Tennis Stroke Classification

Ashwin Prasanth
School of Computer Science
University College Dublin
Dublin, Ireland
ashwin.prasanth@ucdconnect.ie

Abstract

Classifying tennis strokes from raw video is a detailed action recognition task. It is challenging due to limited labeled data, fast human movement, and subtle differences between classes. A 2D CNN trained from scratch gives a strong spatial baseline with 89% accuracy. This shows that static posture and racket position provide important information. However, when I added temporal modeling with a 3D CNN, the accuracy dropped to 80%. This highlights the data needs and instability of spatio-temporal convolutions in low-resource settings. A baseline Pose lstm model that includes pose-estimation features only performed poorly, achieving only 40–50% accuracy. Noisy keypoints from quick strokes disrupted the integration process. My final model solves these issues by using a lightweight R3D-18 backbone, a custom 3D convolutional stem, and significant temporally consistent augmentation. This model achieves 96% accuracy. The study demonstrates that, with limited data, performance relies more on suitable temporal biases, stable feature representations and effective augmentation rather than just on model depth.

1 Introduction

Understanding and classifying fine grained human motion from video presents a significant challenge in computer vision. This is especially true in sports like tennis, where stroke categories have only slight differences in body posture, timing, and racket movement. Accurately recognising strokes is essential. It allows scalable player analysis, automated skill assessment and data driven coaching systems. However, this task is tough because of rapid movement, self occlusions, changing viewpoints and the lack of high quality labeled sports datasets. These challenges emphasise the need to find the best architectural choices, input methods and training strategies under low data conditions.

To explore this I compared four modeling approaches, each focused on a different representational assumption. A 2D CNN trained from scratch on uniformly sampled frames acts as a spatial baseline that relies only on static appearance cues. A 3D CNN, trained on short clips, brings in explicit spatiotemporal modeling, letting the network analyze motion patterns directly. To see if structured human motion cues enhance performance, I also used baseline pose estimation features model. I also created a hybrid method that merges a custom 3D convolutional stem with an R3D-18 residual framework. This method uses a temporally consistent augmentation pipeline to boost motion and appearance diversity. All these approaches cover spatial, temporal, mediapipe and residual spatiotemporal architectures, allowing us to examine the effectiveness of different representations for tennis stroke classification.

Spatial models pick up strong posture level cues, while purely temporal models reveal the data needs of 3D convolutions. Pose lstm points out the instability of noisy pose signals during quick strokes. Hybrid residual architectures highlight the need for matched temporal induction rules and aggressive augmentation for fast sports motion. This study helps clarify which representations are most reliable for precise stroke recognition in low data scenarios.

2 Related Work

Video action recognition has mainly relied on 3D convolutional networks, multimodal fusion models and transformer based architectures. Early 3D CNNs like C3D [10] introduced temporal convolutions for capturing motion directly. I3D [4] improved spatiotemporal representation quality by using inflated 2D kernels and large-scale pretraining. Residual 3D architectures such as R3D [7] made optimising deeper models more stable. A common conclusion across these studies is that 3D CNNs depend heavily on data. They tend to overfit when training sets are small. This aligns with our findings, where a new 3D CNN performed worse than a simpler 2D model. Multimodal and two-stream approaches try to address motion by combining RGB appearance with optical flow or pose-based cues. Classic two-stream models [9] showed strong performance improvements by separating motion and appearance pathways. Pose-based methods that use OpenPose-style estimators [3] have been used in sports analytics to track limb movements. However, earlier studies point out that pose quality declines during quick, non-rigid movements, leading to unstable keypoints. Our multimodal experiments reveal the same issue. Despite theoretical advantages, the RGB+pose architecture performed poorly because of noisy pose extraction during fast tennis strokes.

More recent transformer-based video models like TimeSformer [2] and ViViT [1] achieve top accuracy through space-time attention, but only when using large datasets like Kinetics. This makes them unsuitable for sports scenarios with limited data. In tennis and broader sports-action recognition research, lightweight 2D CNNs, shallow 3D CNNs, and pose-augmented methods have been investigated for fine-grained stroke classification [6, 8]. There is consistent evidence that effective temporal biases and strong augmentation are essential when data is limited. These insights, along with the trends toward compact temporal backbones and aggressive spatiotemporal augmentation seen in methods like SlowFast [5], directly inform our final hybrid design. This design combines a lightweight residual 3D backbone with consistent temporal augmentation to achieve reliable, high-accuracy performance in low-data settings.

