human Orientation in Localization)

Effect of Human Orientation in Localization

For the evaluation of the robustness, we conducted evaluation in different camera positions and human orientations. One participant was located at one of the lattice points in Figure 4.6, and took six types of poses as shown in Figure 4.5. The poses are standing, sitting (ground), sitting (chair), half-sitting, crouching, and one-leg-up. There were two types of camera heights: 3.0m and 4.5m, and the participant faced four orientations: 0°, 90°, 180°, and 270°. The degree increases clockwise with 0°as the orientation when the person's body is facing the camera. Finally, we collected 2536 images.

Localization Comparison with Other Methods

Next, we compared our system with other methods using the same data as in Section 4.4.1. In many related works of close-contact detection, homography transformation is used for human localization. Therefore, we use other coordinates instead of the waist coordinates in homography transformation and compare the localization performance. We use the following two types of coordinates: the midpoint of the coordinates of both ankles [7], the bottom of the bounding box [5,6,49]. There is also a method for detecting close-contacts based on the size of the overlapping area of the bounding boxes [50]. However, since the distance between people is not calculated, we cannot compare our system with it. As the human detector that outputs bounding boxes, we used YOLOv4 [29] as in the Ref. [49] and [6].

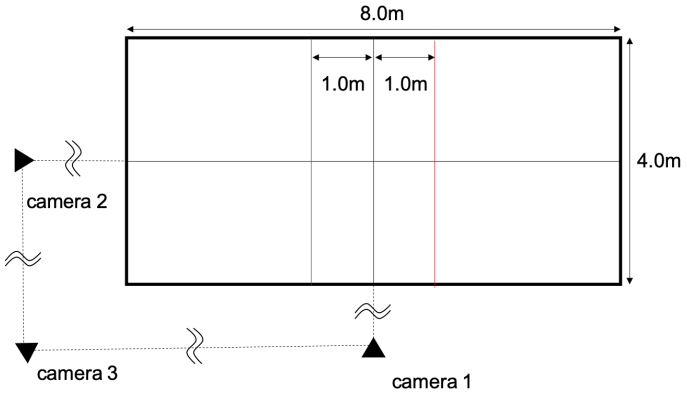

Figure 4.7: Evaluation Area (Close-contact Detection and Tracking)

Close-contact Detection and Tracking Performance

For the evaluation of the close-contact detection and tracking performance, we collected data where the participants moved according to a predefined scenario. The participants moved in an area of 4.0 m × 8.0 m as shown in Figure 4.7. To determine the effect of the orientation of the person and the orientation of the occlusion, the images were taken from three different angles. To eliminate the effect of persons out of the target area, we pre-processed the images by manually specifying the target area.

There are three types of scenarios: (1) conversation, (2) passing each other, and (3) passing through. In all scenarios, each group consisting of one or two subjects moved according to the specified trajectories. In the case of two subjects, the distance between them was always kept within 2.0 m. The movement in each scenario is shown in Figure 4.8. In scenario (1), two groups walked from different starting positions toward the other group's starting position, stopped near the center, turned around, and walked back to their original positions. In scenario (2), two groups walked from different starting positions to the other group's starting position. In scenario (3), one group was stationary at a position, and the other group passed in front of or behind the other group. The stationary group was in one of three poses which are standing, sitting, and crouching. For each combination of a scenario, the number of subjects, and a pose, we recorded videos more than 10 times by randomly changing the subjects. Finally, we collected 278 videos. Because we used three cameras, the total number of videos was 834. Details of the data are listed in Tables 4.2 and 4.3.

Each subject was asked to move at a constant speed on the trajectory specified in the scenario. However, the speed was slightly different for each participant and each trial. Therefore, we recorded the start and end times of the movement as well as the time at the moment of crossing the red lines which are 1.0 m away from the center line as shown in Figure 4.7. The ground truth of the trajectories were then obtained by linear interpolation. We also obtained the ground truth of the close-contact

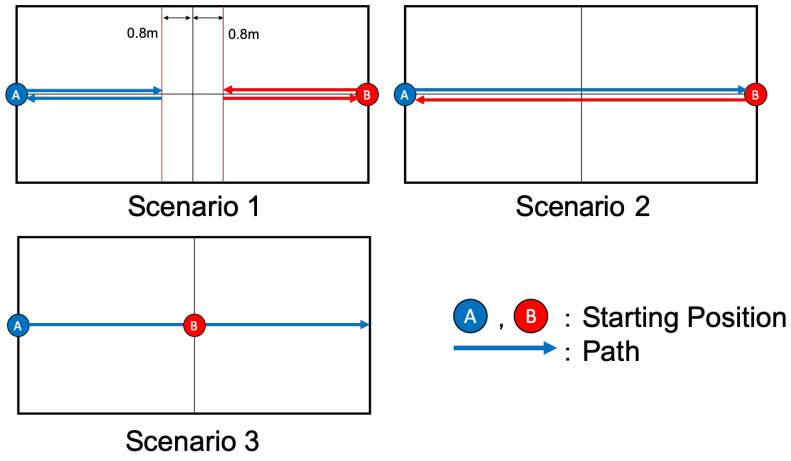

Figure 4.8: Movement Scenarios

Table 4.2: Details of Data Collection Scenarios

Scenario	Participants	Moves	Videos
1	1-1	10	30
	1-2	10	30
	2-2	10	30
2	1-1	10	30
	1-2	10	30
	2-2	10	30
3	(walking)1-1(stationary)	54	162
	1-2	54	162
	2-1	56	168
	2-2	54	162
Total	-	278	834

occurrences from the ground truth of the trajectories.

4.4.2 Results

Effect of Waist Height Correction in Localization

First, we evaluated the localization performance. Table 4.4 lists the mean absolute error distance for each pose. From the results, we observe that we can estimate the position of the standing person with a low error. However, the error increases as the waist height gets closer to the ground. Additionally, we successfully decreased the error by an average of 23 cm.

However, we could not observe significant improvement for the pose of sitting on a chair. This is because the elevation angle of the camera and the angle of the leg were almost equal. When the

Table 4.3: Details of Data Collection in Scenario (3)

Participants (Walking-Stationary)	Pose	Moves	Videos
1-1	standing	18	54
	sitting	18	54
	crouching	18	54
1-2	standing	18	54
	sitting	18	54
	crouching	18	54
2-1	standing	18	54
	sitting	20	60
	crouching	18	54
2-2	standing	18	54
	sitting	20	60
	crouching	16	48
Total	-	218	654

Table 4.4: Mean Absolute Error for Each Pose [m]

Pose	Corrected	Original	Effect of correction
Standing	0.056	0.064	-0.008
Sitting (ground)	0.729	1.338	-0.609
Sitting (chair)	0.716	0.779	-0.063
Half-sitting	0.370	0.641	-0.271
Crouching	0.712	0.915	-0.203
Average	0.517	0.747	-0.231

relative angle of the leg (thigh and lower leg) to the camera is 0°, the leg length appears the shortest in the image while it appears the longest if the relative angle is 90°. In this experiment, the relative angle of the thigh to the camera was close to 0°, which means the appearance of the leg length in the image is shorter than the actual length. To address this problem, we may need to obtain a more accurate leg length using a technique of estimating a 3D pose from a skeleton, for example.

Effect of Human Orientation in Localization

Table 4.5 shows the localization performance for different body orientations and poses. As a result, when facing backwards (180°), the error was large in the poses such as sitting and crouching where the legs were hidden by the chair or the person's body itself because there were many false positives of the skeleton of the legs. In the standing pose, there are little difference of error in all orientations since there is no occlusion. This is due to the relationship between the camera and the legs as discussed in Section 4.4.2. Table 4.6 shows the localization error without the waist height correction. From this

Table 4.5: Mean Absolute Error [m] for Different Poses and Orientations (with Waist Height Correction)

Pose	Orientation			
	0°	90°	180°	270°
Standing	0.214	0.183	0.144	0.182
Sitting(ground)	1.867	0.819	2.523	0.491
Sitting(chair)	1.342	1.096	1.640	1.527
Half-sitting	0.474	0.589	0.553	0.648
Crouching	0.960	0.413	1.170	0.969
One-leg-up	0.316	0.287	0.200	0.310

Table 4.6: Mean Absolute Error [m] for Different Poses and Orientations (w/o Waist Height Correction)

Pose	Orientation			
	0°	90°	180°	270°
Standing	0.219	0.189	0.146	0.186
Sitting(ground)	1.990	2.240	2.247	2.292
Sitting(chair)	1.701	1.882	1.859	1.851
Half-sitting	0.743	0.885	1.002	0.894
Crouching	0.983	1.323	1.228	1.406
One-leg-up	0.313	0.217	0.188	0.248

result, we see that the correction is more or less effective in any orientation and pose. Even if the legs are bent, if they are facing to the side (90°or 270°), the whole lengths of the legs are visible in the image. Therefore, the performance greatly improves by the waist height correction.

Localization Comparison with Other Methods

Table 4.7 shows the localization error compared with the other methods. Since the data was collected in an unobstructed environment, the ground contact part of the body was visible without occlusion. Therefore, in poses such as sitting (chair) and half-sitting where the ankles are clearly visible, using the midpoint of the ankles shows the best result. The bottom of the bounding box showed the best results in poses such as sitting (ground) and crouching where the legs may be invisible depending on the orientation of the person. However, when using the midpoint of the ankles, the error increases significantly in a pose in which one leg is floating in the air. This is important because such a pose occurs frequently during exercise.

In addition, to evaluate the effect of occlusion, we virtually placed a wall at the feet by image processing when the participant takes a standing pose. The values in parentheses in Table 4.7 are the heights of the walls. In the method using the bottom of the bounding box, because the skeleton of the person is not estimated, it is not possible to determine whether the legs are hidden. The error thus

Table 4.7: Comparison of Localization Mean Absolute Error [m] with Other Methods

Pose	Ours (waist)	Midpoint of ankles [7]	Bottom of box [6, 49]
Standing	**0.181**	0.211	0.286
Standing(0.3)	**0.244**	0.352	0.721
Standing(0.5)	**0.210**	0.333	1.344
Sitting(ground)	1.504	0.700	**0.515**
Sitting(chair)	1.475	**0.318**	0.393
Half-sitting	0.577	**0.211**	0.402
Crouching	0.931	**0.289**	0.436
One-leg-up	0.280	0.692	**0.230**

increases significantly as the height of the virtual wall increases. In these cases, the increase of the error is mitigated by our method. However, our method has a larger error in sitting and crouching poses than the result in Section 4.4.2. This is due to false negatives of legs caused by the human detection. Especially, the pattern on the ground increased false positives in leg detection in a larger environment. This problem is due to the accuracy of the skeleton detector, which can be improved by properly using the detected skeletons depending on the situations. For example, we may use the waist key points when standing or moving, while we use the leg key points when sitting or stationary. Therefore, as future work, we leverage the poses of the person from the time-series data of the skeleton.

Close-contact Detection and Tracking Performance

Human Detection For the evaluation of the human detection, we use precision and recall. If the distance between the ground truth and the estimated position is within 1.0 m, we regard the detection result as a true positive.

As a result, the precision and recall are 84.6% and 92.3%, respectively. The F1 score is 88.5%, indicating that many close-contacts are correctly detected spatially and temporally. The precision and recall for each scenario are listed in Tables 4.8 and 4.9, respectively. The number of true positives, false positives, and false negatives are listed in Table 4.10. From these results, we observe that there are almost no false negatives in any scenario. However, there are many false positives despite the scenarios with only two participants. This is because the marker lines on the floor were wrongly recognized as persons. In the scenarios with more people, the floor was hidden by them, leading to less false positives (i.e., increase of precision). To avoid the problem, we may consider the use of a high-resolution camera that can clearly capture the boundary between the floor and a person, or background subtraction to remove the effect of the floor pattern. In scenario 3, when there is a sitting participant, both precision and recall are lower than the other scenarios. This is because the waist height correction did not work well due to the wrong detection of legs. The skeleton detector wrongly recognized the chair as the legs

Table 4.8: Human Detection Result by Scenarios (Precision[%])

# of participants	Scenario				
	1	2	3 standing	3 sitting	3 crouching
1-1	72.4	82.5	91.8	69.4	79.9
1-2	87.1	89.6	88.0	61.6	86.6
2-1			90.5	75.2	85.4
2-2	90.8	89.8	95.4	82.7	94.6

Table 4.9: Human Detection Result by Scenarios (Recall[%])

# of participants	Scenario				
	1	2	3 standing	3 sitting	3 crouching
1-1	98.6	99.0	94.9	81.7	94.3
1-2	97.4	98.5	96.6	70.7	97.3
2-1			97.6	85.2	96.6
2-2	96.9	95.4	95.4	84.5	96.1

of the person, leading to larger position error.

Close-contact Detection For the evaluation of the close-contact detection, we use precision and recall. If the distance between the ground truth and estimated position is within 1.0 m, we regard the detection result as a true positive. As a result, the precision and recall are 83.9% and 83.4%, respectively. The F1 score is 83.6%, indicating that many close-contacts are correctly detected spatially and temporally. The precision and recall for each scenario are listed in Tables 4.11 and 4.12, respectively. The number of true positives, false positives, and false negatives are listed in Table 4.13. From these results, we observe that there are almost no false negatives in any scenario. Especially in the simple scenarios with a small number of people, we could detect close-contacts with higher recall.

However, there are many false positives in the simple scenarios. This is because there are many false positives in human detection in these scenarios. The precision of close-contact detection is lower than that of human detection. This happens when there are multiple people in close proximity. For example, if two true positives and one false positive are close to each other, the close-contact is detected between each pair. This means three close-contacts are detected. However, one of these is the correct close-contact while the other two are the wrong close-contacts. Therefore, in such a case, the number of false positives increases.

Table 4.10: Human Detection Result (TP, FP, FN)

# of participants	Scenario	TP	FP	FN
1-1	1	17179	6534	245
	2	16638	3530	162
	3(standing)	25326	2265	1356
	3(sitting)	21859	9641	4901
	3(crouching)	25709	6486	1567
1-2	1	25975	3832	692
	2	22462	2603	335
	3(standing)	36326	4940	1294
	3(sitting)	25806	16082	10689
	3(crouching)	35934	5579	1002
2-1	3(standing)	36394	3839	902
	3(sitting)	30403	10047	5291
	3(crouching)	38739	6597	1356
2-2	1	33610	3412	1058
	2	29078	3312	1414
	3(standing)	46789	2248	2231
	3(sitting)	37961	7965	6967
	3(crouching)	52650	3022	2118

Table 4.11: Close-contact Detection Result by Scenarios (Precision[%])

# of participants	Scenario				
	1	2	3 standing	3 sitting	3 crouching
1-1	57.6	55.5	83.3	82.6	72.2
1-2	78.4	80.4	92.9	91.5	75.8
2-1	-	-	84.8	85.5	77.5
2-2	85.1	83.8	85.8	88.4	83.2

Close-contact Tracking Finally, we evaluated the close-contact tracking performance. We used the Identification Precision (IDP), Identification Recall (IDR) and Identification F1 (IDF1) proposed in Ref. [90] for the evaluation metrics to focus on the length of correct tracking. This is reasonable because the duration of close-contacts is important for the assessment of the infection risk.

From the results, we confirmed that IDP, IDR, and IDF1 are 67.6%, 67.1%, and 67.3%, respectively. The precision and recall of each scenario are listed in Tables 4.14 and 4.15, respectively. The number of true positives, false positives, and false negatives are listed in Table 4.16. The IDP decreases in the scenario with a small number of subjects because of the false positives by the line markers on the floor as mentioned in Section 4.4.2. However, the IDR decreases with the increase of the number of subjects. This is because an ID frequently switched with another ID when multiple close-contacts occurred at

Table 4.12: Close-contact Detection Result by Scenarios (Recall[%])

# of participants	Scenario				
	1	2	3		
			standing	sitting	crouching
1-1	89.0	88.9	93.3	90.9	77.7
1-2	85.7	87.3	92.9	90.6	72.8
2-1	-	-	90.5	88.9	82.1
2-2	86.9	86.2	81.0	78.1	73.9

Table 4.13: Close-contact Detection Result (TP, FP, FN)

# of participants	Scenario	TP	FP	FN
1-1	1	1831	1347	230
	2	1747	1399	267
	3(standing)	5280	1061	394
	3(sitting)	5309	1121	540
	3(crouching)	4393	1694	1245
1-2	1	11150	3069	2004
	2	9170	2234	1369
	3(standing)	20755	1587	1824
	3(sitting)	19616	1819	2023
	3(crouching)	15704	5001	5652
2-1	3(standing)	19834	3554	2099
	3(sitting)	20984	3565	2707
	3(crouching)	17414	5070	3811
2-2	1	22124	3876	3179
	2	18005	3485	2814
	3(standing)	34000	5636	8074
	3(sitting)	36321	4768	29264
	3(crouching)	28118	5679	9883

the same time. One of the solutions to solve the problem is a Kalman filter for close-contact tracking to predict human and close-contact movement.

