Automatic Scoring in Fencing by using Skeleton Points Extracted from Images

Takehiro Sawahata[a], Alessandro Moro[b], Sarthak Pathak[a], Kazunori Umeda[a]
[a]Department of Precision Engineering, Chuo University, Tokyo, Japan [b]RITECS, Tokyo, Japan

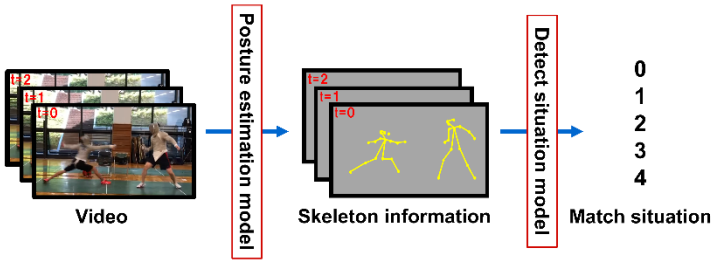


Figure 1. Flowchart of the proposed scoring model. The proposed system is divided into two steps: acquisition of skeleton point information and phrase detection (detecting the match situation).

ABSTRACT

First time spectators of fencing competitions cannot understand the complicated rules, making it difficult for them to enjoy the game. Therefore, in this paper, we propose a system that detects the situation of a fencing match using skeleton points extracted from videos. Players cannot be equipped with sensors or other devices to prevent interference with the match. Consequently, this research proposes a system that detects "phrases" using skeleton point information extracted from videos and displays the game situation. We evaluate actual videos of fencing to confirm the performance.

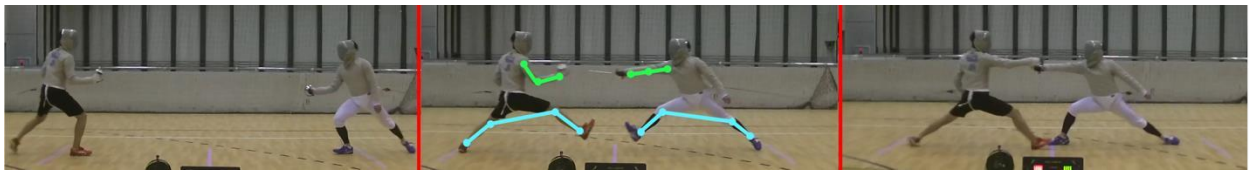
Keywords: Action Recognition, Deep Learning, Sports Video Analysis, Image Processing, Fencing

1. INTRODUCTION

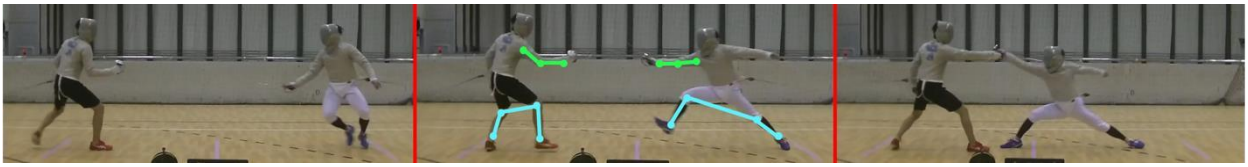
While fencing is well-known worldwide, it is a minor sport with a small population of less than 6,000 in Japan. The Japan Fencing Association has been making various efforts to increase the number of players to improve the competitiveness of the sport. For example, fencing classes, distribution of free tickets to domestic tournaments. The former are held in cooperation with local governments to increase the number of people who experience fencing. However, even if people have the opportunity to watch fencing competitions, the rules are so complicated that they do not understand what is interesting about fencing. On the other hand, those who have experience in fencing have a good understanding of the rules and can understand the game situation sensitively. However, for a first-time spectator, it is difficult to understand the game situation because the game situation changes rapidly, and it is difficult to follow the fast-moving athletes with the eyes. This makes it difficult for the audience to understand the essentials of the sport.

Sports video analysis [1], such as “ball-tracking” and “player detection,” has been actively studied in baseball, soccer, golf, and even fencing to make sports easier to understand and not burdensome. A large amount of data is available in popular sports a large amount of data is available for popular sports conducted in large-scale environments. However minor sports have the opposite case. Therefore, in fencing, it is necessary to consider a sports video analysis method that uses less equipment and a small amount of data. Various approaches have been adopted for fencing video analysis using a single camera. Takahashi et al. [2] developed a system called “Sword Tracer,” which can track the movement of the sword tips and show it in a video. The thin tip of the sword moves at a high speed during competition. Visualization of the trajectory of the sword tip improves the appearance of the match; however, this does not lead to a fundamental understanding of the rules. So we attempt to solve this problem by visualizing the match situation.

