IEEE GCCE 2022 1570818857

Motion Matching Based on Pose Similarity Using Movement of Body Parts

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

Shohei Adachi
Graduate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan
alice-fr@asagi.waseda.jp

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

***Abstract*— The use of high-performance equipment to analyze motion has become popular in the field of sports. However, high-performance systems are expensive and difficult to install for amateur players and teams. Therefore, a method of analyzing motion based on video captured by smartphones and cameras is attracting attention as an easy-to-introduce system. One way for amateur athletes to practice efficiently is to compare their motion videos with the expert's one. However, since different players operate at different speeds, it is difficult to make an accurate comparison by simply playing back the two videos at the same time. Therefore, we propose a method to synchronize the timing of similar motion in two videos with high accuracy. To improve the accuracy of the mapping, we corrected the right and left inversion detection of the skeleton and considered the amount of time variation of the skeleton. Through evaluation experiments, we confirm that the proposed method is effective in improving the performance of motion matching.**

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

I. Introduction

In many sports, athletes' movements have been analyzed using various data taken of athletes during games and practices. Especially in professional baseball, data analysis is conducted using high-performance equipment such as TrackMan [1], Hawk-Eye [2] and dedicated radar systems with high speed and high-resolution cameras. However, these systems require very high costs in terms of the price of the machines and the space required for installation. This makes it difficult for amateur sport teams to use these systems.

Therefore, we focused on an approach to use inexpensive commercial cameras for motion analysis. It is beneficial to compare one's own play with that of a professional play. By using multiple videos and comparing the movements, differences in movement can be easily found.

When using video to compare multiple motions in detail, it is important to match the timing of the motions. Simply playing back multiple videos from the start of a motion may not account for differences in motion speed between players, and the timing of the motion may gradually deviate.

Therefore, we propose a method to synchronize the timing of similar motions in two videos with high accuracy based on the posture similarity of the players.

II. Related Work

A. OpenPose

OpenPose [3] is a method for estimating the skeletons of multiple people from video frames alone. OpenPose detects joints such as shoulders and knees, and facial parts such as eyes and nose for each person in an image or video frame. For each detected body part, the 2D coordinate values on the image and the confidence level of the estimation are output. The coordinates are taken from the upper left point of the input image. The confidence level is given in the range of 0 to 1, with a confidence level closer to 1 indicating that the skeleton estimation is more accurate.

The greatest advantage of using OpenPose is the ability to estimate a person's skeleton from images and videos alone. It can estimate a person's posture even for video images captured by a smartphone, without the need for special sensors on the body or special devices such as a high-performance camera. In particular, the fact that the device does not need to be attached to the body has the advantage that it does not affect the player's performance. Therefore, motion analysis can be performed not only during practice but also during games.

B. Start and end point free dynamic time warping

Start and end point free Dynamic Time Warping (DTW) [4] is a method for mapping elements in a sequence so that the two time series can match with the highest similarity. When it is applied to video data, an element refers to a video frame.

In start and end point free DTW, the difference is first calculated for any element in reference time series and any element in the other time series in a round-robin fashion. The mapping between elements is then determined so that the average of the differences is minimized. In this case, all elements in reference time series are mapped, but not all elements in the other time series are always mapped.

In ordinary DTW, the start point of the reference time series is always matched to the start point of the other time series, and the end point of the reference time series is always matched to the end point of the other time series. Therefore, in the conventional mapping method using ordinary DTW, it is necessary to prepare video clipped from only the motion timing as input video. However, with start and end point free DTW, accurate mapping can be performed even if the input video includes other than motion timing.