Next, Tables 4.17 and 4.18 show the precision and recall when occlusion is not considered (i.e. without the missing point), respectively. The best results in each scenario are shown in bold type. The number of true positives, false positives, and false negatives are listed in Table 4.19. From these results, we have achieved significant performance improvements in many scenarios for both IDP and IDR by continuing to track hidden persons. Therefore, regardless of the frequency of occlusion, considering occlusions has a large effect. It is because the evaluation metrics IDP and IDR consider matching between close-contact IDs of the ground truth and the estimated result. For example, if ID switching occurs, all subsequent ground truth tracks and estimation tracks are considered as false

Table 4.14: Close-contact Tracking Result by Scenarios (IDP[%])

# of participants	Scenario				
	1	2	3		
			standing	sitting	crouching
1-1	54.8	**53.6**	**81.5**	**76.1**	**66.2**
1-2	**74.4**	**77.8**	**74.8**	**77.5**	58.9
2-1	-	-	**77.4**	**76.2**	**70.8**
2-2	**65.5**	**61.1**	**64.2**	**59.5**	**58.2**

Table 4.15: Close-contact Tracking Result by Scenarios (IDR[%])

# of participants	Scenario				
	1	2	3		
			standing	sitting	crouching
1-1	**84.5**	**85.8**	**91.3**	**83.8**	**71.3**
1-2	**81.3**	**84.5**	**74.8**	**76.7**	**56.5**
2-1	-	-	**82.6**	**79.2**	**75.0**
2-2	**66.9**	**62.8**	**60.6**	**52.6**	**51.7**

negatives and false positives, respectively. Therefore, IDP and IDR become better if we keep the same close-contact IDs for each close-contact track for a longer time. This means that when the occlusion occurs is more important than the number of occlusions.

We also evaluated the start and end times of the close-contacts. Figure 4.9 shows the cumulative distribution function (CDF) of the absolute time error. The result shows that 85.1% of the start time error for all the close-contacts were within 30 frames (i.e., 1 s). 84.7% of the end time errors were also within 30 frames. Moreover, 71.0% of the elapsed time error were within 30 frames ,whereas 86.0% of the elapsed time errors were within 60 frames (i.e., 2 s). This is reasonable because both of the start and end time errors are less than 1 s for more than 84.7% of the close-contacts. We also note that there is a little uncertainty in the ground truth of the subject positions (i.e., close-contact positions as well) owing to the manual labeling and linear interpolation. Nevertheless, our system could detect more than 80% of close-contacts with the start and end time errors within 0.83 s.

4.4.3 Use Case

We used our system in a professional tennis tournament for safety management against COVID-19. We used close-contacts longer than 4 s in the following analysis because the normal interval between breaths is approximately 4 s. The results are listed in Table 4.20. In the table, the frequency refers to the number of close-contacts per hour. Because the number of staff (ball persons and line persons) was different in the final match on the third day, we analyzed the final match and other matches separately

Table 4.16: Close-contact Tracking Result (TP, FP, FN)

# of participants	Scenario	TP	FP	FN
1-1	1	1740	1438	347
	2	1686	1460	329
	3(standing)	5168	1173	507
	3(sitting)	4895	1535	949
	3(crouching)	4031	2056	1631
1-2	1	10586	3633	2634
	2	8876	2528	1680
	3(standing)	16716	5626	5882
	3(sitting)	16602	4833	5047
	3(crouching)	12190	8515	9127
2-1	3(standing)	18099	5289	3860
	3(sitting)	18707	5842	5062
	3(crouching)	15909	6575	5340
2-2	1	17027	8973	8182
	2	13124	8366	7677
	3(standing)	25442	14194	16584
	3(sitting)	24449	16640	57671
	3(crouching)	19665	14132	18475

Table 4.17: Close-contact Tracking Result by Scenarios w/o Missing Point(IDP[%])

# of participants	Scenario				
	1	2	3		
			standing	sitting	crouching
1-1	**56.4**	44.6	60.1	54.6	54.1
1-2	67.7	69.5	64.2	53.5	**62.3**
2-1	-	-	64.1	60.9	62.0
2-2	54.6	57.8	58.2	51.8	53.8

on the last day.

On the third day, the frequency of the close-contacts was lower than that of the first and second days. This is because we reported to the tournament management team on the situations (i.e., locations and timing) where close-contacts frequently occurred at the end of the second day. We noted that the frequency slightly increased in the final match owing to the increase of the number of staff. Overall, the frequency of the close-contacts decreased significantly after the report based on our system, highlighting its usefulness for safety management against COVID-19.

Additionally, we analyzed the time and locations of the close-contacts. First, the number of close-contacts over time is shown in Figure 4.10. Based on the analysis, we found that many close-contacts occurred not during the game but in between the games. Next, Figure 4.11 shows a heat map of the

Table 4.18: Close-contact Tracking Result by Scenarios w/o Missing Point(IDR[%])

# of participants	Scenario				
	1	2	3		
			standing	sitting	crouching
1-1	62.1	62.4	50.1	41.6	42.4
1-2	67.1	68.0	57.6	40.0	54.4
2-1	-	-	60.2	54.1	56.2
2-2	46.6	53.2	48.0	35.6	41.3

Table 4.19: Close-contact Tracking Result w/o Missing Point (TP, FP, FN)

# of participants	Scenario	TP	FP	FN
1-1	1	1279	989	787
	2	1226	1521	789
	3(standing)	2837	1882	2822
	3(sitting)	2477	2100	3449
	3(crouching)	2352	1952	3329
1-2	1	8728	4164	4708
	2	7146	3129	3454
	3(standing)	12867	7161	9659
	3(sitting)	11769	7118	9713
	3(crouching)	8578	7456	12864
2-1	3(standing)	13209	7383	8804
	3(sitting)	13282	8131	10608
	3(crouching)	11472	7367	9723
2-2	1	11866	9872	13093
	2	11117	8118	9669
	3(standing)	20160	14493	21845
	3(sitting)	19225	16500	77854
	3(crouching)	13560	12593	24513

close-contact locations. From this result, we can observe that close-contacts occurred mostly in the center of the court and near the referee chair. As a result of checking the video, we found that players often moved around the referee chair at the changes of the ends and new balls were placed behind the referee chair, which is the cause of the frequent close-contacts. Additionally, some ball persons did not maintain a sufficient distance when they waited between games at the center of the court. Our system can support such analysis by providing spatial and temporal trends of close-contacts for safer risk management against COVID-19.

In this use case, we introduced an example of using a close-contact detection system in a singles tennis tournament. There are two players participating in one match, and the moving range of each player is divided into two by the net. Therefore, there is almost no close-contacts between players

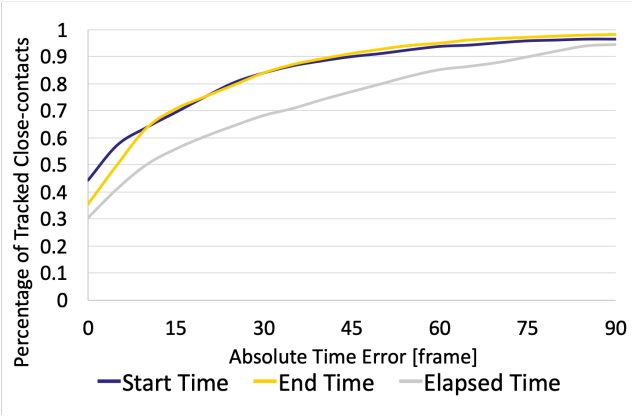

Figure 4.9: Time Error in Close-Contact Tracking

Table 4.20: Close-Contact Occurrence during Tennis Tournament

Day	Time	Close-contact over 4sec.		Close-contact All	
		Quantity	Frequency	Quantity	Frequency
1	5:39:41	184	32.5	806	142.4
2	6:17:45	208	33.0	940	149.2
3	1:51:18	29	15.6	147	79.2
3(Final)	1:23:42	34	24.4	158	113.3
All	15:12:26	455	29.9	2051	134.9

during the match. In addition, the position of an umpire, ball persons, and line persons during the match are fixed. From the above reasons, it is unlikely that close-contacts will occur during the match. Therefore, the frequency of detected close-contacts was significantly reduced just by alerting the staff to the movement routes. As a future work, we would like to confirm whether it is possible to reduce the frequency of close-contacts even in an environment where athletes move freely, such as sports schools. In addition, this result may include false negatives and false positives. Therefore, in order to investigate how false negatives and false positives occur in actual situations other than the scenarios we assumed in this section, we would like to conduct comparison with other localization methods such as using LiDAR and radio signals.

4.5 Conclusion

In this chapter, we proposed a close-contact detection and tracking system using a single camera during sports. We reduced the effect of the pose variation on the position estimation by adjusting the

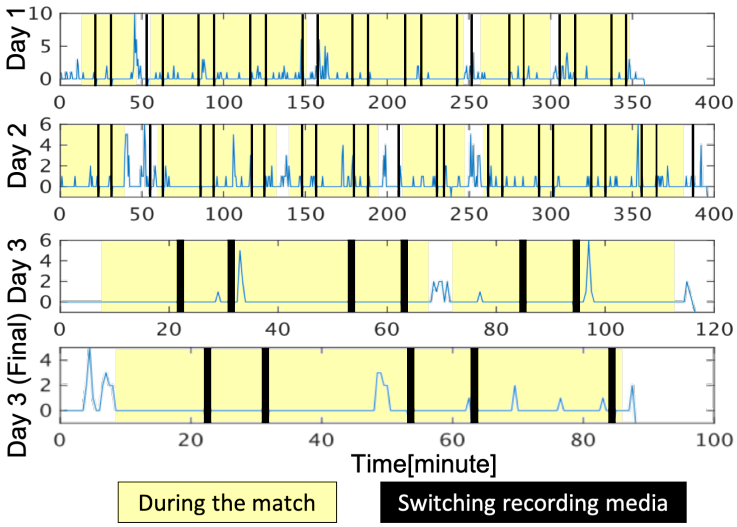

Figure 4.10: Time vs. Number of Close-Contacts

position of the detected person according to the pose of the legs. The evaluation results showed that our system achieved F1 scores of 83.6% and 67.3% for close-contact detection and tracking, respectively. Additionally, we confirmed that the start and end time errors were within 1 for more than 80% of the close-contacts.

One of our future works is to evaluate a method using the upper body skeleton for more robust position estimation. We also plan to deploy our system in various sports schools and gyms for our new lifestyle with COVID-19.

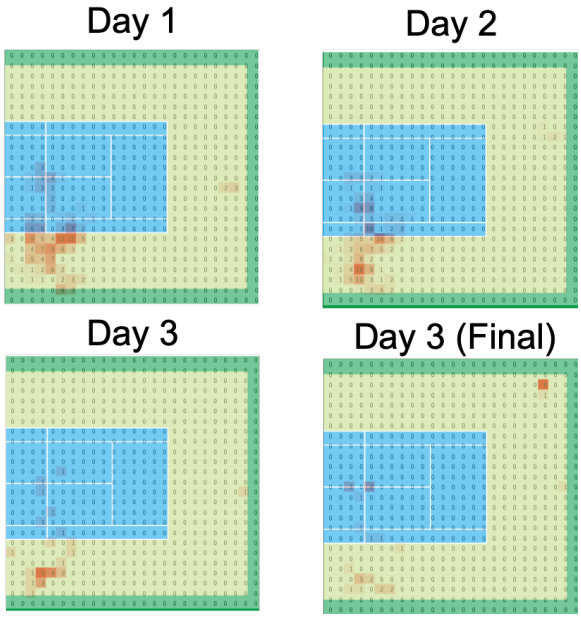

Figure 4.11: Place of Occurrence

Chapter 5

Maneuver Action Recognition and Vehicle Movement Classification in Wheelchair Sports Using Inertial Sensors

5.1 Introduction

Emerging developments in sensing technology have focused the attention of athletes, coaches and fans on applying it to data analysis in sports training, strategies and entertainment [76]. One of major challenges in sports data analysis is to design sophisticated methods suitable for target sports for useful data collection. In major sports such as football, basketball, and baseball, data analysis is already essential since many engineers and researchers have developed practical systems to apply it. However, data analysis in wheelchair basketball (WB) still requires further effort to establish building blocks essential for data analysis.

Therefore, we have been working on the development of a system to support WB data analysis in cooperation with athletes and coaches. In WB, players constantly strive to improve their wheelchair movement techniques since it is important to be able to move the wheelchair quickly and efficiently depending on the time and position. In particular, the wheelchair maneuvering, which is movement of the wheel, is the most basic and important action in all situations. However, the assessment of maneuvering quality is difficult due to the lack of quantitative metrics. To support quantitative analysis of the maneuver quality in this chapter, we propose a method to detect and classify maneuver actions using inertial sensors. We define the target maneuver actions as *PUSH* and *PULL* through discussion with experts because statistics such as strength and interval of these actions are closely related to the quality of the maneuver. *PUSH* is the maneuver of grabbing the rim and pushing it forward to accelerate the wheel, while *PULL* is the maneuver of grabbing the rim and stopping the

wheel or pulling it backwards to decelerate it. Although camera-based approaches are widely used for sports data analysis, they cannot measure such precise motions. To deal with this problem we first of all clamped two inertial sensors to the left and right wheels of the wheelchair to measure the angular velocity of each wheel. Even using inertial sensors, the classification of maneuver actions is still challenging because of various movements of wheels with different speeds and directions. To clarify the maneuvering actions concealed within such complicated movements we employ a segmentation algorithm followed by classification. First, we segment candidates of maneuver periods by the local maximum/minimum of the angular velocity since the rotation of the wheel generated by maneuvering leads to sharp changes in the angular velocity. Then, we classify maneuver actions in each segment based on thresholds.

In addition to the maneuver, wheelchair behavior is also important. In this chapter, we design a method to classify types of turns into *PIVOT* and *TURN*. This is because the pivot turn is one of the most important techniques in WB to push an opponent away by applying power efficiently. For this purpose, similarly to the maneuver classification, we detect any turns by calculating the amount of wheelchair rotation from the angular velocities of the both wheels. We then identify *PIVOT* turns based on thresholds for each wheel based on the typical movement of the pivot turns.

In order to evaluate the performance of the proposed maneuver classification method, we collected data from six players in a WB practice game containing 1005 *PUSH* and 152 *PULL* actions. From the results, we confirmed that the precision and recall of both maneuver classifications are more than 84.6%. We also collected data from 172 pivots and 192 turns to evaluate our turn classification. The result shows our method successfully classifies *PIVOT* and *TURN* with an F-measure of 99.7%. Furthermore, we show the effectiveness of our classification results in assessing maneuver quality through maneuver analysis combined with other information such as player positions.

Our contributions are summarized as follows.

- We developed a system to support the data analysis in wheelchair basketball by using inertial sensors and a camera.
- We designed methods to classify maneuver actions and turns in wheelchair basketball by focusing on specific movement of wheels.
- We evaluated the performance of our methods by collecting data from athletes.
- We showed the potential of the classification results through the analysis of data collected in a practice game.

5.2 System Overview

Figure 5.1 illustrates an overview of our system. Instead of manual video analysis currently used by many teams, we extract statistics from videos and inertial sensors. Figure 5.2 shows a snapshot of

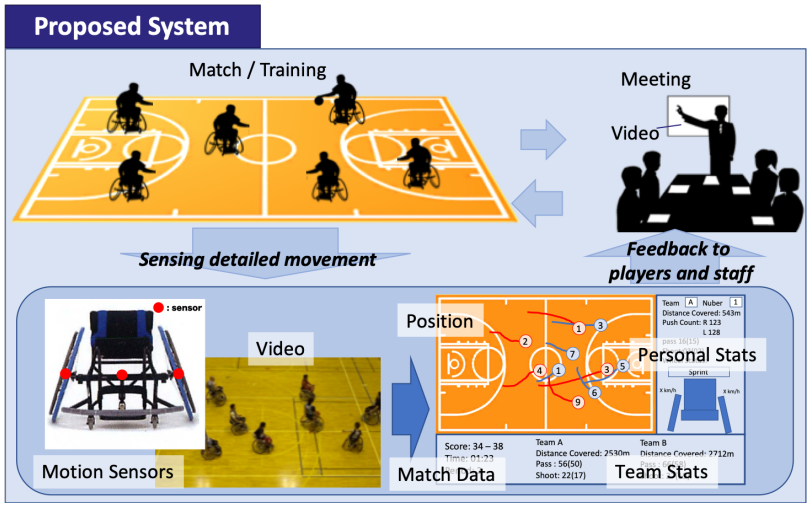

Figure 5.1: System overview

our system developed for players and coaches. Our system provides player tracking and visualization of statistics on wheelchair movement. For player tracking, we have implemented DeePSORT [39, 91] combined with YOLOv3 [28] for object detection. Since YOLOv3 itself cannot detect wheelchairs, we trained the model by using 602 images of wheelchairs cropped from videos of WB.

