BADMINTON SHOT RECOGNITION USING LONG SHORT TERM MEMORY

Avinav Jain, Shubham Agrawal, Gaurav Singh Chauhan, Sai Shruti I, Preety Singh

The LNM Institute of Information Technology, Jaipur, Rajasthan, India

ABSTRACT

In this paper, we classify three shots, *Clear*, *Serve*, and *Smash*, played in the game of badminton. We have created our own dataset for the same. For recognition of the shots, we have trained a Long Short Term Memory deep learning model to learn the movements of keypoints of the player's body when playing a particular shot. We have achieved an accuracy of 89.49% for shot classification in our experiments.

***Index Terms*—** Badminton, Action recognition, Long Short Term Memory, Cross-validation

1. INTRODUCTION

Statistics of a game and the players are of great interest to fans of the game. Automated action recognition in sports can be very useful for the purpose of analyzing the game and identifying the types of shots played [17]. A player's posture is important to enhance the skill level in the game. Determining the correctness of the player's posture while playing various shots can help coaches to detect inadequacies and to plan training schedules accordingly [10, 13]. Action recognition can be utilized to identify the weaknesses and strengths of a player to help him/her improve the game.

The use of deep learning for action recognition and classification in sports has gained popularity in recent years [6, 12, 15]. Recurrent Neural Networks (RNNs) and Long Short Term Memory (LSTM) have been explored extensively in this field [2, 4, 9]. In games like cricket, computer vision has become an integral part, as the outcome of the batsman being marked as *out* can be predicted by tracing the probable trajectory of the ball. Similarly, in tennis, trained machine models are used to tell if the ball played is *fair* or *foul*. In football, it is used to detect the touch between the player and the ball to determine if it is a call for a *penalty* shootout. Match prediction is also one of the major applications of machine learning in the field of sports. By utilizing data from past matches, a trained system can predict the winning probability of each team on the basis of the current match scenario.

Badminton is one of the fastest racket sport and a major sport at global level. However, despite its popularity, there is limited research in this game due to non-availability of a freely available dataset. In this paper, we have used our own dataset for recognition of three different kinds of strokes played in the game of badminton. For the purpose of this study, we have taken three classes of shots and used a sequential model to predict the shot played on basis of body movements of the player. Most of the research related to the game of badminton focuses on the posture of the player and or classifying the game of the player as *offensive* or *defensive*. Our main focus in this paper is to classify the shot played by the sportsperson. The main contributions of our research are as follows:

- Created a new badminton dataset comprising of three classes of strokes, namely *Serve*, *Smash* and *Clear*. The recorded dataset consists of 984 videos (328 videos per class), which has been augmented to a total of 5516 videos.
- Trained a time-distributed sequential model to classify badminton shots based on movements of the player in a sequence of frames.

The rest of the paper is organized as follows: Section 2 presents related literature in the field of sports. Section 3 discusses our proposed methodology. Section 4 presents our experiments and results while Section 5 concludes the paper.

2. RELATED WORK

In this section, we present few research papers in the domain of action recognition in the field of sports [1, 7, 14]. In [5], Ghosh et al. proposed a generic pipeline for the analysis of racket games. In this work, they have provided an analysis of the broadcast videos of badminton matches. Having detected players and strokes for each frame in a match, they have computed the player's reaction time, dominance, positioning and footwork around the court, etc. Their player detection model uses color histogram features to cluster the image into two groups using the Gaussian mixture model, each representing a player. This model has a mAP@0.5 value of 97.85% for the bottom player, and 96.90% for the top player. (The bottom player refers to the player closer to the camera, whose back is visible, and the top player refers to the player who is far away from the camera, whose front is visible).

In [3], Chu et al. proposed a framework to classify the strategy of an individual player in the game of badminton as *offensive* or *defensive*. They have detected court lines

and players by background subtraction, and classify players' postures into six categories of stroke types. Taking the sequence of observations of stroke type and the movements of the player, they predict the strategy of the gameplay. They have reported an average strategy classification accuracy of 70%.

In [18], Fok et al. have presented a framework to recognize various styles of swimming, such as *butterfly*, *freestyle*, etc. by feeding the posture key points to a neural network. They have shown 92.9% test accuracy. In [11], Mora et al. have classified strokes in the game of tennis into four action groups and have trained a 3-layer LSTM model on an independent dataset. They have reported an accuracy of 84.10% when tested on a mixed set of matches of amateurs and professionals.