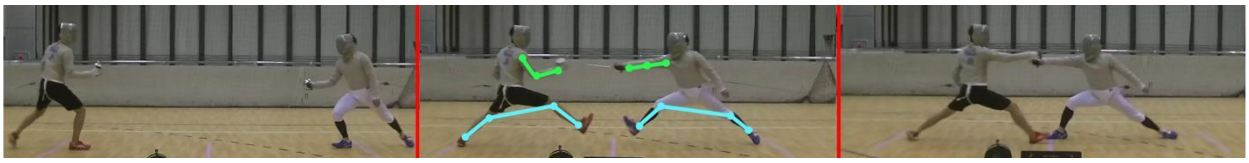
*Email: sawahata@sensor.mech.chuo-u.ac.jp



(a)Simultaneous



(b)Right-Attack



(c)Right-Preparation-Attack

Figure 2. Three types of phrase

Our aim is to display the game status in real-time. Therefore, from the image, we extract the posture and direction of the players' movement to understand and recognize them. Our system visualizes the match situation by using skeleton point information extracted from video images to identify "phrases," which are transitions of the match situation.

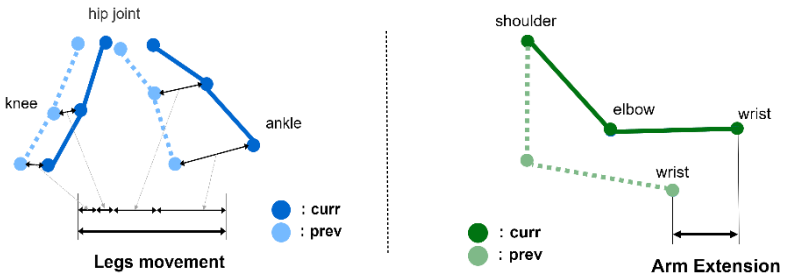
2. METHOD

The process flow of the proposed scoring system is illustrated in Figure 1. The proposed system is divided into two models. First, from a video, skeleton point information is obtained using the "posture estimation model." Second, from skeleton point information, the "Phrase Judging model" determines the situation of a match. Neither player scores a point. Both players start attacking at the same time, and both players make similar moves during the course of the attack. This is the phrase "Simultaneous", meaning that neither player has priority to attack and neither player scores a point. Next, let's look at "Attack". In the middle photo in Figure 2(b), the right player's arms and legs are both extended toward the opponent. Finally, "Preparation-Attack" is a phrase in which one of the players starts an attack at the same time, but one of the players extends his arm first during the attack preparation movement to gain the priority to attack first. The right player's arms are extended earlier than the left player's. This is another example of the "Attack" phrase. In the phrase "Attack," the player who initiates the attack scores a point even if he pokes at the same time. In addition, since "Attack, Preparation-Attack" also identifies the left and right players, the total number of phrases is five, as follows:

Simultaneous
Right-attack
Right-preparation-attack
Left-attack
Left-preparation-attack

2.1 Posture Estimation Model

A light weight open pose implementation [3] is used for skeleton point extraction from the images. It is a lighter version of OpenPose[456] and can run on a CPU at real-time speed. Although OpenPose is not as precise as 3D motion capture systems, it can extract 2D skeletal point information from videos by using only an RGB camera. Under this system, it is not required for the subject to put on any extra equipment for tracking.



(a) Movement value **(b)** Arm extension value

Figure 3. value for Minimum squared error model.

2.2 Phrase Judging Model

We evaluated two different phrase judging models. one is "Minimum Squared Error(MSE)" model and the other is "Deep learning model".

2.2.1 Minimum Squared Error(MSE) model

This model judges phrases by the "movement value" and "arm extension value". These two values were calculated through OpenPose. The difference in each value is taken between players, and the phrase is judged by the behavior in the time series. This method was compared with the deep learning method.