3 Experimental Setup

3.1 Dataset Preprocessing and Augmentation

All experiments were conducted on a custom tennis-stroke dataset made up of 100 short video clips labeled as forehand, backhand, serve, and nostroke. These clips were recorded from a fixed rear court viewpoint with different lighting, players, and backgrounds. Each clip was decoded into RGB frames, resized to 112×112, and sampled to $T = 16$ frames. For shorter clips, I padded them by repeating the final frame. The training/validation, test sets dataset was taken in 80:10:10 ratio, with 80 percent clips to train and 10 each to test and validation set. I normalized all frames using ImageNet statistics, creating standardized tensors for each of the four modeling approaches.The 2D CNN processed a single RGB frame sized at $3 \times 112 \times 112$, while the 3D CNNs used spatiotemporal tensors shaped $3 \times 16 \times 112 \times 112$. The pose only baseline model used normalised pose sequences that had 33 anatomical keypoints per frame taken using MediaPipe.

Table 1: Augmented tennis-stroke dataset

Class	**Train Frame**	**Validation Frame**	**Test Frame**	**Clips**		
				Train	**Validation**	**Test**
Forehand	256	32	32	54	7	7
Backhand	256	32	32	54	7	7
Nostroke	512	64	64	108	15	13
Serve	240	30	30	54	7	7
Total	1264	158	158	270	36	34

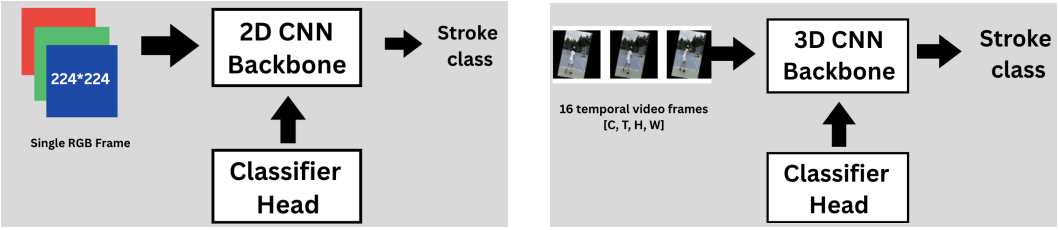
For Approach 4, I added a synthetic spatiotemporal set to meet the data needs of 3D architectures. Synthetic process generated 240 additional clips from real data with 48 for each stroke class and 96 for nostroke bringing the total training samples to 270. Also produced synthetic clips by frame dropping, duplication, temporal jitter, speed changes (0.7–1.3), linear resampling, rotations (±10°),

translations (±5%), scale changes (0.95–1.05), shear (±5°), brightness and contrast adjustments, and low-variance Gaussian noise. I applied all transformations consistently across frames within a clip to keep the temporal flow intact for 3D convolutional networks.

3.2 Model Building and Training

The first approach sets a spatial baseline to measure how much distinct power is present in single frames without using temporal reasoning. A lightweight 2D CNN was created from four Conv2D, BatchNorm, ReLU, and MaxPool blocks with channel widths of 32, 64, 128, and 256. This was followed by global average pooling and a dropout-regularized fully connected classifier. Each video was represented by its center frame, treating samples as separate images. Training used Adam with a learning rate of 10^{-3} and cosine decay, a batch size of 32, and a weight decay of 10^{-4}, with early stopping to avoid overfitting. Hyperparameters were adjusted for kernel sizes (3×3, 5×5), network depth (three or four blocks), and learning rates in $[10^{-2}, 5 \times 10^{-4}]$. This model serves as a base for evaluating the added value of temporal and multimodal cues.

