Pitching Motion Matching Based on Pose Similarity Using Dynamic Time Warping

Ryohei Osawa
Graduate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan
r-osawa@fuji.waseda.jp

Takaaki Ishikawa
Global Information and Telecommunication Institute
Waseda University
Tokyo, Japan
takaxp@ieee.org

Hiroshi Watanabe
Graduate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*—By comparing multiple videos, it is possible to find the differences between the movements by different players or between motions of the same player in separate situations. To compare multiple motions, it is necessary to match the timing of the motions. However, it is difficult to synchronize the timing since the motion speed and usage of the body varying by player and situation. Therefore, in this paper, we propose a matching method for two pitching motions based on pose similarity using Dynamic Time Warping (DTW). This method enables accurate matching of two pitching motions in many cases.**

Keywords—motion analysis, video analysis, dynamic time warping, motion matching, baseball

I. Introduction

In recent years, video data of baseball players during games and practices have been utilized to analyze their movements. Since the movements recorded on video can be confirmed repeatedly, it is possible for player to check the details of the movement. In addition, by comparing the movements, such as pitching motions, in different videos, players can find differences between the multiple motions and clarify the issues they need to address to improve their pitching skills.

In order to compare multiple motions in detail, matching the timing of the motions is important. For example, when comparing pitching motions, it is necessary to synchronize all timings such as the start and end of pitching, the moment of throwing the ball and just before kicking up the pivoting foot. It is, however, difficult to accurately match all timings of multiple motions since the use of the body and the speed of the movement vary depending on player and situation.

Therefore, in this paper, we propose a method to match the timing of the pitching motions in two videos. We use OpenPose [1] to detect the pitcher's body keypoints from pitching video and calculate the pose similarity between frames using body keypoints data. Based on the pose similarity, we match the timing of two pitching motions using Dynamic Time Warping [2].

II. Related Work

A. OpenPose

OpenPose proposed by Cao et al. is a method for detecting human body keypoints in an image or video. OpenPose can estimate the position of 15, 18, or 25 body keypoints per person, depending on the model. An example of estimation result by OpenPose is shown in Fig. 1. For each body keypoint, 2D coordinate values on the image and the confidence score in the range [0,1] are output. If OpenPose fails to detect a keypoint, its 2D coordinate values and confidence score are zero.

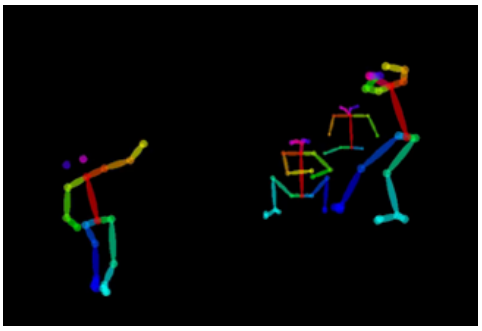

Fig. 1. Example of estimation result by OpenPose.

OpenPose has the advantage that it does not require any device to be attached to the body. It is possible to estimate the position of human body keypoints only from a image or video. OpenPose enables us to analyze the motion and estimate the pose of a person without putting loads on the body.

Another advantage of OpenPose is that it is possible to detect body keypoints from photographs or videos taken by a single camera, without using any special sensors. Since neither specific analysis device nor sophisticated photographic equipment is required, OpenPose can estimate the position of human body keypoints from images captured by smartphones and videos broadcasted on TV, etc.

B. Dynamic Time Warping

Dynamic Time Warping (DTW) is a method of matching elements in two sequences to maximize the similarity between two sequences. DTW calculates the distance between two elements and determines the correspondence so that the sum of distances is minimized. This algorithm can be applied even when the number of elements in two sequences is different. Since DTW can determine the similarity between sequences, it is used for human action recognition [3] and brain activity classification [4].

In DTW, each element in one sequence must be matched to one or more elements in the other sequence. Therefore, if noise, element that should not be matched, is included in one sequence, it may result in an inaccurate correspondence.

