CEUR-WS.org/Vol-2882/paper17.pdf

Leveraging Human Pose Estimation Model for Stroke Classification in Table Tennis

Soichiro Sato[1], Masaki Aono[1]
[1]Toyohashi University of Technology, Japan
s-sato@kde.cs.tut.ac.jp,masaki.aono.ss@tut.jp

ABSTRACT

In this paper, we propose a stroke classification method for table tennis, submitted to MediaEval 2020 Sports Video Classification: Classification of Strokes in Table Tennis. The main focus of this paper is on the exploitation of features extracted from a pose estimation model in stroke classification. Specifically, we first introduce an original method that incorporates PoseNet. Then, we construct a DNN model based on our proposed method. Subsequently we evaluate our stroke classification using unannotated unknown data. Finally, we analyze the proposed method from the classification results. The results demonstrate that the classification accuracy of the proposed method outperforms the baseline by 4.8%.

1 INTRODUCTION

In recent years, research on video action recognition using a DNN model has gained the popularity. Considering this popularity, it is natural to think of applying video action recognition to a variety of sports fields such as athletes action analysis and creation of educational videos for the sports. The datasets used in video action recognition include UCF-101 [8] and Kinetics [1]. These datasets are classified typically by types of sports and human action. On the other hand, the stroke classification of table tennis in MediaEval2020 [3] requires the stroke classification within a single sport. Therefore, it is a difficult task due to the higher degree of similarity between the classes than usual general datasets. RGB and Optical Flow have been often used for the input to the DNN model for video action recognition [2, 7, 9]. We speculate that the features extracted from a DNN model, which enables posture estimation from images and movies, could be used for stroke classification. In this paper, from the above observations and speculation, we propose a stroke classification method for table tennis based on features extracted from the posture estimation model.

2 APPROACH

In this paper, we leverage PoseNet [5], which one of the posture estimation models. PoseNet can estimate a total of seventeen different skeletal coordinates including a person's wrist, elbow, shoulder, and knee by inputting RGB images. By applying this method, it is possible to create a time series data of human skeletal coordinates in the video. It is also possible to determine the position of a human in the video frame based on the estimated coordinates of the skeleton, which can be utilized to determine the crop position. In this paper, we define Pose Time Series Data as time series data representing the transition of human skeletal coordinates in a video.

2.1 Steps to create Pose Time Series Data

First, we input T video frames into the pre-trained PoseNet and extract features with the number of dimensions (T,17,2) that represent the estimated coordinates (x,y) of the seventeen different skeletons. Next, we pre-process the extracted features to input them into the DNN model. The preprocessing of the extracted features uses four methods: transformation from absolute coordinates to relative coordinates, computation of moving average, normalization, and zero padding. Here, the transformation from absolute coordinates to relative coordinates is based on the estimated coordinates of the skeleton in the first frame of the video. If the player is not visible on the first frame of the video, the transformation is based on the center coordinates in the first frame of the video. The Pose Time Series Data created by the above procedure is used as input of the DNN model described in section 2.3.

2.2 Crop based on skeletal coordinates

When a video frame is fed into the DNN model, it is pre-processed by cropping video frame at the size of 120 × 120. For video action recognition using a DNN model, we could employ cropping methods such as Center Crop and Random Crop. However, if these methods are used to crop the video frame, the person who is actually performing the table tennis action will not be included in the cropped area, which will increase the risk of not being able to classify strokes correctly. Therefore, we take advantage of PoseNet ability to estimate seventeen different skeletal coordinates, and compute crops based on the estimated skeletal coordinates. Specifically, after inputting a video frame into PoseNet and obtaining seventeen different skeletal coordinates, the average value of their x-coordinates and y-coordinates is calculated. Then we crop the frame at the size of 120 × 120, with the coordinates calculated as the center position of the crop. The video frame cropped by the above procedure is used as input of the DNN model described in the section 2.3.

2.3 Model

In this task, we have implemented five different DNN models because it allows us to submit up to five runs. First, we have reproduced the SSTCNN [4]. This model served as the baseline model used for performance comparison with the proposed method. Next, we have developed a DNN model in which Pose Time Series Data is inputted and the part that performs 1D convolution is added to SSTCNN. The inputs to this model are three types of data: RGB, Optical Flow, and Pose Time Series Data. RGB used as input for the DNN model is cropped according to the method described in

Copyright 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
MediaEval'20, December 14-15 2020, Online

Table 1: Training Hyperparamater

Hyperparamater	Value
Optimizer Method	SGD
Learning Rate	0.002
Momentum	0.5
Decay	0.005
Loss Function	Categorical Cross Entropy
Epoch	300
Batch Size	8

section 2.2. Optical Flow used as input for the DNN model is a combination of two time-consecutive video frames created by DeepFlow [10] and a background difference proposed by Zivkovic et al [11]. This allows us to filter out only the locations where the change is presumed to have occurred between two consecutive frames in time. In addition, we have enhanced the DNN model so that it incorporates a Depthwise Separable Convolution [6] in the convolution layer of Pose Time Series Data, free of Optical Flow for input. These models aim to reduce the number of parameters in the model, as the number of parameters in the training model increases with the type of data for inputs. The differences between the five DNN models for our submitted runs are denoted by ①, ②, ③ which is also delineated in Table 2. Here, RGB is used for all of DNN models.

① : Include Optical Flow in the input
② : Include Pose Time Series Data in the input
③ : Introduce Depthwise Separable Convolution

2.4 Training and Submission Runs

The models are trained by the hyperparameters shown in Table 1 for five different runs with our DNN models. The dataset [3] consists of short movie clips of table tennis strokes practice. The training dataset includes 755 actions and the test dataset 354 actions. During the training, we have not set up validation data. Instead, we have used all 755 training data solely for training the model. After training, we have fed the test data into the trained model and performed stroke classification.

