Serve, Volley, Pose: Doubles Tennis Pose Estimation for Strategy Analysis

Vanessa Bellotti and Tanay Nistala

1 Introduction

In recent years, computer vision and deep learning have made significant advancements in various fields, including sports analytics and human pose estimation. Accurate pose estimation has various applications, and one of the exciting areas is tennis, a popular sport that relies heavily on body movements for performance analysis and training.

Tennis pose estimation involves tracking and recognizing the key points on a tennis player's body in real-time, providing valuable insights into their form, technique, and movement patterns. This project aims to develop a robust and accurate pose estimation system for tennis players using a Recurrent Neural Network (RNN), and in particular a Long Short Term Memory (LSTM) body in order to capture the temporal aspects of the short term and long term poses in the match.

Pose estimation in tennis has numerous applications, such as player performance analysis, injury prevention, and enhancing coaching strategies for the player's benefit. It can provide data for measuring racket and body angles, speed, and court positioning, enabling a deeper understanding of a player's strengths and weaknesses. Furthermore, it can be leveraged to aid more casual and recreational tennis players who lack access to the sophisticated training and technology that professional players can utilize.

Current implementations of pose estimation for tennis cover the base case of a single player. However, the literature focuses on leveraging classical computer vision techniques to this end, with no current implementations of pose estimation based on input features of key points obtained from those classical techniques to then input into a deep neural network system. The focus of this project then is to to provide a framework by which a single tennis player's tennis stroke and pose can be classified by a deep neural network (DNN) based on the pose keypoints. Further extension of this work then can be expanded to cover the case of doubles tennis where each side of the court has two teammates and to cover every tennis player actively playing in a scene.

2 Background Related Work

Numerous studies in the realm of pose estimation, particularly in sports-related contexts, have significantly contributed to our understanding of this complex field. Within the domain of tennis, research efforts have been directed towards various aspects such as racket tracking, ball tracking, and player tracking. Notable works in this domain include:

"Real-Time Human Pose Detection in Parts from Single Depth Images" by Shotton et al., 2013 []:

This influential work introduced a novel method for human pose estimation utilizing a decision forest approach. The research laid the foundation for pose estimation techniques, providing insights that have transcended into subsequent studies in diverse applications.

"OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields" by Cao et al., 2018:

Widely recognized and utilized in the field of human pose estimation, this work by Cao et al. presents an advanced approach using part affinity fields. The real-time multi-person 2D pose estimation framework introduced in this paper has become a benchmark in the field, offering robustness and efficiency.

3 Technical Approach

To tackle the challenge of classifying a tennis player's pose and current stroke during a match, we employ a Long Short-Term Memory (LSTM) model trained on inputted estimations derived from keypoints capturing the tennis player's pose. To this end, we consider a singles match involving two players, each on opposite sides of the court. To streamline our analysis and enhance the accuracy of our model, we specifically target the player closest to the camera, ensuring a focused perspective that aligns with the current dynamics of the court. This deliberate isolation of key players not only refines the input data but also contributes to the precision of our LSTM-based classification system for accurately identifying both the pose and ongoing stroke in the fast-paced context of a tennis match.

3.1 Dataset

Our dataset comprised of five singles tennis matches from the London Summer 2012 Olympics obtained from [3]. In terms of annotations, the dataset contained information regarding the player closest to the camera in the field of view, information about the tennis players' serves and hits during the match, and the status of that stroke in terms of whether it succeeded or failed by the metrics of the game of tennis. These annotations had been automated in prior work, and one of this work's limiting factors is that automating annotations for further matches to generate more diverse resources would require the usage of paid

software. As such, the provided annotations were a great resource but also a limiting factor to this work's scope.