An example of applying DTW to two sequences with similar shapes is shown in Fig. 2. The elements of sequence A and sequence B are connected by blue and orange lines,

978-1-7281-9802-6/20/$31.00 ©2020 IEEE

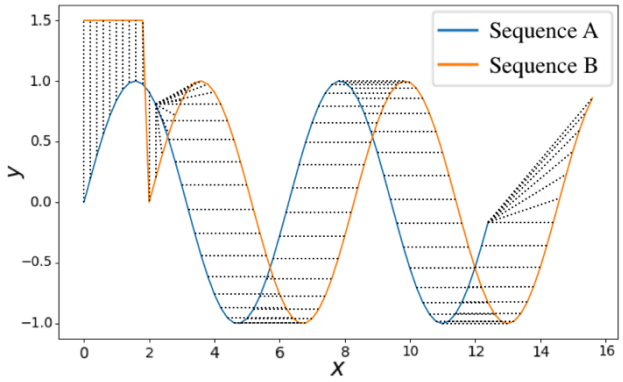

Fig. 2. Example of matching result by DTW.

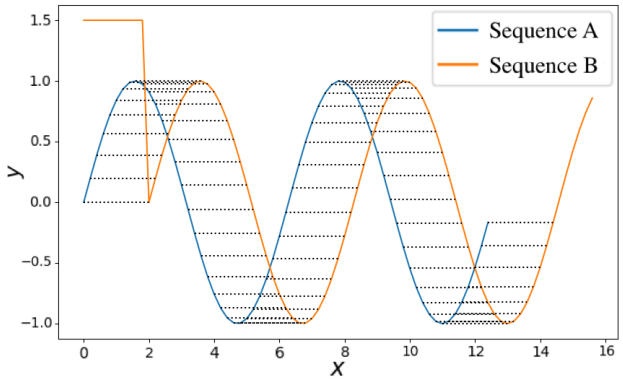

Fig. 3. Example of matching result by DTW that automatically extracts the start and end points.

respectively. Near the beginning of sequence B, i.e. the left end of the orange graph, there are elements that are not included in sequence A. The *y* coordinate differences between any element in sequence A and any element in sequence B are all calculated, and the correspondence is determined to minimize the sum of *y* coordinate differences.

In Fig. 2, the matched elements are connected to each other by a dotted line. The vertical dotted lines, i.e. the correspondences connecting two elements with different *y* coordinates, exist on the left side of the graph. The elements near the beginning of sequence A are not matched with elements in sequence B that have the same *y* coordinate. The correspondence may be inaccurate if elements that should not be matched are contained in one of the two sequences.

C. *Pitching Motion Matching Using DTW*

DTW always corresponds all elements of one sequence to one or more elements of the other sequence. Therefore, the correspondence may be inaccurate if noise, element that should not be matched, is included in one sequence, as in the example shown in Fig. 2. When DTW is used for motion matching, it is necessary to determine the start and end points of the motion in the video subjectively. Yokoi et al. [5] proposed DTW that automatically extracts the start and end points of a motion from a sequence based on the distance between the elements. This method determines the correspondence so that the average of the distances, rather than the sum of the distances, is minimized. The average of the distances is calculated based on the number of matched elements.

An example of applying the method proposed by Yokoi et al. to two similar sequences is shown in Fig. 3. The elements of two sequences A and B are 2D coordinate values and are connected by blue and orange lines, respectively. The *y* coordinate differences between any element in sequence A and any element in sequence B are all calculated, and the correspondence is determined to minimize the average of *y* coordinate differences.

In Fig. 3, the matched elements are connected by a dotted line. Near the beginning and end of sequence B, i.e. the left and right ends of the orange graph, there are elements that are not included in sequence A. In Fig. 3, those elements are not linked to other elements by a dotted line, i.e. they are not matched to any of the elements in sequence A. On the other hand, all elements in sequence A are connected to the elements with the same *y* coordinate in sequence B.

Yokoi et al. proposed an approach for pitching motion matching using DTW that automatically extracts the start and end points of a motion. The approach corresponds the frames in two videos based on the 2D coordinate values of human body keypoints obtained by OpenPose. This method enables accurate matching of two pitching motions in many cases, even if one sequence contains other movements before or after the pitching motion. However, there are two weaknesses in this method.

The first is that inaccurate detection of human body keypoints by OpenPose may reduce accuracy of the matching. The approach proposed in [5] calculates the similarity between video frames based on coordinate data of body keypoints acquired by OpenPose. If the position of the body keypoints is inaccurately estimated, the similarity is not calculated correctly, and inaccurate matchings may occur.

