



Commonality of Motions Having Effective Features with Respect to Methods for Identifying the Moves Made during *Kumite* Sparring in Karate

Keiichi Sato¹, Hitoshi Matsubara² and Keiji Suzuki³

¹National Institute of Technology, Hakodate College, Japan

²Graduate School of Information Science and Technology, The University of Tokyo, Japan

³Faculty of Systems Information Science, Future University Hakodate, Japan


Keywords: Image Recognition, Deep Learning, Convolutional Neural Network.


Abstract: In recent years, active use of imaging technology, sensor technology, and AI (artificial intelligence) has been on the rise to identify various plays that occur in sports competition to assist judges, coaches, and athletes. However, no study result has been reported on this research topic as it pertains to karate sparring competition. The lack of past studies on this topic is attributable to the fact that no viable method has been developed to acquire motion data or to identify athletes' motions at a fundamental level in competition, as no sensors can be worn by athletes on their bodies, and there are a number of blind spots that occur in competition due to fast-paced exchanges that competing athletes engage in, among other factors. Therefore, in this study, footage of *kumite* (sparring) in the practice of karate simulating actual competitive matches was captured using video cameras to conduct a motion identification experiment using a CNN (convolutional neural network). After comparing the result of this study to that of previous motion identification experiments in which subjects wore sensors on their bodies, it has been determined that the motions having effective features are common between the two types of experiments.


1 INTRODUCTION

No published studies have been conducted on the *kumite* competition of karate. One of the main reasons is because no viable method to accurately sense and identify the various motions that occur in karate sparring contests has been developed to date. Another reason is because a method is yet to be invented that enables acquisition of appropriate motion data for motion identification in a manner that does not interfere with the contests that are taking place. These are the two main objects. Therefore, in this study, a video-camera-based motion data acquisition method is suggested that won't affect the karate sparring matches themselves, and also a deep-learning-powered method that will form a basis for identifying the various motions of karate sparring, to achieve the aforementioned two objectives. In related studies,

deep-learning-based play identification methods have been proposed, in which use of CNNs (convolutional neural networks) is the basic approach to achieving a certain level of precision, where the output from the CNNs is fed to other computer systems that are capable of performing the play identification tasks even at a more advanced level. Therefore, in this study, motion identification experiments using a CNN was conducted in order to achieve a certain level of identification precisions that is needed of the CNN that gets connected to a system that performs advanced identification of various motions. The study also compared its findings to the result of some previously conducted motion identification experiments that tested the subjects wearing sensors on their bodies, and also examined the efficacy of the methods it suggests.

^a <https://orcid.org/0000-0002-6712-4505>

^b <https://orcid.org/0000-0002-3104-3025>

^c <https://orcid.org/0000-0002-8599-3712>

2 RELATED STUDIES

When categorizing published studies conducted on methods for identifying the motions of humans in various sports and martial arts, they can be roughly divided into two groups, i.e, studies on methods that involve the subjects wearing sensors on their bodies(Kwon and Gross,2005;Kwon et al.,2008), and studies where the subjects don't wear any sensors. As it's impossible for the subjects to wear any sensors in the case of *kumite* competition in karate, this study mainly deals with published studies on methods that don't involve the subjects wearing any sensors. Such methods can be further broken down into subcategories based on the types of sensor technology they use, which are the method that uses RGB-D sensors, the method that employs LIDAR (light detection and ranging), and the method that utilizes RGB cameras.

Concerning the method that uses RGB-D sensors, one notable published study that was conducted using the Microsoft Kinect sensors to identify the motions of karate (Hachaj et al., 2015). However, this particular study only measured the basic kicking and defending motions of one subject from the front, and so should be deemed basically the same as a number of other research papers that only dealt with simple motions.