Building on the spatial baseline, the second approach adds explicit temporal modeling using 3D convolutions. The network consists of Conv3D, BatchNorm3D, ReLU, and MaxPool3D blocks with channel widths of 64, 128, and 256. This is followed by global 3D average pooling and a fully connected classifier. Inputs were fixed-length clips with the shape of $3 \times 16 \times 112 \times 112$. Training used Adam with a learning rate of 10^{-4}, batch sizes between 8 and 16 and gradient clipping set to 5.0 to stabilize temporal gradients. To reduce overfitting in this higher-capacity model, we used temporal jittering, spatial dropout, and random cropping. I further tested different 3D kernel sizes (3, 5), clip lengths (8, 16, 32), variations in channel widths, and temporal strides. This approach isolates the effects of direct motion modeling and highlights the challenges 3D CNNs face when working with small datasets.



(a) Approach 1: 2D CNN baseline operating on a single RGB frame (224×224).

(b) Approach 2: 3D CNN baseline processing a short video clip of 16 frames.

Figure 1: Comparison of spatial and spatiotemporal baseline architectures.

The third approach examines whether skeletal motion alone is enough for classifying tennis strokes. For each video, MediaPipe Pose extracts 33 anatomical keypoints per frame, creating a 66-dimensional vector. Sixteen frames are sampled evenly to form a temporal keypoint sequence. This sequence goes through a two-layer LSTM with a hidden size of 128 units, and the final hidden state is used to predict the class. Unlike methods based on RGB, this approach focuses only on body joint movement, without relying on a 3D CNN backbone or fusion techniques. Training employs Adam with cross-entropy loss and a batch size of 4. This method acts as a pose only baseline to evaluate how well pure skeletal motion captures stroke patterns.

Table 2: Hyperparameters and final choices for each approach.

Approach	Model	Hyperparameters Tuned	Final Setting
1	2D CNN	LR, batch size, dropout	LR=1e^{-3},batch=32, dropout=0.5
2	3D CNN	LR, batch size, clip length	LR=1e^{-3}, batch=4, clip=16
3	Pose + LSTM	LR, hidden size, layers	LR=1e^{-3}, hidden=128, layers=2
4	Hybrid R3D-18	LR, decay,clip, augment	LR=3e^{-4}, WD=1e^{-3}, clip=16

The final approach combines insights from the previous models, merging lightweight temporal biases with a more expressive yet regularized spatiotemporal backbone. A custom Conv3D, BatchNorm3D,

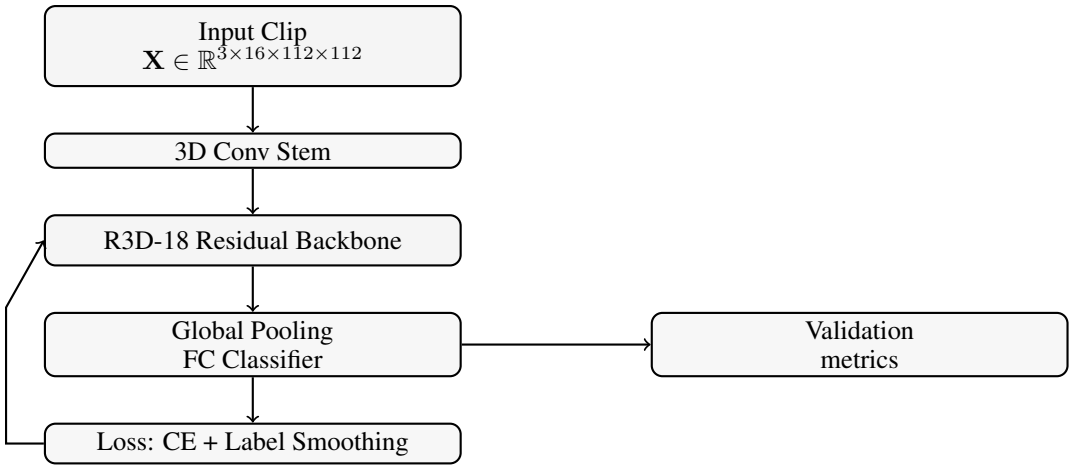
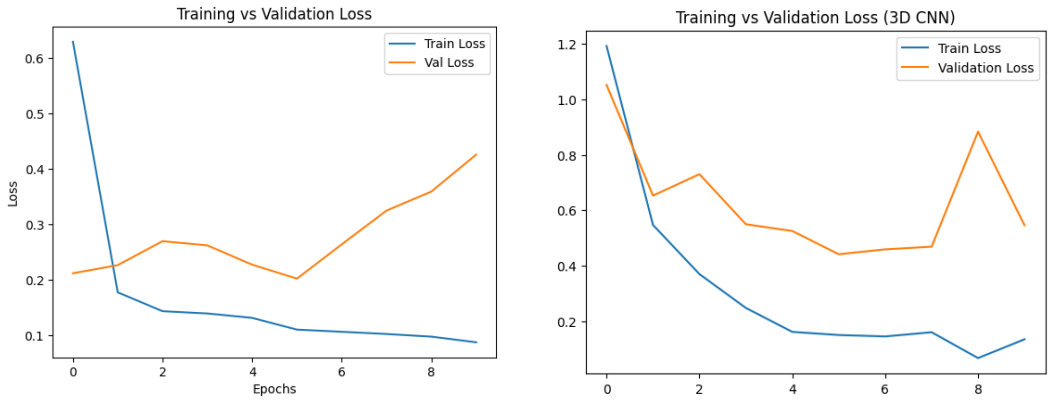