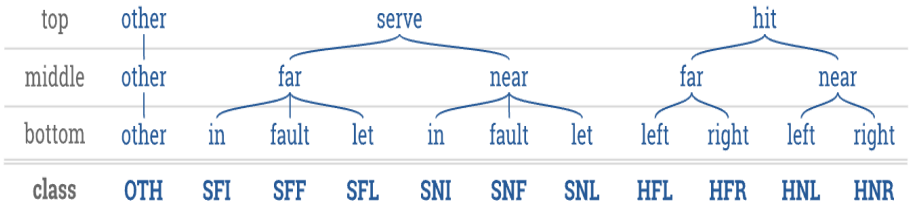

Figure 1: Hierarchical view of classes for tennis hits and serves for dataset [3]

The dataset achieved a balanced male-to-female ratio of 50:50, enabling us to evaluate the neural network in a fair manner across these gender categories. All annotated players in the dataset were right-handed.However, recognizing the prevalence of left-handed players, exemplified by the likes of the renowned Rafael Nadal, our dataset's applicability can be extended to encompass left-handed cases. This extension involves incorporating a straightforward preprocessing step for frames featuring left-handed players. By mirroring the image, this preprocessing step aligns the left-handed player's pose with the perspective our neural network has been trained to recognize. This flexibility enhances the dataset's adaptability, accommodating a broader range of tennis players and ensuring the neural network's efficacy in capturing pose and stroke dynamics, irrespective of the player's dominant hand.

In terms of lighting conditions and background, the videos were fairly diverse as some of the matches were fairly long, meaning the lightning naturally changed, and also the five matches happened at different times of the day. Furthermore, the background did pan and include some different angles, though for consistency, the most popular tennis camera position was targeted for our purposes as that is what is broadcast during matches and what recreational players would also use to record their own matches.

3.2 Algorithm

3.2.1 Data Collection and Annotation:

Gather a diverse dataset of tennis players in various playing scenarios. Leverage the annotated keypoints of tennis player body pose such as wrist, elbow, shoulder, hip, knee, and ankle joints utilizing the Tensorflow MoveNet Lightning Model [6].

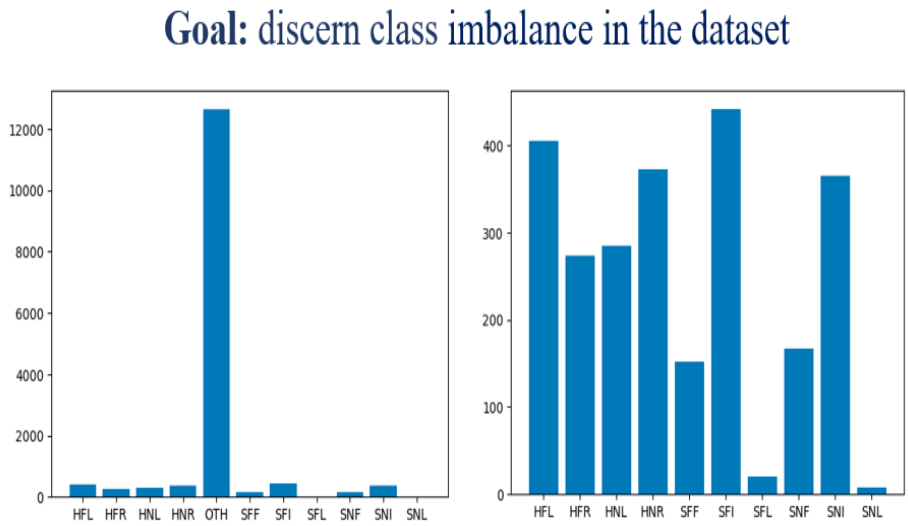

Figure 2: Class imbalance captured before and after the "Other" label was excluded from the datset before the training and testing split

3.2.2 Data Preprocessing:

Data preprocessing steps included isolating the tennis hit and serve stroke information, excluding the commentary annotations pulled in from the original dataset, isolating frames to target only the player nearest to the camera, BGR to RGB color space transformation to adjust the image display after the pose detection and classification were run, exclude the 'Other' label to only include frames relevant to our use case, reducing noise. The colorspace transformation was computed using the OpenCV library's functionality.

3.2.3 Network Architecture:

This paper proposes a two-fold system by which tennis player's hits, serves, and strokes can be classified into two predetermined labels. For the human pose detection component of the system, the Tensorflow MoveNet Lightning model was leveraged to take advantage of its incredibly fast performance, which enabled the real time performance of the system overall, as opposed to prior approaches which experienced such low frame rates that real time tracking and certainly real time classification would have been impractical and nearly impossible. Given that this application is intended to be deployed for real time tennis pose and stroke classification, speed was paramount to this project so long as a certain degree of accuracy is maintained.

From this pre-trained pose detector model, transfer learning was leveraged to extract the keypoints in the form of joints, giving us a resulting matrix composed of 17 joints and their positions in terms of x-y-z spatial coordinates.