As for the studies that focused on the method employing LIDAR, one reports its application to a scoring system used in gymnastics (Sasaki, 2018;Tomimori,2020). More specifically, the study improved LIDAR technology so it could be applied to 3D laser sensors for use in gymnastics competition, which enabled precise measurement of human motions. This system developed in the study was able to use machine learning to recognize a skeletal model off of depth images and automatically scored each performance put on by a gymnast according to scoring criteria that utilized the skeletal model.

In regard to the studies conducted on the method using RGB cameras, in one published study, captured video data was divided up into still images, and the pose estimation library OpenPose was used to generate a human skeletal model from those still images. Then, data on the features of the skeletal model was fed to a neural network for learning, and then motion identification was performed (Nakai et al.,2018;Takasaki et al., 2019). In addition, there are other published studies that used deep learning to recognize various types of plays made in athletic competition, including a study done on tennis (Mora and Knottenbelt, 2017), a study done on ice hockey (Tra et al., 2017), a study done on volleyball (Ibrahim

et al.,2016) and a study done on football (Tsunoda et al., 2017). When using a method that utilizes deep learning to identify different plays that occur in athletic contests, it's important to achieve a certain level of precision using a CNN, where the output from the CNN is connected to a system capable of identifying various plays at a more advanced level.

The foregoing is a list of key published studies on motion identification that don't involve the subjects wearing any sensors on them. Meanwhile, when it comes to the research methods that involve use of RGB-D sensors and LIDAR or OpenPose, motion data is measured based on a human skeletal model. However, in the case of *kumite* contests in karate, the nodes (joints) of each human skeletal model being used to measure the subjects' motions might get lost with high frequency, which is a major issue. To address it, this study adopted a method that identified the subjects' various motions based on images captured in video data that were acquired with RGB cameras.

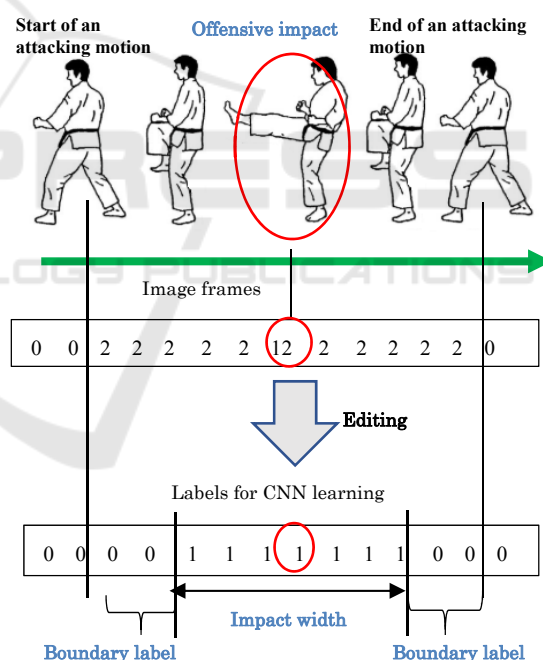


Figure 1: Base Labels Edited into Learning Labels for a CNN.

3 EXPERIMENT SYSTEM

3.1 Labelling and Suggested Method

Labels 0 through 14 and 99 as specified are assigned to the corresponding image frames extracted from the

captured video data. These label values are not ones to be fed to a CNN for learning but are created for the purpose of recording the subjects' motions, which are referred to as "base labels" in this study. As shown in the illustration of a front kick in Figure 1 below, and so on. The moment either subject's attacking limb lands on the opponent is referred to as an "offensive impact" in this study. After the base labels are created as described above, they are edited to suit the purpose of each experiment, and then the label values used for CNN learning are created.

