Player Pose Analysis in Tennis Video based on Pose Estimation

Ryunosuke Kurose
Keio University
Department of Electronics and Electronical Engineering
Kanagawa, Japan
rkurose@aoki-medialab.jp

Masaki Hayashi
Keio University, Recreation Lab
Department of Electronics and Electronical Engineering
Kanagawa, Japan
mhayashi@aoki-medialab.jp

Takeo Ishii
Matsudo Orthopedics Hospital
Chiba, Japan
spotake009@ybb.ne.jp

Yoshimitsu Aoki
Keio University
Department of Electronics and Electronical Engineering
Kanagawa, Japan
aoki@elec.keio.ac.jp

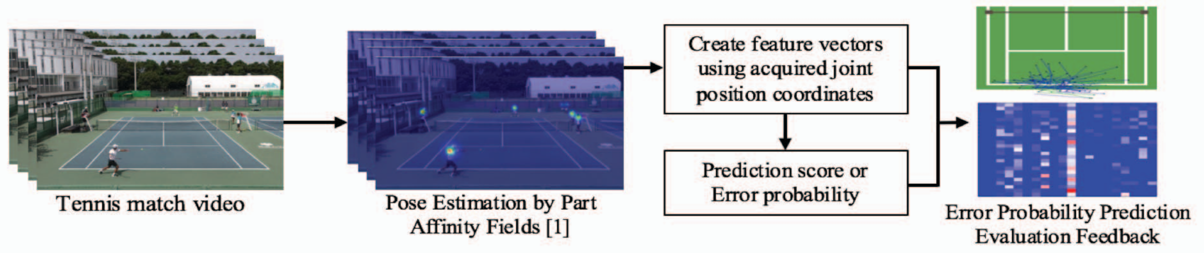
Tennis match video
Pose Estimation by Part Affinity Fields [1]
Create feature vectors using acquired joint position coordinates
Prediction score or Error probability
Error Probability Prediction Evaluation Feedback

Fig.1 Overview of the proposed method

***Abstract*—The demand for sports video analysis is expanding. Sports video analysis is used to analyze their own play and play of opponent players, and visualize movement and ability of players. Scientific and objective analysis become possible by incorporating video analysis. It is desired to improve the competition level by feeding back the analysis results to the players. Therefore, in this research, in the case of feeding back the analysis result to the athlete, research purpose is to realize a method which can understand and evaluate the details of the form hitting the ball in detail. First, joint position coordinates are estimated from input RGB tennis images by a posture estimation method. The joint position coordinates in each frame at the time of shot are classified using unsupervised method and represented by BoW. The feature vector is designed by combining this with the shot position. The probability of shot success is predicted using this feature vector. By visualizing BoW of the shot with the high probability of success and the high failure probability, it is possible to extract and compare poses that are likely to appear in each case without giving correct labels.**

Keywords—tennis; video analysis; unsupervised method; feedback; pose estimation; joint coordinates; visualization;

I. INTRODUCTION

In recent years, the demand for sports video analysis has greatly expanded. By conducting video analysis of players and teams, it is possible to conduct scientific and it becomes possible to evaluate individuals or teams with the same index. It is desired to improve the competition level by feeding back to player analysis results visualizing the player's movement and ability.

As a research on which sports video is being analyzed, there is research to predict that the ball will be hit next time from the movement of the athlete and the orbit of the ball [2]. In this kind of research, we predict from the various information such as the movement of the player, the trajectory of the ball, the place where the play was performed, etc., predicted by the skilled top-class players forecasting intuitively, visualization doing.

Sports video analysis has also been widely used not only for analysis of players' movement and play but also assistance for judgment. An example used for assisting judgment is the system called "The Hawk-Eye Officiating System" by Hawk-Eye Company. This is introduced as a system that automatically judge in/out in tennis, and as a goal judgement system in soccer.

Video analysis required in the competition environment is a detailed analysis of play by myself or opponent. However, it has not been proposed to finely analyze the play performed focusing on posture state of the athlete in the video. This is because it was difficult to acquire the joint position information of the athlete from the image with only RGB information. Therefore, it is considered that a system that can understand the details of the athlete by conducting play analysis of the player from the sports video is necessary.

Therefore, in this research, we proposed a method for objective form analysis without using teacher labels by expert analysis. Realize a method that can evaluate the quality of play performed using observable information such as estimated joint position of the player and result of play.

In the rest of this paper, we will first describe the research related to this research in section 2, the processing procedure and method of the proposed method in section 3 in detail, the experiment that we performed in section 4 and summarize this paper in section 5.

