BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026
http://excel.fit.vutbr.cz

Badminton Action Recognition for Video-Based Performance Analysis

Martin Pribylina*

Abstract

Video analytics are increasingly common in professional sport, yet manually reviewing match footage remains time-consuming for coaches and players. This paper presents a pipeline, which uses pose estimation to track body movement and then classifies actions with a Spatial-Temporal Graph Convolutional Network (ST-GCN), achieving 89.5 % test accuracy across ten action classes on 5,545 labeled clips from 21 players. Recognized strokes are post-processed into non-overlapping action segments and displayed in a dedicated viewer. The system requires no wearable sensors, shuttlecock detectors or broadcast-quality footage, demonstrating that skeleton-based graph neural networks offer a practical foundation for automated badminton video analytics.

*xpriby19@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Reviewing match footage manually is a time consuming task for coaches and analysts [1]. A system that automatically labels every stroke in an unedited recording would substantially speed up this process.

The proposed system detects the player of interest, segments motion into 21-frame windows, and classifies each segment into ten action categories (CLEAR, DRIVE, DROP, IDLE, LIFT, LONG SERVE, NET KILL, POSITIONING, SHORT SERVE, SMASH), while highlighting the player and predicted action.

Vision-based approaches [2] apply CNNs directly to video frames, achieving 85.7% accuracy[1], but they remain sensitive to variations in camera angle and lighting. Sensor-based methods can reach higher accuracy—ranging from 83.97% to 93.66%[2] [3] and up to 98%[3] [4]—yet the required hardware and resulting player discomfort limit their practicality. The most relevant skeleton-based approaches, TemPose [5] (90.7% accuracy[4]) and BST [6] (65.17% accuracy[5]), combine joint keypoints with shuttlecock trajectory features, but both rely on a dedicated shuttlecock detector.

This work proposes a vision-based, equipment-free pipeline that operates on ordinary video from any camera angle. The contributions are:

1. Comparison of custom and standard ST-GCN architectures under various training settings.
2. An ST-GCN on whole-body skeletons reaching 89.5 % test accuracy on 10 classes without auxiliary detectors.
3. Complete inference pipeline running at approximately 73 ms per frame on an NVIDIA RTX 3060.

2. Methodology

2.1 Pipeline

Player detection and tracking are handled by YOLOv8 (≈5 ms per frame): all persons are detected, the target player is selected by a user-specified court position (Top, Bottom, Left, or Right), and tracked across frames.

Pose estimation is performed by RTMW3D-X [7] (≈98 M parameters, ≈65 ms per frame), chosen for its keypoint accuracy on variable-quality footage. Any compatible estimator can be substituted without retraining the classifier.

An ST-GCN then classifies the resulting spatiotemporal skeleton sequence (≈3.7 ms). Before classification,

[1] 5 classes: clear, drop, lift, net shot, smash

[2] 12 classes: forehand/backhand service, clear overhead, clear underarm, net shot underarm, drop overhead, smash overhead

[3] 9 classes: clear, dab, drive, short serve, lob, drop, smash, displacement, rest

[4] 13 classes: top/bottom player forehand/backhand, smash, lob, react strokes, and a none class

[5] 18 classes: top/bottom player block, clear, drive, drop, net kill, net lift, net shot, serve, smash

skeletons are normalized: the origin is placed at the ankle midpoint, the Y axis is reoriented upward, the shoulders are aligned to the X axis, and joint positions are scaled by body height, illustrated in Figure 1. Raw per-frame predictions are post-processed into non-overlapping action segments. Detections are ordered by confidence, those below a threshold are suppressed. Remaining windows are built greedily and conflicting windows are discarded, as shown in Figure 2. A viewer then overlays current, past, and upcoming action labels on the video, illustrated in Figure 3.

2.2 Dataset

The dataset combines a public badminton video repository [8] with recordings provided by the project supervisor. Public clips are pre-segmented with the stroke occurring in the few frames. For the additional data, the frame before racket–shuttlecock contact was manually identified and used to construct 31-frame clips aligned with this format. The final dataset contains 5,545 clips, split 70/15/15 into training, validation, and test sets.

2.3 Model and Training

Rather than using the reference ST-GCN [9] directly, a lighter custom architecture was developed to match the dataset size and real-time constraints. The reference model stacks ten graph-convolutional blocks with channel widths up to 256 and temporal downsampling, totalling approximately 3 M parameters (≈11 ms per inference).