The second weakness is that it is difficult to accurately match two pitching motions of right-handed pitcher and left-handed pitcher. The method proposed in [5] corresponds two pitchers' pitching motions based on the location similarity between same body keypoints, e.g., right shoulder and right shoulder or left knee and left knee, etc. However, right-handed pitchers and left-handed pitchers use their bodies differently, i.e. the arm that releases a ball and the pivoting foot at the time of pitching are opposite to each other. Therefore, it is hard to exactly match the pitching motions of right-handed and left-handed pitcher.

III. Proposed Method

A. *Vector Data Creation*

In this study, we apply the OpenPose model which can detect 25 keypoints to a pitching video and obtain body keypoints data of a pitcher in the video. Among the 25 keypoints, we used 12 keypoints, neck, shoulders, elbows, hips, the middle of the waist, knees and ankles. This is because we have confirmed that the 12 keypoints are stably detectable in the videos used in this study. The length and angle of the vector connecting any two keypoints among the 12 keypoints are calculated from the 2D coordinates of each keypoint. In addition, we compute the confidence value of the vector using the confidence score of keypoints obtained by OpenPose. The confidence value is the smaller of the confidence scores of two keypoints.

For each frame, 66 vectors data are created using the body keypoints data obtained by OpenPose. Each vector is

978-1-7281-9802-6/20/$31.00 ©2020 IEEE

composed of three elements: length, angle, and confidence value.

B. Calculation of Pose Similarity

The purpose of this method is to match the timing of two pitching motions in different videos. Therefore, it is important to correspond the frames in two videos based on the pose similarity of two pitchers. The pose similarity between two frames is calculated using 66 vectors data.

We compute the corresponding vectors difference. If we match two pitching motions of right-handed pitchers or left-handed pitchers, we calculate the difference between vectors connecting same two keypoints such as "a vector connecting neck and right shoulder" and "a vector connecting neck and right shoulder". On the other hand, when matching two pitching motions of right-handed pitcher and left-handed pitcher, the corresponding vectors do not necessarily mean vectors connecting same two keypoints, e.g., "a vector connecting neck and right shoulder" and "a vector connecting neck and left shoulder". This is because the arm that releases a ball and the pivoting foot at the time of pitching are opposite to each other. It is not suitable for computing the pose similarity between the right-handed pitcher and left-handed pitcher to invariably calculate the difference of vectors connecting same two keypoints. To accurately calculate the pose similarity of pitching motions, we change some of the corresponding vectors based on whether the hands that release a ball for the two pitchers being compared are opposite to each other.

The corresponding vectors difference is the product of the angular difference and the length difference between two vectors. As an exception, if at least one confidence values of two vectors is less than a threshold, the corresponding vectors difference is the constant. The reason for this process is to reduce the effect of inaccurate detection by OpenPose on the pose similarity. The sum of 66 corresponding vectors differences is the pose similarity between two frames.

C. Matching by Dynamic Time Warping

To correspond frames in two videos, we use the DTW that automatically extracts the start and end points of a motion proposed in [5]. The two videos to be compared are considered as the model sequence and the input sequence, respectively. For the model sequence, the start and end points are predetermined by subjectivity. On the other hand, the start and end points of the motion in the input sequence are determined based on the pose similarity. Therefore, the model sequence contains only one pitching motion and the input sequence may include another motion in addition to one pitching motion.

The following is an overview of the frame matching between the model sequence with M elements and the input sequence with N elements. M and N are both natural numbers and represent the number of video frames in each sequence.

First, we prepare the matrices D and E of $(M + 1) \times (N + 1)$. The initial states of matrix D and matrix E are shown in Fig. 4 and Fig. 5, respectively. In Fig. 4 and Fig. 5, the white cells indicate each element of the matrices D and E. The top left white cell in Fig. 4 is denoted as $D(0, 0)$ and the bottom right cell in Fig. 5 is denoted as $E(M, N)$.

