HCMUS at MediaEval 2020: Ensembles of Temporal Deep Neural Networks for Table Tennis Strokes Classification Task

Hai Nguyen-Truong*[1,3], San Cao*[1,3], Khoa N. A. Nguyen*[1,3], Bang-Dang Pham*[1,3], Hieu Dao*[1,3], Minh-Quan Le*[1,3], Hoang-Phuc Nguyen-Dinh*[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
{nthai18,ctsan18,nnakhoa18,pbdang18,dhieu}@apcs.vn
{lmquan,ndhphuc,nhdang}@selab.hcmus.edu.vn,tmtriet@fit.hcmus.edu.vn

ABSTRACT

The Sports Video Classification Tasks in the Multimedia Evaluation 2020 Challenge focuses on classifying different types of table tennis strokes in video segments. In this task, we - the HCMUS Team - perform multiple experiments, which includes a combination of models such as SlowFast, Optical Flow, DensePose, R2+1, Channel-Separated Convolutional Networks, to classify 21 types of table tennis strokes from video segments. In total, we submit eight runs corresponding to five different models with different sets of hyper-parameters in each of our models. In addition, we apply some pre-processing techniques on the dataset in order for our model to learn and classify more accurately. According to the evaluation results, one of our team's methods out-performs all other teams. In particular, our best run achieves 31.35% global accuracy.

1 INTRODUCTION

In the Multimedia Evalutaion Challenge 2020 (MediaEval2020), one of the tasks is classification table tennis strokes in video segments [8]. In the task, the authors conduct experiments on the TTStroke-21 dataset [9]. The dataset consists of 20 table tennis stroke classes, combining 8 kind of services, 6 offensive strokes, and 6 defensive strokes. In addition, there is a class named "Unknown" for identifying the video segments without any activity or stroke.

We implemented five runs independently in order to benchmark different methods, and we conduct experiments on distinct sets of our augmented / pre-processed dataset. Thus, we describe the five runs in Section 2 and 3.

2 APPROACH

By examining the videos of training and test set, we realize that the context around the table tennis player in each video is not important, and we desire our models to solely concentrate on learning the action of the player. Therefore, we propose a way to remove the background around the player by the following technique.

2.1 Data pre-processing with DensePose for Background Removal

Particularly, for both train and test set, we utilize the DensePose model [1] to extract the mask from the person in each frame of the video sequence. Then, we extend the mask to its local region to capture minor context around using binary dilation, and we blur the mask inside-out by the Gaussian filter with suitable parameters. After that, we multiply the created mask with the original frame to get a new frame showing just the "biggest" player. In case the mask obtained from DensePose from a frame is too small in area (smaller than a pre-defined threshold - 5 percent of the area of the image in our experiment), we do not modify that frame. After this step, we have videos that only concentrate on showing the players. We are still unable to process the case when DensePose detects more than one player in a frame. Besides, we also employ simple data augmentation methods on the video segments such as rotation, translation, flip to get more relevant data. The background removal process is shown in the Figure 1.

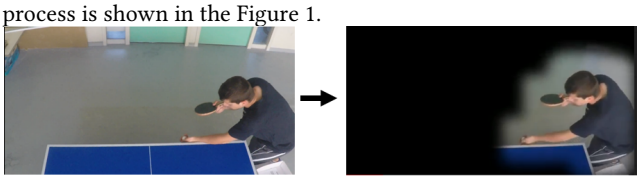

Figure 1: Background Removal Process

2.2 Late Temporal Modeling in 3D CNN Architectures with BERT

Late Temporal Modeling in 3D CNN Architecrtures (LateTemporal3DCNN) with BERT for Action Recognition [6] is a method combining 3D convolution with late temporal modeling for action recognition. The paper replaces the conventional Temporal Global Average Pooling (TGAP) [7] layer at the end of 3D convolutional architecture with the Bidirectional Encoder Representations from Transformers (BERT) [3] layer in order to better utilize the temporal information with BERT's attention mechanism.

2.3 Channel-Separated Convolutional Networks (CSN)

Channel-Separated Convolutional Networks [11] was first introduced by Facebook AI in ICCV 2019. The paper emphasized the important role of the amount of channels interaction in the accuracy of 3D group convolutional networks. All of convolutional operations are separated into either pointwise $1 \times 1 \times 1$ or depthwise $3 \times 3 \times 3$ convolutions. That change not only reduces the computational cost but also improves the accuracy significantly.

2.4 Twin Spatio-Temporal Convolutional Neural Networks (TSTCNN)

In this task, we also use the Twin Spatio-Temporal Convolutional Neural Networks (TSTCNN) [10] and conduct experiments on it with our minor adjustments to classify the fine-grained sports actions. To extract the useful information, we compute the optical flow values of each video frame by Lucas-Kanade method, then

MediaEval'20, 14-15 December 2020, Online

Classifier number	Videos with labels
(1)	Forehand, Backhand
(2)	Defensive, Offensive, Serve
(3)	Offensive Hit, Offensive Flip, Offensive Loop
(4)	Defensive Push, Defensive Block, Defensive Backspin
(5)	Serve Forehand Topspin, Serve Forehand Sidespin, Serve Forehand Loop, Serve Forehand Backspin
(6)	Serve Backhand Topspin, Serve Backhand Sidespin, Serve Backhand Loop, Serve Backhand Backspin

Table 1: Six classifiers for six corresponding set of labels

utilize that information for estimating the Region Of Interest (ROI). The idea is for our model to concentrate on the stroke motion region for learning the unique feature of each class. And then, we feed the sequence of cropped RBG images and its corresponding normalized (by min-max normalization) optical flow values into the TSTCNN, which consists of a spatio-temporal CNN siamese network. We inherit the TSTCNN architecture with 3 spatio-temporal convolutional layers and a fully connected layer in each branch.