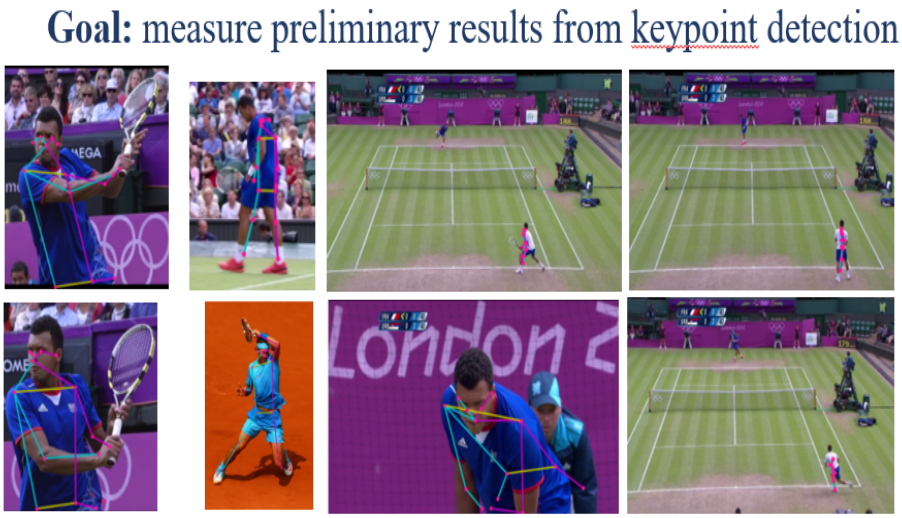

Figure 3: Samples of the key point detector on frames from the dataset and also external images of tennis players

This matrix was captured at each time, allowing for the flattened keypoints to be one-hot encoded and then input as features to the LSTM neural network that defined the second part of our two-fold system.

LSTM layers are used to capture temporal relationships, and fully leverage the sequential nature of the keypoint data. The flattened keypoints, representing the spatial coordinates of 17 joints at each time step, are fed into the LSTM neural network for the temporal analysis of the tennis player's pose dynamics. The LSTM layers enable the model to discern patterns and dependencies over time, allowing it to understand the sequential flow of movements during various tennis strokes. The hidden states within the LSTM cells retain information from previous time steps, contributing to the network's ability to capture and learn the nuances of the player's body pose evolution throughout a given sequence of frames. Additionally, the one-hot encoded keypoints serve as meaningful features, preserving the spatial relationships between joints while facilitating the extraction of temporal patterns critical for accurate stroke classification. This approach not only enhances the model's interpretability but also exploits the inherent temporal dependencies in the data, resulting in a more robust and effective system for real-time tennis pose and stroke classification.

3.2.4 Training:

Train the LSTM model using the annotated dataset. Utilize a loss function that penalizes the model for the discrepancies between the predicted label classifications for the tennis strokes and ground truth. Model performance was enhanced by leveraging the pretrained pose detector, which had been trained on a diverse

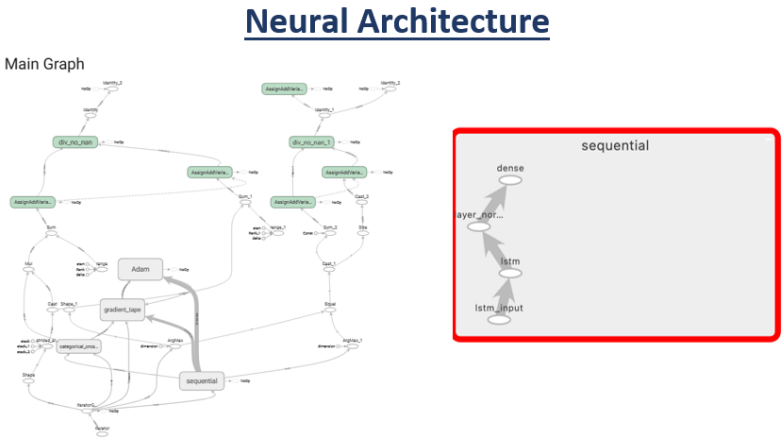

Figure 4: Neural architecture of LSTM trained to classify tennise hits and serves based on kind of stroke and ball placement

and large dataset in the past, to serve as an input feature to our own model.

