Predictive Pre-Training of Tennis Shot Embeddings

Final Project Presentation

Project/Team ID: TECO
George Tamer (gyt2107)
Flor Sanders (fps2116)
Tawab Safi (as7092)

EECS E6691 Advanced Deep Learning, 2024 Spring

Outline

1. Introduction
2. Methodology
3. Automated Video Annotation
4. Deep Learning Model Architecture
5. Model Pre-Training
6. Downstream Task Fine-Tuning
7. Conclusion
8. Future Work
9. References

Introduction

Tennis = Skill + Strategy

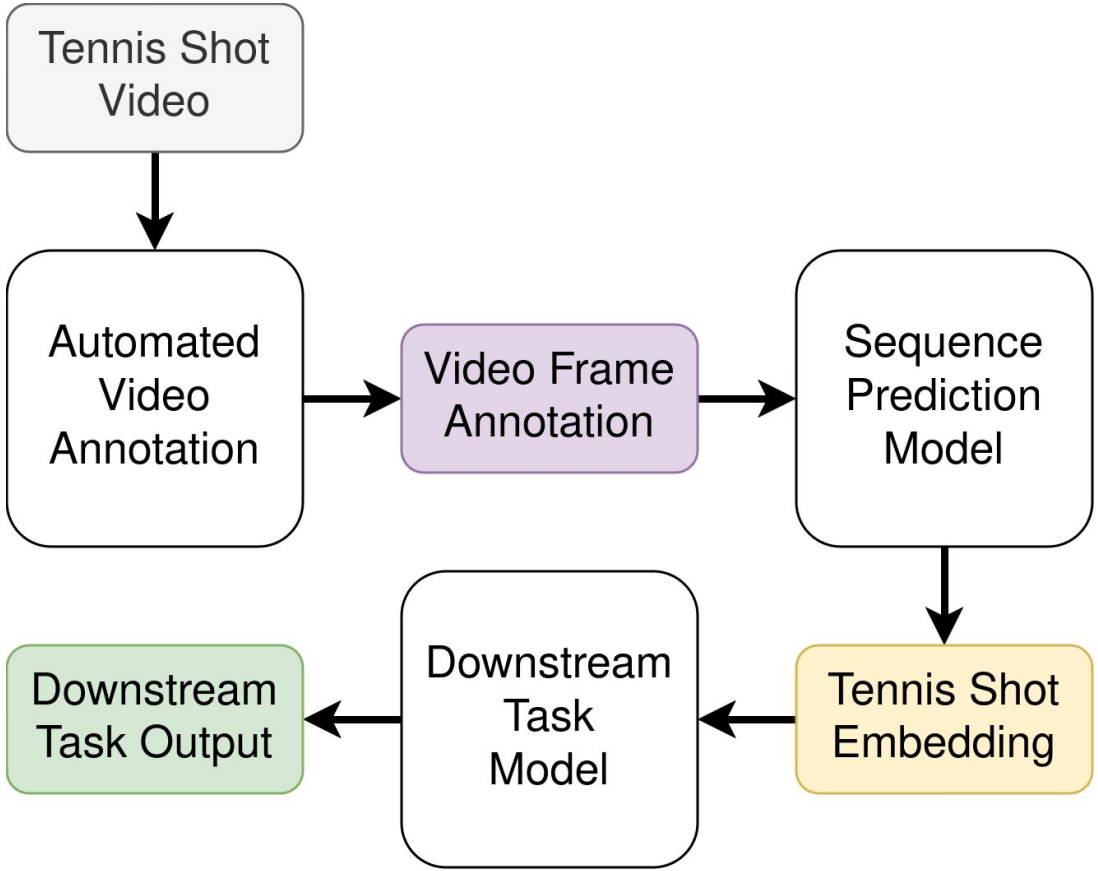
- Lots can be learned from:
  - player movements
  - ball trajectories
- Game analysis and coaching are big businesses
- Prohibitively expensive for most people

→ **Solve using Deep Learning**

- Predictive pre-training of shot embeddings
- Use embeddings as features in downstream tasks

Tennis Game [Wikipedia]