3.2 Overview of the Experiment System

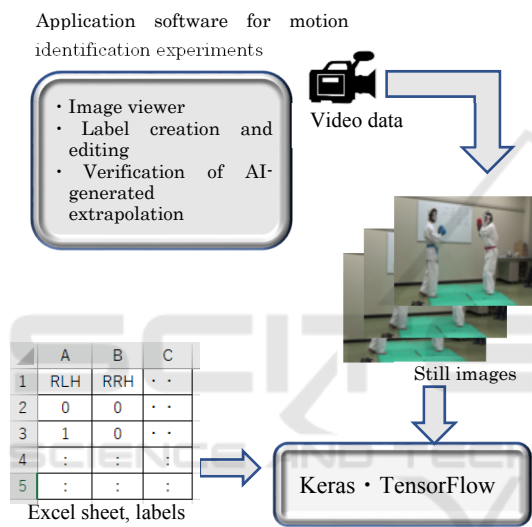


Figure 2: Outline of the Experiment System.

An overview of the experiment system used in this study is shown in Figure 2. Various karate motions are captured using a low-cost video camera for home use, after which the video data is divided into still images at a rate of 30 frames / sec. Then, the still images are assigned the base labels that match them as specified, using the application software developed by our laboratory for motion identification experiments.

Then, those base labels are converted into matching label values through editing for CNN learning in a format suitable for the purpose of each motion identification experiment as shown in Figure 1, and then the label values and the still images are fed to a CNN for learning. In terms of how the CNN was set up in this study, Keras, which has TensorFlow on the back end, was used. Keras is a library that allows creation of neural networks with relative ease. As for

the design of the CNN used in this experiment, it has a structure typically seen in image recognition, with the first part of the neural network being comprised of convolutional layers and pooling layers, with the latter part consisting of fully connected layers and dropout layers. As for the algorithm for updating the neural network weights, Adam was used in this study, while ReLU was used as the activation function.

4 RESULT

4.1 Punch-only *Yakusoku Kumite* (Pre-arranged Sparring)

To test the efficacy of the method suggested in this study, a motion identification experiment was conducted, in which the subjects engaged in *yakusoku kumite* (pre-arranged sparring) that only allowed punching. This experiment had two subjects, one of which had no experience with karate, and the other a skilled karate practitioner. In this experiment, a total of two video cameras were used to capture the subjects' images from the front, from stationary positions at a 45-degree angle from both sides. The two subjects then took turns throwing punches at each other, and traded their left-right positions from one round to the next, as viewed from the cameras. When any of the punches either subject threw so much as touched the opponent (i.e., "skin touch"), it was deemed effective. Meanwhile, if a punch thrown by either subject was met with a defensive technique of the opponent or if the opponent avoided the punch by distancing, it's deemed ineffective. In the experiment, a total of 1,600 attacking motions occurred, as broken down in formula (1) specified below.

$$\begin{aligned}
 & 20 \text{ reps} \times 5 \text{ sets} \\
 & \times 2 \text{ (left and right hands)} \\
 & \times 2 \text{ (effective/ineffective motions)} \\
 & \times 2 \text{ (positions)} \times 2 \text{ (subjects)} \\
 & = 1,600 \text{ attacking motions (1)}
 \end{aligned}$$

The test data considered in the experiment was sampled by taking one set's worth of data from each round that took place under different conditions as shown in formula (2) below, which resulted in a sample size of 20% of all data.

$$\begin{aligned}
 & 20 \text{ reps, 1 set} \times 16 \text{ rounds} \\
 & = 320 \text{ test motions (2)}
 \end{aligned}$$

The experiments as indicated in Tables 1 through 3 were all conducted with an epoch size of 1,500. In terms of what the label values specified in Tables 1

and 2 mean, numbers 1.1 through 1.6, 2.1, and 2.2 having a label value of 0 means it's not an attacking motion, and if any of their label values is 1, it means it's an attacking motion. So in these experiments, each motion is determined either as an attacking one or not, based on inference. If numbers 3.1 and 3.2 have a label value of 1, it's deemed an 'effective strike (hit the opponent)' while them having a label value of 2 means they are 'ineffective strikes (not hitting the opponent).' In these experiments, the subjects' attacking motions were identified in further detail.

Table 1: Accuracy Rate in Relation to Impact Widths.