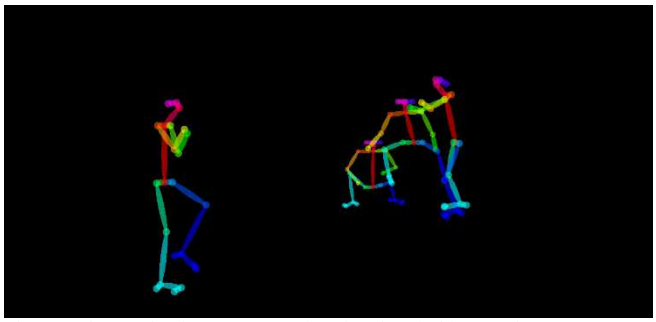

Fig. 1. The result of OpenPose estimation for pitching motion

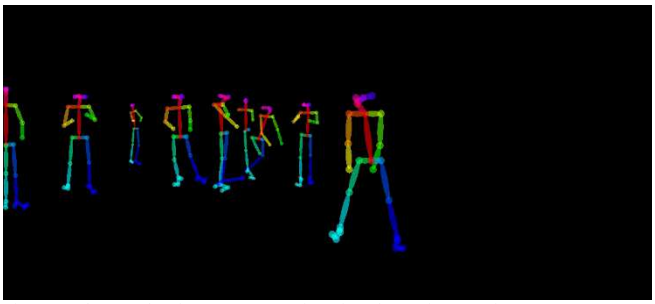

Fig. 2. The result of OpenPose estimation for swing motion

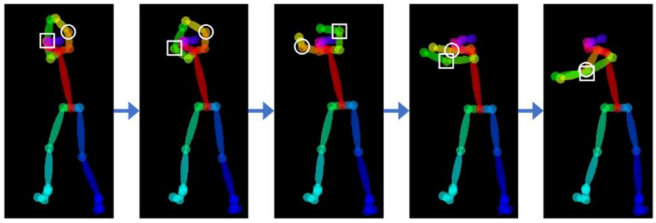

Fig. 3. Example of right and left inversion detection

III. Proposed Method

A. Obtaining skeltal information

The OpenPose is applied to the video containing the sports motion to be synchronized, and the positions of the body parts of the players in the video are estimated. As a result, for each body part detected in each video frame, the 2D coordinate values on the image and the confidence level of the estimation are obtained. In this study, 12 body parts that can be detected by OpenPose are used: both shoulders, elbows, wrists, hips, knees, and ankles.

The results of OpenPose estimation for the video containing the pitching motion are shown in Figure 1 and the results of OpenPose estimation for the video containing the swing motion are shown in Figure 2. In Figures 1 and 2, the skeleton is displayed on a black image of the same size as the input video frame to make the estimated skeleton more visible.

In the video used in this study, there are cases in which people other than the player performing the action to be synchronized are shown. In Figure 1, the player performing the pitching motion is the leftmost person. In Figure 2, the player performing the swing motion is the rightmost person. We focused on the ankle coordinates to identify the player performing the action to be synchronized. Since all the video data used in this study were shot from the same camera angle, the person with the lowest ankle coordinates is the person to be mapped.

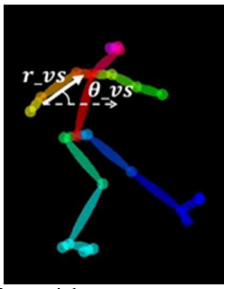

Fig. 4. Example of a spatial vector

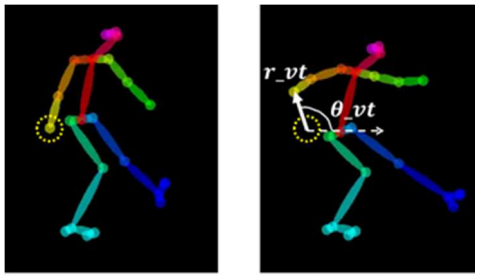

Fig. 5. Example of a temporal vector

B. Correction of right and left inversion detection

In this study, skeletal data obtained by OpenPose is used to calculate posture similarity. However, OpenPose has the problem that wrong detections often occur due to the overlap of the right and left limbs. An example of right and left inversion detection is shown in Figure 3. In Figure 3, right and left elbow inversion detection occurs in the third frame. If this problem occurs, it has a negative impact on the calculation of the posture similarity. As a result, the accuracy of the mapping may be reduced.

We propose a process to correct the right and left limbs inversion detection of consecutive frames. For the right and left limbs, let D_Same be the sum of the differences in the coordinates of the same skeleton in consecutive frames. Similarly, let $D\_Symmetry$ be the sum of the coordinate differences of the symmetric skeleton. Here, a symmetric skeleton is a combination of right and left wrists, right and left knees, and so on. When right and left inversion detection occurs, D_Same will be large and $D\_Symmetry$ will be small. Therefore, if $D\_Same - D\_Symmetry$ is positive and exceeds a certain value, swapping the right and left skeletal coordinates will correct the inversion .

C. Creation of vector data

Using the estimated results of the 12 body parts of the player obtained by OpenPose, two types of vector data are created for each frame: a spatial vector and a temporal vector.

Spatial vector data is a set of vectors vs created based on the estimation results of any two of the 12 body parts. The elements of each spatial vector vs include the length r_vs and angle θ_vs. The length r_vs and angle θ_vs are shown in Figure 4 as an example of spatial vector whose component sites are the right elbow and right shoulder.

Temporal vector data is a set of vectors vt representing the motion of each body part in two neighboring frames. Here, neighboring frames mean that the frame spacing is three. The elements of each temporal vector vt include the distance r_vt and direction θ_vt of movement. Figure 5 shows the distance r_vt and direction θ_vt of movement for the right wrist in the past and target frames with an interval of three frames.