3 EXPERIMENTS AND RESULTS

3.1 First run - Run 03

For this run, we use CSN method (mentioned in Section 2.3) without modified as the baseline to demonstrate for the method. We use Resnet3D architecture as our backbone, with I3D heads [2] as classification part. We also disable batch norm operations because it leads to a higher accuracy in overall. At the result, we achieve 86.9% on our validation dataset and 28.81% on the test dataset.

3.2 Second run - Run 04

In this run, we use LateTemporalModeling3DCNN method (mention in Section 2.2) combined with several models to inspect the effectiveness of the method. The used methods are RGB ResNeXt101 [5] and RGB ResNeXt101 with BERT, RGB SlowFast50 [4] (derived from ResNet50) and RGB SlowFast50 with BERT, and RGB R(2+1)D [12]. All models use 64-frame length except the RGB R(2+1)D uses 32-frame length because we want to keep the configuration from [6]. Initially, we accidentally set the number of classes to be 51 since we try to configure the dataset to be the same as the HMDB51 and the RGB ResNeXt101 gives the best result of this run (we achieve 87.9% on our validation dataset and 25.42% on the test dataset). However, when we fix the number of classes to be 20 - the actual one - and use the more complex backbones (even with BERT architecture), the results are not as good as the initial one.

3.3 Third run - Run 06

In run 6, we use multi-video classification models based on the SlowFast Network [4] on the background removal video frames. Particularly for the training phase, we train six different classifiers with six different sets of videos, shown in Table 1.

All of the six classifiers are SlowFast Network with the ResNet50 backbone. In the inference phase, we first predict the person playing Forehand or Backhand stroke, then Serve, Offensive, or Defensive stroke. After that, based on this prediction, we choose the model to infer the remaining part of the stroke. Experiments show that with this method our models can recognize forehand, backhand, and serve with high precision.

3.4 Fourth run - Run 07

Inherited from the impressive performance of CSN method (run 03), we modify the model to solve multi-label classification task. By our observation, the 20 classes can be split into three separate labels as following, Offensive/ Defensive/ Serve, Forehand/ Backhand, and Loop/ Backspin/ Sidespin/ Topspin/ Hit/ Push/ Flip/ Block. That idea makes our model learn the partial labels and reduce the confusion of the similarity in the 20 classes. Instead of using Cross Entropy Loss, we use Binary Cross Entropy with Logits Loss to demonstrates the score of each class.

After the training phase, we post-process the predictions in each video. Among the predictions which have positive scores, if the labels of the top 3 highest scores exist (present in the 20 classes), then we use 3 of them as the *reliable results*. In case the predicted label is not in the 20 labels, we take top 5 (either positive or negative), and the combination which has the highest total score is chosen. The results is *unreliable* and need to take into consideration in the *ensemble process*. Using multi-label classification, we achieve 97% mAP ($\approx$ 89.51% top 1 score) on the validation dataset. Another key idea of the run 07 is that, we try to ensemble it with *run 03* on the *Serve* activities.

3.5 Fifth run - Run 08

In this run, we consider this task as a multi-label classification problem and we design our pipeline to classify each of the videos to multi-labels. We split the combined original label into multi-label as in run07 but there is a minor difference, for instance, Defensive Backhand Backspin is split into Defensive Backhand and Backspin. Our pipeline consists of three modified TSTCNN models (mentioned in Section 2.4) with the same architecture and their outputs are two spitted labels and the original label, respectively.

3.6 Results

Table 2 shows the results of our 5 runs in term of accuracy.

Run ID	Run 3	Run 4	Run 6	**Run 7**	Run 8
Accuracy	28.81%	25.42%	27.96%	**31.35%**	25.42%

Table 2: HCMUS Team Submission results for Table Tennis Stroke Classification Task

4 CONCLUSION AND FUTURE WORKS

In conclusion, we benchmark many different approaches on the manipulated TTStroke dataset during the MediaEval Challenge 2020. One of our submissions achieve the best result in term of global accuracy, which is 31.35%, compared to the submissions of all other teams. For the future work, we aim to extract human 3D mesh-based from each frame of the videos in order to have better classification results. The mesh could be rotated in different angles which helps our model to learn more efficiently.

ACKNOWLEDGMENTS

Research is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA19. We would like to give a special thank to Mr. Huu-Quoc Hoang (Ho Chi Minh city University of Technology), who supports us in examining the dataset and gives us advices.

REFERENCES

[1] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7297–7306.

[2] Joao Carreira and Andrew Zisserman. 2018. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. (2018). arXiv:cs.CV/1705.07750

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[4] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*. 6202–6211.

[5] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6546–6555.

[6] M Kalfaoglu, Sinan Kalkan, and A Aydin Alatan. 2020. Late Temporal Modeling in 3D CNN Architectures with BERT for Action Recognition. *arXiv preprint arXiv:2008.01232* (2020).

[7] Min Lin, Qiang Chen, and Shuicheng Yan. 2013. Network in network. *arXiv preprint arXiv:1312.4400* (2013).

[8] 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*.

[9] Pierre-Etienne Martin, Jenny Benois-Pineau, Boris Mansencal, Renaud Péteri, Laurent Mascarilla, Jordan Calandre, and Julien Morlier. 2019. Sports Video Annotation: Detection of Strokes in Table Tennis task for MediaEval 2019. In *MediaEval 2019 Workshop*.

[10] Renaud Péteri Julien Morlier Pierre-Etienne Martin, Jenny Benois-Pineau. 2020. Fine grained sport action recognition with Twin spatio-temporal convolutional neural networks. (2020).

[11] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. 2019. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 5552–5561.

[12] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6450–6459.