On the other hand, precise motions such as maneuver of wheelchairs are extracted from inertial sensors. In this chapter, we aim at designing a method to detect and classify *PUSH* and *PULL* actions of wheelchair maneuvering. We also design a method to detect and classify *PIVOT* and *TURN*. This leads to the quantification of the maneuver quality by analyzing statistics related to the detected actions. Furthermore, the detection results can be used for the analysis of strategies and performance assessment in combination with other information such as players positions. In WB, the basic strategy is to screen and block the defending opponent and help a team member with making shots. This is very effective because a wheelchair needs a large area to turn. Therefore, it is important to analyze how the wheelchair is manipulated to move to the proper position to allow shooting and blocking. Figure 5.3 illustrates an example situation in which a pivot turn is more efficient than a spin turn. When the position in front of a player is blocked by an opponent, a pivot turn is more efficient for moving forward because the player can move along a straighter path toward the target position by changing the direction of the opponent. If the player changes direction without a pivot turn, the opponent can easily screen that player by moving back and forth.

As shown in Fig. 5.4, inertial sensors are fixed to the axles of the left and right wheels. We use DSP wireless 9-axis motion sensors manufactured by SPORTS SENSING Co., LTD (Figure3.2). The sensor is capable of measuring 3-axis acceleration, 3-axis angular velocity, and 3-axis geomagnetic data at a sampling rate of 200 Hz. The measurement ranges of the sensor are shown in Table 3.1. Hereafter,

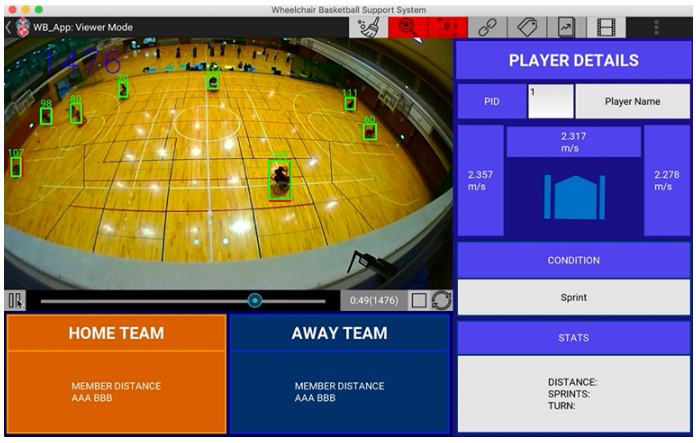

Figure 5.2: Snapshot of support system

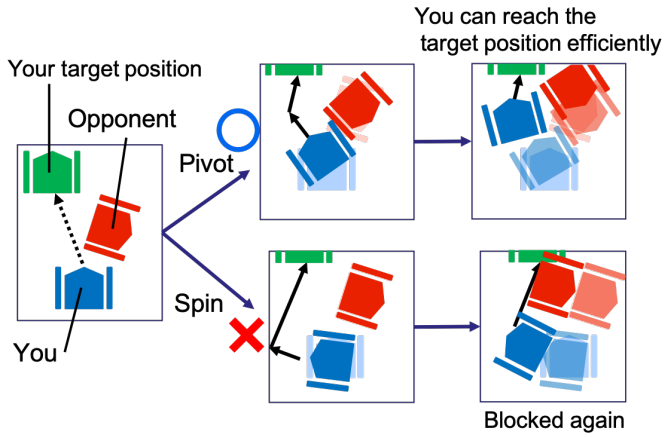

Figure 5.3: Effectiveness of pivot turn

we use $[radians/second]$ as the unit of angular velocity unless otherwise stated.

5.3 Maneuver Classification

5.3.1 Overview

The overview of the proposed method is illustrated in Figure 5.5. The maneuver actions in WB are instantaneous movements consisting of independent movements of left and right wheels. This means we need an approach different from activity recognition for continuous motions such as walking. Therefore, our method firstly segments the time series of the angular velocity to extract candidate periods of *PUSH* and *PULL* motions without any fixed window size. We then classify the maneuver

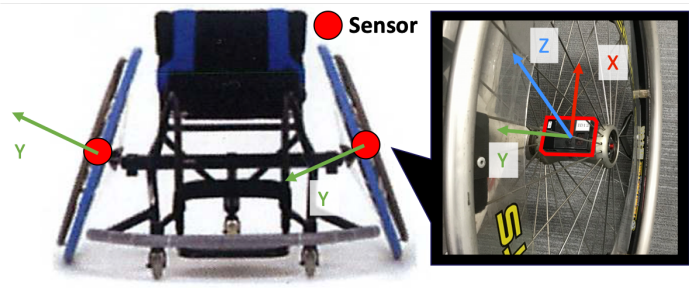

Figure 5.4: Sensor equipment

Figure 5.5: Method overview

actions for each segment. Our target actions are *PUSH* and *PULL* since they are frequently observed in WB. *PUSH* is the motion to apply force the wheel to the forward direction while the *PULL* is the motion to apply force in the reverse (backward) direction. Since the segmented periods are still the candidates of *PUSH* and *PULL*, there is a possibility of other actions. We define the other actions as *OTHERS* and design a classification method for the three maneuver actions. The classification is performed by thresholds for the angular velocity. Finally, we remove the segment classified *PULL* when wheelchairs collide with each other because the change in angular velocity is greatly affected by collision rather than *PULL* .

5.3.2 Preprocessing

Since the raw sensor data contains noise, we apply a Chebyshev type I filter [92] which is a low pass filter using the Chebyshev polynomials. The Chebyshev polynomials of the first kind are defined by the recurrence relation.

$$T_0(x) = 1 \tag{5.1}$$

$$T_1(x) = x \tag{5.2}$$

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \tag{5.3}$$

The ordinary generating function for T_n is

$$\sum_{n=0}^{\infty} T_n(x)t^n = \frac{1-tx}{1-2tx+t^2} \tag{5.4}$$

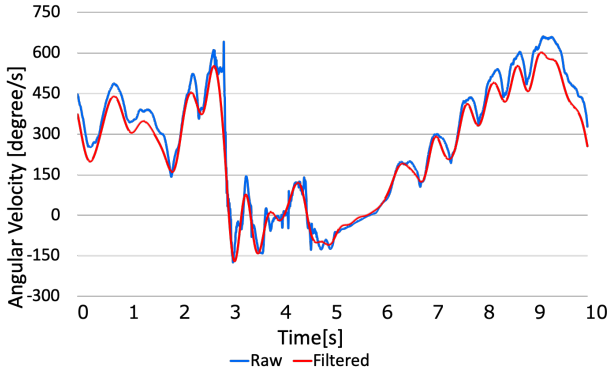

Figure 5.6: Example of filtered angular velocity

We let $G_n(\Omega)$ be a function of the angular frequency Ω of the n-th order low-pass filter as below.

$$G_n(\Omega) = \frac{1}{\sqrt{1 + \varepsilon^2 T_n^2(\frac{\Omega}{\Omega_0})}} \tag{5.5}$$

Where ε, Ω_0, and T_n are a ripple factor, a cutoff frequency, and a Chebyshev polynomial of the n-th order, respectively. We empirically set the above parameters as $\varepsilon = 1.0$, $\Omega_0 = 0.03$, and $n = 6$. Figure 5.6 illustrates an example of the filtering. We see that the raw data is smoothed by filtering the noise.

5.3.3 Segmentation

Each maneuver action consists of grip, move, and release. The time from the grip to the release widely varies even for the same action, which means that a sliding window with a fixed size does not work well. Therefore, we apply peak/valley detection in order to extract candidate segments of the maneuver actions from the time series of the angular velocity. We remove small peaks/valleys due to noise in addition to peaks/valleys detected within an extremely short interval.

The peak detection is performed as follows. We denote the angular velocity at time t as $\omega(t)$. We then determine the time t that satisfies the following condition (5.6) of a local maximum (peak).

$$\omega(t-1) < \omega(t) > \omega(t+1) \tag{5.6}$$

Similarly, we also find the time t that satisfies the condition (5.7) of a local minimum (valley).

$$\omega(t-1) > \omega(t) < \omega(t+1) \tag{5.7}$$

To remove the peaks and valleys due to noise, we further apply noise filtering based on a prominence [93]. The prominence is used in signal processing to measure how much the peak is prominent

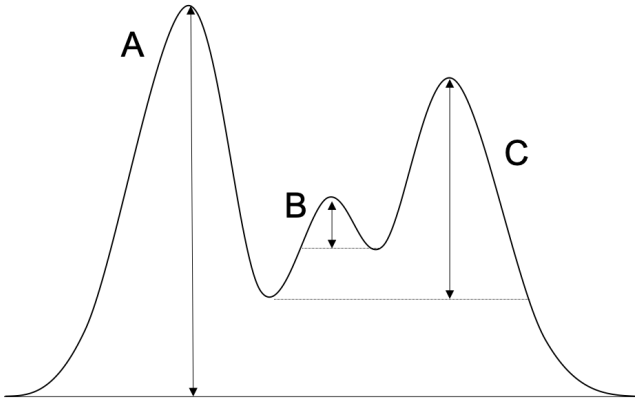

Figure 5.7: Example of prominence

Table 5.1: Minimum maneuver interval [s] of players

	PUSH		*PULL*	
Player ID	# of *PULL*	Min. Interval	# of*PULL*	Min. Interval
1	113	0.33	40	1.13
2	72	0.30	31	0.86
3	63	0.33	35	0.90
4	90	0.33	32	1.53
5	105	0.30	36	1.00
6	80	0.40	38	1.06

considering the relative height to its surrounding peaks. An example of prominence is shown in Figure 5.7. Vertical arrows show the prominence of three peaks on a prominence island which is the reference level of the prominence illustrated by the dashed horizontal lines in Figure 5.7. The prominence island is defined as follows. First, we extend a horizontal line from a peak to the left and right until the line crosses the signal due to a higher peak or the end of the signal. Then, we find the minimum of the signal in each of the two intervals. Finally, the higher of the two intervals minimal specifies the reference level. The height of the peak above the reference level is its prominence. We analyzed the characteristics of the prominence of angular velocity peaks. Figure 5.8 shows the prominence during practice and Figure 5.9 shows the height distribution of the prominence. We found there are many peaks with low prominence due to noise. We therefore chose to exclude peaks with a prominence of less than 20. If there are multiple peaks/valleys within T_{min}, all peaks/valleys except for the one with the largest/smallest value have also been removed since such extremely fast actions are impossible.

Table 5.1 shows the minimum interval between *PUSH* and *PULL* for six players during the game. From this result, the threshold of the minimum interval T_{min} was set to 0.3.

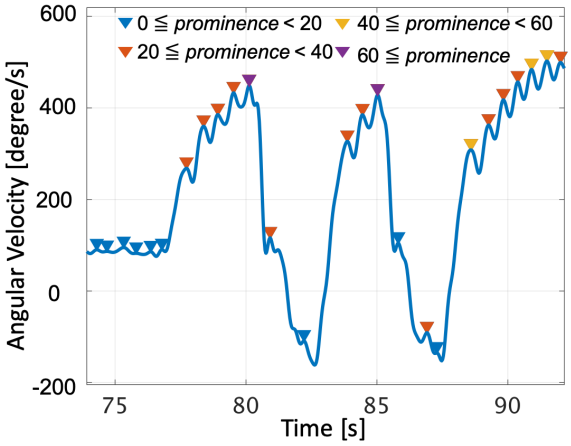

Figure 5.8: Prominence of angular velocity peaks during practice

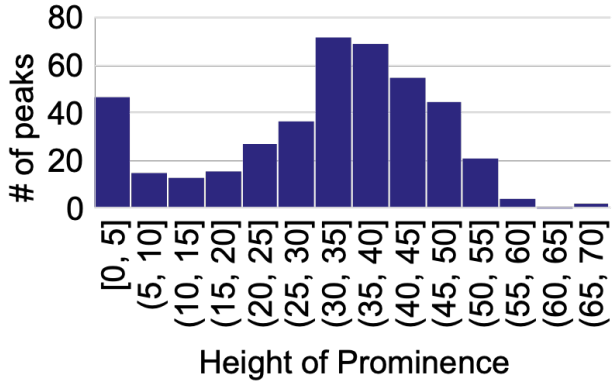

Figure 5.9: Distribution of the prominence

Finally, we segment the time series of the angular velocity by the detected peaks and valleys. We let d^i denote the i-th detected peak or valley for the time series of the detected peaks and valleys. Then, the i-th segment s^i $(i > 0)$ is defined as $(t(d^{i-1}), t(d^i)]$ where $t(d^i)$ is the time when d^i is observed. Since there is no zero-th peak or valley, $t(d^0)$ is defined as 0 which is the start of the measurement.

5.3.4 Classification

We classify each segment into *PUSH*, *PUSH*, or *OTHERS*. However, the waveform greatly differs depending on the speed of the wheelchair even for the same *PUSH* actions. For example, Figure

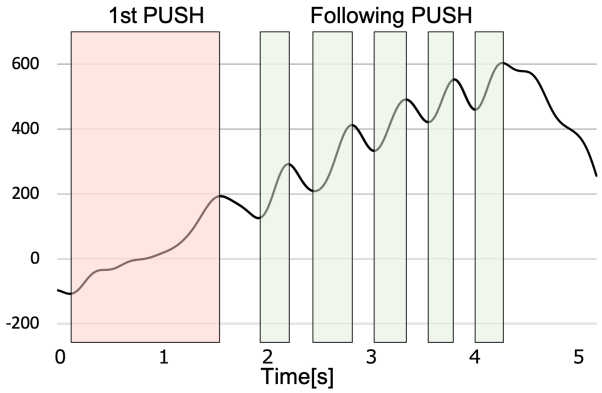

Figure 5.10: Example of *PUSH* angular velocity

5.10 shows the waveform of the angular velocity during a sprint. The first *PUSH* segment and the following *PUSH* segments are clearly different. A large velocity change occurs at the first *PUSH* while the velocity change in the following *PUSH* is not as large as the change in the first *PUSH*. This is due to the player's own ability and the speed just before *PUSH*.

The classification is performed as follows based on the above observation. Each segment s^i is classified into *PUSH* if the following three conditions (5.8), (5.10), and (5.11) are satisfied. The *PUSH* action maximizes the speed in the short term. Therefore, the first condition is that s^i ends at a peak. This is expressed as below.

$$\omega(t(d^i) - 1) < \omega(t(d^i)) > \omega(t(d^i) + 1) \tag{5.8}$$

The second condition is that there is a large speed change within the segment period. Since the degree of the speed change depends on the player's ability, we determine the threshold T_{height} for the amount of speed change considering the player's ability and the speed before the action. We consider the player's ability as the highest speed in a game or practice. When the maximum speed is ω_{max} and the angular velocity at the end of the previous segment is $\omega(t(d^{i-1}))$, the threshold T_{height} of the speed change is defined as given below.

$$T_{\text{height}} = \begin{cases} \omega_{max}/16 & (\omega(t(d^{i-1})) >= \omega_{max}/4) \\ (\omega_{max} - 3\omega(t(d^{i-1})))/4 & (\omega_{max}/4 > \omega(t(d^{i-1})) > 0) \\ \omega_{max}/4 & (0 >= \omega(t(d^{i-1}))) \end{cases} \tag{5.9}$$

Since the amount of speed change within the segment periods must exceed this threshold, the second condition is given below.

$$|\max_{t \in (t(d^{i-1}), t(d^i))} \omega(t) - \min_{t \in (t(d^{i-1}), t(d^i))} \omega(t)| \geq T_{\text{height}} \tag{5.10}$$

The third condition is the rapid speed increase. The increase in speed occurs due to not only *PUSH* but also weight shifting and movement of the opposite wheel. On the other hand, in a *PUSH* action, it is necessary to grab the rim, leading to a slight instantaneous decrease in speed, before the speed increase. This leads to a significantly rapid increase of the angular velocity. Therefore, by using the threshold for the rapid speed increase T_δ^{PUSH}, the third condition is represented as follows:

$$\max_{t\in(t(d^{i-1}),t(d^i)]} \omega'(t) \geq T_\delta^{PUSH}, \tag{5.11}$$

where $\omega'(t)$ is the time derivative of $\omega(t)$. When the above three conditions (5.8), (5.10), and (5.11) are satisfied, segment s^i is classified into *PUSH*.