2.2.1.1 Arm extension value

The shoulder, elbow, and wrist of the side holding the sword are used. As shown in Figure 3(a), the 2D coordinates of the shoulder $(x_{shoulder}, y_{shoulder})$ and 2D coordinates of the wrist are used to calculate the elongation value the initial value of elongation is obtained, and the difference $\delta_{n:arm} = |x_0 x_n| - |x_{wrist} - x_{shoulder}|$ at each frame is the elongation value of the arm.

2.2.1.2 Movement value

Four points on the knees and ankles of both feet are used. The movement value is the sum of 4 points $\delta_{n:all:foot} = \delta_9 + \delta_{10} + \delta_{12} + \delta_{13}$ with the difference $\delta_n = \sqrt{(x_{curr} - x_{prev})^2 + (y_{curr} - y_{prev})^2}$ from the previous frame as shown in Figure 3(b).

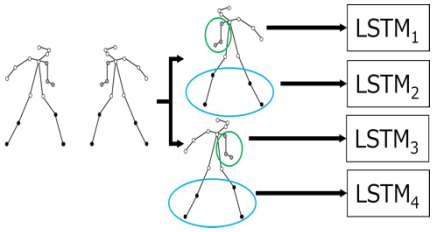


Figure 4. The input of neural network of deep learning model.

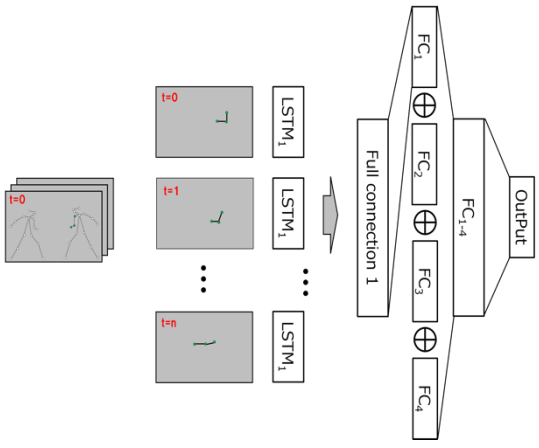


Figure 5. Neural network of deep learning model.

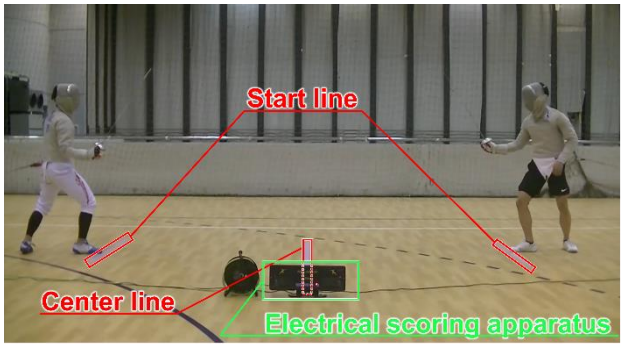


Figure 5. The environment of the experiment at Chuo University.

2.2.2 Deep learning model

Long-short-term memory (LSTM) is used for deep-learning-based phrase decision models. Figure 4 shows an overview of the network. The input (Figure 3) is the time-series skeletal point information for seven points and the two players used in Minimum-squared error model section. The output is the class names from 0 to 4 for the five phrases.

3. EXPERIMENT

To evaluate the accuracy of the two proposed methods described in Methods Section, we conducted a comparison experiment between the deep method and the method that does not use learning. A total of "Five phrases" were judged and evaluated basis of the percentage of correct responses.

-Phrases-

0 Simultaneous
1 Right-attack
2 Right-preparation-attack
3 Left-attack
4 Left-preparation-attack

3.1 Environment

Data capture was conducted in the environment shown in Figure 5. The shooting location was the Arena A court on the third floor of the first gymnasium in the Chuo University Tama Campus.

The experimental environment is shown in Figure 3 was set up in front of the piste to prevent players from hiding from the referee. The players were deployed such that they would not step on the starting line. The distance between the starting line and the center line was 4 m (as required by the rules), and the center has a center line. The video was taken from the first signal of the referee to the moment the house (thrust) decision was displayed on the console of the judges. It was about approximately 1 second.

