

Tennis Strokes Recognition from Generated Stick Figure Video Overlays

Boris Bačić¹^a and Ishara Bandara²^b

¹*Auckland University of Technology, Auckland, New Zealand*

²*Robert Gordon University, Aberdeen, U.K.*

Keywords: Computer Vision, Deep Learning, Spatiotemporal Data Classification, Human Motion Modelling and Analysis (HMMA), Sport Science, Augmented Broadcasting.

Abstract: In this paper, we contribute to the existing body of knowledge of video indexing technology by presenting a novel approach for recognition of tennis strokes from consumer-grade video cameras. To classify four categories with three strokes of interest (forehand, backhand, serve, no-stroke), we extract features as a time series from stick figure overlays generated using OpenPose library. To process spatiotemporal feature space, we experimented with three variations of LSTM-based classifier models. From a selection of publicly available videos, trained models achieved an average accuracy of between 97%–100%. To demonstrate transferability of our approach, future work will include other individual and team sports, while maintaining focus on feature extraction techniques with minimal reliance on domain expertise.

1 INTRODUCTION

Automated video indexing of recognised motion patterns and human motion activity have broad application. For example, in the last decade, we have seen enhancements in augmented video broadcasting with real-time game statistics. Overlaid statistics can help commentators to share strategic information which often only competitive-level athletes, coaches and domain experts would intuitively consider. Quantifying motion events has also become pervasive in other contexts such as augmented video coaching, surveillance, elderly care, activity monitoring and various mobile and smartwatch apps development.

1.1 Vision and Motivation

In the authors' view, identifying task-specific motion events will enable further advancements in areas such as rehabilitation, smart cities, and to improve privacy, safety and usability of spaces where human activity occurs. To illustrate the demand for quantifying events as part of strategic video analysis, an online search will return numerous examples of commercial


software (e.g. LongoMatch, Coach Logic, Nacsport, Metrica Sports, and Sports Code).


LongoMatch, as one of the earliest open source coaching software, was originally designed for video replay analysis of team sports with customisable and manual event indexing (during live video capture or in post-match analysis) with rudimentary overlay capabilities (Bačić & Hume, 2012). Today, it is a high-end annual-subscription licensed commercial product, reflecting the opportunities in this area.

Thanks to recent advancements in computer vision, deep learning, recurrent neural networks and human pose estimation, the underlying development of automated indexing of human motion events and sport-specific movement patterns has become less labour intensive and less dependent on expertise-driven feature extraction techniques than in the past.

To quantify aspects of the game relying on Computer Vision and Artificial Intelligence (AI), the research questions associated with our work are:

1. Can we develop recognition of sport-specific movement patterns such as tennis strokes?
2. Can we reduce dependence on expert-driven feature engineering and produce simplified

^a <https://orcid.org/0000-0003-0305-4322>

^b <https://orcid.org/0000-0002-7346-248X>

feature extraction techniques relying on common-sense visual observation?

3. If so, can we develop a multi-stage video processing and modelling framework that is transferable to other sports?

1.2 Background and Prior Work

Advancements in motion pattern indexing can not only be evaluated by improving classification performance for a specific task, low-cost real-time computing and extending the number of labelled events of interest, but also on their universal applicability to various sources such as 3D motion data (Bačić & Hume, 2018), video (Bloom & Bradley, 2003; D. Connaghan, Conaire, Kelly, & Connor, 2010; Martin, Benois-Pineau, Peteri, & Morlier, 2018; Ramasinghe, Chathuramali, & Rodrigo, 2014; Shah, Chockalingam, Paluri, Pradeep, & Raman, 2007), and sensor signal processing (Anand, Sharma, Srivastava, Kaligounder, & Prakash, 2017; Damien Connaghan et al., 2011; Kos, Ženko, Vlaj, & Kramberger, 2016; Taghavi, Davari, Tabatabaee Malazi, & Abin, 2019; Xia et al., 2020).