On the other hand, each segment s^i is classified into *PULL* if conditions (5.10), (5.12), and (5.14) are satisfied. Contrary to the *PUSH* action, the *PULL* action minimizes the speed in the short term. Therefore, the first condition is that s^i ends at a valley. This is expressed as below.

$$\omega(t(d^i)-1) > \omega(t(d^i)) < \omega(t(d^i)+1) \tag{5.12}$$

The second condition is that there is a large speed change within the segment period. This is same as the condition (5.10) in *PUSH*. However, the threshold T_{height} of the speed change is defined as given below.

$$T_{\text{height}} = \begin{cases} \omega(t(d^{i-1})) - \omega_{max}/4 & (\omega(t(d^{i-1})) > \omega_{max}/2) \\ \omega_{max}/4 & (\omega_{max}/2 >= \omega(t(d^{i-1}))) \end{cases} \tag{5.13}$$

The third condition is the rapid speed decrease occurs. The decrease in speed occurs due to not only *PULL* but also to friction and weight shifting. On the other hand, a *PULL* action needs to grab the rim, resulting in a rapid decrease in speed. Therefore, by using the threshold for the rapid speed decrease T_δ^{PULL}, the third condition is represented as follows.

$$\min_{t\in(t(d^{i-1}),t(d^i)]} \omega'(t) \geq -T_\delta^{PULL} \tag{5.14}$$

Finally, all of the other segments are classified as *OTHERS*. An example of the classification result using the proposed method is shown in Figure 5.11.

5.3.5 Remove Noise by Collision

Wheelchair collisions frequently occur during games. At that moment, a maneuver to stabilize the wheelchair may be performed. However, the wheel angular velocity decreases rapidly regardless of the occurrence of maneuvers. Such rapid decreases of the angular velocity are wrongly classified as *PULL*. To solve this problem, we detect collisions and change *PULL* labels within a fixed period from the collisions to *OTHERS*. We use acceleration to detect collisions. We also determined the threshold for

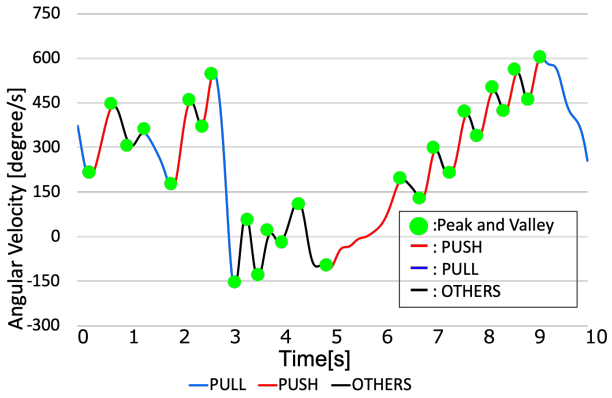

Figure 5.11: Example of classification result

collision detection and the duration of the period causing wrong *PULL* labels based on the statistics as follows.

The magnitude of 3 axis acceleration $a[G]$ can be expressed by the following equation.

$$a = \sqrt{a_x^2 + a_y^2 + a_z^2}\,[\mathrm{G}] \tag{5.15}$$

The waveforms of the acceleration in sprint and collision are shown in Figure 5.12. We see that the acceleration is obviously higher during the collision than the other cases. Figure 5.13 shows the range of the maximum acceleration for 24 collisions observed during one minute of the preliminary experiment. The minimum value of the maximum acceleration was 10.17. Therefore, we detect a collision when a exceeds $10[G]$. Figure 5.12 shows there are several peaks in addition to the maximum peak at collisions. In order to investigate the effect of the vibration caused by the collisions, the number of peaks that exceed $5[G]$, which is half of the acceleration threshold of 10, is analyzed within 0.1 to 0.6 seconds from the maximum peaks. As shown in Figure 5.14, the number of peaks around the maximum peak increases as the range of time is expanded. We also found the increase is small around 0.5 seconds. Therefore, we determined the duration of the period causing wrong *PULL* labels as 0.5 seconds.

We note that the speed of the wheelchair does not increase upon collision. This means that collisions do not cause wrong *PUSH* labeling. Therefore, *PUSH* actions are not filtered since *PUSH* soon after the collisions typically shows a significant increase in the angular velocity which is totally different from wrong *PULL* labels due to collisions.

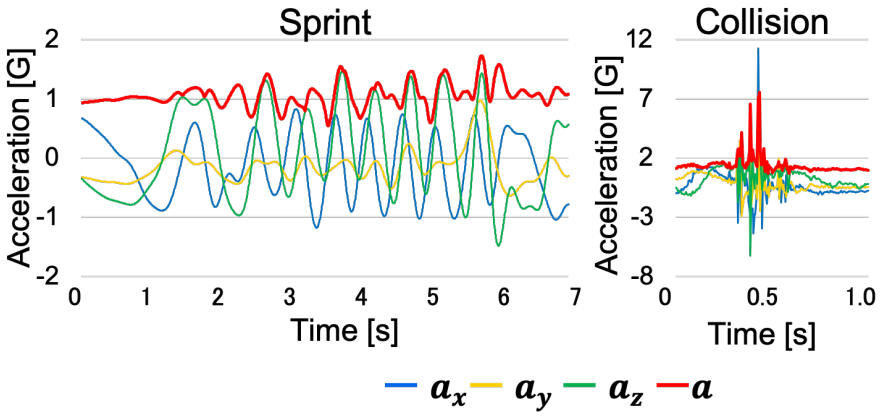

Figure 5.12: Example of acceleration during sprint and collision

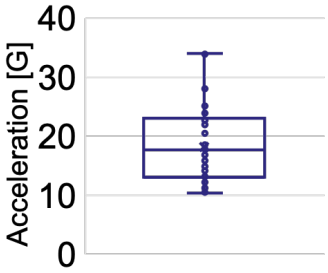

Figure 5.13: Distribution of maximum acceleration at collisions

5.4 Turn Classification

5.4.1 Detection

We also classify turns of wheelchairs into *PIVOT* and *TURN* (the other turns). As Figure 5.15 shows, *PIVOT* is a change of direction with one fixed wheel while *TURN* is any other change in direction. *TURN* has two types of motions, curve and spin. A curve is an action where both wheels move in the same direction when changing the direction of the wheelchair, while spin moves the wheels in the opposite directions. From observations and discussions with players and coaches, we define turns as the movement with the rotation of wheelchairs of more than 60 degrees within 1.5 seconds.

To extract periods that meet the above definition, we calculate the rotation degree of the wheelchair based on the model of the two-differential wheeled robot [94]. From the inertial sensors, the angular velocities $\omega_{right} and \omega_{left}$ around the axles of the left and right wheels are obtained. Let r denote the length of the radius of the wheel. The speeds of the left and right wheels v_{right} and v_{left} are then

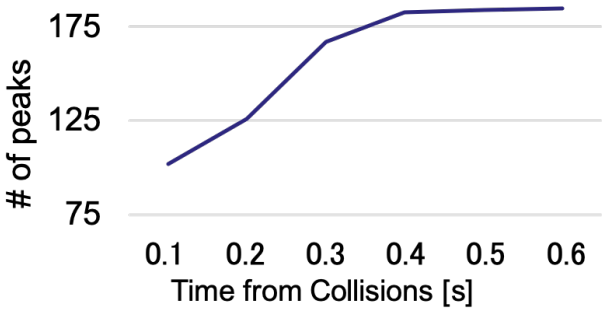

Figure 5.14: Number of peaks around collisions

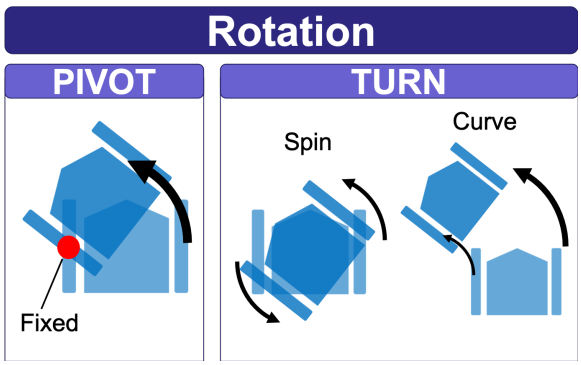

Figure 5.15: Type of rotation

respectively expressed as below.

$$v_{right} = r * \omega_{right} \tag{5.16}$$

$$v_{left} = r * \omega_{left} \tag{5.17}$$

Next, we assume that the wheelchair is making a motion around the center of the rotation. As shown in Figure 5.16, if the angular speed of turning is ω_{turn} and the radius of the turn is ρ, the speed at the center of the wheelchair v is represented as shown below.

$$v = \rho\omega_{turn} \tag{5.18}$$

On the other hand, if the distance from the center to the wheel is d, the turn radii of each wheel increase or decrease by d, and the speeds of the left and right wheels are as follows.

$$v_{right} = (\rho + d)\omega_{turn} \tag{5.19}$$

$$v_{left} = (\rho - d)\omega_{turn} \tag{5.20}$$

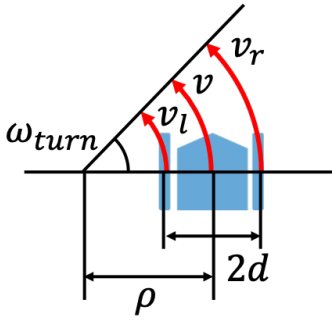

Figure 5.16: Wheelchair in turn

Solving for the above formulas yields the following formula:

$$\omega_{turn} = (v_{right} - v_{left})/2d \tag{5.21}$$

$$v = (v_{right} + v_{left})/2 \tag{5.22}$$

$$\rho = d(v_{right} + v_{left})/(v_{right} - v_{left}), \tag{5.23}$$

where ω_{turn}, v and ρ denote the rotation speed, forward speed and radius of rotation of the wheelchair, respectively. We use ω_{turn} to detect turns as we defined. The center of the rotation is on the left side of the direction of movement when ρ is positive, and vice versa.

5.4.2 Classification

Next, for each detected turn, we classify whether it is a pivot or not. During a pivot, one of the wheels is stationary. However, it is difficult to completely stop the wheel. We therefore use the amount of rotation of the wheel with lower speed during the turn. We represent the amount of rotation of the wheel with lower speed during the i-th turn as θ^i_{lower} defined as:

$$\theta^i_{lower} = \min \left[\sum_{t \in turn^i} \omega_{right}(t), \sum_{t \in turn^i} \omega_{left}(t) \right], \tag{5.24}$$

where $turn^i$ denotes the period of the i-th turn. If θ^i_{lower} is less than $\pi/4$, it is classified as *PIVOT*. In addition, if θ^i_{lower} exceeds $\pi/4$, the i-th turn is classified as *TURN*.

5.5 Evaluation

5.5.1 Maneuver Classification

Settings

We collected real data in a practice game for evaluation. The game duration was 317 seconds. We attached inertial sensors to wheelchairs of six players. The maximum speed observed by each player

Table 5.2: Max speed [degree/s] of players

Player ID	Max Speed(Left)	Max Speed(Right)
1	675.7	667.8
2	764.4	725.7
3	630.6	624.6
4	779.3	802.7
5	640.3	693.0
6	636.8	678.6

Table 5.3: Leave-one-person-out cross validation

	Threshold Setting		F-measure		
ID	T_δ^{PUSH}	T_δ^{PULL}	*PUSH*	*PULL*	Both
1	1.5	1.5	0.90	0.80	0.89
2	1.5	1.5	0.88	0.81	0.86
3	1.5	1.5	0.83	0.77	0.82
4	1.5	1.5	0.92	0.81	0.90
5	1.5	1.5	0.92	0.86	0.91
6	1.5	2.0	0.91	0.75	0.88
All	-	-	0.87	0.85	0.88

during the game is listed in the Table 5.2. We manually labeled the maneuver actions by recording a video. However, due to occlusion and image quality, the labeling was sometimes difficult. To deal with such ambiguity in manual labeling, we established the criteria of ground truth. We judged that the wheel was pushed when it was obvious that the player gripped the rim with the movement going forward based on the player's arm motion. We also identified *PULL* when it was obvious that the wheel suddenly decelerated or moving backward while the player gripped the rim based on the player's arm motion.

After labeling, a total of 1157 maneuver actions were performed, consisting of 1005 *PUSH* and 152 *PULL*. Since the ground-truth is labeled manually, we allow 1.5 seconds difference for the detection time or in other words the detected class is regarded as correct if the same ground-truth label exists within 1.5 seconds.

Results

Threshold Configuration In our method, we need to configure the thresholds appropriately. To see the difference in the thresholds for different players, we conducted a leave-one-person-out cross validation. The thresholds T_δ^{PUSH} and T_δ^{PULL} are set from 1.0 to 3.5 with increments of 0.25 as below.

$$(T_\delta^{PUSH}, T_\delta^{PULL}) \in [(1.0, 1.0), (1.0, 1.25), ...(3.5, 3.5)] \tag{5.25}$$

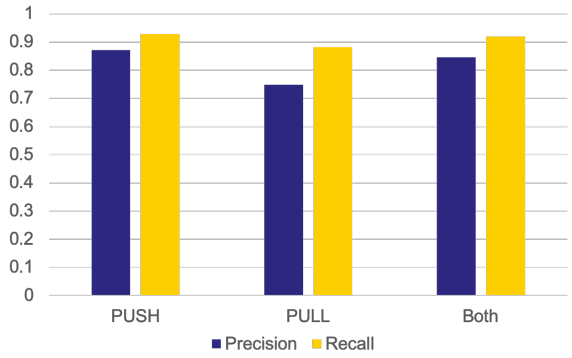

Figure 5.17: Average maneuver classification result

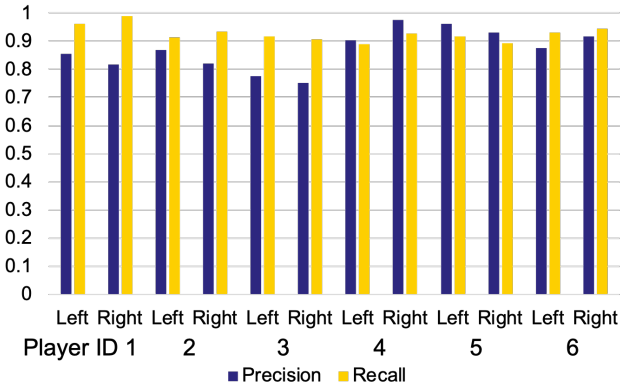

Figure 5.18: Classification performance of left and right hands for each player($PUSH$)

From the result shown in Table 5.3, we confirm the optimal threshold setting is the same among 5 of the 6 cases in the cross validation. However, we found that the F-measure of the player 3 is slightly worse than the others. This is mainly because the wheel size of the player 3 was larger than the others due to the wheelchair configuration. Therefore, we may adjust the thresholds according to the wheelchair configuration to improve the performance. In the following evaluation, the thresholds are set as $T_\delta^{PUSH} = 1.5$ and $T_\delta^{PULL} = 1.5$.

Maneuver Classification Performance Figure 5.17 shows the result of maneuver classification. From the results, we confirm precision and recall of $PUSH$ are more than 87.1%. We also confirm the precision and recall of $PULL$ are more than 74.8%. Figures 5.18 and 5.19 show the maneuver classification performance of the left and right hand for each player. From the results, we see that the $PUSH$ classification performance is independent of hands in most cases. However, we also see the $PULL$ classification performance is different between the left and right hands for players 2 and 6. Furthermore,

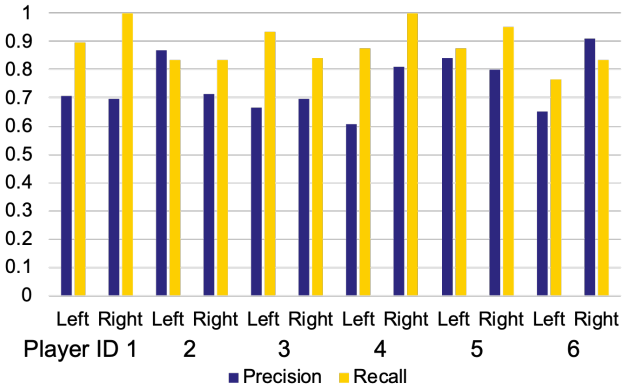

Figure 5.19: Classification performance of left and right hands for each player(*PULL*)

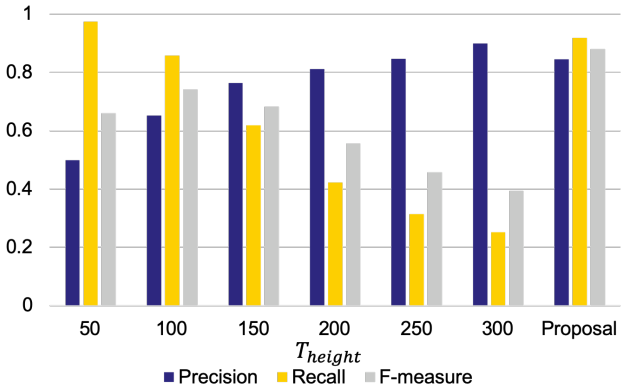

Figure 5.20: Performance when T_{height} is fixed

as for player 4, recall of the left side is the lowest in all the players. Conversely, the precision of player 4's right side is the highest among the other players. This implies that the characteristics of the maneuver actions may differ slightly depending on hands and/or players. The results also mean our method can potentially assess the quality of maneuver. To improve the performance, in addition to the maximum speed, we may investigate the factors that can estimate the ability of individual players.

