HCMUS at MediaEval 2021: Ensembles of Action Recognition Networks with Prior Knowledge for Table Tennis Strokes Classification Task

Trong-Tung Nguyen[1,3], Thanh-Son Nguyen[1,3], Gia-Bao Dinh Ho[1,3], Hai-Dang Nguyen[1,3], Minh-Triet Tran[1,2,3]
[1]University of Science, VNU-HCM, [2]John von Neumann Institute, VNU-HCM
[3]Vietnam National University, Ho Chi Minh city, Vietnam
{ntrtung17,dhgbao}@apcs.fitus.edu.vn,{nthanhson,nhdang}@selab.hcmus.edu.vn,tmtriet@fit.hcmus.edu.vn

ABSTRACT

The SportsVideo task of the MediaEval 2021 benchmark is made up of two subtasks: stroke detection and stroke classification. For the detection task, participants are required to find some specific frame intervals broadcasting the strokes of interest. Subsequently, this can be utilized as a preliminary step for classifying a stroke that has been performed. This year, our HCMUS team engaged in the challenge with the main contribution of improving the classification, aiming to intensify the effectiveness of our previous method in 2020. For our five runs, we proposed three different approaches followed by an ensemble stage for the two remaining runs. Eventually, our best run ranked second in the Sports Video Task with 68.8% of accuracy.

1 INTRODUCTION

In the Multimedia Evaluation Challenge 2021, there are two main sub-tasks: detection and classification. Specifically, the latter specifies video boundaries as inputs to perform classifying stroke categories. About the dataset, strokes are categorized into the same 20 classes as that of last year, with an addition of new and more diverse samples [6].

We conducted three experiments with different model architectures and submitted five runs in total. Generally, the first, second, and fifth runs were independent methods. Turning to the other versions, the third run is the ensemble of the first and the fifth runs, while the first and second runs are used for ensembling the fourth run. For the first run, we employed a rudimentary method to handle video classification by spatially stacking images in a video sequence to form a super image, as such a simple idea is proven to be efficient in [8]. The second run was delegated to a more systematic approach. We decomposed the problem into three branches of classification problem with the help of multi-task learning. This aims to inject relevant features and human biases into each branch independently. For the fifth run, we continued to employ our previous approaches [7] with some modifications. Our post-processing stages were modified to a more general scenario with the help of conditional probabilities and prior knowledge to eliminate the sensitive outcomes of classification models.

2 METHOD

2.1 Run 01

In this run, we stacked images in sub-clips spatially to create a super image with size $N \times N$ as a representation for the full clip and treat the video classification task as an image classification problem. After that, a classification head was used for making prediction about stroke categories.

2.2 Run 02

In this run, we decomposed the original classification problem into three sub-classification branches, with the sub-categories split for each classifier described in Table 1. This mechanism was based on our motivation to disentangle the existing ambiguity of the raw labels. It would be more relevant to discriminate among serve, offensive, and defensive strokes rather than serve, forehand, and backhand types. Moreover, our proposed sub-categories classification method by breaking the raw labels into many sub-classes can supplement more training samples for each category in the classifiers, as the collection of some strokes in table tennis are still limited. Eventually, each classifier utilized both shared and exclusive features useful for the corresponding tasks.

The first and third components utilized the shared features f_{shared_13} which were constructed by performing concatenation between the temporal visual features and temporal pose features. A 3D-CNN architecture implemented by [5] was employed for extracting the temporal visual features $f_{visual\_3DCNN_1}$, given an image with shape $H \times W \times C$. On the other hand, the temporal pose features $f_{temporal\_pose}$ were the results of providing 17 human key points of multiple frames successively to an LSTM architecture. Initially, we performed sampling F frames with a strategy for ensuring the consistency of keypoint extracted in video sequences. Key points were represented by two coordinate values, which results in 34 different values for a specific pose. The first and third components were paired to use similar features due to their similarity in visual appearance and might use the same sources of information for predicting sub-categories.