To our knowledge, tennis shots or strokes action recognition relying on computer vision started in 2001, by combining computer vision and hidden Markov model (HMM) approaches, before HD TV-broadcast resolution became available (Petkovic, Jonker, & Zivkovic, 2001). After Sepp Hochreiter and Jürgen Schmidhuber invented Long Short Term Memory (LSTM) in 1997, LSTMs have been used in action recognition (Cai & Tang, 2018; Liu, Shahroudy, Xu, Kot, & Wang, 2018; Zhao, Yang, Chevalier, Xu, & Zhang, 2018). In 2017, inertial sensors with Convolutional Neural Networks (CNN) and bi-directional LSTM networks were used to recognise actions in multiple sports (Anand et al., 2017). In 2018, an LSTM with Inception v3 was used to recognise actions in tennis videos achieving 74% classification accuracy (Cai & Tang, 2018).

For prototyping explainable AI in next-generation augmented coaching software, which is expected to capture expert's assessment and continue to provide comprehensive coaching diagnostic feedback (Bačić & Hume, 2018), we can rely on multiple data sources including those operating beyond human vision.

Prior work on 3D motion data is categorised as: (1) traditional feature-based swing indexing based on sliding window and thresholding (Bacic, 2004) and expert-driven algorithmic approach in tennis shots and stance classification (Bačić, 2016c); (2) featureless approach for accurate swing indexing using Echo State Network (ESN) (Bačić, 2016b) and

(3) further sub-event processing i.e., *phasing analysis* via produced ensemble of ESN (Bačić, 2016a).

Prior work on video analysis applied Histograms of Oriented Gradients (HOG), Local Binary Pattern (LBP) and Scale Invariant Local Ternary Pattern (SILTP) for human activity recognition (HAR) in surveillance (Lu, Shen, Yan, & Bačić, 2018). A pilot case study on cricket batting balance (Bandara & Bačić, 2020) used recurrent neural networks (RNN) and pose estimation to generate classification of batting balance (from rear or front foot). This prior work on privacy-preservation filtering is aligned with privacy-preserving elderly care monitoring systems and with extracting diagnostic information for silhouette-based augmented coaching (Bačić, Meng, & Chan, 2017; Chan & Bačić, 2018). It is also generally applicable to usability and safety of spaces where human activity occurs such as smart cities, future environments and traffic safety (Bačić, Rathee, & Pears, 2021).

2 METHODOLOGY

Considering past research, our objective is to produce a relatively simple and generalised initial solution and a human motion modelling (HMMA) framework for video indexing applicable to tennis. The tennis dataset was created from both amateur and professional players' videos. It is also expected that the produced framework may be easily transferable to other sport disciplines and related contexts such as rehabilitation and improving safety and usability of spaces where human movement occurs. As part of movement pattern analysis, we focused on expressing features as spatiotemporal human movement patterns from faster moving segments (e.g., dominant hand holding a racquet) relative to the more static trunk segment.

2.1 Stick Figure Overlays as Initial Data Preprocessing

To retrieve player's motion-based data from video, we generated stick figure overlays using OpenPose (<https://github.com/CMU-Perceptual-Computing-Lab/openpose>) and 25 key point estimator COCO+Foot (Figure 1 and Figure 2).

Figure 2 shows an example of data format representing the key points coordinate of a player in each video frame recorded as multi time series data. As video overlays, animated stick figure topology of generated key points (Figure 3) represents a way of extracting information from video to facilitate human

motion modelling and analysis (HMMA) and assist in the feature extraction process.

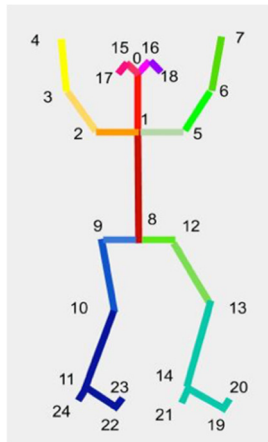


Figure 1: Stick figure overlay topology: 25 key points COCO+Foot model. Reproduced from (Bandara & Bačić, 2020), copyright permission (IEEE No. 5170730636786).