3.2 Procedure

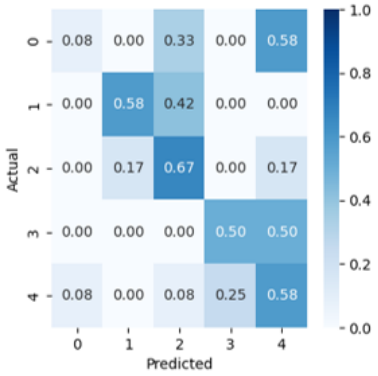
Accuracy test

Two subjects were captured twice for each of the five phrases. Subsequently, the two subjects were replaced, and this procedure was then repeated. This was performed for all combinations, and a total of 12 videos were taken for each phrase, for a total of 60 videos. The recognition rates of two methods, one without learning and the other with deep learning, were evaluated on these 60 images.

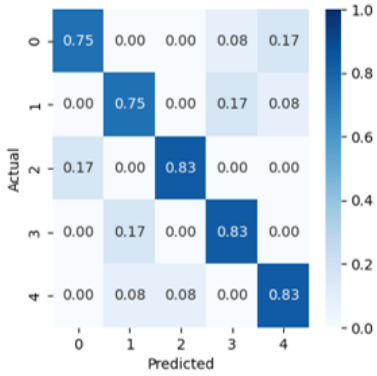
Table 1. Recognition accuracy between Minimum squared error model and Deep learning model[%].

	0	1	2	3	4	Avg.
MSE	8.3	58.3	66.7	50.0	58.3	48.3
Deep Learning	75.0	75.0	83.3	83.3	83.3	80.0

0: Simultaneous 1:attack-right 2:preparation-attack-right 3:attack-left 4:preparation attack left. There are much of differences in 0: simultaneous between the Minimum squared error model and the Deep learning model.



(a)MES model



(a)Deep Learning model

Figure 6. Confusion matrices at prediction and true phrase.

Table 2. Processing speed of proposed system.

	Pose Estimation [ms]	Phrase detection [ms]	Whole System [ms]	Whole System (frame rate)[fps]
MSE model	36.5	121.8	160.7	6.2
Deep learning model	36.5	266.1	305.8	3.3

Speed test

We describe the verification of the processing speed of the proposed model. In order to verify the speed of the proposed model, a total of 360 frames were inferred, and the statistics were calculated using the results. Table 2 shows the results of the processing speed test.

3.3 Result And Discussion

Table 1, and Figure 6 present the experimental results. The average recognition rate for the unlearning method was 48.3%, whereas that for the deep learning method was 80.0%. A 31.7% difference in recognition rate was observed between the two methods. This is due to the misrecognition and missing detection of skeleton points. The recognition rate was significantly affected by large outlier values. Meanwhile, the deep learning method, which learned even when the data was flawed, was not too affected by missed detections and thus had a higher recognition rate than the method that did not use deep learning. In addition, the process of change of skeletal points, for example, the arm suddenly extends from the middle of the body, and the arm slowly extends from the beginning. It is difficult to distinguish between phrases in methods that do not use learning with MSE model error, and deep learning can distinguish these differences. Thus, the recognition rate of the deep learning method is higher than that of the other methods.

4. CONCLUSION

In this study, we developed a scoring system for fencing competitions using skeleton point information extracted from images. Experimental results confirmed the effectiveness of the deep learning method. In the future, we plan to improve the deep learning method, enhance the system, and improve the processing speed improvement, phrase diversity, expansion of shooting range, and so on. It is also planned to construct a scoring system that can be used in an actual game environment.

REFERENCES

[1] Shih, H.-C., "A Survey of Content-Aware Video Analysis for Sports," IEEE Transactions on Circuits and Systems for Video Technology 28(5), 1212–1231 (2018).

[2] Takahashi, M., Yokozawa, S., Mitsumine, H., Itsuki, T., Naoe, M. and Funaki, S., "Sword tracer: visualization of sword trajectories in fencing," ACM SIGGRAPH 2018 Talks, 1–2, ACM, Vancouver British Columbia Canada (2018).

[3] Osokin, D., "Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose," arXiv:1811.12004 [cs] (2018).

[4] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E. and Sheikh, Y., "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(1), 172–186 (2021).

[5] Cao, Z., Simon, T., Wei, S.-E. and Sheikh, Y., "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1302–1310 (2017).

[6] Núñez, J. C., Cabido, R., Pantrigo, J. J., Montemayor, A. S. and Vélez, J. F., "Convolutional Neural Networks and Long Short-Term Memory for skeleton-based human activity and hand gesture recognition," Pattern Recognition **76**, 80–94 (2018).