However, another significant feature should be incorporated when handling with the third classifier (Forehand, Backhand). We first performed cropping the original image based on the boundaries of the hands' region, which can be extracted by selecting the coordinates of key points that satisfy a plausible position for human hands. After that, the concatenation of two hand images of shape $H_1 \times W_1 \times C$ were supplied into a different 3D-CNN branch

MediaEval'21, December 13-15, 2021, Online

Classifier type	Categories	# Prediction Heads
First Component	Serve, Offensive, Defensive	3
Second Component	Forehand, Backhand	2
Third Component	Backspin, Loop, Sidespin, Topspin, Hit, Flip, Push, Block	8

Table 1: Three splitted sub-categories for three classifier types

to produce another temporal visual hand feature for the third classification branch

After that, three multi-layer perceptrons MLP_i (1) were designed for each branch of classification with different number prediction heads shown in Table 1. The loss function of each branch was then aggregated for serving the final multi-task learning loss $\mathcal{L}$.

$$\hat{p_i} = Softmax(MLP_i(f_i)) \tag{1}$$

Finally, we formulated the joint probabilities $P(c_1, c_2, c_3)$ (2) of predicting three independent sub-categories using prior knowledge. By conducting a thorough analysis about the co-existence of three sub-categories, we concluded that the existence of the second component label was independent of the first and third component label. On the other hand, it was possible to narrow down plausible labels of the third component given the prior knowledge about the categories of the first component. In Table 2, we summarize the relation of existence between the first and third components that we have investigated so far.

$$\begin{aligned} P(c_1, c_2, c_3) &= P(c_3, c_1 | c_2) \cdot P(c_2) \\ &= P(c_3, c_1) \cdot P(c_2) \\ &= P(c_3, c_1) \cdot \hat{p}_{2c_2} \end{aligned} \tag{2}$$

The second term can be referred to the c_2^{th} value of $\hat{p_2}$ (1) of the second classifier. Meanwhile, the first term $P(c_3, c_1)$ (3) was factorized into two terms.

$$\begin{aligned} P(c_3, c_1) &= P(c_3 | c_1) \cdot P(c_1) \\ &= \hat{p}_{refined_3 c_3} \cdot \hat{p}_{1c_1} \end{aligned} \tag{3}$$

Given the prior knowledge tables, we first construct a binary referenced matrix $M \in R^{3\times 8}$, which encodes the co-existence of labels between the first and third component. Then, we perform Hadamard product on the two vectors $M_{g(c_1)}$ $\in$R$^{1\times 8}$ (where $g(c_1)$ = {0, 1, 2} represents the true index of c_1) and $\hat{p_3}$ $\in$R$^{1\times 8}$ to produce the refined probability $\hat{p}_{refined_3}$ $\in$R$^{1\times 8}$ (4). Finally, it is normalized before being multiplied with the c_1^{th} value of $\hat{p_1}$ (1):

$$\hat{p}_{refined_3} = M_{g(c_1)} \bigodot \hat{p_3} \tag{4}$$

$$\begin{aligned} P(c_3, c_1) &= P(c_3 | c_1) \cdot P(c_1) \\ &= \frac{\hat{p}_{refined_3 c_3}}{\sum_{i=1}^{n=8} \hat{p}_{refined_3 i}} \cdot \hat{p}_{1c_1} \end{aligned} \tag{5}$$

Prior Knowledge about First Component	Possible sets of labels for Third Component
Serve	Backspin, Loop, Sidespin, Topspin
Offensive	Hit, Loop, Flip
Defensive	Push, Block, Backspin

Table 2: Prior Knowledge tables

2.3 Run 05

We made a small modification on the second run by replacing our designed classifier with a more powerful model architecture for the action recognition problem, which we have utilized last year [1, 7]. Similarly, three different classifiers produced the outputs independently which were then combined to get the final results with our conditional probability using the prior knowledge mechanism demonstrated.