978-1-5386-2615-3/18/$31.00 ©2018 IEEE

II. Related Work

As a research on sports image analysis that has been conventionally done, motion recognition method by moving region extraction using tennis image interframe difference has been proposed [3]. In this method, it is possible to detect shots such as forehand and backhand from positional information by detecting players and balls in the video using interframe difference. However, since the player is detected by performing the labeling process on the dynamic area, it is not possible to consider the attitude state of the athlete at all and the stage of evaluating the form when hitting the ball has not been reached.

As a research focusing on the posture state of athlete, research is also conducted to prepare a plurality of cameras to convert the posture state of the player to 3D, and to analyze what kind of motion each part performs when hitting the serve [4]. By measuring the speed of movement of the wrist or elbow, it is possible to classify the type of serve, such as flat serve and slice serve. However, in this research, the posture state is represented in three dimensions by measuring the players using eight cameras, and calibration is also required depending on the environment, so in order to understand the posture state of the athlete, it takes time and effort to arrange, there is a problem that the environment where the experiment can be done is restricted.

Therefore, in this paper, we estimate the posture state of players from tennis images with only RGB image information, and focus on attitude state transition and perform form analysis.

III. Proposed Method

In this research, we estimate the pose of the player in the video from the tennis image, and create the feature vector using the estimated joint position information. Using this feature vector, we predict the probability of play success. In addition, classification is performed by unsupervised classification method, and detailed form analysis is enabled by comparing appearance features appearing for each success probability.

A. Pose Estimation

As a method of estimating the joint position of a person existing in the image, there is a method of cascading classifiers [5,6]. However, in these methods, there is a constraint condition that there is only one person in the rectangle to be estimated, and it was necessary to combine with the object detection method such as [7,8]. In this research, we use the pose estimation method using Part Affinity Fields as a method to estimate the pose of the athlete [1]. In this method, it is possible to obtain the joint position coordinates of all the persons present in the image at high speed and accurately by detecting candidate parts of a person for all the regions in the image and considering their positional relationship. The parts that can be estimated are eye, ear, nose, neck, shoulder, elbow, wrist, hip, knee, ankle. Fig 2 shows the result of estimating the joint position of the athlete.

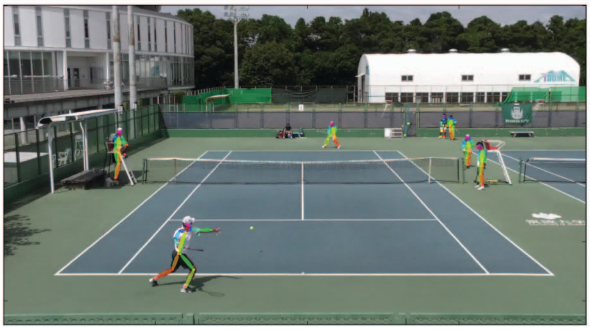

Fig 2. Pose Estimation Result

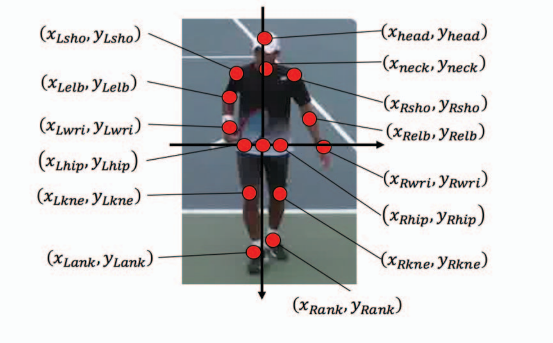

Fig 3. Relative Joint Position Coordinates

B. Creating Feature Vectors

We create feature vectors using joint position coordinates obtained by posture estimation. In order not to depend on the place where the athlete is located, the coordinates of the joint position of the player uses relative joint position coordinates with the pelvis position as the origin. At this time, the pelvic position is assumed to be the center point of the hip positions. Relative joint position coordinates with the pelvic position as the origin are as shown in Fig.3. Also, assuming that the athlete's height is the same for all players, we assume that the length of the spine from the pelvis to the neck is constant in all situations. The relative joint position coordinates are normalized by the length of the spine, and the feature vector at time t simply aligned as it is can be expressed as $\boldsymbol{j}_t$.

$$\boldsymbol{j}_t = [x_{head}, y_{head}, x_{neck}, \cdots, x_{Lank}, y_{Lank}]$$