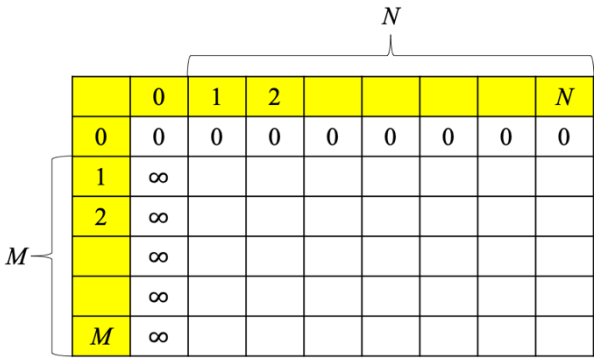

Fig. 4. The initial state of matrix D.

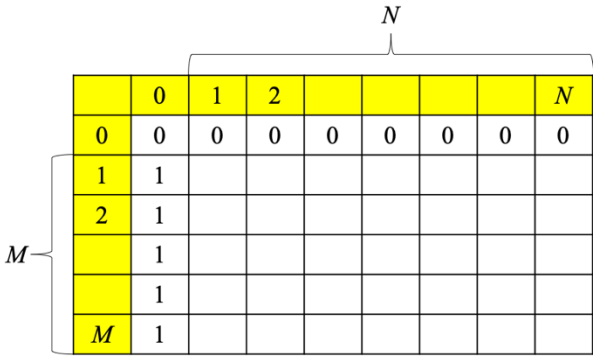

Fig. 5. The initial state of matrix E.

Second, we compute the values of the elements $D(m, n)$ and $E(m, n)$ $(1 \le m \le M, 1 \le n \le N)$ in the ascending order of m and n. Note that both m and n are natural numbers. The value of $D(m, n)$ is calculated based on the three elements $D(m - 1, n)$, $D(m, n - 1)$ and $D(m - 1, n - 1)$. As a preparation for determining $D(m, n)$, we compute $f(p, q)$ using the components of matrices D and E,. The combination of p and q can be either $(m - 1, n)$, $(m, n - 1)$ or $(m - 1, n - 1)$. We define $f(p, q)$ using the equation,

$$f(p, q) = \frac{D(p, q) \times E(p, q) + s(m, n)}{E(p, q) + 1}, \quad (1)$$

where $s(m, n)$ is the pose similarity between the m frame of the model sequence and the n frame of the input sequence. $D(m, n)$ is determined as follows:

$$D(m, n) = \min(f(m - 1, n), f(m, n - 1), f(m - 1, n - 1)). \quad (2)$$

The value of $E(m, n)$ is determined according to $D(m, n)$, i.e. the smallest among the three values shown in (2), $f(m - 1, n)$, $f(m, n - 1)$ and $f(m - 1, n - 1)$. When the minimum value is written as $f(s_m, t_n)$, $E(m, n)$ is calculated by,

$$E(m, n) = E(s_m, t_n) + 1. \quad (3)$$

For example, if the smallest value is $f(m - 1, n)$, then $E(m, n)$ is $E(m - 1, n)$ plus 1. Note that s_m and t_n are always recorded.

Finally, we correspond the frames in two sequences based on the components of matrix D. We compare all the values of $D(M, n)$, the bottom column elements of the matrix D shown in Fig. 4, and find the minimum. Using recorded s_m and t_n shown in (3), the correspondence is determined.

978-1-7281-9802-6/20/$31.00 ©2020 IEEE

IV. Experiments

A. Experiment Outline

The experiment dataset used in this study are 101 pitching videos of right-handed pitchers and one pitching video of left-handed pitcher. Each video includes one pitching motion. One of the right-handed pitchers' videos is considered as model sequence 1 and the others are the input sequences. Also, the left-handed video is considered as model sequence 2. For the two model sequences, we determined the start and end points of the pitching motion by visual observation. The start point is the moment when the foot opposite to the pivoting foot leaves the ground, and the end point is just before the kicked pivoting foot touches the ground. In the case of right-handed pitcher, the pivoting foot is the right foot.

We applied this method to all sequences. We judged the success or failure of each matching between frames in the two sequences, a model sequence and an input sequence. The success or failure of each matching was judged visually based on whether there were more suitable frames for matching.