TABLE I. DETAILS OF TARGET VIDEO

	Baseball video	Golf video
Number of videos	122	83
Number of average frames	172.1	340.5
Dominant hand	Right	Right

TABLE II. DETAILS OF MODEL VIDEO (PITCHING)

	Model1	Model2	Model3	Model4
Number of frames	84	80	69	77
Dominant hand	Right	Right	Left	Left

TABLE III. DETAILS OF MODEL VIDEO (SWING)

	Model1	Model2	Model3	Model4
Number of frames	59	65	48	57
Dominant hand	Right	Right	Left	Left

D. Calculation of posture similarity

Based on the created vector data, the similarity of the player's posture between the frames in the two videos is calculated. Pose similarity is defined as the sum of the differences between the corresponding spatial vectors and the corresponding temporal vectors. Here, the corresponding spatial vectors and the temporal vectors depend on whether the dominant hands of the two players are the same during the motion.

The difference in spatial vectors is defined as the average of the difference in distance r_vs multiplied by the difference in angle θ_vs. The difference in temporal, vectors is defined as the average of the difference in distance r_vt multiplied by the difference in angle θ_vt multiplied by the weight. The weight is the ratio of the distance of each vector to the total distance of the time vectors within a frame. Vectors with a confidence level of 0 are not used in the difference calculation because of their negative impact on the posture similarity.

E. Mapping by start and end point free dynamic time warping

The start and end point free DTW is used to determine the correspondence between the two video frames. The posture similarity between frames calculated in 3-D is used for the mapping. Assuming that the first and last frames of reference video coincide with the start and end point of the motion, the start and end point of the motion in the targeted video are determined based on the similarity of the player's posture.

IV. EXPERIMENTS

A. Experimental Methods

The proposed method is applied to the reference video in which the start and end times of motion coincide with the start and end point of the video, and to the target video in which the start and end point do not match. We prepared 122 baseball videos containing one pitching motion per video and 83 golf videos containing one swing motion per video as target videos. These videos include sections other than the pitching or swing motion section. Details of the target video are shown in Table I. The first row indicates the number of videos for each action, the second row indicates the average number of frames in the video, and the third row indicates the dominant hand of the person in the video.

In addition, four model videos were prepared for each motion as reference videos. Models 1 and 2 include the right-handed motion, and Models 3 and 4 include the left-handed motion. These videos contain only motion segments. In this study, the start of the pitching motion was defined as the moment when the off-axis foot left the ground, and the end was defined just before the kicked-up axis foot reached the ground. The start of the golf swing motion was defined as the moment when the takeback begins, and the end of the golf swing motion was defined as the moment when the elbow on the dominant hand side was at the highest position between hitting the ball and the moment when the body rotation ceased. Takeback is the action of pulling the golf club backward from the stance position. The details of the model videos are shown in Tables II and III, with the number of frames in each model video in the first row and the dominant hand of the person in the video in the second row.

TABLE IV. ACCURACY OF CORRECTION OF INVERSION DETECTION

	Arm (frames)	Foot (frames)
Corrected Frame	39	105
Inverted Frame	30	105
Accuracy [%]	76.9	100

TABLE V. KENDALL'S RANK CORRELATION COEFFICIENT (PITCHING)

Method	Kendall's rank correlation coefficient			
	Model1	*Model2*	*Model3*	*Model4*
Proposed	**0.8746**	**0.8192**	**0.8267**	**0.8014**
w/o Inversion	0.8744	0.8191	0.8266	0.8014
w/o Temporal Vector	0.8111	0.7327	0.7365	0.6956

TABLE VI. KENDALL'S RANK CORRELATION COEFFICIENT (SWING)

Method	Kendall's rank correlation coefficient			
	Model1	*Model2*	*Model3*	*Model4*
Proposed	**0.8817**	**0.8966**	0.9109	**0.9176**
w/o Inversion	0.8804	0.8961	**0.9115**	0.9167
w/o Temporal Vector	0.7674	0.7782	0.7026	0.7653

B. Assessment of corrections to right and left inversion detection

The correction of right and left inversion detection is applied to the pitching and swing datasets. The results are shown in Table IV. The number of frames corrected by this method is shown in the first row, the number of frames actually inverted is shown in the second row, and the accuracy of correction is shown in third row.

In Table IV, the accuracy of the arm inversion correction was 76.9%. This method uses the difference in coordinates of body parts between the target frame and the previous frame to determine the inversion detection. Therefore, if the arms and torso overlap in the previous frame and an incorrect skeleton estimation is made, an incorrect correction of inversion detection is likely to occur.