2.2 Data Collection and Analysis

Data collection for experimental work was carried on Google cloud platform using publicly available videos of tennis practice matches, also available via YouTube (12kgp-Tennis, 2019; Back of the line tennis, 2018; Page, 2020; Tenfitmen Tennis Impulse, 2020; Tennis Legend TV, 2019; Top Tennis Training - Pro Tennis Lessons, 2014; TV Tennis Pro, 2020).

```
{
  "version": 1.3,
  "people": [
    {
      "person_id": -1,
      "pose_keypoints_2d": [
        681.995, 351.872, 0.658373, 633.28, 364.934, 0.864891, 641.162, 364.985, 0.756001, 646
        .987, 404.082, 0.258643, 607.867, 415.895, 0.189584, 621.588, 366.773, 0.892353, 615
        .781, 398.228, 0.872284, 588.283, 415.792, 0.911615, 629.372, 445.185, 0.785502, 631
        .361, 445.175, 0.714422, 598.183, 494.228, 0.871724, 584.25, 547.133, 0.815248, 623
        .584, 445.199, 0.819114, 681.982, 583.946, 0.848848, 658.98, 543.172, 0
        .892837, 0, 0, 683.837, 343.324, 0.712675, 0, 0, 613.655, 343.371, 0.812629, 643
        .11, 574.514, 0.855431, 658.962, 572.519, 0.897959, 664.687, 541.271, 0.944661, 652
        .938, 558.854, 0.815353, 654.969, 556.846, 0.82357, 588.282, 556.886, 0
        .911153],
      "face_keypoints_2d": [],
      "hand_left_keypoints_2d": [],
      "hand_right_keypoints_2d": [],
      "pose_keypoints_3d": [],
      "face_keypoints_3d": [],
      "hand_left_keypoints_3d": [],
      "hand_right_keypoints_3d": []
    }
  ]
}
```

Figure 2: Example of 25 key points extracted from stick figure overlay in a single frame in JSON format. Each key point consists of three variables (x,y,confidence).

The dataset included a balanced distribution of 150 extracted tennis strokes (30 forehands, 30 backhands and 30 serves, 60 no stroke play), which were labelled manually into their corresponding output classes. All videos were of the same framerate (30fps) and of duration between 0.8 – 1.0 seconds (28-31 frames).

2.3 Feature Extraction Technique

Our approach to Feature Extraction Technique (FET) is based on visual analysis between faster and slower moving body segments (Figure 4 and Table 1).



Figure 3: A mosaic of stick figure overlays generated during the serve, forehand and backhand strokes as intermediate computer vision processing steps.

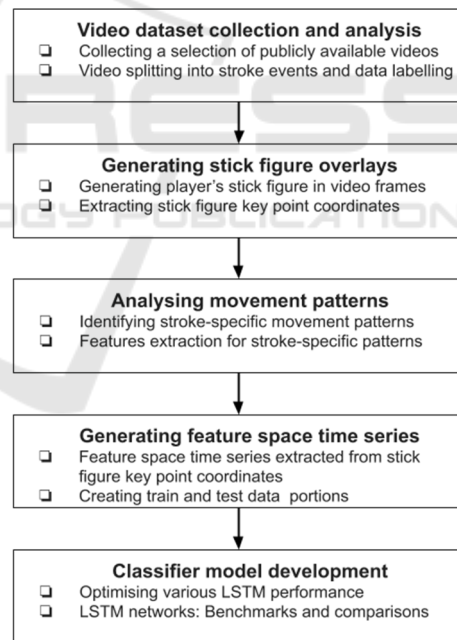


Figure 4: Human motion data processing, analysis and modelling framework for video indexing. Adapted from the initial concept (Bandara & Bačić, 2020).

In Table 1, a function $d(P_j, P_k)$ represents Euclidean distance measure (1) between the key points P_j and P_k of overlaid stick figure calculated for each video frame.

$$d(P_j, P_k) = \sqrt{(x_j - x_k)^2 + (y_j - y_k)^2} \quad (1)$$

Pixel coordinates of a stick figure’s key point P_n are denoted as: $(x_n, y_n) \in P_n$; $1 \leq n \leq 25$.

2.4 Feature Space: Visualisation Remarks

Based on visual movement pattern analysis, we identified nine distance-based features that collectively produce the best results (Table 1). As such, the 25 key-points time series data representing the human body were further reduced to 36% representing the feature space.