B. Experiment Results

A part of results of pitching motion matching between two right-handed pitchers is shown in Fig. 6 and Fig. 7. Each image in Fig. 6 and Fig. 7 is the result of arranging two frames matched by conventional method in [5] and proposed method, respectively. Also, for pitching motion matching between a right-handed pitcher and left-handed pitcher, the results using proposed method is shown in Fig. 8. In order to compare two pitchers' poses easily, only the

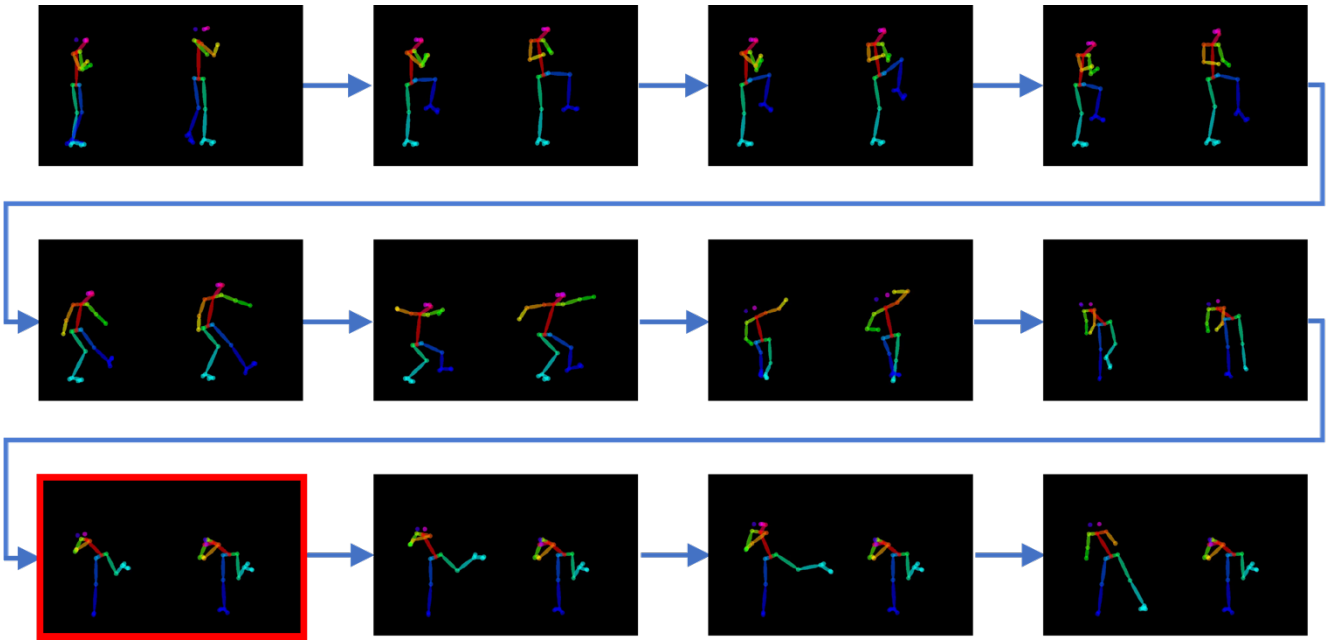

Fig. 6. A part of matching results between two right-handed pitchers using conventional method.