No	No. of frames comprising impact width	Rate of accuracy of label value per image (%)	
		0	1
1.1	1	99.3	34.2
1.2	3	98.7	60.3
1.3	5	98.2	73.7
1.4	7	96.9	76.8
1.5	9	96.4	72.9
1.6	11	95.4	71.9

Table 2: Accuracy Rate in Relation to Boundary Label Values.

No	Boundary label value	Rate of accuracy of label value per image (%)			Rate of accuracy on attacking motions (%)
		0	1	2	
2.1	99	98.3	89.6		93.1
2.2	0	98.0	71.3		87.8
3.1	99	98.6	90.3	85.1	93.1
3.2	0	97.6	81.8	63.6	84.3

Table 3: Distance Until Offensive Impact.

No	Distance until offensive impact	
	Method 1	Method 2
2.1	0.2	1.7
2.2	0.5	1.6
3.1	0.2	1.6
3.2	0.4	1.6

While a label value representing each attacking motion was assigned based on its offensive impact in a manner that maintained front-back symmetry, such

width of frames as shown in Figure 1 is referred to as "impact width" in this study.

When the value of impact width was 7 in the experiment described in Table 1, the rate of accuracy on the motions assigned a label value of 1 was the highest. Therefore, the following experiments were all conducted with an impact width of 7.

As shown in Figure 1, the labels that are assigned to the image frames extending from the edges of the impact width to the start and end of each motion are referred to as "boundary labels" in this study. With this in mind, an experiment was conducted on motions having boundary label values of 0 (non-attacking motion) and 99 (outside the learning scope) as specified in Table 2. When motion data with a boundary label of 99 were compared to those with a boundary label value of 0 were compared, the accuracy rate on label value per image improved by 18% with the label value being 1 as specified in 2.1, and by 22% with the label value being 2 as specified in 3.2.

The accuracy rate was close to 100% on the motion data having a label value of 0 (non-attacking motion) in each experiment. This could be attributed to the fact that all attacking motions could be detected almost entirely within their impact width. When the extrapolation process occurring on the video data was checked using the viewer, the extrapolation result on each attacking motion made could be confirmed near the offensive impact, which is quantitatively expressed in the columns "Method 1" and "Method 2" under "Distance until offensive impact" in Table 3. Based on these data, it's possible to know when an extrapolation is made on each attacking motion in terms of how close it is to the offensive impact.

Method 1

Distance=

$$\left| \text{Avg. frame numbers accurately extrapolated by the CNN} - \text{frame numbers corresponding to the offensive impact} \right| \quad (3)$$

Method 2

$$\text{Distance} = \text{Avg. of } \left\{ \left| \text{frame numbers accurately extrapolated by the CNN} - \text{frame numbers corresponding to the offensive impact} \right| \right\} \quad (4)$$

Method 1 is used as a means by which to extrapolate the offensive impact using the CNN. It calculates the difference between the arithmetic mean value of the frame numbers that the CNN accurately

extrapolated within the impact width, and the frame numbers corresponding to the offensive impact. As the result of the extrapolation performed by Method 1 does not exceed 0.5 as shown in Table 3, it's discernible that it almost perfectly captured the frames of each offensive impact observed. However, there are some instances where the CNN-inferred frame numbers deviated from the intended offensive impact by large margins, on either side, while there are other instances where the calculated mean values happen to fall on the frame numbers that are in the centers of the offensive impacts observed. Meanwhile, in the case of Method 2, the mean values are calculated after calculating the distance between the subjects by each inferred frame number, the distance until offensive impact is accurately represented. From the result of the extrapolation made using Method 2 as specified in Table 3, it's discernible that Method 1 < Method 2. Based on this observed relation between the two methods, it's apparent that the CNN-inferred non-zero label values representing the impacts are distributed being centered on the offensive impacts.

Concerning the "rate of accuracy on attacking motions (%)" as specified in Table 2, each instance where the CNN's inference was accurate within the impact width is deemed as an accurate inference for each attacking motion concerned. This adjustment was made because there were many instances where the rate of accuracy of label value per image was low but the offensive impact was recognized correctly even if only one frame was inferred accurately.