R.H. to C. Euc. (4-1)	L.H. to C. Euc. (7-1)	R.H. to R.L. Euc. (4-11)	L.H. to L.L. Euc. (7-14)	H. to H. Euc. (4-7)	R.H. to L.Hip Euc. (4-12)	L.H. to R.Hip Euc. (7-9)	M.P. to R.H. X-axis (8-4)	M.P. to L.H. X-axis (8-7)	CAT
53.818722	56.701332	168.545705	190.663235	24.615104	47.239837	87.311691	-4.104	-64.608	0
63.969882	60.303565	153.839331	199.131221	35.504819	35.438996	88.055195	-36.683	-67.510	0
63.969472	60.186808	154.221090	207.577791	38.201435	36.457432	91.968582	-38.232	-70.425	0
62.699508	57.084061	158.967006	208.872326	27.032017	49.298864	92.600330	-49.881	-69.025	0
63.812838	56.066555	159.505392	209.108462	26.218830	53.707721	93.335948	-54.307	-70.506	0
...
38.437590	43.429309	117.814112	145.211123	48.115745	34.247332	56.357347	-1.458	-48.461	3
34.693504	49.301733	123.990834	142.542978	45.584684	33.825000	55.258556	-2.961	-45.573	3
34.892195	52.589202	118.169935	136.510967	34.964308	35.481064	56.748124	-16.152	-48.506	3
43.500591	55.890792	114.829095	122.688746	15.874021	39.923449	56.151473	-35.247	-48.495	3
46.659229	57.216277	117.026114	117.082427	13.747250	40.319421	57.508792	-36.816	-48.441	3

Figure 5: Data snippet showing produced feature space and output class (in last column) based on calculated distances.

Figure 5 shows feature space as the distance dataset and Figure 6 indicates characteristic spatiotemporal

patterns as the distance values variations during a serve, forehand, a backhand stroke and a no stroke play.

From Figure 6, it can be clearly observed that the feature values sub plot of the backhand is similar to the forehand. However, the directions of the forehand values and backhand values are different so the approach is also robust to e.g., inside-out forehands (executed from the player’s backhand side of the court). Plots of features like hand-to-hand distance are almost identical in both forehand and backhand stroke plots. The serve is always a one-handed stroke and the dominant hand reaches above the head during a serve. Therefore, in the feature value variation plot of the serve (Figure 6), the dominant hand to dominant leg pixel distance value increases before contact with the ball and decreases after contact with the ball.

Regarding multi-class classification problem investigated here, there are four output classes representing three common stroke categories, and no stroke players’ activity (e.g., walking or running). Hence, there is no visible distinctive pattern that we could immediately associate with no stroke output class. Another example of no stroke activity is during the game breaks, where players can be taking courtside rest.

Table 1: Feature space description and design rationale.

	Distance Definition	Distance Measure	Design Rationale
1.	Dominant hand to chest	$d(P_4, P_1)$	To improve separation of serves from forehand and backhand
2.	Non-dominant hand to chest	$d(P_7, P_1)$	To improve separation of forehand and backhand strokes
3.	Dominant hand to dominant side foot	$d(P_4, P_{11})$	To improve separation of serves and strokes starting from dominant hand side
4.	Non-dominant hand to non-dominant hand side foot	$d(P_7, P_{14})$	To identify strokes starting from the non-dominant side
5.	Hand to hand	$d(P_4, P_7)$	To identify one-handed strokes and serves
6.	Dominant hand to non-dominant side hip	$d(P_4, P_{12})$	To identify the circular motion around the hip in ground strokes and to identify strokes starting from the dominant side
7.	Non-Dominant hand to dominant side hip	$d(P_7, P_9)$	To identify strokes starting from the non-dominant side or circular motion around the hip in ground strokes (forehand and backhand)
8.	Body to dominant hand x-axis distance	$P_4(x) - P_8(x)$	To identify strokes starting from the dominant side over body’s vertical (symmetrical) axis
9.	Body to non-dominant hand x-axis distance	$P_7(x) - P_8(x)$	To identify strokes starting from the non-dominant side over body’s vertical axis.