TABLE VII. FRAME ERRORS (PITCHING)

Method	Start or End	Average of frame errors			
		Model1	*Model2*	*Model3*	*Model4*
Proposed	Start	**6.36**	**4.77**	**7.21**	**5.05**
	End	**7.04**	8.27	6.75	**4.92**
w/o Inversion	Start	**6.36**	**4.77**	**7.21**	**5.05**
	End	**7.04**	8.43	6.88	5.02
w/o Temporal Vector	Start	7.68	8.26	8.16	9.85
	End	9.65	**7.57**	**6.45**	5.75

TABLE VIII. FRAME ERRORS (SWING)

Method	Start or End	Average of frame errors			
		Model1	*Model2*	*Model3*	*Model4*
Proposed	Start	**4.94**	3.95	**3.36**	**6.71**
	End	7.57	**6.33**	9.28	**10.78**
w/o Inversion	Start	**4.94**	**3.89**	3.37	**6.71**
	End	**7.29**	6.74	9.43	10.88
w/o Temporal Vector	Start	12.84	38.13	34.06	97.51
	End	7.77	6.56	**8.72**	11.83

C. *Evaluation by Kendall's rank correlation coefficient*

Kendall's rank correlation coefficient [5] is used to evaluate the process of calculating the similarity between frames in each method. Kendall's rank correlation coefficient indicates that the closer to 1, the higher the rating. Table V shows the results of the pitching motion and Table VI shows the results of the swing motion. The results of the proposed method are shown in the first row. The results without the correction of inversion detection are shown in the second row. The results without the temporal vector are shown in the third row.

From Tables V and VI, the correction of the right and left inversion detection and the use of temporal vectors can improve the accuracy, and the proposed method including both can achieve the highest accuracy in most cases.

D. *Evaluation by Frame Error*

To evaluate the accuracy of the mapping results, we measured and compared the average frame error between the start and end points of motions determined by start and end point free DTW and the start and end points of motions determined by visual inspection. The smaller the average frame error, the more accurate the mapping. Table VII shows the results for the pitching video and Table VIII shows the results for the swing video. For each method, the Start row shows the average error for the start frame and the End row shows the average error for the end frame.

The results in Tables VII and VIII show that the proposed method using the correction of right and left inversion detection and the temporal vectors produces the fewest error frames. It can also be confirmed that the use of temporal vectors makes a particularly significant contribution to the mapping performance.

In Table VIII, the results for swing videos, the start frame error is much larger than the other results when the temporal vector is not used. This may be due to the fact that the golf swing has a very similar initial posture and posture at the moment of hitting after takeback. However, by using the temporal vector, it is possible to determine chronologically whether the posture is before or after the takeback, resulting in improved mapping accuracy for the start frame.

In addition, when we look at some results of Model 3 and Model 4, which are videos of left-handed players, we can confirm that the average frame error is smaller than the results of Model 1 and Model 2, which are videos of right-handed players. Since the target videos for the mapping are all of right-handed players, this method is also effective for mapping the motions of right-handed and left-handed players.

The above comparison of the average error frames at the start and end points confirms the superiority of the proposed method in terms of similar motion mapping.

V. CONCLUSION

In this study, we propose a method to synchronize the timing of similar motions in two videos with high accuracy. The first step of the method is to estimate the skeleton of the player performing the action from the video. Then, based on the obtained skeletal coordinates, two types of posture data are calculated for each frame. Finally, the frames of the two videos are mapped based on the posture similarity between the frames obtained from the posture data.

To evaluate the effectiveness of the proposed method, we also investigated the effects of correction of right and left inversion detection during skeletal estimation and the use of posture data considering the time variation of each body part on the performance of the mapping.

Evaluation experiments using Kendall's rank correlation coefficient confirmed that correction of right and left inversion detection and the use of temporal vectors improved performance in the calculation of posture similarity. Comparison of the errors between the start and end frames of the motion segment determined visually and the start and end frames determined by start and end point free DTW confirms that the correction of right and left inversion detection and the use of temporal vectors also improve the performance of the motion segment mapping.

ACKNOWLEGEMENT

This work was supported by JSPS KAKENHI Grant Number JP20K11344.

REFERENCES

[1] TRACKMAN, "TrackMan Baseball," (Last viewed on 4th Aug. 2022), https://trackmanbaseball.com/

[2] SONY, "Hawk-Eye Making Sports More Exciteing Through Visualization Technology," (Last viewed on 4th Aug. 2022), https://www.sony.com/en/SonyInfo/technology/stories/Hawk-Eye/

[3] Z. Cao, T. Simon, S. Wei and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinitiy Fields," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1302-1310, Jul. 2017.

[4] Shinya Yokoi, "Alignment Method for Body Parts Coordinates Obtained from Sport Video," Master thesis of Waseda University School of Fundamental Science and Engineering Department of Computer Science and Communications Engineering, Feb. 2019. (in Japanese)

[5] Maurice George Kendall, "A New Measure of Rank Correlation," Biometrika Vol.30 No. 1/2, pp. 81-93, Jun. 1938.