This is also supported by the fact that the values indicating the distance until offensive impact have been small. The rates of accuracy on attacking motions for 2.1 through 3.2 specified in Table 2 were highly precise, ranging between 84.3% and 93.1%, indicating the efficacy of the method suggested in this study.

4.2 *Yakusoku Kumite* (Pre-arranged Sparring) Simulating Actual Competition

An experiment was conducted to identify the subjects' motions in *yakusoku kumite* (pre-arranged sparring) simulating actual competitive matches, in which both punches and kicks were allowed. The subjects' images were captured using four units of overhead cameras, at a height of 5.75 m above the floor. In this report, the motion identification experiment was conducted based on two cameras' worth of data, so that it could be compared to another experiment done on punch-only *kumite* matches, which was recorded

using two cameras also. Recording the video data from overhead positions eliminated the possibility of any blind spots occurring due to the main referee blocking the view, preventing all motions of the contestants to be seen clearly in official karate matches. Such camera positioning would also reduce the likelihood of contestants blocking the view to ensure accurate judging.

For this experiment, the parts of the subjects' bodies that would be involved in the moves they would make on each other and the order of the moves were decided in advance. The subjects were also told beforehand that they could throw punches in one-two combinations, and also kick both middle and high. The subjects were allowed to move freely inside the designated court, and could initiate their attacking motions at any time they would like. As per the Olympic rules, each effective attacking motion had to be either a light touch or a non-contact strike stopping several centimeters short of contact. Meanwhile, an attacking motion was deemed ineffective when it was successfully defended by the opponent by deflection, evasion, or distancing. As for the composition of the subjects, a total of six high school students all having a 1st-degree black belt took part in *kumite* matches in three pairs, the first pair being subjects A (red) and B (blue), the second being subjects C (red) and D (blue), and the third being subjects E (red) and F (blue). The motions of the subjects that were supposed to occur within one round of *kumite* are described in (a) and (b) below.

- (a) Left and right punches had to be thrown in that order. They could be thrown in series or in sporadic bursts. Such one-two punch combination had to be thrown for a total of 20 times, and subject 'red' had to go first, followed by subject 'blue' each time.
- (b) Kicks also had to be thrown by the subjects following the same rules applied to punches.

The total number of attacking motions executed by the subjects for both (a) and (b) = 160 / round.

While following the round-by-round rules described above, the subjects sparred for the numbers of rounds as specified below.

Effective attacking motions (non-contact): 10 rounds

Non-effective attacking motions (misses): 10 rounds

Total number of attacking motions attempted

$$= 160 / \text{round} \times 20 \text{ rounds}$$

$$= 3,200 \tag{5}$$

(no. of punches: 1,600; no. of kicks: 1,600)

Table 4: Rate of Accuracy on Attacking Motions.

No	Subject	Rate of accuracy of label value per image (%)			Rate of accuracy on attacking motions (%)
		0	1	2	
4.1	AB	97.9	66.8		94.2
4.2	CD	98.5	46.5		85.6
4.3	EF	95.0	70.6		92.6
5.1	AB	97.9	63.9	62.2	91.5
5.2	CD	99.1	35.2	40.7	77.7
5.3	EF	94.0	64.5	65.0	89.1

Each pair was assigned a total of six rounds, consisting of three rounds of effective attacking motions and three rounds of non-effective attacking motions, plus some extra work α depending on the pair.

Extra work α : The pair A and B was assigned one additional round of effective attacking motions while the pair C and D was assigned one additional round of non-effective attacking motions.

Table 5: Distance until Offensive Impact.