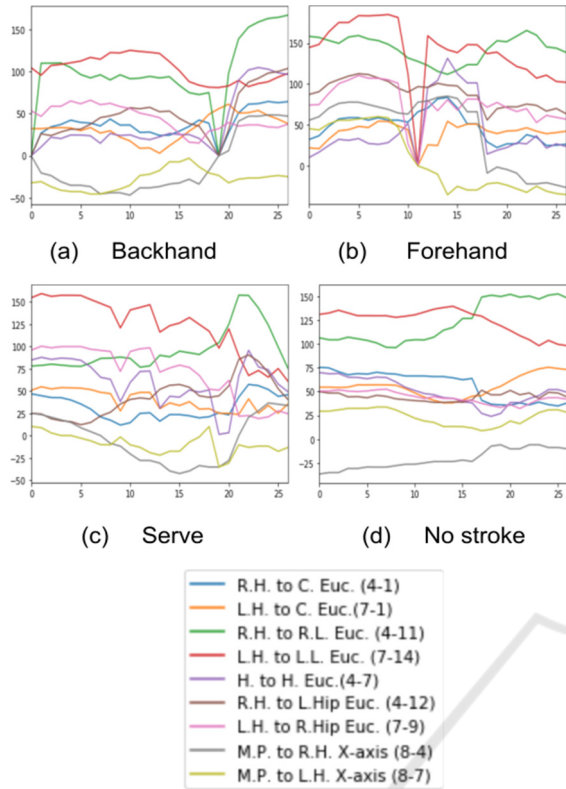


Figure 6: Spatiotemporal multi plot depicting output class patterns from feature value variations. Top-to-bottom subfigures: A serve, forehand and backhand patterns (x-axis: Frames, y-axis: Distance values). Nine colour-coded timeseries in legend are arranged by Table 1 row order.

3 CLASSIFIER MODELLING AND RESULTS

As part of data filtering process, visual inspection of collected videos, revealed that one video footage had to be removed from the dataset due to being recorded from a substantially different camera position and where the majority of the frames did not show visible player's figure.

The dataset was randomly divided into a training dataset and testing portion. 119 strokes (approx. 80%) were considered as the training dataset (24 serves, 24 forehands, 24 backhands and 47 no stroke play), and 30 strokes (approx. 20%) were considered as a testing dataset (6 serves, 6 forehands, 6 backhands, 12 no stroke play). Next, the created time series training dataset was classified by experimenting with three different LSTM networks (LSTM, Bidirectional LSTM and CNN+LSTM network). Table 2. provides the summary of produced LSTM model variations used in our experiments.

Table 2: Model and parameter summaries.

Classifier	Layer	Output Shape	Parameters	
LSTM	LSTM	(None,100)	44.000	
	Dropout	(None,100)	0	
	Dense	(None,100)	10.100	
	Dense	(None,4)	404	
Total params.			54.504	
Bidirectional LSTM	Bidirectional LSTM	(None,27,200)	88.000	
	Bidirectional LSTM	(None,200)	240.800	
	Dropout	(None,200)	0	
	Dense	(None,100)	20.100	
	Dropout	(None,100)	0	
	Dense	(None,4)	404	
Total params.			349.304	
CNN + LSTM	Time Distributed Conv 1D	(None,3,23,64)	1.024	
	Time Distributed Conv 1D	(None,3,19,64)	20.544	
	Time Distributed Dropout	(None,3,19,64)	0	
	Time Distributed Max Pooling 1D	(None,3,9,64)	0	
	Time Distributed Flatten	(None,3,576)	0	
	LSTM	(None,200)	621.600	
	Dropout	(None,200)	0	
	Dense	(None,4)	804	
	Total params.			643.972

Table 3 shows the classification results with the LSTM, Bidirectional LSTM and CNN+LSTM networks. CNN+LSTM network consists of both CNN layers and LSTM layers.

Table 3: Classification performance.

Classifier	Output class	Precision	Recall	F1-score
LSTM	Backhand	0,83	1,00	0,93
	Forehand	1,00	1,00	1,00
	Serve	1,00	1,00	1,00
	No Stroke	1,00	0,92	0,96
F1 MCS				0,97
Bidirectional LSTM	Backhand	1,00	1,00	1,00
	Forehand	1,00	1,00	1,00
	Serve	1,00	1,00	1,00
	No Stroke	1,00	1,00	1,00
F1 MCS				1,00
CNN+LSTM	Backhand	1,00	1,00	1,00
	Forehand	1,00	0,83	0,91
	Serve	1,00	1,00	1,00
	No Stroke	0,92	1,00	0,96
F1 MCS				0,97

Note: F1 multi-class score (F1 MCS). For reader's convenience, all values are rounded to two decimal points. Achieved classification performance for both LSTM and CNN+LSTM networks reached 96,67%, which were rounded to 97% (shown as 0,97), while Bidirectional LSTM network performance reached 100% for the validation dataset.