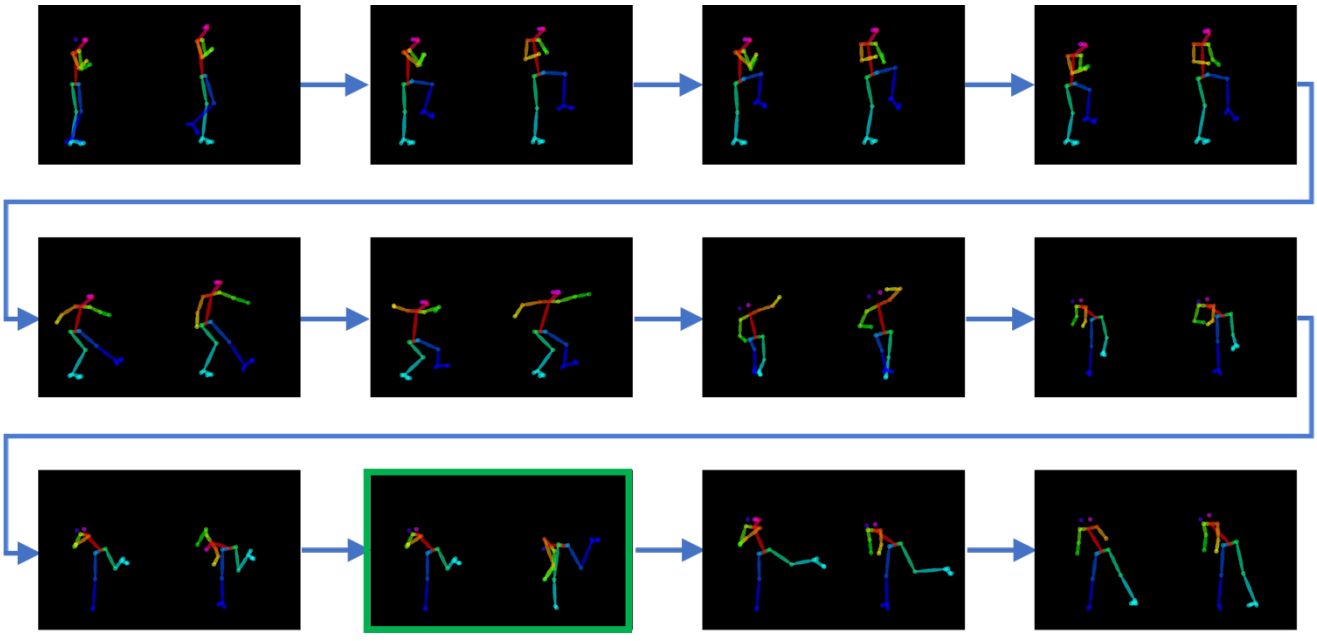

Fig. 7. A part of matching results between two right-handed pitchers using proposed method.

978-1-7281-9802-6/20/$31.00 ©2020 IEEE

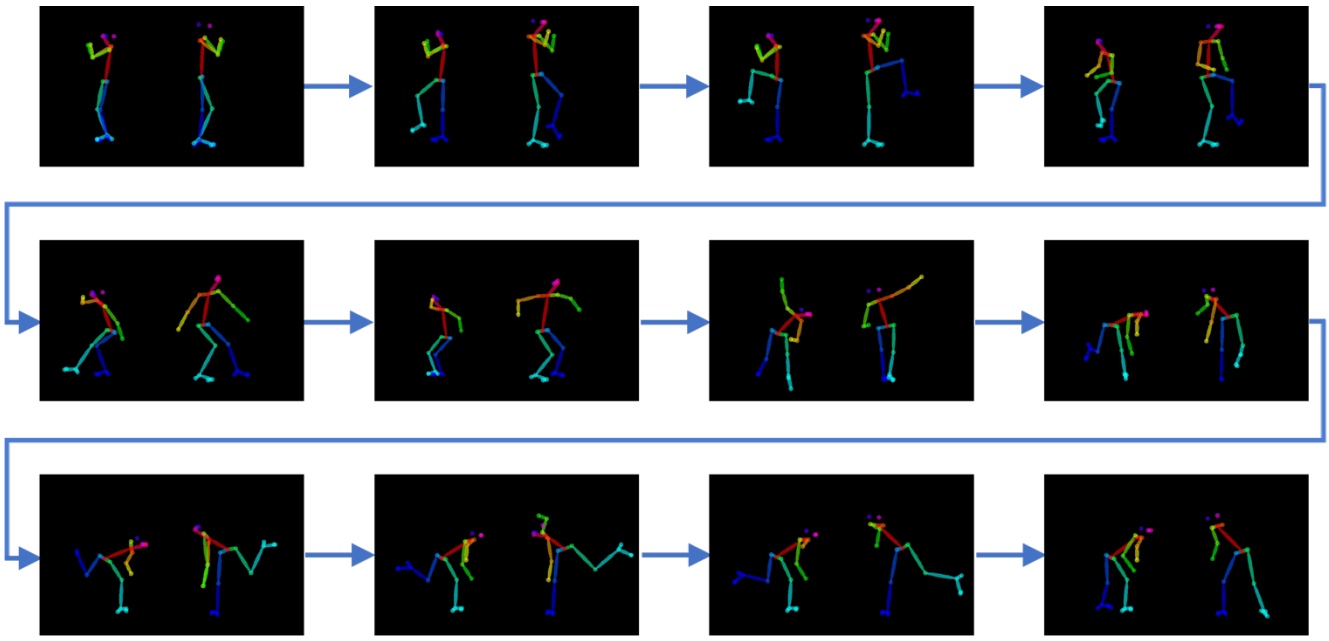

Fig. 8. A part of matching results between a right-handed pitcher and left-handed pitcher using proposed method.