The custom model uses three ST-GCN blocks with no temporal downsampling, replacing the temporal convolution with a multi-scale TCN (kernel sizes 1, 3, 5, 7) to capture both fine-grained and longer-range motion. A node-wise attention layer learns a softmax-weighted scaling over joint features, upweighting diagnostically important joints such as wrists and elbows. These changes reduce parameters to approximately 400 K and inference time to ≈3.5 ms.

Training used cross-entropy loss with label smoothing (0.1), learning rate and weight decay of 1×10^{-3} with a CosineAnnealingLR schedule, batch size 64, and early stopping (patience 40). Training converged at epoch 87 (of a 117-epoch run), yielding validation accuracy of 91.14 % and validation loss of 0.4057.

3. Results

The custom model reaches **89.5 % test accuracy** with a macro-average F1-score of 0.894 on 865 test samples. The reference ST-GCN, trained identically on the same split, achieves comparable accuracy at approximately 3× higher inference cost. The full breakdown is presented in Figure 4.

Serve classes achieve the strongest results: SHORTSERVE (F1 = 0.958) and LONGSERVE (0.956) are well-separated. The most challenging classes are DROP (F1 = 0.814) and LIFT (0.841), which share similar motion patterns. SMASH and CLEAR are also very similar, as both involve an overhead swing that only diverges in the follow-through. This highlights a key limitation of skeleton-based features: actions that differ primarily in shuttlecock interaction or late-stage execution are difficult to distinguish without explicit shuttlecock information, as shown in the per-class F1 scores in Figure 5 and the confusion matrix in Figure 6.

The complete pipeline runs at approximately 73 ms per frame, corresponding to about 2.2× slower than real-time for 30 fps video.

4. Conclusions

This work demonstrates that reliable badminton stroke recognition can be achieved using only pose-based features extracted from ordinary video. The proposed pipeline reaches 89.5% accuracy across ten classes without requiring specialized hardware or controlled recording conditions. Although the current implementation is not yet real-time, the classification stage is lightweight, and real-time operation appears achievable with a faster pose estimator.

The custom ST-GCN architecture achieves comparable accuracy to the reference model at roughly one third of the computational cost, making it a practical choice for deployment. At the same time, the results reveal clear limitations: actions with similar body motion, such as DROP/LIFT and SMASH/CLEAR, remain difficult to distinguish, and the system can produce fragmented or occasionally false positives during simple movements.

These findings suggest that pose alone is not always sufficient. The most promising direction for improvement is the integration of additional context, particularly shuttlecock information, which would help resolve ambiguities that cannot be captured from joint trajectories alone.

Acknowledgements

I would like to thank my supervisor prof. Ing. Adam Herout Ph.D. for his guidance and for providing the training recordings used in this work.

References

[1] Tica Lin, Alexandre Aouididi, Zhutian Chen, Johanna Beyer, Hanspeter Pfister, and Jui-Hsien Wang. Virdi: Immersive match video analysis for high-performance badminton coaching, 07 2023.

[2] Nur Azmina Rahmad and Muhammad Ar As'ari. The new convolutional neural network (CNN) local feature extractor for automated badminton action recognition on vision based data. *Journal of Physics: Conference Series*, 1529(2):022021, 2020.

[3] Indrajeet Ghosh, Srijan Ramasamy Ramamurthy, and Nirmalya Roy. StanceScorer: A data driven approach to score badminton player. In *Proceedings of the IEEE PerCom Workshops*, pages 1–6, March 2020.

[4] Thomas Steels, Bram Van Herbruggen, Jarne Fontaine, Toon De Pessemier, David Plets, et al. Badminton activity recognition using accelerometer data. *Sensors*, 20(17):4685, 2020.

[5] Magnus Ibh, Stella Grasshof, Dan Witzner, and Pascal Madeleine. TemPose: A new skeleton-based transformer model designed for fine-grained motion recognition in badminton. In *Proceedings of the CVPR Workshops (CVSports)*, pages 5199–5208, 2023.

[6] Jing-Yuan Chang. BST: Badminton stroke-type transformer for skeleton-based action recognition in racket sports. *arXiv preprint arXiv:2502.21085*, 2025.

[7] Tao Jiang et al. RTMW3D-X: Extra-large whole-body 3-d pose estimation model. HuggingFace model hub, https://huggingface.co/rbarac/rtmpose3d, 2024. Checkpoint: `rtmw3d-x_8xb64_cocktail14-384x288-b0a0eab7_20240626.pth`.

[8] HyperAI. Badminton video dataset. https://beta.hyper.ai/en/datasets/30582, 2024.

[9] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.