In [8], Kumar et al. have worked on the game of cricket. They have used the optical flow technique to identify whether the ball was hit by a batsman. Because of the small dataset, they have used the k-fold cross-validation spitting technique and obtained an accuracy of 80%. In [16], Tsunoda et al. have used multi-person centered features and temporal dynamics of features to recognize various futsal plays in the game of soccer recorded from multiple cameras in multi-view. They have integrated one more state in the general LSTM, called *Keeping state*, which is externally controlled. Their proposed method with RGB image and meta-information inputs yielded an accuracy of 70.90%.

3. PROPOSED METHODOLOGY

The block diagram of our proposed methodology is given in Figure 1. We have created a dataset for our experiments. The dataset was augmented using various techniques. From each frame of a shot video, we have extracted keypoints of the player to track his/her movements. The dataset is split into two: the training dataset is used to train a deep neural network which is then evaluated with the test dataset.

3.1. Dataset Collection

We have created our dataset by recording shots of three classes, namely, *Serve*, *Smash* and *Clear* in the badminton game. We have collected 328 videos for each class of shot. We used the Microsoft Xbox Kinect One Sensor for recording. The videos have been recorded at 30fps. To simulate behind the baseline view, we placed the sensor around 80cm above the ground, behind the player, and parallel to the net. The *bottom* player (whose back is visible) is more in view as compared to the *top* player across the net. To record the front view, we placed the sensor directly under the net at a similar height. Five files are available for each sample of a shot - a color video, a depth video, a raw depth video, a skeleton points video, and a JSON file with information about the skeleton joints coordinates. Few samples are shown in Figure 2.

Shot	No. of Videos	Average frames per video
Serve	1859	11.94
Smash	2109	11.49
Clear	1710	10.77

Table 1. Details of augmented dataset.

The Kinect contains two cameras with different resolutions: a color camera and a depth camera. The color camera has a resolution of 1920×1080. The depth camera has a resolution of 512×424. While the Kinect sensor provides both color and depth videos, we have used only the color videos so that the experimental results can be extended to videos recorded from regular video cameras without depending on any special camera sensor. Moreover, as usually the whole game is shown from behind the baseline in a professional match, we have considered the recordings of the bottom player only for our experiments.

3.2. Video Augmentation and Labelling

To increase the number of videos in our dataset, we have applied augmentation on videos by applying contrast changes, rotations, flipping the frames horizontally, and smoothening the frames. The total number of videos after augmentation is 5516. The total number of videos for each shot and average number of frames per video are shown in Table 1. The augmented dataset is available publicly and is published at the link: https://github.com/sam-2200/Badminton-Dataset/.

3.3. Keypoint Detection

To predict the type of shot played, we have extracted 33 key points from the body posture of the bottom player in each frame (refer Figure 3) by applying a keypoint detection model on the frames of the videos. Each keypoint comprises of the three-dimensional coordinates of the human joint. This results in a total of 33×3 feature points per frame.

3.4. Classification Model

We have used the Long Short Term Memory (LSTM) architecture as our deep learning classification model to capture the important features of the stroke play and to classify the shots. Our dataset is divided into two parts: approximately 90% for training, and 10% for testing, using a stratified split. During training of the model we have used the 5-fold cross-validation method. The number of videos used for training and testing are given in Table 2.

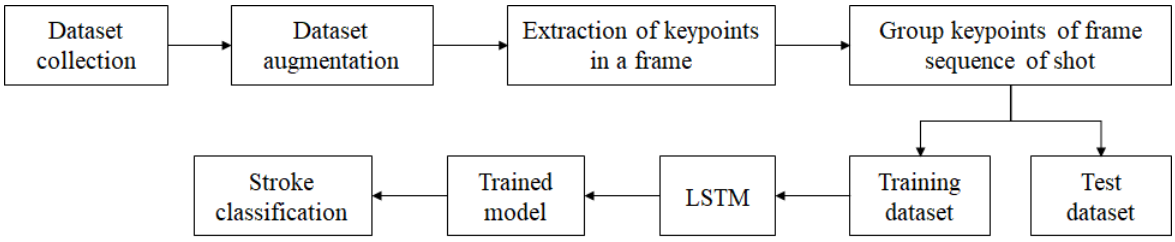

Fig. 1. Block diagram of proposed methodology.