skeleton of the pitcher, which is estimated by OpenPose, is shown.

The images shown in Fig. 6 and Fig. 7 are the results of matching the model sequence 1 to the same input sequence including frames in which the position of the pitcher's legs was incorrectly detected by OpenPose. In Fig. 6, the skeleton of the pitcher on the right side of each picture is identical in the four images from the one surrounded by the red box to the bottom right one. In the lower right image, the two pitchers' poses are very different, which is an inaccurate matching. On the other hand, in Fig. 7, the skeletons of the two pitchers are similar in the bottom right image. In the green-framed image, we confirmed that both legs of the pitcher on the right side were detected opposite to each other. However, even though the sequence included false detections by OpenPose, the proposed method corresponded to similar poses of the two pitchers.

The accuracy of pitching motion matching between model sequence 1 and input sequences, i.e., two right-handed pitchers, is shown Table I. As shown in Table I the matching accuracy of our method was higher than that of the conventional method. We speculate that this is because the proposed method includes a process to reduce the effect of inaccurate detections by OpenPose in the calculation of pose similarity. The accuracy of the matching between two right-handed pitchers and that between left-handed and right-handed pitchers were both 92% or more.

Table II shows the result of pitching motion matching by proposed method between model sequence 2 and input sequences, i.e. left-handed pitcher and right-handed pitcher. It is difficult for the conventional method to accurately match two pitching motions of right-handed pitcher and left-handed pitcher because it only targets the correspondence between two right-handed pitchers. On the other hand, our method enabled the correct matching of two pitching motions with more than 92% accuracy.

TABLE I. THE ACCURACY OF PITCHING MOTION MATCHING BETWEEN TWO RIGHT-HANDED PITCHERS

Method	Number of Matched Frames	Accuracy [%]
Yokoi et al. [3]	9484	94.24
Ours	10586	95.06

TABLE II. MATCHING RESULT OF LEFT-HANDED PITCHER AND RIGHT-HANDED PITCHER BY PROPOSED METHOD

	Correct Matching	Incorrect Matching
Number of Frames	8478	726
Percentage [%]	92.11	7.89

V. CONCLUSIONS

In this paper, we propose the method of matching two pitching motions based on the pose similarity using DTW. First, we create vectors using body keypoints data obtained by OpenPose. Second, the pose similarity between frames is calculated from vectors data. Then, based on the pose similarity, we correspond frames using DTW that automatically extracts the start and end points of a pitching motion. The experimental results showed that the correspondence between frames was accurate in many cases.

REFERENCES

[1] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," IEEE Conference on Computer Vision and Pattern Recognition, No.121, pp.1302-1310, Jul. 2017.

[2] P. F. Felzenszwalb and R. Zabih, "Dynamic programming and graph algorithms in computer vision," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No. 4, pp. 721-740, Apr. 2011.

[3] S. Sempena, N. U. Maulidevi and P. R. Aryan, "Human action recognition using dynamic time warping," Proceedings of the 2011 International Conference on Electrical Engineering and Informatics, pp.1-5, Jul. 2011.

[4] W. A. Chaovalitwongse and P. M. Pardalos, "On the Time Series Support Vector Machine Using Dynamic Time Warping Kernel for Brain Activity Classification," Cybernetics and Systems Analysis, Vol. 44, No. 1, pp.125-138, Jan. 2008.

[5] S. Yokoi, T. Ishikawa and H. Watanabe, "Alignment for skeleton coordinates data obtained from sports video," IEICE General Conference, D-12-59, Mar. 2019. (in Japanese)

978-1-7281-9802-6/20/$31.00 ©2020 IEEE