Figure 5.20 shows the result when the threshold T_{height} is fixed. When the threshold T_{height} is low, the recall is high because it can detect small *PUSH* and *PULL*. However, the peaks and valleys due to noises are wrongly recognized as maneuvers, leading to low precision. On the other hand, when the threshold T_{height} is high, precision becomes high while recall becomes low. This is because only maneuver actions with large movement are recognized. Our method achieves the highest F-measure, which means adjusting the threshold T_{height} based on the speed works well.

Table 5.4: Turn dataset

Condition	Angle[degree]	# of *PIVOT*	# of *TURN*
Moving	180	27	33
	-180	25	39
Stationary	270	20	20
	180	20	20
	90	20	20
	-90	20	20
	-180	20	20
	-270	20	20
Total		172	192

Table 5.5: Turn classification result

		Predicted Class		
		PIVOT	*TURN*	Recall
True Class	*PIVOT*	172	0	1
	TURN	1	191	0.9947
	Precision	0.9942	1	

5.5.2 Turn Classification

Settings

In order to measure the turns while moving, we collected the data by repeating a turn after moving forward. We asked the player to make either a pivot turn or a spin turn in a specified direction (i.e. left or right). Also, in order to evaluate turns while stationary, we asked the player to make a turn with a specified angle from -270 to 270 [degree]. The leftward rotation is considered positive and the rightward rotation is considered negative. The summary of the collected data is as shown in Table 5.4. We observed 172 *PIVOT* and 192 *TURN*.

Turn Classification Result

The results are shown in Table 5.5. The results show almost all the turns were correctly classified except only one *TURN* which is a spin turn during forward movement. To investigate the reason for the wrong classification, Figure 5.21 shows the amount of rotation of a low-speed wheel during a turn. This figure shows that it is difficult to change the moving direction of the wheel suddenly. As a result, the minimum amount of rotation of *TURN* during forward movement becomes closer to the maximum amount of rotation of *PIVOT* during stop. This leads to a wrong classification. This problem may be solved by considering the speed before a turn.

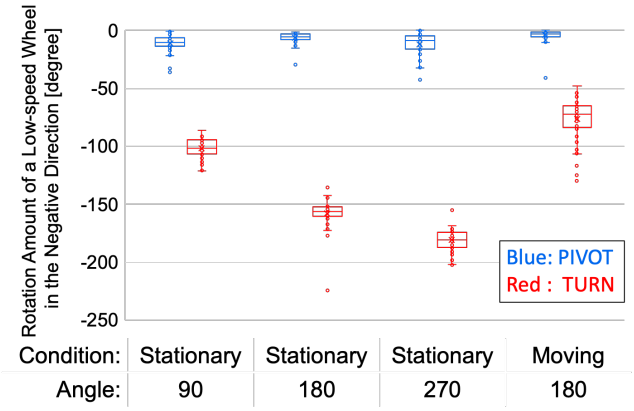

Figure 5.21: The amount of rotation of a low-speed wheel during a turn

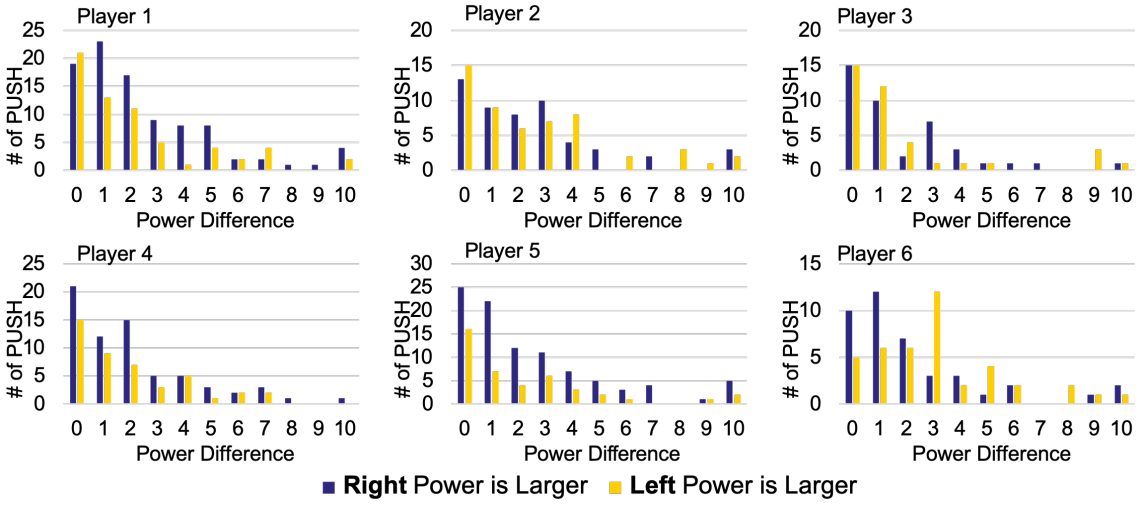

Figure 5.22: Power Difference of Left and Right Hands

5.5.3 Use Cases on Data Analysis

Difference between left and right hands

To see the difference between the left and right hands, we define the power of *PUSH* action in segment s^i as the difference in the angular velocity (i.e. $\max_{t \in s^i} \omega(t) - \min_{t \in s^i} \omega(t)$). Then, we calculate the difference in the power between the left and the right wheels when *PUSH* is recognized for both of the wheels simultaneously. We assume the left and the right wheels are pushed simultaneously if both of the peaks at the end of the segments are detected within 0.5 seconds. Figure 5.22 shows the

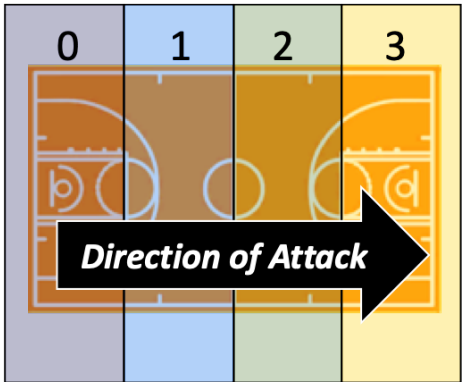

Figure 5.23: Basketball court divided into 4 areas

Table 5.6: Percentage of maneuver actions in each area[%]

Area\Action	*PUSH* (Low)	*PUSH* (High)	*STOP*
0	30.1	38.6	31.3
1	16.2	68.7	15.1
2	13.5	71.3	15.1
3	25.7	44.6	29.8

distributions of the difference in power between the left and the right wheels. It is clear that 3 out of 6 players (players 1, 4, and 5) push the right wheel more strongly than the left wheel. This result implies that some players tend to rely on their dominant hands and to make turns in the same direction.

Relationship between Maneuver Motions and Positions

We analyzed the relationship between the recognized maneuver actions and the positions. The players' positions are obtained from the video. For the analysis, the basketball court is divided into four areas by the foul lines and the center line as shown in Figure 5.23. In this analysis, *PUSH* is categorized into two types depending on whether the angular velocity of the previous segment is above the threshold $\omega_{max}/4$ (high speed) or not (low speed).

Table 5.6 shows the percentage of the maneuver types in each area. We see that high-speed *PUSH* exceeds 68.7% around the areas 1 and 2 which are the center of the areas. In particular, the percentage of *STOP* is extremely low in area 2 which is the first area of the opponent's court. This is because the players tend to accelerate rapidly for good movements when they are attacking. In addition, the result indicates that various maneuver types are mixed near the goals (i.e. the areas 0 and 3) because sophisticated movements are required to avoid or to interfere with opponents.

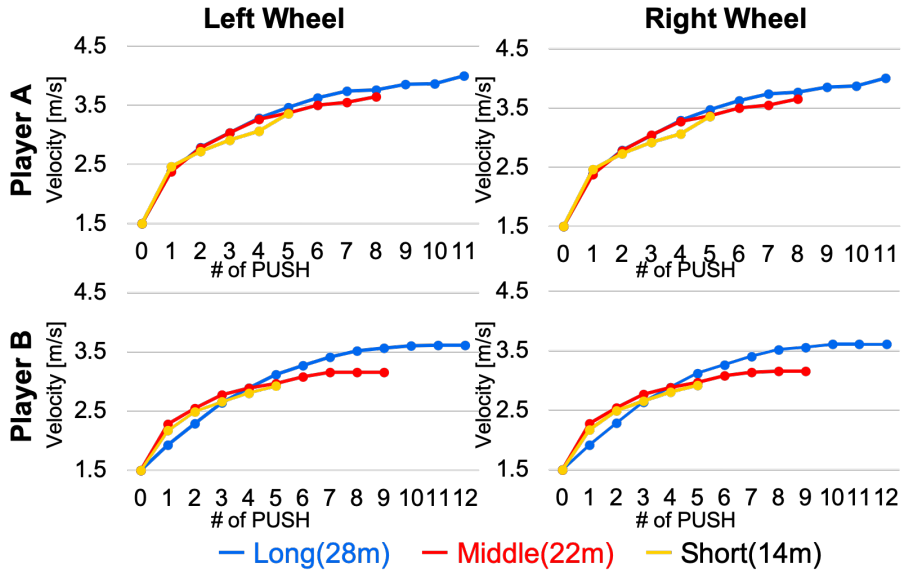

Figure 5.24: Speeds of 2 players over time in sprints at 3 different distances

Sprint Comparison

In training, sprints are often practiced. Figure 5.24 shows the speeds of two players over time in three trials of sprints at different distances. The left side is the speed of the left wheel and the right side is the speed of the right wheel. The x-axis is the number of *PUSH* actions. For example, the upper right figure shows player A reached 3.5m/s at the fifth right hand *PUSH* for all the trials.

As seen from Figure 5.24, in the short and long distance sprints, both players achieved almost the same speed with the same number of pushes. However, in the middle distance sprint shown in the red color, player B after the third *PUSH* shows a smaller increase in speed than yellow, which is long-distance sprint. This means that player A always achieved high performance in terms of speed regardless of distance.

5.6 Discussion

In this chapter, we classify maneuver actions and turns for the purpose of assessing maneuvering quality. Our system allows confirming whether the players to confirm whether the players moved quickly and/or efficiently. It also helps them to improve their handling technique. For example, if a player tries to push a wheel with a strong force, that player can greatly accelerate all at once. At the same time, the force to grip the rim can become stronger, leading to larger deceleration of the wheel. By collecting practice data for sprints, players can understand how to efficiently reach the maximum

speed without wasting force. Furthermore, we are planning to analyze the relation between the degree of disability and maneuver statistics measured by our system.

We note that, "better" actions often depend on situations. This means maneuver actions and turns recognized by our system may be not enough for analysis in games. To enable game analysis, we may integrate a video tracking system to use player positions that reflect situations in a game. Such information about player positions enables us to consider what the best action is at the moment. For example, a player should conduct a pivot turn if an opponent is blocking the player's forward movement.

We assume our system will be used in training and practice games. In official games, the current rules for wheelchair basketball do not allow the use of sensors [95]. However, since data analysis in sports is becoming more common, the rules may change in the future.

5.7 Conclusion

In this chapter, we propose two methods to detect and classify maneuver actions and turns of wheelchair basketball by using inertial sensors. Our design of the proposed method focuses on the specific movement of wheels. The evaluation results showed that our method achieves an F-measure of 88.1% for classification of maneuver actions. Also, our method achieves an F-measure of 99.7% for the classification of turns. Furthermore, we have shown usage cases for data analysis by using the classification results combined with other information such as player positions.

One topic for future work is applying in order to support technical improvement of maneuvering. For example, it is possible to achieve efficient training by quantifying the wheelchair maneuver actions through feedback to the players. In addition, visualization of changes in the maneuver actions over time may motivate the players. Furthermore, in cooperation with athletes and coaches, we are planning to develop a system to support data analysis in wheelchair basketball.

Chapter 6

Maneuver and Play Action Recognition in Wheelchair Sports Using a Single Camera

6.1 Introduction

Recently, data utilization in sports has been attracting attention in all over the world, and it is used for various purposes such as advanced strategic analysis and effective training. In sports such as soccer and basketball, the number of actions that utilize balls such as passes and shots are collected and the data is opened to the public as one of the stats showing the performance of players, and that is used for entertainment. Wheelchair basketball, which is one of the sports for people with disabilities, is also required to support technical improvement and strategy planning using data, and several attempts have been made so far.

Data such as passes and shoots in sports are mainly manually tagged by experts. Many tools have been developed to make tagging easier, but even at big sports data collecting company *Opta Sports* [96], which collects data in 30 different sports in 70 countries, three experts manually tag each match. Therefore, it is difficult to collect data for analysis even in amateur sports such as universities, and automation of data collection is required. As a method of automatically collecting data in sports, an approach using an inertial sensor or a camera are considered. In the approach using inertia sensors, by attaching a sensor to the player's own body or equipments, it is possible to directly obtain player's movements such as acceleration and angular velocity of the limbs, equipments and etc. However, there is a risk of injury when using in sports with contact. There is also concern that the player's performance is deteriorated by bad feelings of attaching sensors. In addition, since one or more sensors are required for each person, there is a problem that workloads such as mounting and charging sensors is high to preparation for every time of practices or games. On the other hand, in the camera-based approach, it is not necessary to attach a sensor to player's body or equipments, so the effect on the performance

is small. When using a camera, it is necessary to recognize objects such as hands and legs for each person in the image, and to recognize how each object moves. To achieve it, many methods use CNNs for spatial feature extraction in images. Recently, a network model *Slowfast* involving a Slow pathway, operating at low frame rate, to capture spatial semantics, and a Fast pathway, operating at high frame rate, to capture motion at fine temporal resolution has been proposed [11], and it achieved to recognize daily actions with high accuracy. If the entire court is photographed with a camera, we can collect much data, which greatly reduces the time and effort required for preparation such as mounting sensor on each player. Therefore, we consider action recognition method using a camera.

When recognizing actions using a camera in sports, there are problems that actions in sports are different from daily actions. Even if the movements of person are very similar, these actions sometimes have different meanings in sports. For example, in sports, it is necessary to classify actions such as passes and shoots, but these actions are similar and it is difficult to classify. This is because existing methods including *Slowfast* focus on movements of the limbs without considering positions in the field. Therefore, in the existing methods, it is difficult to classify different actions with similar movements such as passes and shoots. They can recognize the movement of releasing the ball from their hands, but they do not know whether the ball is going to the goal or the ally. Therefore, in this chapter, we consider that when taking an action, a player's judge is affected by the position in the court, and propose a action recognition method that takes into account players' positions. In our method, we use *Slowfast* model to capture human centric movements (gesture) and extract the players' position characteristics which is court centric movements by localization. By using these two outputs, we classify the actions.

For evaluation, we compared our method with the *Slowfast*-only model. As a result, our method using location information achieved better accuracy than the *Slowfast*-only model in all four types of classifiers. Using the configs that achieved the best accuracy, the accuracy reached 78.7 %, which was 9.6 points higher than the best accuracy using the *Slowfast*-only model. From these result, it can be seen that the players' positions are important in the action recognition task in sports.

6.2 Data Preparation

As far as we know, there is no datasets with tagged actions of wheelchair basketball. Therefore, we create new datasets. We got a video from actual practice games. The recorded videos' sampling rate was 60 frames per second. Wheelchair basketball has ball actions such as passes and shoots, as in basketball and maneuver actions such as pushing the wheel. Therefore, we defined the 6 types of action labels which includes 3 ball actions (pass, shoot, and dribble) and 3 maneuver actions (both-handed pushing, left-handed pushing, and right-handed pushing). In labeling, the type of action, the start/end frame of the action, and the player ID are recorded for one action. We collected data from 14 of the players who participated in two games. Wheelchair maneuver actions can be obtained from

Table 6.1: Details of Data

Label	Clips	Time[s]
Pass	179	168.3
Shoot	99	137.5
Dribble	198	208.4
Both Push	366	209.6
Left Push	200	167.7
Right Push	226	197.1
Total	1268	1088.5

players who do not have the ball, but since there is only one player having the ball during the game, the data labeled with ball actions is very few compared to the data labeled with maneuver actions. Therefore, we collected data of ball actions from videos of eight games.