3 EXPERIMENTS AND RESULTS

In the first run, the final score is the average score of two sub-clips in the video. All of the images were resized to shapes of 224×224. We passed the super image to the ResNet-50 [4] backbone, followed by a global average pooling layer to get a 2048-dimension vector. For each video, we sampled two sub-clips separated by five frames, with 16 images per sub-clips. Random flip, color jittering, and random augmentation [3] are also used with the default settings in MMAction2 [2]. We trained our model in this run using the focal loss [9] to handle the data imbalance problem. In the second run, we passed video sequences with 30 samples of frame interval with a shape of 120 × 120 to the shared network (the first and third classification branch). Meanwhile, two hand images were cropped and concatenated as shape of 120 × 240 before feeding the third classifier. In the fifth run, we utilized the parameters similar to our previous methods [7] for each classifier. For the ensemble versions, highest confidence scores were returned as final results.

Run ID	Run 1	Run 2	**Run 3**	Run 4	Run 5
Accuracy	61.99%	44.80%	**68.78%**	60.63%	67.87%

Table 3: HCMUS Team Submission results for Table Tennis Stroke Classification Task

4 CONCLUSION AND FUTURE WORKS

Conclusively, we benchmarked various different approaches on the video classification task for table tennis at MediaEval benchmark 2021. Furthermore, one of our submissions achieved the second-best result in terms of global accuracy, which is 68.78%. Future works should be considered on the analysis of features selection and semantic of the raw labels for modeling the action in the table tennis domain, with the help of human pose and prior knowledge information.

ACKNOWLEDGMENTS

This work was funded by Gia Lam Urban Development and Investment Company Limited, Vingroup and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2019.DA19.

REFERENCES

[1] MMPose Contributors. 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose. (2020).

[2] MMAction2 Contributors. 2020. OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. https://github.com/open-mmlab/mmaction2. (2020).

[3] Jonathon Shlens Quoc V. Le Ekin D. Cubuk, Barret Zoph. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 702–703.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. 770–778. https://doi.org/10.1109/CVPR.2016.90

[5] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Peteri, and Julien Morlier. 2020. Fine grained sport action recognition with Twin spatio-temporal convolutional neural networks: Application to table tennis. *Multimedia Tools and Applications* 79 (07 2020). https://doi.org/10.1007/s11042-020-08917-3

[6] Pierre-Etienne Martin, Jordan Calandre, Boris Mansencal, Jenny Benois-Pineau, Renaud Péteri, Laurent Mascarilla, and Julien Morlier. 2021. Sports Video: Fine-Grained Action Detection and Classification of Table Tennis Strokes from videos for MediaEval 2021. (2021).

[7] Hai Nguyen-Truong, San Cao, N. A. Khoa Nguyen, Bang-Dang Pham, Hieu Dao, Minh-Quan Le, Hoang-Phuc Nguyen-Dinh, Hai-Dang Nguyen, and Minh-Triet Tran. 2020. HCMUS at MediaEval 2020: Ensembles of Temporal Deep Neural Networks for Table Tennis Strokes Classification Task. In *Working Notes Proceedings of the MediaEval 2020 Workshop, Online, 14-15 December 2020 (CEUR Workshop Proceedings)*, Steven Hicks, Debesh Jha, Konstantin Pogorelov, Alba García Seco de Herrera, Dmitry Bogdanov, Pierre-Etienne Martin, Stelios Andreadis, Minh-Son Dao, Zhuoran Liu, José Vargas Quiros, Benjamin Kille, and Martha A. Larson (Eds.), Vol. 2882. CEUR-WS.org. http://ceur-ws.org/Vol-2882/paper50.pdf

[8] Rameswar Panda Quanfu Fan, Chun-Fu (Richard) Chen. 2021. An Image Classifier Can Suffice For Video Understanding. (06 2021).

[9] Ross Girshick Kaiming He Piotr Dollar Tsung-Yi Lin, Priya Goyal. 2017. Focal Loss for Dense Object Detection. In *ICCV*.