No	Distance Until Offensive Impact	
	Method 1	Method 2
4.1	0.7	1.9
4.2	1.0	1.9
4.3	0.7	1.9
5.1	0.8	1.9
5.2	1.1	1.8
5.3	0.8	1.9

Test data on a total of two rounds contested by each pair, consisting of one round of effective attacking motions and one round of non-effective attacking motions, were used. While a total of 20 rounds took place, the number of rounds that made up 10% of all rounds were used to extract test data from. The result of the test executed is shown in Table 4. The number of epochs was set at 500.

The rate of accuracy of label value per image on all motions of which the label value was 0 (non-attacking motions) for numbers 4.1 through 5.3 was fairly high, ranging between 94.0% and 99.1%. As for the distance until offensive impact as specified in Table 5, it turned out be less than two frames. This means that had either a label value of 1 or 2 centering on the offensive impact frames occurring within the impact width were detected, which was also confirmed during an observation that used the viewer.

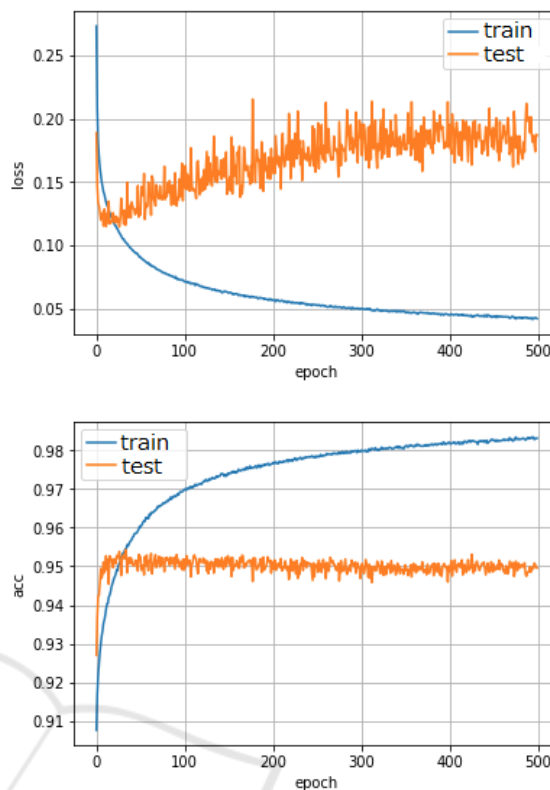


Figure 3: No 4.1, Losses During the Learning Process, and Motion Identification Accuracy.

Meanwhile, the rate of accuracy of label value per image for the label values 1 and 2 turned out to be slightly low. However, it must be noted that this rate of accuracy does not indicate the rate of accuracy on the attacking motions attempted. As far as the method suggested in this study is concerned, what's important is to accurately capture each offensive impact that occurs. Therefore, even in instances where only one image frame within any given impact width is accurately inferred, the motion identification is deemed to have functioned correctly on those attacking motions per se. Hence, in this experiment, if the CNN extrapolation is done correctly within the impact width of a motion, the extrapolation is deemed accurate on the motion, since its offensive impact is captured almost completely, as was the case with the punch-only *kumite*. The average rate of accuracy on attacking motions for numbers 4.1 through 5.3 turned out to be quite high at 88.5%.

Figure 3 contains graphs indicating the losses that occurred during the learning process of number 4.1 specified in Table 4, along with the corresponding motion identification accuracy. 10% of the data was unknown data not part of the training data, which was intended for testing use. The losses apparently

increased as the learning process progressed, while there is a sign of overtraining. Such patterns could also be discerned from the experiment rounds numbered 4.2 through 5.3.

5 DISCUSSION

In the experiment covering numbers 1.1 through 1.6 specified in Table 1, the rate of accuracy is the highest on the motions having a label value of 1 (attacking motions) when the impact width is 7. In addition, as evident in the results shown in Table 2, the rate of accuracy of label value per image improved by roughly 20% when the boundary label was 99. This might mean that there were motions with features that would prove effective for motion identification that exit within a certain range centering on the offensive impact, while there were those other motions immediately preceding and following the aforementioned range having features that lowers the motion identification accuracy. Phenomena similar to these were also encountered in a separate experiment conducted previously that used optical motion capture (Sato and Kuriyama, 2011). It's believed that the motions having effective and non-effective features might be the same as the motion identification that uses the joint position data generated with a human skeletal model, and also as the motion identification based on image data.