Next, for each labeled action, we create a video clip that trimmed the video from the start to the end of the action. However, it is very difficult to manually annotate each person by a bounding box for each frame. Therefore, we use human detection and tracking technologies. By using them, we need only record the actual player ID corresponding to tracking ID which is given by tracking technology, the workload was greatly reduced. In this Chapter, the latest technologies, A and B, were used to detect and track people, respectively. However, human tracking may fail due to occlusion or going out of the image. In that case, a complete human trajectory was obtained by linear interpolation until the next tracking ID was observed. By cutting out the bounding box of the tracking ID corresponding to the player ID from the start to the end of the action, a video clip of one person performing one action is generated. As a result, 1268 videos were obtained for about 1089 seconds in total. The average length per video clip was about 0.86 seconds. The details of the datasets are listed in the Table 6.1.

6.3 Method

The overview of our method is shown in Figure 6.1. Because the human detection and tracking is out of scope of the action recognition, our method assumed that the person ID and the position of the bounding box in the video are given. Our action recognition method consists of three blocks: (1) localization, (2) gesture recognition, and (3) action classification. First, the positions of the players in input video are estimated. In localization, the coordinates in the image are transformed from the pixel axis to the world axis by homography [48]. By using this, the place where the action was performed and the trajectory during the action is obtained. Also, in parallel with localization, human gesture recognition using *Slowfast* model is performed. As a result, it outputs the confidence score for each action label. Finally, these outputs are input to the classifier to determine an action label. For example, if a running person appears in a video, in localization, we extract the global features such as the speed and moving direction of the person. On the other hand, in gesture recognition, we extract the local

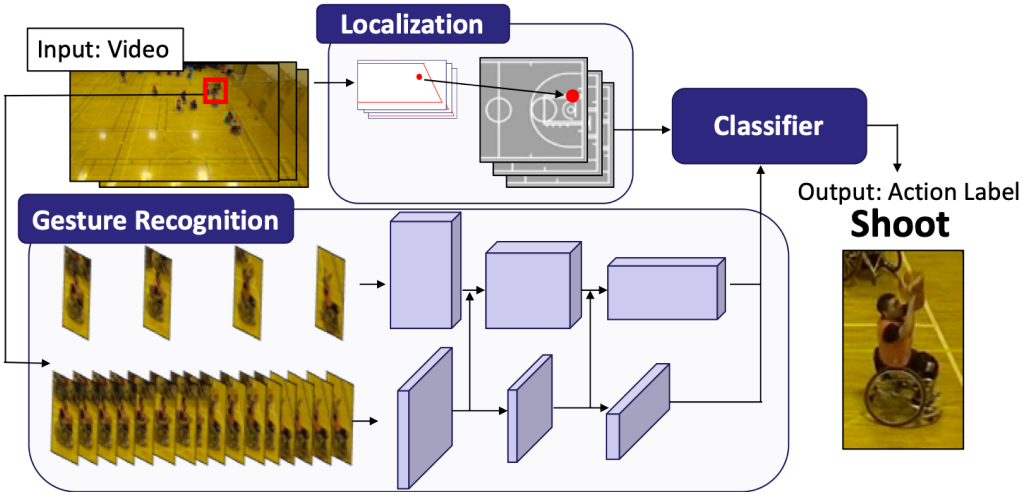

Figure 6.1: Method Overview

features such as the crossing of the person's arms and legs. We then classify the action using these global and local features. In this way, we realized the action recognition model considering the position information.

6.3.1 Extraction of Positional Features by Localization

First, the player's position is estimated for the purpose of extracting the features of the movement in the court. The coordinates of the center of the bottom of the bounding box detected in the video are used as the position of the person, and the pixel axis is transformed to the world axis by homography. Homography requires that the coordinates of the actual world corresponding to the four points in the image are known. The position data is converted into a form in which the data for input to the classifier. Specifically, three types of data are obtained: the position pos where the action was performed, the displacement d during the action, and the change in orientation θ during the action. Because the size and position of the detected bounding box changes between frames, we define the average position during the action as the position where the action was performed. Assuming that the number of frames in the action period is F and the position in the frame f is (x^f, y^f), the position pos in where the action was performed is calculated as follows.

$$pos = (\sum_{f=1}^{F}(x^f, y^f))/F \tag{6.1}$$

In addition, we divide the action period into three equal parts, and the average position in each period is defined as the action start position (x_{start}, y_{start}), the action occurrence position (x_{action}, y_{action}), and the action end position (x_{end}, y_{end}), respectively. Displacement during the action period is defined

as the distance from the action end position to the action start position. Therefore, the displacement d during the action period is as follows.

$$d = \sqrt{(x_{end} - x_{start})^2 + (y_{end} - y_{start})^2} \tag{6.2}$$

In addition, we defined the moving direction from the action start position to the action occurrence position as the orientation before action and we then defined the moving direction from the action start position to the action end position as the orientation after action. At that time, these orientation is absolute. Therefore, the relative change in orientation θ during the action is the difference between them, and it it as follows.

$$\begin{aligned} \theta = &\text{atan2}(x_{end} - x_{start}, y_{end} - y_{start}) \\ &- \text{atan2}(x_{action} - x_{start}, y_{action} - y_{start}) \end{aligned} \tag{6.3}$$

Figure 6.2 shows an example of data for a position during a action.

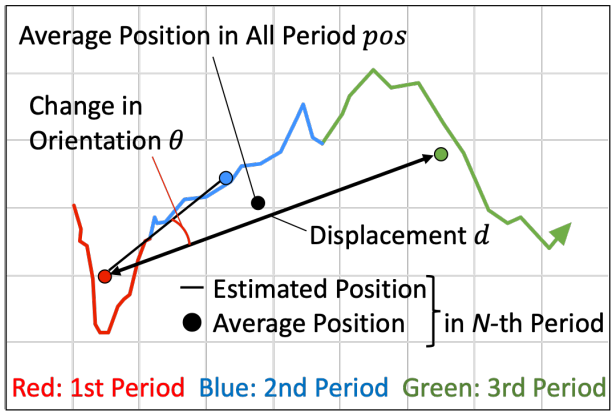

Figure 6.2: Example of Position Data During an Action

The position of the action for each label is shown in Figure 6.3. Since the attack direction is reversed depending on the team, in order to unify the right direction as the attack direction in the figure, the position coordinates of the players of the team whose left direction is the attack direction are converted point-symmetrically with the center of the court as the axis. From this figure, it can be seen that there is a correlation between each action and the place where the action is performed. For example, because the goal position and attack direction is fixed the shot can only be seen on the right court. Also, while dribbling is performed throughout the court, passes are more frequent in front of each goal. This result shows the characteristics of wheelchair basketball, which is looking for an ally who can shoot in good condition and many passes are made in front of the goal. In addition, it can be seen that a pass occurs even outside the court. It is because when the ball has gone out of court, passes to put inside is performed. On the other hand, in wheelchair maneuver actions, pushing

by both hands are performed uniformly throughout the court. However, pushing by only left or right hand is more frequent in front of the goal of each court. This is because the offense and defense are switched immediately after shoots, and the direction is changed accordingly. In this way, it can be seen that each action is related to the position where the action was performed.

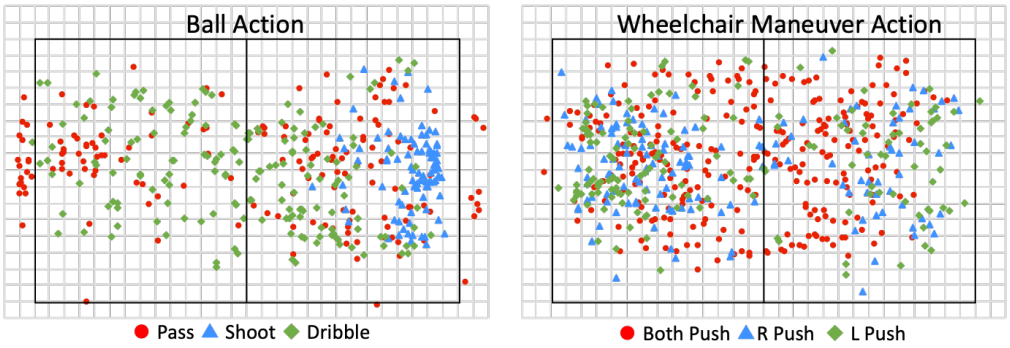

Figure 6.3: Relationship between Action Labels and Positions

Next, the movement vector $(\vec{x}, \vec{y})$ is defined as follows.

$$(\vec{x}, \vec{y}) = (d * \cos\theta, d * \sin\theta) \tag{6.4}$$

At that time, the movement vector of each label is shown in Figure 6.4. The positive direction of the X-axis in this figure is the orientation before action. For example, if a marker is in positive positions on both the X-axis and Y-axis, it can be seen that the player corresponding to the marker turns to the left while moving forward during the action was performed. From this figure, it can be seen that passes and shoots are likely to be performed at low speeds or when stopped, and dribbling is likely to be performed while moving forward. In some cases, shoots such as layup are performed with a larger movement than dribbling. On the other hand, the movement vector in terms of wheelchair maneuver actions has different clearly. It is because pushing by one hand is performed to change the orientation of the wheelchair. Therefore, pushing by the left hand is changing the orientation to the right, and conversely, pushing by the right hand is changing the orientation to the left. That tendency can be seen from this figure.

6.3.2 Gesture Recognition

Gesture recognition is performed in parallel with localization. The purpose is to extract the local features which human centric action such as hand and foot movements have. We use *Slowfast* for gesture recognition. *Slowfast* is a network model with a two-pathway for video recognition. One pathway is designed to capture semantic information that can be given by images or a few sparse frames, and it operates at low frame rates and slow refreshing speed. In contrast, the other pathway

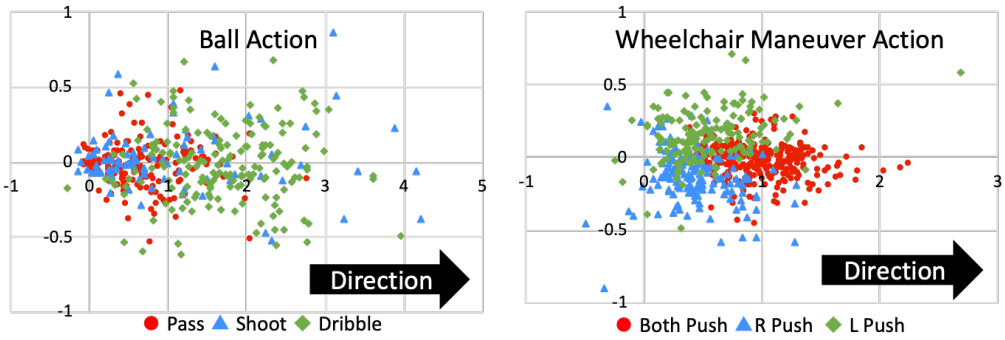

Figure 6.4: Relationship between Action Labels and Movement Vector

is responsible for capturing rapidly changing motion, by operating at fast refreshing speed and high temporal resolution.

For model training, the data were divided into training, validation, and test. For the actions of handling the ball, data of one of eight games was divided for the test and the others for training and validation, and for the wheelchair maneuver actions, data of three out of 14 players were divided for the test and the others for training and validation. Furthermore, the training and validation data were divided so that the number of training data and validation data was 9 : 1 for each label, and model training was performed. The details of data division are shown in Table 6.2 The progress of learning is shown in Figure 6.5. The average error rate in the training and validation data continued to gradually decrease to about 100 epochs, but after that, the model fitting is progressed only to the training data, and the decrease in the error rate of validation data has stopped. Finally, the error rate of validation data converged at about 30%. We changed the way to choose batch and trained with 100 epochs, but the average error rate in the validation data did not increase or decrease. Therefore, in the training after 100 epochs, there is no big change in the ranking of the confidence score for each action label, but it is presumed that the training progressed so that the difference in confidence score between the predicted label and the other labels became large.

6.3.3 Action Classification

Finally, we classify action in video using the result of localization and gesture recognition. From the result of the localization, we obtain four values which are the X and Y coordinate of the position where the action was performed, displacement during the action, and the change in orientation during the action. In gesture recognition, we get a confidence score for each class. In this chapter, there are 6 types of action labels, so 6 values are outputted from *Slowfast* model. Therefore, we input a total of 10 variables to the classifier. Because the output of gesture recognition is required for the training of the action classifier, the gesture recognition output is obtained by applying the training and validation

Table 6.2: Details of Training and Validation Data

Label	Train		Val		Test	
	Clips	Time	Clips	Time	Clips	Time
Pass	145	136.3	17	14.8	17	17.2
Shoot	81	112.4	9	12.5	9	12.7
Dribble	155	162.3	18	18.3	25	27.8
Both Push	248	138.0	28	15.3	90	56.3
Left Push	140	119.1	16	13.4	44	35.3
Right Push	162	141.1	19	16.0	45	40.1
Total	931	809.0	107	90.3	230	189.3

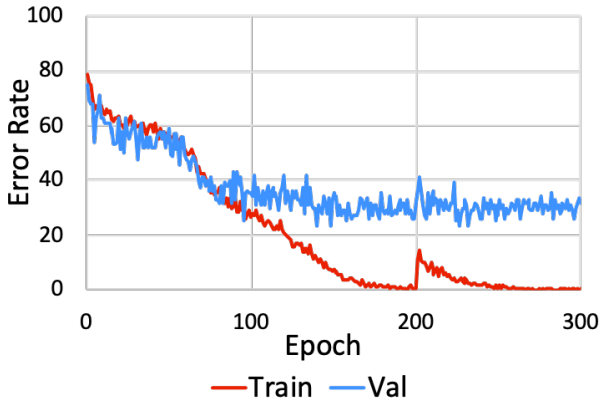

Figure 6.5: Training Progress of Gesture Recognition Model

data among the data divided in Section 6.3.2.

6.4 Evaluation

6.4.1 Evaluation Setup

For evaluation, the test data divided in Section 6.3.2 is used. In order to show the effect of our method more clearly, we compare our method with *Slowfast*-only. We prepared *Slowfast* models which is trained 100 and 200 epochs. We use 4 types of classifiers, which includes Random Forest (RF), Gaussian Process Classifier (GPC), Neural Network (NN), and Support Vector Machine (SVM). To determine the hyper-parameters in each classifier, we performed a grid search on the training and validation data and we used the hyper-parameters that achieved the best accuracy.

Table 6.3: Classification Accuracy by Model

Model		
Slowfast_epoch	Classifier	Accuracy
Slowfast_100	-	68.3
	RF	70.0
	GPC	76.1
	NN	71.7
	SVM	76.5
Slowfast_200	-	69.1
	RF	74.3
	GPC	78.7
	NN	77.4
	SVM	72.2

6.4.2 Comparison Results of Action Classification Models

Table 6.3 shows the accuracy of action classification by each model. From this result, it can be seen that even if it use any classifier, our method can classify actions with higher accuracy than *Slowfast*-only model by setting appropriate parameters. The highest accuracy in *Slowfast*-only was 69.1 % when using a model trained 200 epochs, but the highest accuracy in our method was 78.7 %, and the accuracy improved 9.6 points. In addition, by combining a classifier with a model trained 100 epochs, we obtained better results than with a *Slowfast*-only model trained 200 epochs. From these result, it can be seen that the position, displacement, and change in orientation of movement of players are important factors in classifying sports actions.