The above-expected experimental results suggest that the improved solution would include modelling and analysis of additional output classes (e.g., volleys, drop shots, serve variations) requiring (sub)phasing movement analysis. Similar to prior work on 3D kinematic data (Bačić, 2016a), the ensemble orchestration control would not only rely on a weighted probabilistic equation but also on expert's knowledge captured in a state automata machine. Such approach allows ensemble modelling on small and large dataset, where parameter optimisation and human-labelling efforts can be further reduced by transfer learning and adaptive system design.

4 IMPLEMENTATION FOR VIDEO STREAMING

Trained model can be used to classify strokes in and display overlaid text for video streaming. Model input is a spatiotemporal dataset of nine features (Table 1).

Spatiotemporal dataset subsample should be imputed to a classifier as a block of experimentally determined size of 27 frames (Figure 7). Buffering of 27 frames (of approx. 1 second) represents rolling window concept in time-series analysis, in which key points from 2D pose estimation skeleton overlay were converted into the 9 distance-based features generating a 9×27 size buffered data block.

Therefore, after 27 frames of data were buffered, a trained model (i.e. classifier) was used to detect a stroke and to classify the stroke. If a stroke is not

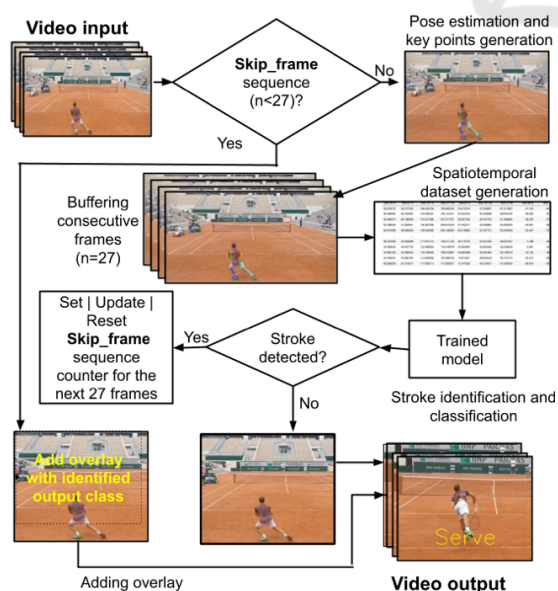


Figure 7: Strokes classification and overlay annotation as video processing workflow concept.

detected, the next rolling-window of 27 frames are buffered and supplied to the classifier. If a stroke is detected, the overlay with the identified stroke will be displayed over the next 27 frames (which are skipped from feature processing, considering minimum times between shots e.g., for the opposing player's stroke).

5 DISCUSSION

For a prototype, the classification performance results exceeded expectations for the collected dataset (with approx. 80:20% split used for model training and testing). We expect that expanding the dataset may reduce classification performance, justifying a follow-up investigation into achieving an improved solution that will generalise well on future data. Another limitation is that occasionally, OpenPose fails to generate the correct stick figure, warranting further investigations to improve overall robustness and accuracy. Further improvement is intended by using additional videos taken from other vantage points e.g. in front of the player. Considering the computational performance of pose estimation, we will look at implementation on lower cost platforms, including tablets and mobiles. In coaching scenarios, the intended platform would also process video feeds from fixed camera positions akin to the dataset used in this paper. Unlike carrying and managing inertial sensors, video is considered: (1) an unobtrusive data source not interfering with the player's feel; and (2) to minimise the possibility of motion data interpretation being contested.

During match situations, players may move closer to the net. When players are close to the net, they will perform stroke exchange in higher frequency than compared to producing strokes behind the baseline. Therefore, time between strokes may be sometimes less than a second. For the scope of this research and the proof-of-concept, one second (or longer time) splits between the strokes in video have been considered as sufficient for stroke identification and classification. Future work will involve modelling of increased number of output classes including faster strokes exchange (e.g., containing further information such as: direction, depth, drop shot and lob volleys).

6 CONCLUSION

This paper contributes to video indexing and human activity recognition by applying a multidisciplinary combination of computer vision, pose estimation and

recurrent neural networks. Related contributions to sport analytics, broadcasting and general information retrieval from low-cost video were also motivated by prior work in golf and tennis relying on sensors and 3D kinematic data. Aligned with prior work, the presented tennis stroke recognition from monocular video is also aimed at contributing to how machines can quantify, assess and diagnose aspects of human movement and provide comprehensive feedback – all contributing to the area of interpretable AI.