As the fairly high rate of accuracy of about 90% could be achieved on attacking motions as specified in Table 4, and given how accurately the offensive impacts that occurred in each experiment could be captured, it could be surmised from a comprehensive viewpoint that the desired level of CNN-assisted motion identification accuracy, which is the objective of this study, might have been achieved, which is necessary for connecting to a system capable of identifying various motions at a more advanced level.

6 CONCLUSION

As there has been no published study on basic methods for acquiring motion data and for identifying various motions in karate *kumite* competition, this study conducted CNN-assisted motion identification experiments that acquired data using overhead video cameras placed above the contestants so that they wouldn't interfere with the contests in progress. CNNs such as one used in this study are connected to advanced sports-specific extrapolation systems such

LSTM, on which there have been published studies focusing on other sports (Tsunoda et al., 2017). While each CNN to which the aforementioned connection is made must possess a certain level of motion identification accuracy, sufficient results have been achieved during the experiments conducted in this study.

Explained above is the basic method that this study suggests for effective data acquisition and motion identification applicable to karate *kumite* contests. In terms of what could be improved in the future, it will be necessary to check the efficacy of the method when the number of cameras is increased, and also to conduct experiments on the identification of motions that more closely resemble the contestants' actual movements in official competitive matches.

REFERENCES

- Hachaj, T., Marek R. O., and Katarzyna K. (2015). Application of Assistive Computer Vision Methods to Oyama Karate Techniques Recognition. *Symmetry*, 1670-1698.
- Ibrahim, M. S., Muralidharan, S., Deng, Z., Vahdat, A., and Mori, G. (2016). A Hierarchical Deep Temporal Model for Group Activity Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1971-1980.
- Kwon, D. Y., and Gross, M. (2005). Combining Body Sensors and Visual Sensors for Motion Training. *The 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*.
- Kwon, T., Cho, Y., Park, S. Il., and hin, S. Y. (2008). Two-Character Motion Analysis and Synthesis. *IEEE Transactions on Visualization and Computer Graphics*, 14(3), 707-720.
- Nakai, M., Tsunoda, Y., Hayashi, H., and Murakoshi, H. (2018). Prediction of Basketball Free Throw Shooting by OpenPose. *JSAI International Symposium on Artificial Intelligence*, 435-446.
- Mora, S. V., and Knottenbelt, W. J. (2017). Deep Learning for Domain-Specific Action Recognition in Tennis. *2017 IEEE Conference on CVPRW*, 170-178(online).
- Sasaki, K. (2018). 3D Sensing Technology for Real-Time Quantification of Athletes' Movements. *Fujitsu*, 13-20 in Japan.
- Sato, K., and Kuriyama, S. (2011). Classification of karate motion using feature learning. *2011 by Information Processing Society of Japan*, 75-80 in Japan.
- Takasaki, C., Takefusa, A., Nakada, H., and Oguchi, M. (2019). A Study on Action Recognition Method with Estimated Pose by using RNN. *2019 Information Processing Society of Japan in Japan*.
- Tomimori, H., Murakami, R., Sato, T., and Sasaki, K. (2020). A Judging Support System for Gymastics Using 3D Sensing. *Journal of the Robotics Society of Japan in Japan*.

- Tra, M. R., Chen, J., and Little, J. J. (2017). Classification of Puck Possession Events in Ice Hockey. *2017 IEEE Conference on CVPRW*, 147-154 (online).
- Tsunoda, T., Komori, Y., Matsugu, M., and Harada, T. (2017). Football Action Recognition using Hierarchical LSTM. *IEEE Conference on CVPRW*, 155-163.