3.2.5 Inference:

Deploying the trained model for real-time pose estimation during tennis matches or practice sessions would enable us to further determine the generalizability of the model on different courts and in different conditions.

4 Evaluation

We evaluated our model using categorical accuracy as the training metric for the optimizer, and we computed precision, recall, and F1 scores after the fact to verify the model's performance. Precision was chosen as it indicates the quality of positive predictions made by the deep neural network as it refers to the number of true positives divided by the total number of positive predictions. In this case, it is important that classification of tennis strokes are precise as coaches or tennis umpires would want high confidence that positive predictions are correct. Recall was chosen as, being the true positive rate, it indicates percentage of data samples that the model correctly identified as belonging to a class of interest out of the total samples. The harmonic mean of precision and recall, F1 Score, enabled us to see whether the trade off between the two was well balanced, which in turn enabled us to see a measure of how robust the system was to noise in the data and to uncertain or unanticipated outcomes. Given our two-fold system to this problem involving both pose detection and classification by a LSTM, this trade-off came into play often and was important to understand with how our dataset and model were behaving.

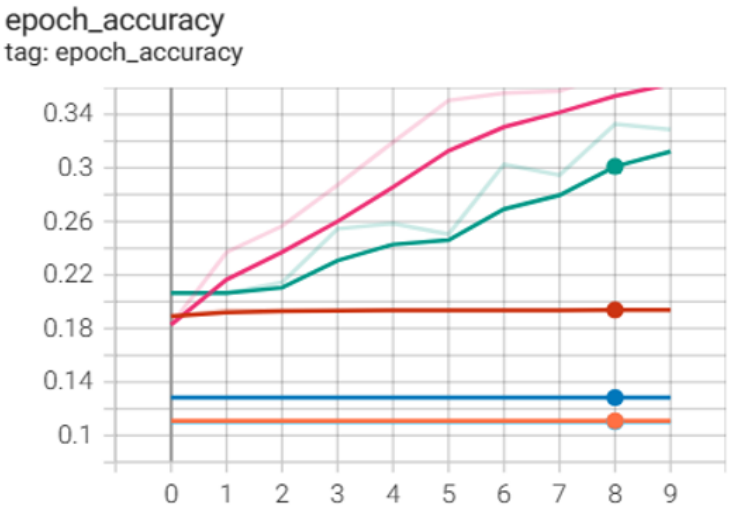

Figure 5: Classification accuracy of preliminary models, showcasing the known issue of LSTMs with only one class learned, thus having a flat accuracy plot. This was crucial to our evaluation understanding as it highlighted a known issue with LSTM training and encouraged us to reconsider class imbalance, making model training iterative

5 Results

On the categorical accuracy, precision, recall, and F1 scores, all metrics remained balanced around 58%, confirming that the model produced was robust to noise in the input videos as well as the underlying pose detector. This was especially important as the faster MoveNet Lightning model used to maintain real-time performance also came with the trade-off of lower accuracy than if we had used the MoveNet Thunder model, which meant that the model had to be able to handle unreliable pose data. The model's ability to perform in real-time was also verified by ensuring that processing times remained under a second per 60 frames even after running our model on top of the pose detection layer.

The results from this work, shown below, highlight the difference between consistency and raw performance in the context of machine learning models. As shown below our preliminary model, which utilized two LSTM layers, performed better than our final model that used an additional LSTM layer, with a validation accuracy higher by about 8 percentage points. However, the shallower model also had much more variable performance, and had a larger gap between training and validation accuracy which suggested it was overfit. In contrast, the final model's training and validation accuracies are much closer over each epoch.