6.4.3 Classification Accuracy for Each Action Label

Next, we evaluate the classification accuracy for each action label. In our model, we use Slowfast_200 and GPC, which are combinations that achieve the highest accuracy for the gesture recognition model and the classification model. Tables 6.4 and 6.5 show the results of Slowfast_200 and our model (Slowfast_200 + GPC), respectively. From these results, it can be seen that both the recall and the precision have increased for many action labels in our method. In particular, in the model with only gesture recognition Slowfast_200, there are frequently occurred that pushing by both hands and by the right hand are misclassified pushing by the left hand. Also, pushing by the left hand is misclassified as another type of pushing. In contrast, our model achieved to significantly reduce the results of misclassification as pushing by the left hand. On the other hand, those that were correctly classified as pushing by the left hand were also classified into several different labels, and as a result, the number that could be correctly classified was reduced by 4, but the number of incorrect classification results could be reduced by 22. Therefore, the recall increased. Although there was no increase or decrease in the number of correct answers for dribbling and pushing by both hands, the recall increased because

Table 6.4: Classification Result of Slowfast_200

		Predicted						
		Pass	Shoot	Dribble	Both Push	Right Push	Left Push	Recall
Ground Truth	Pass	16	1	0	0	0	0	94.1
	Shoot	3	6	0	0	0	0	66.7
	Dribble	2	0	22	0	1	0	**88.0**
	Both Push	0	1	0	73	1	15	**81.1**
	Right Push	1	0	2	5	10	26	22.7
	Left Push	0	2	0	4	7	32	**71.1**
	Precision	72.7	**60.0**	91.7	89.0	52.6	43.8	

Table 6.5: Classification Result of Our Model (Slowfast_200 + GPC)

		Predicted						
		Pass	Shoot	Dribble	Both Push	Right Push	Left Push	Recall
Ground Truth	Pass	17	0	0	0	0	0	**100.0**
	Shoot	2	7	0	0	0	0	**77.8**
	Dribble	2	0	22	0	0	1	**88.0**
	Both Push	0	4	0	73	5	8	**81.1**
	Right Push	0	0	1	3	34	6	**77.3**
	Left Push	0	3	0	3	11	28	62.2
	Precision	**81.0**	50.0	**95.7**	**92.4**	**68.0**	**65.1**	

the number of misclassified results decreased. The reason why the number of correct answers did not change is that dribbling and pushing by both hands are performed evenly in various places on the court, there is little change in orientation, and the action is performed at both low speed and high speed, so positional information such as position and displacement is not very important in these actions. This leads the weight of the classification model is concentrated on the confidence score that is the output of gesture recognition and our model achieved similar results with the *Slowfast*-only.

6.5 Conclusion

In this chapter, we propose action recognition method using a single camera. First, we localize players in input video, which purpose is to extract global feature. Next, in parallel with localization, human gesture recognition is performed, which purpose is to extract local feature. Finally, these outputs are input to the classifier to determine an action label. In this way, we realized the action recognition model considering the position information.

For evaluation, we collected data that is different from the data used for model training and validation, and we then compared our method with the *Slowfast*-only model. As a result, our method

using location information achieved better accuracy than the *Slowfast*-only model in all four types of classifiers. Using the configs that achieved the best accuracy, the accuracy reached 78.7 %, which was 9.6 points higher than the best accuracy using the *Slowfast*-only model. From these result, it can be seen that the players' positions are important in the action recognition task in sports.

One topic for future work is to apply our method to other sports. We believe our method can apply not only to wheelchair basketball but also to sports such as basketball and soccer where a game field size is fixed. Because, in this case, it is considered that the position and orientation of the players are closely related to the action decision. Therefore, we would like to collect data on other sports and show the versatility of our method. In addition, the design of the model that recognizes the gesture used in our method is optimized for daily action. However, actions in sports are different from daily actions, and large and sharp changes occur in a shorter period. Therefore, we would like to evaluate the accuracy when the setting parameters of the model are changed and create a model that is more specialized for sports movements.

Chapter 7

Discussion

7.1 Sensor Selection

We selected inertial sensors and a single camera for localization and action recognition. However, we may also be able to design other approaches using other sensors as shown in Tables 7.1 and 7.2. For example, we can estimate positions of players by deploying a LiDAR (Light Detection And Ranging) sensor. Because light has high straightness, the distance to the object has little effect on the accuracy. However, with a single LiDAR sensor, occlusion becomes a problem. The approach using a single camera also has the same problem. In the image-based approaches, we can recognize a person from the color information. On the other hand, the raw data obtained by the LiDAR sensor is the depth information, which requires person identification from the shape and the movement. Furthermore, since players in many sports take various poses, it is difficult to recognize a person or a person's arms or legs by using LiDAR sensors. In contrast, color information obtained by cameras is more suitable for recognizing people because we can use important features of a person such as their clothes and exposed skin.

Another approach is sensing based on radio signals. The distance and direction to the transmitter can be estimated from the arrival direction, arrival time, and intensity of the received radio signal at the receiver. Therefore, by attaching a single device to the athlete and attaching multiple devices outside the court, we can estimate the athlete's position by triangulation. However, since radio waves are reflected, refracted, and attenuated by people and walls, there is concern about vulnerability to the dynamic environment. In addition, since only the position of the attached part is estimated, we need to attach multiple devices to the arm and foot of the person, for example. Refs. [97, 98] proposed device-free methods for people identification based on radio signals. However, because there are restrictions on the distance between the person and the sensor and the orientation of the person, it is difficult to recognize the action of the person on the court.

In this way, we can design approaches using light or radio signals for localization although they are not suitable for recognition of the movements of the arms and legs. On the other hand, when an

inertial sensor is used, unlike radio signals and light, data can be collected simply by attaching the sensor to the athlete without deploying sensors to the environment. Accumulation of errors is a major problem, however, movements in a short period are less affected by this problem. Therefore, inertial sensor-based approaches are suitable for action recognition. In addition, we can directly obtain data such as acceleration and angular velocity that indicate motion performance. When using videos, there is no additional device because data can be collected from the camera of a smartphone. It is also more suitable for recognizing people and limbs than other device-free approaches. For the above reasons, we selected the inertial sensor and the camera for our design.

Table 7.1: Approaches by Attaching Devices to Players

Sensor Type	Advantage	Disadvantage
Inertial Sensors	•Precise movement can be obtained •Directly acquire data such as acceleration and angular velocity that indicate the performance of actions	•Accumulation of error •Sensing only the position of the attached part
LiDAR	•High accuracy	•Occlusion •Large device size •High power consumption
Radio Signals	•Accuracy depends on radio frequency •Penetrate obstacles	•Interference •Sensing only the position of the attached part •Requires multiple devices around the court

Table 7.2: Device-free Approaches

Sensor Type	Advantage	Disadvantage
LiDAR	•High accuracy	•Occlusion •Hard to recognize limbs from depth information
Radio Signals	•Use of widespread devices	•Affected by reflection, refraction, and attenuation •Low accuracy
RGB Camera	•Use of widespread devices •Color information is suitable for limbs recognition	•Occlusion

7.2 Versatility

The algorithms proposed in this dissertation have the potential to be applied to other sports and environments.

First, the localization algorithm using the inertial sensor in Chapter 3 can be used for two-wheel-drive robots and vehicles. In wheelchair sports, errors tend to increase due to movements such as wheel slippage and floating one wheel. On the other hand, if it is a low-speed moving robot or a normal wheelchair, it is expected that the error becomes smaller. Even if there are three or more wheels but two drive wheels, the algorithm of our method can be applied by considering the vehicle design such as the size of the wheels. However, it cannot be applied when wheels are not used for movement.

Next, since the localization method using the camera proposed in Chapter 4 considers various poses during exercise, it can also be used in daily life. However, since this method assumes the people in the image to be at the same height, it cannot be used in places where the height of the ground plane is not constant, such as stairs and slopes. Additionally, the close-contact detection algorithm assumes an environment where there are no large obstacles other than people, such as sports schools and gyms. If the tracking of a person in the video is interrupted, we assumed that the person is hidden by another person. Therefore, it is difficult to apply our algorithm in an environment where a person is hidden by a large obstacle. When monitoring the occurrence of close-contact in such an environment, it is necessary to use multiple cameras to avoid occlusion.

Next, when recognizing the action with the inertial sensor, it is necessary to determine the mounting location of the sensor depending on what kind of actions are needed for analysis. In Chapter 5, inertial sensors are installed on the axes of the left and right wheels to recognize wheelchair maneuvers and turns. Therefore, in wheelchair sports, it is possible to recognize maneuvers and turns using our algorithm. However, it is difficult to recognize other actions such as shoots and passes that are independent of wheel movements with the same sensor settings. When recognizing a shoot or a pass, it is necessary to additionally attach inertial sensors to the player's arms and/or a ball. In this dissertation, we focused on action recognition of wheelchair maneuvers and turns in wheelchair basketball. Because the sensor equipment and algorithm were designed for this purpose, our method cannot be applied to other sports or other actions as it is. However, we believe that we showed important ideas for designing an action recognition system using inertial sensors in the future. For example, we showed that it is necessary to attach a sensor to a position that acts in common with actions of interest. Additionally, we need to consider individual-dependent actions in designing an action recognition method. In our maneuver recognition method, we regard peaks and valleys of the angular velocity as a common feature and sharpness and amount of its change an individual-dependent feature. We believe that this idea can be applied to behavior recognition tasks in other sports and daily life.

Finally, in the action recognition using the camera proposed in Chapter 6, we have designed a

method that considers not only the movement of the limbs of the person but also the movement of the athlete in the court. Therefore, it can be applied to other sports such as soccer and basketball where a play area is regulated in the rule. In addition, if the position does not significantly affect the judgment of actions, we cannot expect the improvement of the accuracy even if the position information is considered. In this dissertation, we worked on action recognition in wheelchair basketball. As a result, we have found that actions such as dribbling and pushing wheels with both hands were executed regardless of the players' positions on the court. Even if position information is added, there is not a large improvement in the accuracy of these actions compared to other actions. We believe that our method can be applied to actions other than sports. For example, by replacing a goal, which is the object of action in sports, with a television and a computer, it will be possible to classify whether the person is watching television or a computer's display.

7.3 Recommendations for Data Collection

In this dissertation, we have proposed methods that collect position and action data based on two types of devices (i.e., inertial sensors and a single camera). In order to collect position data, localization using a camera that can estimate absolute positions is more appropriate. When using inertial sensors, it is necessary to correct absolute positions in addition to the inertial sensors due to the accumulated error. In order to collect action data, it is necessary to install the device at an appropriate position, however, the approach using an inertial sensor can classify actions with higher accuracy. In addition, it is suitable for the analysis of actions because it can directly obtain the movement such as acceleration and speed. When using a camera for action recognition, the accuracy was not as good as the inertial sensors. However, since it can capture the entire body of a person, it is possible to recognize many types of actions.

For the above reasons, we recommend using a single camera for the initial data collection. This approach does not require additional equipment because it uses widespread devices, such as smartphone cameras. Its workload is also low because we can start data collection only by deploying it around the court. However, if the accuracy of action recognition by a camera is not sufficient, we need to use inertial sensors because they are not affected by occlusion and can capture precise movements directly. Inertial sensors are also required when analyzing each action precisely. There are many uses by fusion of two types of sensors. For example, the performance of localization and action recognition will be further improved. Sensor fusion is also helpful for person identification by matching trajectories in videos and short-term trajectories estimated by inertial sensors.

Chapter 8

Conclusion

In this dissertation, we introduced a method for collecting data easily, which is specialized for the sports environment. The main purpose of our research is to design data collection methods which can be used with low workload for daily practices and games even when the number of people and/or time are limited. To mitigate the effort for deployment, designed single modal methods which use a single type of sensor(s). In the proposed methods, we can start to collect data simply by attaching or deploying devices to the player's body, vehicle, or outside the court. Another advantage is ease of device collection after data collection, which is important for reducing workload. In order to collect data, it is effective to attach the inertial sensor to the place where we want to analyze the motion. Therefore, we proposed a data collection method using an inertial sensor. Also, in sports where hard contact occurs, we may need to avoid attachment of inertial sensors because of the risk of injury. Therefore, we also proposed a method to collect data only from videos. Therefore, we proposed data collection methods using inertial sensors or a camera that can be applied to different environments. Another purpose is to collect the data for the actual analysis using inertial sensors or a camera. Therefore, we proposed localization methods to obtain players' position data which is the basis of data analysis and action recognition methods to obtain the action data for further analysis. The main challenge is that each existing method does not consider the use in sports. Therefore, we proposed methods that consider features of sports such as positions of players in the court and pose variations. Through this dissertation, we have elaborated on four primary contributions as follows.

First, we focus on wheelchair sports and proposed a localization method by dead reckoning using three inertial sensors attached to wheelchair wheels axles and under the chair. We also proposed three types of correction methods to handle the accumulation localization error. From the result, we confirmed how localization errors accumulate in wheelchair sports and the effect of position correction frequency on position estimation accuracy.

Secondly, we proposed a localization method that is robust to the pose during movement using a single camera. We correct a human position using the skeleton of the lower body for various poses when the target does not move. As a result, we successfully decreased the error by an average of 23

cm and mitigated the increase of the error by occlusion and pose variations.

Thirdly, we proposed a maneuver action recognition method using inertial sensors for wheelchair sports. In our method, we firstly segments the time series of the angular velocity to extract candidate periods of actions without any fixed window size. We then classify actions for each segment. Furthermore, by comparing the movements of the left and right wheels, wheelchair movement recognition such as sprints and turns is performed. As a result, our method achieves an F-measure of 88.1% for classification of maneuver actions and an F-measure of 99.7% for the classification of turns.

Fourthly, we proposed an action recognition method for sports. In existing action recognition methods, they focus on actions without considering positions of players in the field. Therefore, we proposed an action recognition method considering player's position. As a result, the accuracy of our method reached 78.7 %, which was 9.6 points higher than the best accuracy of the method without considering the positional data. This means players' positions are important in the action recognition task in sports.

Through these contributions, we have shown that it is possible to automatically collect sports data easily and with a low workload in various environments. Our study leaves potentials for further studies for improving the performance of localization and action recognition. For example, it is expected that the performance of localization and action recognition will be further improved by combining the method using an inertial sensor and a camera. Especially for action recognition, it is expected that the classification accuracy for handling ball actions will be improved by embedding an inertial sensor in the ball or detecting and tracking a ball from images to measure the motion of a ball. We believe that such an idea can be applied with tools other than balls that are used in sports. In addition, by collecting bio-metric data such as heart rate, it is possible to analyze differences in performance with respect to the degree of individual fatigue. This dissertation has established the foundation of a data collection and analysis system, which can be used even in environments with limits to the number of people in the team or time to use the facilities for sports.

Acknowledgement

I would like to express my sincerely appreciation to continued support of my supervisor Professor Hirozumi Yamaguchi of Osaka University through trials and tribulations of this Ph.D thesis. Again I express my heartiest gratitude to him for his encouragement and invaluable comments in preparing this thesis.

I am very grateful to Professor Masayuki Murata, Professor Takashi Watanabe, Professor Toru Hasegawa and Professor Morito Matsuoka of Osaka University for their invaluable comments and helpful suggestions concerning this thesis.

I am heartily grateful to Associate Professor Akira Uchiyama for the precious advices and technical discussions provided through out the research.

I would like to thank Professor Teruo Higashino of Kyoto Tachibana University for his valuable comments.

Thanks go to everyone of Yamaguchi laboratory for their feedback, encouragement and support.

Finally, I would like to thank my family and my friends for their help and understanding.

Bibliography

[1] Exasol, "Data analytics credited for victory at the fifa world cup 2014," https://www.exasol.com/resource/data-scientists-the-unsung-heroes-of-the-football-team-2, Accessed: Dec. 12, 2021.

[2] Sportie, "Ragubi nihon daihyo no daiyakusin ga hituzen datta riyu [the reason why the japanese national rugby team made great achievement]," https://sportie.com/2016/04/rugby-analytics, Accessed: Feb. 1, 2019. (in Japanese).

[3] L. Ciabattoni, G. Foresi, A. Monteriù, L. Pepa, D. P. Pagnotta, L. Spalazzi, and F. Verdini, "Real time indoor localization integrating a model based pedestrian dead reckoning on smartphone and ble beacons," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 1, pp. 1–12, 2019.

[4] M. Hachimoto, F. Oba, Y. Fujikawa, K. Imamaki, and T. Nishida, "Position estimation method for wheeled mobile robot by integrating laser navigation and dead reckoning systems," *Journal of the Robotics Society of Japan*, vol. 11, no. 7, pp. 96–106, 1993, (in Japanese).

[5] P. Khandelwal, A. Khandelwal, and S. Agarwal, "Using computer vision to enhance safety of workforce in manufacturing in a post covid world," 2020, arXiv preprint.

[6] D. Yang, E. Yurtsever, V. Renganathan, K. A. Redmill, and Ü. Özgüner, "A vision-based social distancing and critical density detection system for covid-19," 2020, arXiv preprint.

[7] M. Aghaei, M. Bustreo, Y. Wang, G. Bailo, P. Morerio, and A. Del Bue, "Single image human proxemics estimation for visual social distancing," in *Proceedings IEEE/CVF Winter Conference Applications of Computer Vision (WACV)*, 2021, pp. 2785–2795.

[8] L. Bai, C. Efstratiou, and C. S. Ang, "wesport: Utilising wrist-band sensing to detect player activities in basketball games," in *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. IEEE, 2016, pp. 1–6.

[9] D. Connaghan, P. Kelly, N. E. O'Connor, M. Gaffney, M. Walsh, and C. O'Mathuna, "Multi-sensor classification of tennis strokes," in *SENSORS, 2011 IEEE*. IEEE, 2011, pp. 1437–1440.