Colour Depth Skeleton

Fig. 2. Snapshots of the colour, depth and skeleton views obtained for a *Smash* shot.

Fig. 3. Skeleton of the player for keypoint extraction.

	Serve	Smash	Clear
Training	1668	1895	1401
Test	192	215	145

Table 2. Number of videos in training and test datasets.

4. EXPERIMENTS AND RESULTS

Our badminton dataset consists of three classes of shots, *Clear*, *Serve*, and *Smash*. From the colour video of each shot video, we have extracted keypoints from the skeleton of the bottom player for each shot played using the MediaPipe Keypoints Holistics Model. As the number of frames vary in different videos, we have fixed the number of frames to 15, after carefully observing that in almost all the shots are captured in 15 frames. Those videos which exceeded 15 frames have been trimmed as they generally contained some redundant frames post playing the shot. For videos which had frames less than 15, we have applied post-padding with zero value frames.

The keypoint sequences extracted from each of the 15 frames are fed to an LSTM model to train it to classify the strokes. The architecture of the model is shown in Table 3. To evaluate the performance of our model on the test set, we have used Precision, Recall, Accuracy, $F1$-score and Support as our evaluation metrics. On evaluating our test set on the trained model, we obtained an overall accuracy of 89.49% and support of 552. The other evaluation metrics for the dif-

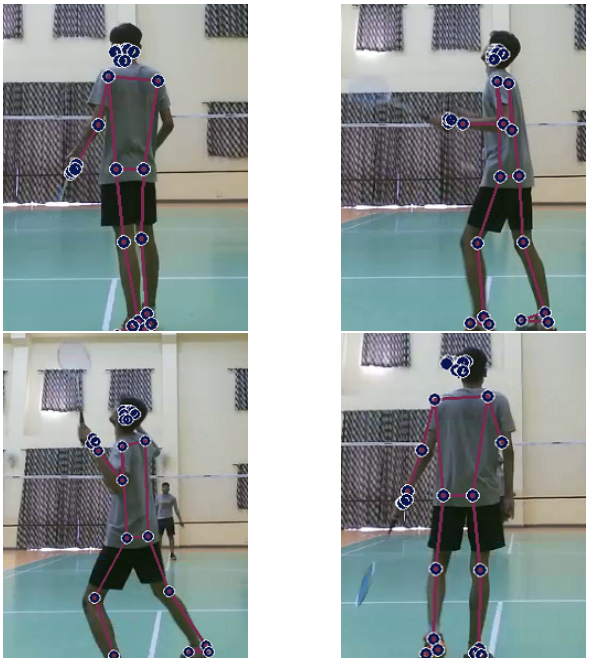

Fig. 4. Player keypoints on frames from a video of *Clear* shot

Layer Type	Output Shape	Param Count
Time Distributed	(None, 15, 225)	0
LSTM	(None, 15, 512)	1511424
LSTM	(None, 15, 256)	787456
LSTM	(None, 128)	197120
Dense	(None, 64)	8256
Dense	(None, 32)	2080
Dense	(None, 3)	99
Total Parameters	–	2,56,435

Table 3. Architecture of LSTM model.

ferent shots are shown in Table 4. The confusion matrix is shown in Figure 5.

As seen in the confusion matrix, our model works best for the *serve* shot, followed by the *clear* shot. Some of the *smash* shots are getting classified as *clear*. This could be due to the close resemblance of the posture of the player while playing these shots. This accuracy can be improved by incorporating the trajectory of the shuttle cock along with the player's body movements as input to the model.

5. CONCLUSION

In this paper, we have introduced a labeled dataset containing three types of shots played in the game of badminton. We have used a time-distributed deep learning model to classify the strokes into three classes, *Clear*, *Serve*, and *Smash*. We have achieved 89.49% accuracy on the test set. In future,

	Precision	Recall	F1-score	Support
Serve	0.95	0.94	0.94	194
Smash	0.82	0.95	0.88	186
Clear	0.93	0.78	0.85	172

Table 4. Evaluation metrics for different shots of the game.