3 RESULTS AND ANALYSIS

The classification results of the submitted runs are shown in Table 2. The left column shows the name of the submitted runs. The middle column shows the differences in the models corresponding to the submitted runs by means of a checklist. The right column shows the classification results corresponding to the submitted runs. In addition to the results of the 20 classes of table tennis strokes, the results of a rough classification of strokes are shown in the columns 'Hand', 'Serve' and 'H&S' (Hand and Serve). Table 2 demonstrates that Run 5 turned out to be the model with the highest Global Accuracy, but Run 2 turned out to be the model with the highest classification accuracy in the case of a rough stroke classification. The Confusion Matrix of Run 5 is shown in Figure 1. From Figure 1, it is observed that there are several classes that could be accurately categorized in the test data, such as 'Offensive Forehand Flip'. On the other hand, it is also observed that the detailed stroke classification is not exactly accurate. In particular, the classification of details in table tennis strokes has been often misclassified, such as the difference in spin on the table tennis ball when a player performs a stroke. It is also possible that the lack of data augmentation when training the model may lead to an inaccurate stroke classification due to overfitting.

Table 2: Runs Results

	Components			Accuracy [%]			
	①	②	③	Hand	Serve	H&S	Global*
Run 1	✓			81.64	57.34	53.67	11.86
Run 2		✓		**81.92**	**62.99**	**57.06**	16.10
Run 3		✓	✓	81.36	56.78	52.26	14.12
Run 4	✓	✓		79.66	57.91	52.54	13.56
Run 5	✓	✓	✓	79.66	58.19	52.82	**16.67**

* Global : Global Accuracy

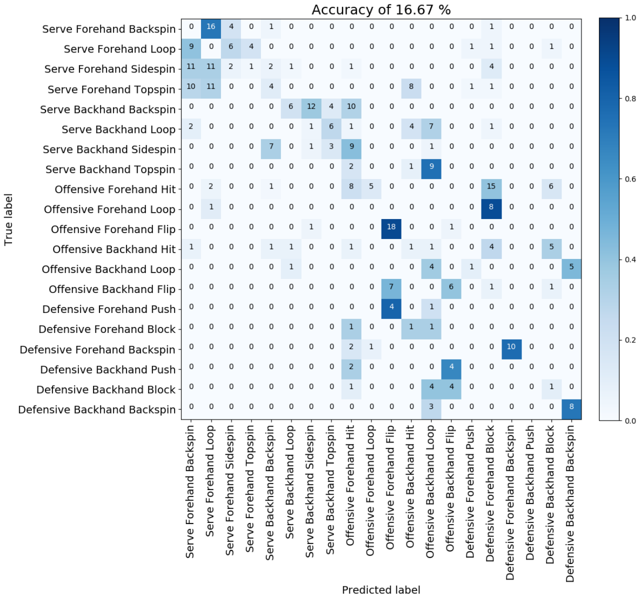

Figure 1: Confusion Matrix (KDEME Run 5)

4 DISCUSSION AND OUTLOOK

In this paper, we proposed a stroke classification method based on PoseNet. We have implemented five different DNN models and trained them to classify table tennis strokes with the test data. The results exhibited that the classification accuracy of the proposed method was up to 4.8% higher than the baseline. However, we have not been able to classify them accurately and found that there is still a room for improvement. In the future, we would like to verify the accuracy of data augmentation and explore methods to improve the accuracy of table tennis stroke classification from various perspectives, such as data preprocessing method and DNN model architecture.

REFERENCES

[1] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 4724–4733. https://doi.org/10.1109/CVPR.2017.502

[2] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional Two-Stream Network Fusion for Video Action Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 1933–1941. https://doi.org/10.1109/CVPR.2016.213

[3] Pierre-Etienne Martin, Jenny Benois-Pineau, Boris Mansencal, Renaud Péteri, Laurent Mascarilla, Jordan Calandre, and Julien Morlier. 2020. Sports Video Classification: Classification of Strokes in Table Tennis for MediaEval 2020. In *Proc. of the MediaEval 2020 Workshop, Online, 14-15 December 2020*.

[4] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. 2020. Fine grained sport action recognition with Twin spatio-temporal convolutional neural networks. *Multim. Tools Appl.* 79, 27-28 (2020), 20429–20447. https://doi.org/10.1007/s11042-020-08917-3

[5] Dan Oved, Irene Alvarado, and Alexis Gallo. 2018. Real-time Human Pose Estimation in the Browser with TensorFlow.js. (2018). https://blog.tensorflow.org/2018/05/real-time-human-pose-estimation-in.html.

[6] Laurent Sifre and Stéphane Mallat. 2014. Rigid-motion scattering for image classification. *Ph. D. thesis* (2014).

[7] Karen Simonyan and Andrew Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 568–576.

[8] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *CoRR* abs/1212.0402 (2012). arXiv:1212.0402 http://arxiv.org/abs/1212.0402

[9] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII (Lecture Notes in Computer Science)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.), Vol. 9912. Springer, 20–36. https://doi.org/10.1007/978-3-319-46484-8_2

[10] Philippe Weinzaepfel, Jérôme Revaud, Zaïd Harchaoui, and Cordelia Schmid. 2013. DeepFlow: Large Displacement Optical Flow with Deep Matching. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. IEEE Computer Society, 1385–1392. https://doi.org/10.1109/ICCV.2013.175

[11] Zoran Zivkovic and Ferdinand van der Heijden. 2006. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognit. Lett.* 27, 7 (2006), 773–780. https://doi.org/10.1016/j.patrec.2005.11.005