Methodology



Video Annotation - Dataset

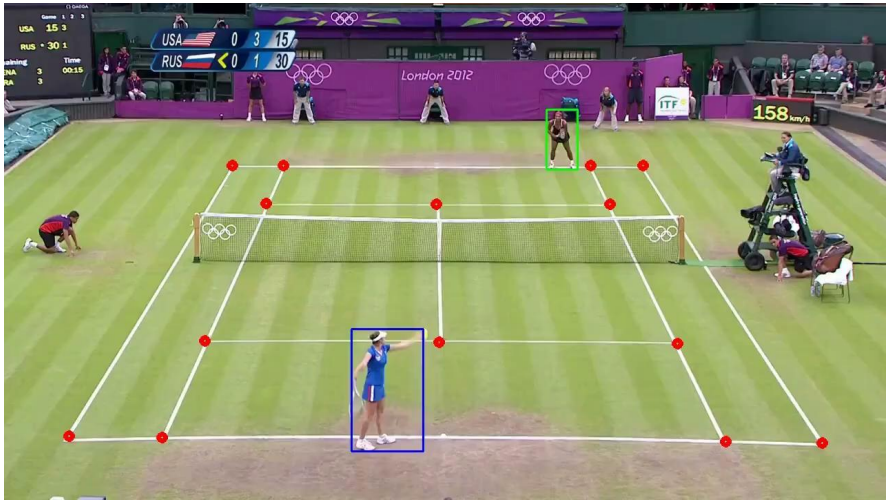
Tenniset

- 2012 London Olympic Games
- Five full-length tennis matches
  - (V006 - V010)
- Start/end frame annotations
- Serve in/out annotations
- Hit type annotations
  - flat/topspin/slice/unsure
  - forehand/backhand

Video Annotation - Ball, Court & Player Detection

TennisProject

- Ball Detection
  - Tracknet
- Court Detection
  - Modified Tracknet
  - 14 Court Points
- Player Detection
  - Faster R-CNN trained on COCO
  - Heuristics to identify playes



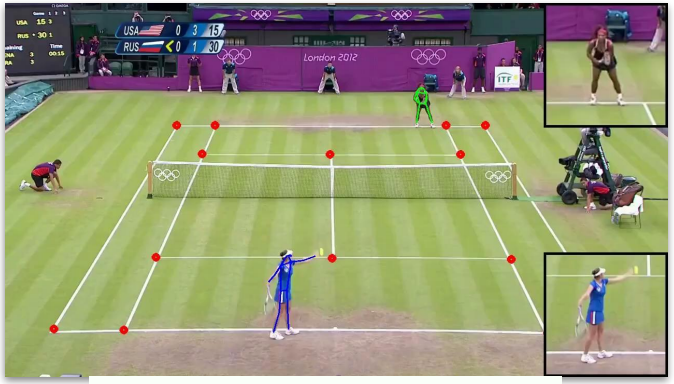
→ **Extract 2D player position on court using homography.**

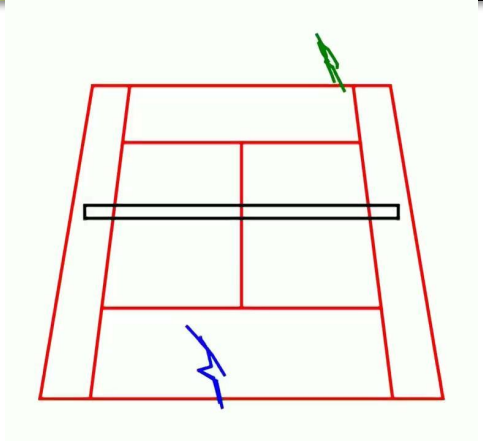
Video Annotation - 2D & 3D Player Pose Detection