Figure 2: hybrid 3D CNN (Conv3D stem + R3D-18) training pipeline.

and ReLU stem improve early temporal sensitivity before connecting to an R3D-18 network. This network features residual 3D blocks that offer stable multi-scale temporal modeling, suitable for small datasets. It concludes with global spatiotemporal pooling and a fully connected classifier. Training used AdamW with a learning rate of 3×10^{-4}, a warmup followed by cosine annealing, batch sizes from 8 to 16, weight decay of 10^{-3}, and label smoothing (0.1) along with mixed-precision training. Hyperparameter tuning assessed deeper variants (R3D-34), clip lengths of 16 and 32, augmentation intensity, and learning rates in $[10^{-4}, 10^{-3}]$. This hybrid architecture provided the most robust and generalizable performance, showing that well-designed temporal biases and strong augmentation surpass deeper or multimodal options in low-data sports scenarios.

4 Results

The models were tested on a four-class tennis stroke classification task. I used top-1 accuracy on a held-out test set, along with cross-entropy loss curves and validation accuracy trends to check optimization stability and generalization. Accuracy fits well for this balanced categorical task. The loss trajectories point out overfitting behavior and issues with specific modalities. Per-class precision and recall, reported for Approach 4, highlight differences at the class level. The validation curves were crucial for understanding the effects of pose noise in Approach 3 and the impact of temporal augmentation in Approach 4.



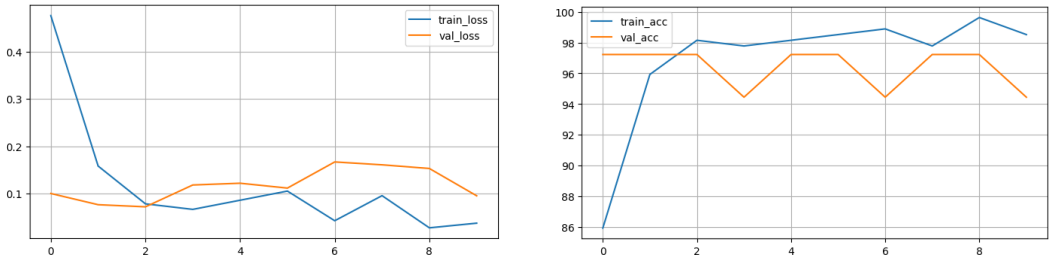
(a) 2D CNN: Train vs Validation Loss

(b) 3D CNN: Train vs Validation Loss

Figure 3: Training loss comparison between 2D and 3D CNN models.

The 2D CNN baseline showed strong performance at 89.87%. This indicates that single-frame appearance cues, such as posture and racket orientation, are very effective for identifying tennis strokes. In contrast, the 3D CNN trained from scratch only reached 80%. This reflects the known sample complexity of 3D convolutions. Without pretraining or significant augmentation, spatiotemporal models quickly overfit and show unstable gradients. I noticed this in the noisy validation curves and the model's sensitivity to hyperparameters.