[10] E. E. Cust, A. J. Sweeting, K. Ball, and S. Robertson, "Classification of australian football kick types in-situation via ankle-mounted inertial measurement units," *Journal of Sports Sciences*, pp. 1–9, 2021.

[11] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[12] H. Duan, Y. Zhao, K. Chen, D. Shao, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," 2021, arXiv preprint.

[13] B. McNaughton, L. Chen, and E. Markus, " "dead reckoning," landmark learning, and the sense of direction: a neurophysiological and computational hypothesis," *Journal of cognitive neuroscience*, vol. 3, no. 2, pp. 190–202, 1991.

[14] K. Nguyen-Huu, K. Lee, and S.-W. Lee, "An indoor positioning system using pedestrian dead reckoning with wifi and map-matching aided," in *2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 2017, pp. 1–8.

[15] P. Tu, J. Li, H. Wang, K. Wanga, and Y. Yuan, "Epidemic contact tracing with campus wifi network and smartphone-based pedestrian dead reckoning," *IEEE Sensors Journal*, 2021.

[16] J. Li, M. Guo, and S. Li, "An indoor localization system by fusing smartphone inertial sensors and bluetooth low energy beacons," in *2017 2nd International Conference on Frontiers of Sensors Technologies (ICFST)*. IEEE, 2017, pp. 317–321.

[17] H. Xing, L. Shi, K. Tang, S. Guo, X. Hou, Y. Liu, H. Liu, and Y. Hu, "Robust rgb-d camera and imu fusion-based cooperative and relative close-range localization for multiple turtle-inspired amphibious spherical robots," *Journal of Bionic Engineering*, vol. 16, no. 3, pp. 442–454, 2019.

[18] P. Nazemzadeh, D. Fontanelli, D. Macii, and L. Palopoli, "Indoor localization of mobile robots through qr code detection and dead reckoning data fusion," *IEEE/ASME Transactions On Mechatronics*, vol. 22, no. 6, pp. 2588–2599, 2017.

[19] C. Wu, F. Zhang, B. Wang, and K. R. Liu, "mmtrack: Passive multi-person localization using commodity millimeter wave radio," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 2400–2409.

[20] R. Zhou, M. Hao, X. Lu, M. Tang, and Y. Fu, "Device-free localization based on csi fingerprints and deep neural networks," in *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 2018, pp. 1–9.

[21] M. Abbas, M. Elhamshary, H. Rizk, M. Torki, and M. Youssef, "Wideep: Wifi-based accurate and robust indoor localization system using deep learning," in *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom*. IEEE, 2019, pp. 1–10.

[22] J. Chen, D. Steinmetzer, J. Classen, E. Knightly, and M. Hollick, "Pseudo lateration: Millimeter-wave localization using a single rf chain," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2017, pp. 1–6.

[23] S. Sen, J. Lee, K.-H. Kim, and P. Congdon, "Avoiding multipath to revive inbuilding wifi localization," in *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, 2013, pp. 249–262.

[24] C. K. Seow and S. Y. Tan, "Non-line-of-sight localization in multipath environments," *IEEE Transactions on Mobile Computing*, vol. 7, no. 5, pp. 647–660, 2008.

[25] Y. Jiang, J. Yang, P. Li, H. Si, X. Fu, and Q. Liu, "High energy lidar source for long distance, high resolution range imaging," *Microwave and Optical Technology Letters*, vol. 62, no. 12, pp. 3655–3661, 2020.

[26] M. Hasan, J. Hanawa, R. Goto, H. Fukuda, Y. Kuno, and Y. Kobayashi, "Tracking people using ankle-level 2d lidar for gait analysis," in *International Conference on Applied Human Factors and Ergonomics*. Springer, 2020, pp. 40–46.

[27] Y.-T. Wang, C.-C. Peng, A. A. Ravankar, and A. Ravankar, "A single lidar-based feature fusion indoor localization algorithm," *Sensors*, vol. 18, no. 4, p. 1294, 2018.

[28] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, arXiv preprint.

[29] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020, arXiv preprint.

[30] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," 2021, arXiv preprint.

[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[32] Y. Raaj, H. Idrees, G. Hidalgo, and Y. Sheikh, "Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields," in *Proceedings IEEE Conference Computer Vision and Pattern Recognition*, 2019, pp. 4620–4628.

[33] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[34] A. Bulat, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "Toward fast and accurate human pose estimation via soft-gated skip connections," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 8–15.

[35] Y. Li, H. Zhao, X. Qi, L. Wang, Z. Li, J. Sun, and J. Jia, "Fully convolutional networks for panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 214–223.

[36] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[37] Z. Li, W. Wang, E. Xie, Z. Yu, A. Anandkumar, J. M. Alvarez, T. Lu, and P. Luo, "Panoptic segformer," 2021, arXiv preprint.

[38] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3464–3468.

[39] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[40] N. Mahmoudi, S. M. Ahadi, and M. Rahmati, "Multi-target tracking using cnn-based features: Cnnmtt," *Multimedia Tools and Applications*, vol. 78, no. 6, pp. 7077–7096, 2019.

[41] H. Sheng, Y. Zhang, J. Chen, Z. Xiong, and J. Zhang, "Heterogeneous association graph fusion for target association in multiple object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 11, pp. 3269–3280, 2018.

[42] G. F. Welch, "Kalman filter," *Computer Vision: A Reference Guide*, pp. 1–3, 2020.

[43] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[44] M. Šarić, H. Dujmić, V. Papić, and N. Rožić, "Player number localization and recognition in soccer video using hsv color space and internal contours," in *International Conference on Signal and Image Processing (ICSIP 2008)*, 2008.

[45] L. Ballan, M. Bertini, A. D. Bimbo, and W. Nunziati, "Soccer players identification based on visual local features," in *ACM International Conference on Image and Video Retrieval*. ACM, 2007, pp. 258–265.

[46] W.-L. Lu, J.-A. Ting, K. P. Murphy, and J. J. Little, "Identifying players in broadcast sports videos using conditional random fields," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 3249–3256.

[47] M. Kytö, M. Nuutinen, and P. Oittinen, "Method for measuring stereo camera depth accuracy based on stereoscopic vision," in *Three-Dimensional Imaging, Interaction, and Measurement*, vol. 7864. International Society for Optics and Photonics, 2011, p. 78640I.

[48] D. Capel and A. Zisserman, "Computer vision applied to super resolution," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 75–86, 2003.

[49] M. Rezaei and M. Azarmi, "Deepsocial: Social distancing monitoring and infection risk assessment in covid-19 pandemic," *Applied Sciences*, vol. 10, no. 21, p. 7514, 2020.

[50] R. Keniya and N. Mehendale, "Real-time social distancing detector using socialdistancingnet-19 deep learning network," 2020, SSRN preprint.

[51] I. H. Lopez-Nava and A. Muñoz-Meléndez, "Human action recognition based on low-and high-level data from wearable inertial sensors," *International Journal of Distributed Sensor Networks*, vol. 15, no. 12, p. 1550147719894532, 2019.

[52] C. Shen, Y. Chen, and G. Yang, "On motion-sensor behavior analysis for human-activity recognition via smartphones," in *2016 Ieee International Conference on Identity, Security and Behavior Analysis (Isba)*. IEEE, 2016, pp. 1–6.

[53] S. Dernbach, B. Das, N. C. Krishnan, B. L. Thomas, and D. J. Cook, "Simple and complex activity recognition through smart phones," in *2012 eighth international conference on intelligent environments*. IEEE, 2012, pp. 214–221.

[54] T. T. Ngo, Y. Makihara, H. Nagahara, Y. Mukaigawa, and Y. Yagi, "Similar gait action recognition using an inertial sensor," *Pattern Recognition*, vol. 48, no. 4, pp. 1289–1301, 2015.

[55] C. Lian, R. Ma, X. Wang, Y. Zhao, H. Peng, T. Yang, M. Zhang, W. Zhang, X. Sha, Z. Wang, *et al.*, "Ann enhanced iot wristband for recognition of player identity, and shot types based on basketball shooting motion analysis," *IEEE Sensors Journal*, 2021.

[56] G. Brunner, D. Melnyk, B. Sigfússon, and R. Wattenhofer, "Swimming style recognition and lap counting using a smartwatch and deep learning," in *Proceedings of the 23rd International Symposium on Wearable Computers*, 2019, pp. 23–31.

[57] F. Haider, F. A. Salim, D. B. Postma, R. Van Delden, D. Reidsma, B.-J. van Beijnum, and S. Luz, "A super-bagging method for volleyball action recognition using wearable sensors," *Multimodal Technologies and Interaction*, vol. 4, no. 2, p. 33, 2020.

[58] J. Fan, S. Bi, G. Wang, L. Zhang, and S. Sun, "Sensor fusion basketball shooting posture recognition system based on cnn," *Journal of Sensors*, vol. 2021, 2021.

[59] N. Shahar, N. Ghazali, M. As'ari, and T. Swee, "Wearable inertial sensor for human activity recognition in field hockey: Influence of sensor combination and sensor location," *Journal of Physics: Conference Series*, vol. 1529, no. 2, p. 022015, 2020.

[60] Y. Gao and G. Ma, "Human motion recognition based on multimodal characteristics of learning quality in football scene," *Mathematical Problems in Engineering*, vol. 2021, 2021.

[61] Y. C. Vanlandewijck, C. Evaggelinou, D. J. Daly, J. Verellen, S. Van Houtte, V. Aspeslagh, R. Hendrickx, T. Piessens, and B. Zwakhoven, "The relationship between functional potential and field performance in elite female wheelchair basketball players," *Journal of Sports Sciences*, vol. 22, no. 7, pp. 668–675, 2004.

[62] C. De Lira, R. Vancini, F. Minozzo, B. Sousa, J. Dubas, M. Andrade, L. Steinberg, and A. Da Silva, "Relationship between aerobic and anaerobic parameters and functional classification in wheelchair basketball players," *Scandinavian Journal of Medicine & Science in Sports*, vol. 20, no. 4, pp. 638–643, 2010.

[63] R. van der Slikke, M. Berger, and D. Bregman, "Wheelchair mobility performance only supports the use of two classes in wheelchair basketball," *ISBS Proceedings Archive*, vol. 35, no. 1, p. 254, 2017.

[64] H. M. Logan-Sprenger and L. R. Mc Naughton, "Characterizing thermoregulatory demands of female wheelchair basketball players during competition," *Research in sports medicine*, pp. 1–12, 2019.

[65] K. Hollander, S. Kluge, F. Glöer, H. Riepenhof, A. Zech, and A. Junge, "Epidemiology of injuries during the wheelchair basketball world championships 2018: A prospective cohort study," *Scandinavian journal of medicine & science in sports*, vol. 30, no. 1, pp. 199–207, 2020.

[66] R. M. Van Der Slikke, A. M. De Witte, M. A. Berger, D. J. Bregman, and D. J. H. Veeger, "Wheelchair mobility performance enhancement by changing wheelchair properties: What is the effect of grip, seat height, and mass?" *International journal of sports physiology and performance*, vol. 13, no. 8, pp. 1050–1058, 2018.

[67] A. M. de Witte, F. S. Sjaarda, J. Helleman, M. A. Berger, L. H. Van Der Woude, and M. J. Hoozemans, "Sensitivity to change of the field-based wheelchair mobility performance test in wheelchair basketball," *Journal of rehabilitation medicine*, vol. 50, no. 6, pp. 556–562, 2018.

[68] B. S. Mason, M. Lemstra, L. H. van der Woude, R. Vegter, and V. L. Goosey-Tolfrey, "Influence of wheel configuration on wheelchair basketball performance: Wheel stiffness, tyre type and tyre orientation," *Medical engineering & physics*, vol. 37, no. 4, pp. 392–399, 2015.

[69] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6047–6056.

[70] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," 2018, arXiv preprint.

[71] P. W. Dempsey, M. E. Allison, S. Akkaraju, C. C. Goodnow, and D. T. Fearon, "C3d of complement as a molecular adjuvant: bridging innate and acquired immunity," *Science*, vol. 271, no. 5247, pp. 348–350, 1996.

[72] G. Chéron, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features for action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3218–3226.

[73] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Improved multiscale vision transformers for classification and detection," 2021, arXiv preprint.

[74] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[75] X. Song, C. Lan, W. Zeng, J. Xing, X. Sun, and J. Yang, "Temporal-spatial mapping for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 3, pp. 748–759, 2019.

[76] MARY ANN LIEBERT, Inc., "How is big data impacting sports analytics?" https://www.eurekalert.org/pub_releases/2018-12/mali-hib122018.php, Accessed: Feb. 1, 2019.

[77] K. Urano and T. Sekiguchi, "Nihon saka-kai ha deta bunseki de tuyokunaru [japanese national football team will become stronger with data analysis]," https://markezine.jp/article/detail/28169, Accessed: Dec. 12, 2019. (in Japanese).

[78] R. Van Der Slikke, M. Berger, D. Bregman, A. Lagerberg, and H. Veeger, "Opportunities for measuring wheelchair kinematics in match settings; reliability of a three inertial sensor configuration," *Journal of Biomechanics*, vol. 48, no. 12, pp. 3398–3405, 2015.

[79] Cabinet office, "Overview of the quasi-zenith satellite system (qzss)," http://qzss.go.jp/index.html, Accessed: Dec. 12, 2021.

[80] K. Kato, "Sakka ni okeru deta bunseki to timu kyouka [data analysis and team training in football]," *IEICE Communications Society Magazine*, vol. 10, no. 1, pp. 29–34, 2016, (in Japanese).

[81] Qoncept, Inc., "Tracking solution," https://qoncept.co.jp, Accessed: Dec. 12, 2021.

[82] R. Fan, S. Lam, E. Lin, O. Artemenko, Y. Lu, and M. Gerla, "Localizing a wheelchair indoors with magnetic sensors," in *Annual Mediterranean Ad Hoc Networking Workshop*, 2013, pp. 141–147.

[83] Ministry of Environment and Ministry of Health, Labor and Welfare in Japan, "Fiscal year 2020 heatstroke prevention actions," https://www.otit.go.jp/files/user/docs/200615-5.pdf, Accessed: Oct. 21, 2021.

[84] V. Vlacha, G. Feketea, A. Petropoulou, and S. D. Trancá, "The significance of duration of exposure and circulation of fresh air in sars-cov-2 transmission among healthcare workers," *Frontiers in Medicine*, vol. 8, 2021.

[85] Ministry of Environment and Ministry of Health, Labor and Welfare in Japan, "Example of practicing new lifestyle," https://www.mhlw.go.jp/content/10900000/000632485.pdf, Accessed: Oct. 21, 2021.

[86] R. Hasegawa, A. Uchiyama, F. Okura, D. Muramatsu, I. Ogasawara, H. Takahata, K. Nakata, and T. Higashino, "Human localization using a single camera towards social distance monitoring during sports," in *Proceedings EAI International Conference Mobile and Ubiquitous Systems*, 2021.

[87] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[88] D. Fu, D. Chen, J. Bao, H. Yang, L. Yuan, L. Zhang, H. Li, and D. Chen, "Unsupervised pre-training for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 750–14 759.

[89] G. Wang, J. Lai, P. Huang, and X. Xie, "Spatial-temporal person re-identification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8933–8940.

[90] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on computer vision*. Springer, 2016, pp. 17–35.

[91] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 748–756.

[92] L. R. Rabiner, J. H. McClellan, and T. W. Parks, "Fir digital filter design techniques using weighted chebyshev approximation," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 595–610, 1975.

[93] MathWorks, Inc., "Prominence," https://www.mathworks.com/help/signal/ug/prominence.html, Accessed: Sep. 23, 2020.

[94] N. E. N. Rodríguez, *Advanced Mechanics in Robotic Systems*. Springer Science & Business Media, 2011.

[95] International Wheelchair Basketball Federation, "2018 official wheelchair basketball rules," https://iwbf.org/wp-content/uploads/2019/03/2018_IWBF_rules-Ver-2_Final.pdf, Accessed: Sep. 23, 2020.

[96] G. M. Arastey, "Opta sports: the leading sports data provider — sport performance analysis," https://www.sportperformanceanalysis.com/article/opta-leading-sport-data-provider, Accessed: Dec. 6, 2021.

[97] Z. Hao, Y. Duan, X. Dang, Y. Liu, and D. Zhang, "Wi-sl: Contactless fine-grained gesture recognition uses channel state information," *Sensors*, vol. 20, no. 14, p. 4025, 2020.

[98] H. F. T. Ahmed, H. Ahmad, and C. Aravind, "Device free human gesture recognition using wi-fi csi: A survey," *Engineering Applications of Artificial Intelligence*, vol. 87, p. 103281, 2020.