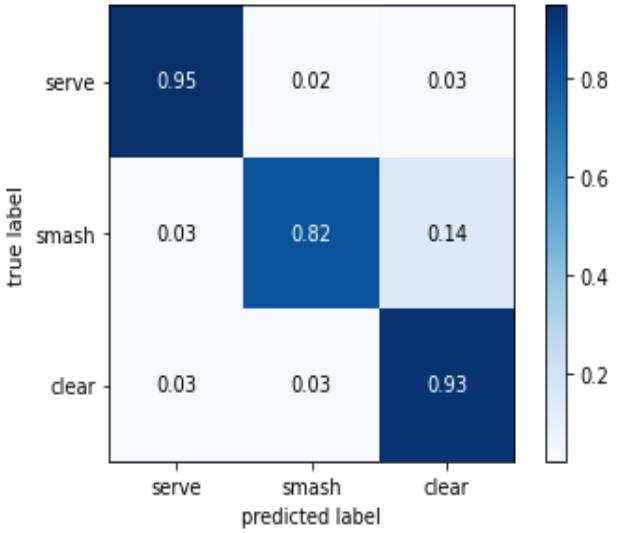

Fig. 5. Confusion matrix.

this study can be extended by introducing more classes of strokes in the dataset. Classification of strokes over frames of a given game video can help in segmenting videos for the various shots for analysis of the game.

6. REFERENCES

[1] K.W. Ban, J. See, J. Abdullah, and Y. P. Loh, "BadmintonDB: A Badminton Dataset for Player-specific Match Analysis and Prediction," in *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*. ACM, 2022, pp. 47–54.

[2] B. L. Bhatnagar, S. Singh, C. Arora, and C. V. Jawahar, "Unsupervised learning of deep feature representation for clustering egocentric actions," In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 2017, pp. 1447-–1453.

[3] W. T. Chu and S. Situmeang, "Badminton Video Analysis based on Spatiotemporal and Stroke Features," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. ACM, 2017, pp. 448–451.

[4] A. Fathi and J. M. Rehg, "Modeling Actions through State Changes," *Conference on Computer Vision and Pattern Recognition*. IEEE 2013, pp. 2579–2586.

[5] A. Ghosh, S. Singh, and C. V. Jawahar, "Towards Structured Analysis of Broadcast Badminton Videos," *IEEE Winter Conference on Applications of Computer Vision*. 2018, pp. 296–304.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 770–778.

[7] K. M. Kulkarni and S. Shenoy, "Table Tennis Stroke Recognition Using Two-Dimensional Human Pose Estimation," *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2021, pp. 4571–4579.

[8] A. Kumar, J. Garg, and A. Mukerjee, "Cricket activity detection," *International Image Processing, Applications and Systems Conference*. 2014, pp. 1–6.

[9] C. Lea, A. Reiter, R. Vidal, and G. D. Hager, "Segmental Spatiotemporal CNNs for Fine-grained Action Segmentation," in *Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV*. Springer, Cham, vol. 99077, 2016.

[10] M. Mlakar and M. Luštrek, "Analyzing tennis game through sensor data with machine learning and multi-objective optimization," in *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. ACM, 2017, pp. 153-–156.

[11] S. V. Mora and W. J. Knottenbelt, "Deep Learning for Domain-Specific Action Recognition in Tennis," *Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2017, pp. 170–178.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39(6), pp. 1137–1149, June 2017.

[13] V. Renò, N. Mosca, M. Nitti, T. DOrazio, C. Guaragnella, D. Campagnoli, A. Prati, and E. Stella, "A technology platform for automatic high-level tennis game analysis," *Computer Vision and Image Understanding*, vol. 159, 2017, pp. 164–175.

[14] M. Sharma, M. Lamba, N. Kumar and P. Kumar, "Badminton match outcome prediction model using Naïve Bayes and Feature Weighting technique," *Journal of Ambient Intelligence and Humanized Computing*, pp. 8441-–8455, 2021.

[15] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *ICLR*. 2014 arXiv.

[16] T. Tsunoda, Y. Komori, M. Matsugu, and T. Harada, "Football Action Recognition Using Hierarchical LSTM," *Conference on Computer Vision and Pattern Recognition Workshops*. 2017, pp. 155-163.

[17] L. Wang, Y. Qiao, and X. Tang, "Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors," *Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 4305–4314.

[18] Wilton W. T. Fok, Louis C. W. Chan, and Carol Chen, "Artificial Intelligence for Sport Actions and Performance Analysis using Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM)," in *Proceedings of the 2018 4th International Conference on Robotics and Artificial Intelligence*. ACM, 2018. pp. 40-–44.