The created feature vectors are clustered for each frame by GMM. Assuming that one shot is while the players is swing the racket, the posture class in the shot motion is represented by Bag of Words. When classified as K class by GMM, BoW is expressed by the following expression $\boldsymbol{x}$.

$$\boldsymbol{x} = [x_{k=0}, x_{k=1}, \cdots, x_{k=K}]$$

Next, in addition to the feature amount of the shot motion created above, the position information and movement amount are taken into consideration. For the position information and the movement amount of the athlete, the coordinate of the foot of the joint position coordinate acquired by the posture estimation is used. Regarding the position information, the play area that was at the moment of shot is represented by one-hot vector $\boldsymbol{l}$. The play area divides the tennis court into six areas as shown in Fig 4. In this paper, we divide it only in the horizontal direction

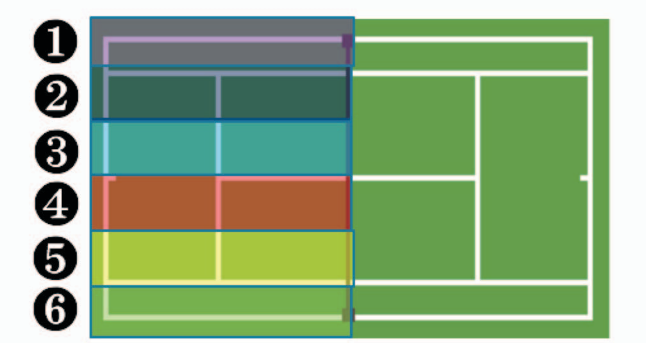

Fig 4. Divided Figure of Tennis Court

and don't distinguish shot positions in the front and back direction.

$$\boldsymbol{l} = [l_{p=1}, l_{p=2}, \cdots, l_{p=6}]$$

The amount of movement from start of the shot to the moment of the shot is represented by a vector $\boldsymbol{m}$.

$$\boldsymbol{m} = [\, x_{movement}, y_{movement} \,]$$

Based on BoW on posture information, the position information and the vector of the movement amount created above, a vector representing the feature of the shot is created for each shot. This feature vector X is expressed by the following equation.

$$\boldsymbol{X} = [\, \boldsymbol{x} \mid \boldsymbol{l} \mid \boldsymbol{m} \,]$$

C. Predict shot success probability

We predict the shot result from the created BoW. The shot result is assumed to be a score, a losing point, or a rally continuation. The probability of the shot result was predicted using SVM. For BoW of each shot, learning data was obtained by using one-hot vector of shot result as teacher data. When BoW representing the characteristic of the shot is input, the probability that the shot result becomes any can be predicted.

D. Visualization

The evaluation of each shot is visualized based on the prediction result of the shot result predicted above. The results of visualization are two comparisons of movement paths and posture classes appearing by prediction results.

Regarding the comparison of movement paths, how much the position of the ankle moved was mapped on the tennis court based on the acquired joint position information of the athlete. By doing this, we confirm the relationship between the success rate of the shots and the distance moved until hitting the ball.

Regarding the comparison of the appearing posture classes, BoW indicating the characteristics of shots is listed for each prediction result, and the difference in appearance frequency of posture classes for each shot is represented by a colormap. By doing this, we confirm the relationship between the success rate and composition of posture class in shot.