The presented video processing and modelling framework, using selected publicly available tennis videos, was implemented in Python on Google cloud platform. The framework uses generated trajectories of key points (represented as human stick figure video overlays) which were further transformed into the spatiotemporal feature space. Multi time data series from the feature space were processed using three variations of LSTM classifiers. As a multi-class classifier, the developed tennis shots recognition system exceeded expected performance (96,67% – 100%), and did not rely on specialist expertise or insights for developed feature extraction techniques.

Using video-based feature extraction techniques to provide diagnostic information without redundant data: (1) minimises reliance on domain expertise; (2) enables interaction with and visualisation of intermediate preprocessing operations (via animated stick figure overlays), which is also important for initially small dataset modelling, and transparent and comprehensive feature engineering process; and (3) maximises the role of AI, computer vision and pose estimation for human motion modelling and analysis (HMMA), and for advancements of sport science.

Our multi-class approach is transferable to signal processing and has been evaluated in prior work on indexing and analysis of two-class classification in cricket. Future work will include application to other sports, alongside broader contexts involving privacy-preserving filtering and data fusion from wearable and equipment-attached sensors.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the contributions of the OpenPose team, in sharing and maintaining their code and documentations.

REFERENCES

- 12kpg-Tennis. (2019, 28 Aug. 2020). Nick Kyrgios - Jeremy Chardy (4k 60fps). Retrieved from <https://www.youtube.com/watch?v=KQI6ZvE14nw>
- Anand, A., Sharma, M., Srivastava, R., Kaligounder, L., & Prakash, D. (2017). Wearable motion sensor based analysis of swing sports. In *ICMLA, 16th IEEE International Conference on Machine Learning and Applications*. Cancún, Mexico.
- Bačić, B. (2004). Towards a neuro fuzzy tennis coach: Automated extraction of the region of interest (ROI). In *FUZZ-IEEE, International Conference on Fuzzy Systems*. Budapest, Hungary.
- Bačić, B. (2016a). Echo state network ensemble for human motion data temporal phasing: A case study on tennis forehands. In *Neural Information Processing* (Vol. 9950). Springer.
- Bačić, B. (2016b). Echo state network for 3D motion pattern indexing: A case study on tennis forehands. In *Image and Video Technology. Lecture Notes in Computer Science* (Vol. 9431). Springer.
- Bačić, B. (2016c). Extracting player's stance information from 3D motion data: A case study in tennis groundstrokes. In *PSIVT 2015 Image and Video Technology*. Springer.
- Bačić, B., & Hume, P. (2012). *Augmented video coaching, qualitative analysis and postproduction using open source software*. In *ISBS, 30th International Conference on Biomechanics in Sports*. Melbourne, VIC.
- Bačić, B., & Hume, P. A. (2018). Computational intelligence for qualitative coaching diagnostics: Automated assessment of tennis swings to improve performance and safety. *Big Data*, 6(4). doi:<https://doi.org/10.1089/big.2018.0062>
- Bačić, B., Meng, Q., & Chan, K. Y. (2017). Privacy preservation for eSports: A case study towards augmented video golf coaching system. In *DeSE, 10th International Conference on Developments in e-Systems Engineering*. Paris, France.
- Bačić, B., Rathee, M., & Pears, R. (2021). Automating inspection of moveable lane barrier for Auckland harbour bridge traffic safety. *Neural Information Processing* (Vol 12532). Springer.
- Back of the line tennis. (2018). Tennis practice match. Retrieved from <https://www.youtube.com/watch?v=nqC0K4yGdxM>
- Bandara, I., & Bačić, B. (2020). *Strokes classification in cricket batting videos*. In *CITISIA, 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications*. Sydney, NSW.
- Bloom, T., & Bradley, A. (2003). Player tracking and stroke recognition in tennis video. In *APRS Workshop on Digital Image Computing*. Brisbane, QLD.
- Cai, J.-X., & Tang, X. (2018). RGB video based tennis action recognition using a deep weighted long short-term memory. Retrieved from <https://arxiv.org/abs/1808.00845v2>