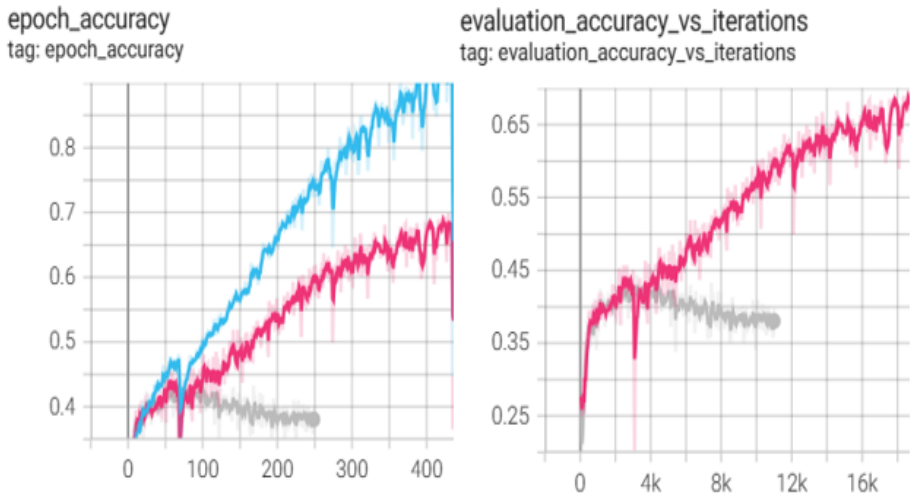

Figure 6: Epoch accuracy for the highest performing model, which had variable performance

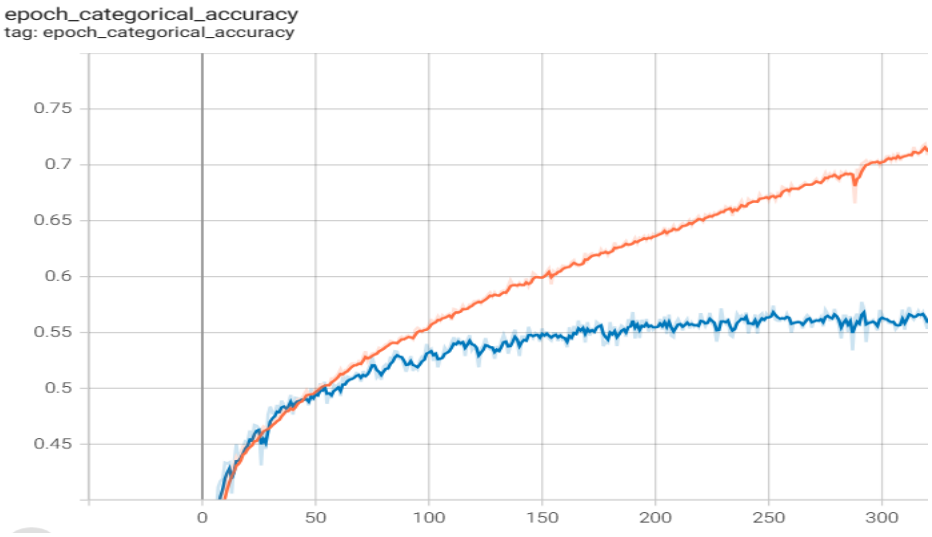

Figure 7: Epoch accuracy for the most consistent model, which had lower validation accuracy but less variability in performance over time and over situations

6 Future Work

With a classifier trained to accurately detect what serve or hit a tennis player is performing at a given moment given their image (and given the back-end computation of pose detector extracted body keypoints) having been successfully trained, there are multiple ways in which this neural network can be leveraged in future work as a backbone upon which further deep learning models can be trained.

One area for future work involves applying the LSTM architecture we have outlined to a multi-person tracking system. That is, our current methodology focuses on only classifying the actions of one person - the player closest to the camera. In the future however, this model could be leveraged to track both opponents in a singles match or, in additional extension, to cover the case of a doubles match wherein two pairs compete against each other. In the second case, an additional neural network could be applied that receives as input the classification of each player's individual poses and strokes on a team and then classify the pair's current collective behavior. This was initially proposes for this work, but the lack of any annotations publicly available for doubles matches made it an impractical problem to consider within the current scope. Nevertheless, our developed pose estimation and LSTM combined system could contribute to this field.

7 Collaboration Note

This work was authored by two individuals in a collaborative effort that was shared equally. From ideation to experimentation to preprocessing to training to evaluation to paper writing, all responsibility and credit should be shared equally.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[2] Cao et al. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2018.

[3] Hayden Faulkner and Anthony Dick. Tenniset: A dataset for dense fine-grained event recognition, localisation and description. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE.

[4] Soo Min Kang and Richard P. Wildes. Review of action recognition and detection methods, 2016.

[5] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *CVPR 2011*, pages 1297–1304, 2011.

[6] Ronny Votel and Na Li. Next-generation pose detection with movenet and tensorflow.js. blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html.

[7] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking, 2018.