The pose lstm baseline achieved 50% accuracy and showed overfitting. Validation accuracy fluctuated between 30% and 50%. This was caused by noisy pose estimates during rapid motion and occlusion. These poor keypoints led the lstm to overfit pose artifacts instead of useful kinematic structure, making it unreliable for fusion models.



(a) Approach 4: Training vs. Validation Loss

(b) Approach 4: Training vs. Validation Accuracy

Figure 4: Hybrid 3D CNN + R3D-18 model metric with synthetic spatiotemporal augmentation.

The final hybrid model, which included a 3D convolutional stem, an R3D-18 backbone, and a synthetic augmentation pipeline, produced the best results with 97.06% accuracy. The training curves indicated smooth convergence and a small generalization gap. The classification report showed consistently high precision and recall across all classes, with a macro F1 score of 0.9733. The confusion matrix had perfect recognition for three classes and only one misclassification in the no-stroke class. This performance came from the effective combination of parameter-efficient residual 3D modeling and consistent augmentation, including speed changes, frame jitter, and affine transforms. The results demonstrate that in low-data sports video situations, lightweight residual 3D models combined with strong augmentation outperform deeper 3D CNNs and multimodal fusion methods.

Table 3: Comparison of performance across approaches.

Approach	**Model**	**Input Modality**	**Test Acc**	**Macro F1**
1	2D CNN	Single RGB frame	89.87%	—
2	3D CNN from scratch	RGB clip (16 frames)	80.00%	—
3	Pose LSTM base	33-joint pose	50.00%	∼0.45
4	Hybrid 3D Stem + R3D-18	RGB clip + augmentation	97.06%	0.9733

5 Conclusion and Future Work

This work compared four modeling approaches for classifying tennis strokes. It moved from static spatial models to multimodal fusion and spatiotemporal architectures. The findings reveal several important points. The 2D CNN baseline achieved a strong performance of 89.87%. This indicates that using single-frame posture and racket cues is very effective for this dataset. When the model was extended to a 3D CNN, it improved temporal expression but underperformed at 80%. This was due to high data demands and instability with spatiotemporal convolutions trained from scratch. Adding pose information lowered performance further to 50%. Noisy keypoints during fast strokes caused the fusion model to overfit on specific errors related to the different modalities. The final hybrid architecture, which included a 3D convolutional stem, an R3D-18 backbone, and a synthetic augmentation pipeline, achieved the best results at 97.06%. This shows that lightweight temporal models combined with effective augmentation work well in low-data sports-video settings.

Future work has the potential to improve both performance and reliability. Pretraining on large video datasets or using self-supervised temporal pretraining could enhance motion understanding and lessen the need for synthetic data. The pose-based model might benefit from more accurate keypoint extraction, higher-resolution inputs or spatiotemporal pose encoders that take uncertainty into account, potentially stabilized with cross-modal attention. There are additional opportunities to test transformer-based video architectures, investigate contrastive or generative augmentation methods, and include finer temporal details like ball-contact events or stroke-phase boundaries. If the project were repeated, the focus would be on gathering a wider range of real videos, integrating motion representations like optical flow, and applying domain adaptation to ensure the model works well across different players, lighting, and camera angles. These extensions would improve the system's scalability and usefulness for real-world coaching and performance analysis.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846. IEEE, 2021.

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR, 2021.

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2021.

[4] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308. IEEE, 2017.

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6202–6211. IEEE, 2019.

[6] Andrew Gilbert, Robert Churchill, and Adrian Hilton. Tennis stroke recognition using convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2301–2310. IEEE, 2019.

[7] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d CNNs retrace the history of 2d CNNs and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555. IEEE, 2018.

[8] Zhihong Jiang, Xiaoli Ma, and Yu Wang. Fine-grained sports action recognition via spatiotemporal convolutional networks. *Pattern Recognition Letters*, 138:348–355, 2020.

[9] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, pages 568–576, 2014.

[10] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497. IEEE, 2015.