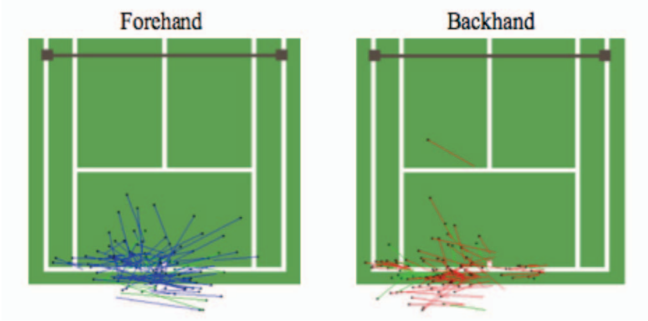

Fig 5. Movement Path during Play

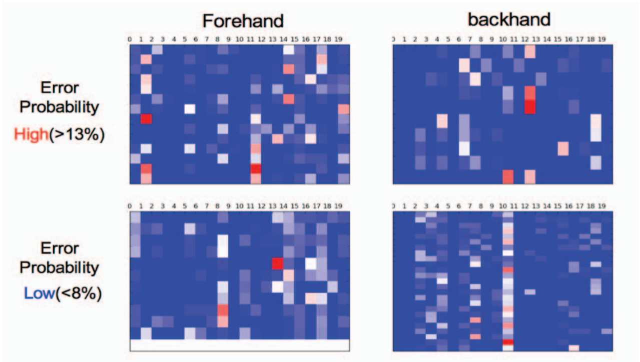

Fig 6. Form Characteristics Visualization Results

IV. Experimental Results

A. Outline of Experiment

The dataset used in the experiment is a match video of the athlete recorded with the cooperation of the University of Tsukuba tennis team. The resolution of the match video is 1920 × 1080, and the frame rate is 30 fps. There are images in which multiple players are performing serve, forehand, backhand, and standby state in this dataset. In this time, we will focus on only the singles if the right-handed player in the dataset and carry out the experiment. The total number of frames of the images is about 44000 frames. The number of shots performed is about 160 times.

We extracted the joint position coordinates of the athlete by posture estimation to those video. When classifying the joint position coordinates by GMM, we decided experimentally the number of classes as 20 classes. Also, one of labels Error, Winner, Continuation is given to the feature vector showing the characteristic of one shot. We divided this dataset into learning data and test data, and conducted experiments to evaluate the form.

B. Experimental Result

Figure 5 shows the route that the player moved before the shot and position at the time shot. The path when the blue and red lines show a path when the probability that a shot result becomes Error is less than 15%, the green lines show path with probability of Error of 15% or more, and the black circles show position at shot. From this result, it is possible to visualize what kind of features are in the route when mistakes are likely to occur. It can be confirmed that mistakes are likely to occur, such as when the amount of movement is large or when moving back and forth in the case of backhand is large.

In Figure 6, BoW representing posture characteristics of shots is listed for each prediction result, and it is possible to confirm in what posture the shot is performed. The blue parts show posture classes that don't appear, the classes that appear slightly are white, and the most appearing posture classes are shown in red. In this figure, in the forehand and backhand, the failure probability of the shot is high when it is 13% or more, and 8% or less, respectively. Also, the vertical axis shows the number of shots, and the horizontal axis shows posture classes classified by GMM.

Looking at the table when the Error rate of forehand is high, there are about two posture classes that appear with high probability in every shot. When you fail the shot, you can see that these postures are appearing a lot. Also, when looking at the case where the backhand Error rate is low, posture classes with obvious appearance frequencies exist. This class hardly appears when the Error rate is high. In other words, in order to hit a shot that is difficult to mistake in backhand, it is important to hit in this posture class.

As described above, shot results are predicted from BoW created by unsupervised classification of joint position coordinates, and the experiment of evaluating the shot from the comparison of BoW was performed based on the results. From this experiment, it was confirmed that there is a difference in posture appearing depending on the success rate of shots, and it can be extracted from only observable information.

V. Conclusion

In this paper, we evaluated shots played by athletes in tennis video. By analyzing using only observable information such as joint position and shot results, we could analyze without qualitative opinion of experts. From these analysis results, it was confirmed that there is a tendency for posture appearing due to the difference in success rate of shots to be different.

Future prospect is to expand not only to form analysis but also to tactical analysis by considering the relation with the shots before and after. By proposing an extended method, it is thought that it will become possible to practically analyze play of athletes.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", arXiv preprint arXiv:1611.08050, 2016

[2] Xinyu Wei, Patrick Lucey, Stuart Morgan, Sridha Sridharan, "Forecating the Next Shot Location in Tennis Using Fine-Grained Spatiotemporal Tracking Data", IEEE Transaction on Knowledge and Data Engineering 28.11 (2016): 2988-2997.

[3] Chihiro Antoku, Masayuki Kashima, Kiminori Sato, Mutsumi Watanabe, "Research for automatic tennis play recognition and recording based on motion analysis", IPSJ SIG Technical Report, 2013.

[4] Alison L. Sheets, Geoffrey D. Abrams, Stefano Corazza, Marc R. Safran, Thomas p. Andrriacchi, "Kinematics Differences Between the Flat, Kick, and Slice Serves Measured Using Markerless Motion Capture Method", Annals of biomedical engineering 39.12 (2011): 3011-3020.

[5] Varun Ramakrishna, Daniel Munoz, Martial Hebert, J. Andrew Bagnell, Yaser Sheikh, "Pose Machine: Articulated Pose Estimation via Inference Machines", European Conference on Computer Vision. Springer International Publishing (2014).

[6] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, Yaser Sheikh, "Convolutional Pose Machines", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016).

[7] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", arXiv preprint arXiv: 1506.01497 (2015).

[8] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg, "SSD: Single Shot MultiBox Detector", European Conference on Computer Vision (2016).