- Chan, K. Y., & Bačić, B. (2018). Pseudo-3D binary silhouette for augmented golf coaching. In *ISBS, XXXVI International Symposium on Biomechanics in Sports*. Auckland, New Zealand
- Connaghan, D., Conaire, C. Ó., Kelly, P., & Connor, N. E. O. (2010). Recognition of tennis strokes using key postures. In *ISSC, 21st Irish Signals and Systems Conference*. Dublin, Ireland.
- Connaghan, D., Kelly, P., O'Connor, N., Gaffney, M., Walsh, M., & O'Mathuna, C. (2011). *Multisensor classification of tennis strokes*. In IEEE Sensors, Limerick, Ireland.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8). doi: <https://doi.org/10.1162/neco.1997.9.8.1735>
- Kos, M., Ženko, J., Vlaj, D., & Kramberger, I. (2016). Tennis stroke detection and classification using miniature wearable IMU device. In *IWSSIP, International Conference on Systems, Signals and Image Processing*. Bratislava, Slovakia.
- Liu, J., Shahroudy, A., Xu, D., Kot, A., & Wang, G. (2018). Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12). doi:<https://doi.org/10.1109/TPAMI.2017.2771306>
- Lu, J., Shen, J., Yan, W. Q., & Bačić, B. (2018). An empirical study for human behavior analysis. *International Journal of Digital Crime and Forensics*. IGI Global. <http://doi.org/10.4018/IJDCF.2017070102>
- Martin, P.-E., Benois-Pineau, J., Peteri, R., & Morlier, J. (2018). Sport action recognition with Siamese spatio-temporal CNNs: Application to table tennis. In *CBMI, International Conference on Content-Based Multimedia Indexing*. La Rochelle, France.
- Page, S. (2020). Tennis practice match points - NTRP 4.5 vs 5.0. Retrieved from <https://www.youtube.com/watch?v=dfrec4pjnI0>
- Petkovic, M., Jonker, W., & Zivkovic, Z. (2001). Recognizing strokes in tennis videos using hidden Markov models. In *IASTED, International Conference on Visualization, Imaging and Image Processing*. Marbella, Spain.
- Ramasinghe, S., Chathuramali, K. G. M., & Rodrigo, R. (2014). Recognition of badminton strokes using dense trajectories. In *7th International Conference on Information and Automation for Sustainability*. Colombo, Sri Lanka.
- Shah, H., Chockalingam, P., Paluri, B., Pradeep, S., & Raman, B. (2007). Automated stroke classification in tennis. In *ICIAR, 4th international conference on Image Analysis and Recognition* (Vol. 4633). Springer.
- Taghavi, S., Davari, F., Tabatabaee Malazi, H., & Abin, A. A. (2019). Tennis stroke detection using inertial data of a smartwatch. In *ICCKE, 9th International Conference on Computer and Knowledge Engineering*. Mashhad, Iran.
- Tenfitmen Tennis Impulse. (2020). Tennis match - player vs coach (Tenfitmen - episode 123). Retrieved from <https://www.youtube.com/watch?v=uSoD2yyzRgY>
- Tennis Legend TV. (2019, 28 Aug. 2020). Roland-Garros 2019: Federer - Schwartzman practice points (court level view) Retrieved from <https://www.youtube.com/watch?v=vkGwyke5jDU>
- Top Tennis Training - Pro Tennis Lessons. (2014, 28 Aug. 2020). Tsonga vs Anderson training match 2014-court level view. Retrieved from <https://www.youtube.com/watch?v=RHokxoEsFsc>
- TV Tennis Pro. (2020, 28 Aug. 2020). Alexander Zverev practice match vs Andrey Rublev court level view tennis. Retrieved from <https://www.youtube.com/watch?v=mcR3d9jnWal>
- Xia, K., Wang, H., Xu, M., Li, Z., He, S., & Tang, Y. (2020). Racquet sports recognition using a hybrid clustering model learned from integrated wearable sensor. *Sensors*, 20(6). doi:<https://doi.org/10.3390/s20061638>
- Zhao, Y., Yang, R., Chevalier, G., Xu, X., & Zhang, Z. (2018). Deep residual Bidir-LSTM for human activity recognition using wearable sensors. *Mathematical Problems in Engineering*, 2018. doi:<https://doi.org/10.1155/2018/7316954>