MMPose

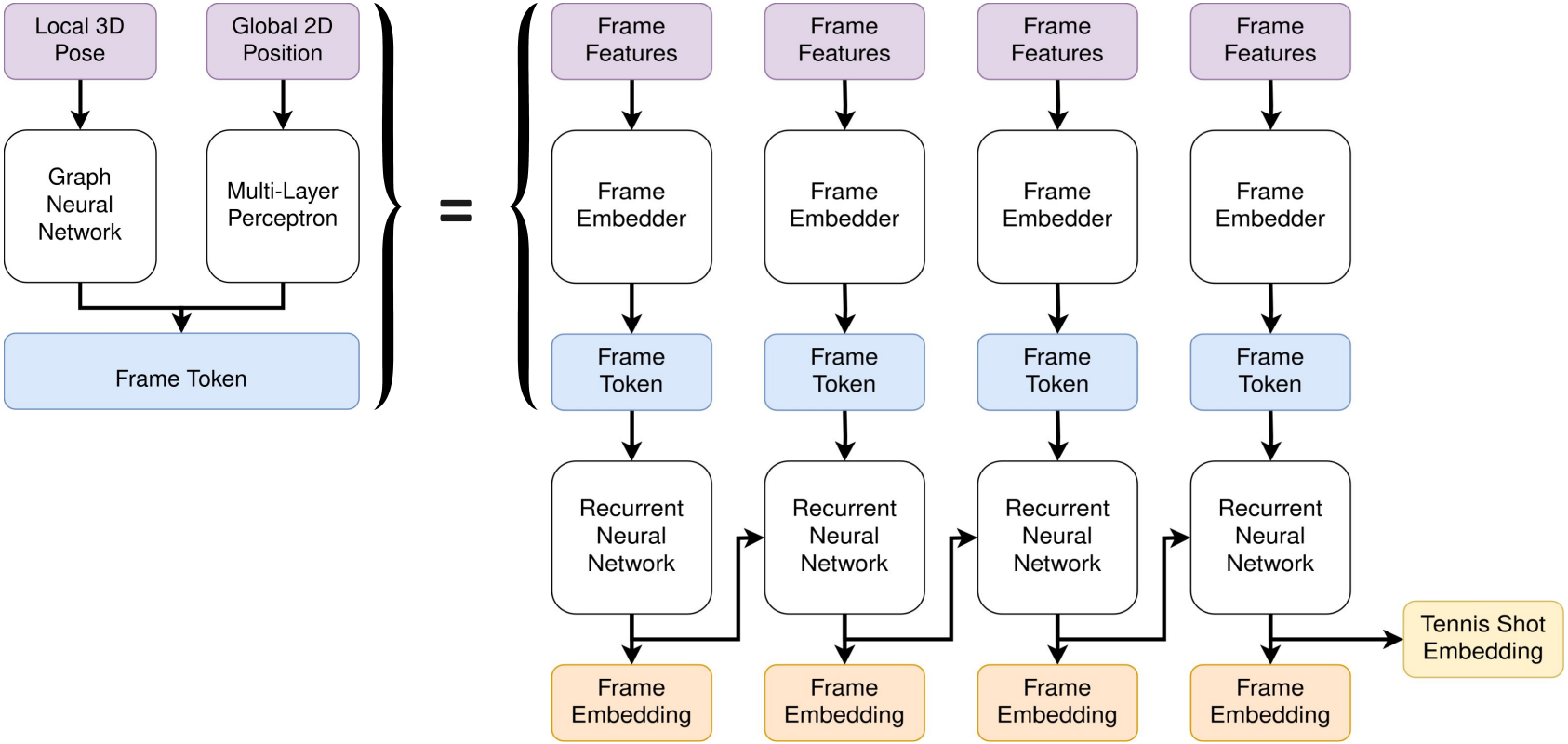
- Framework for pose detection models
- 2D Pose Detection
  - RTMDet + RTMPose
  - Refine bounding box estimates
- 3D Pose Detection
  - MotionBERT
  - Fix orientation by matching 3D to 2D poses
- Total of 3,183 processed shot segments

→ **Reduced frame dimensionality from 1280x720x3 pixels to 2x17x3 coordinates**



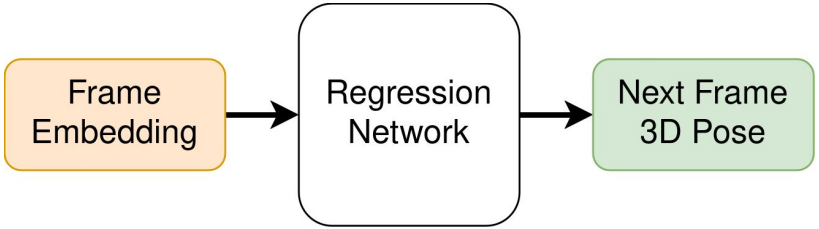


Deep Learning Model Architecture



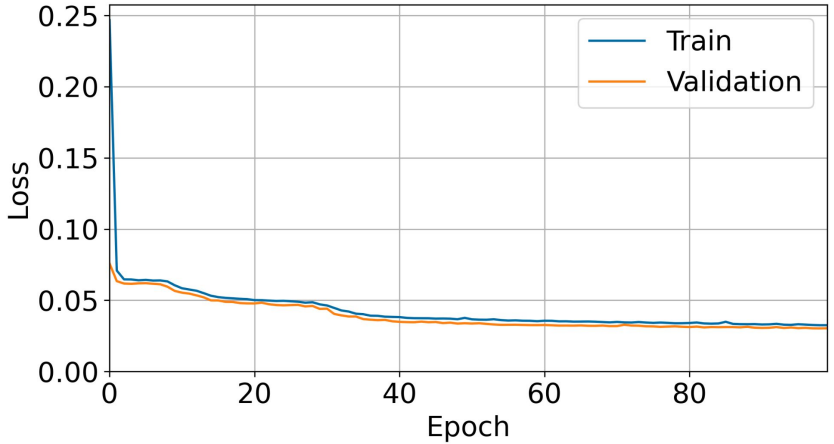
Model Pre-Training

Pre-Training Task



- Goal: Learn meaningful embedding
- Task: Predict next 3D pose
  - Inspired by ELMo
- Train model end-to-end
- Test Loss: 0.0331

Training History

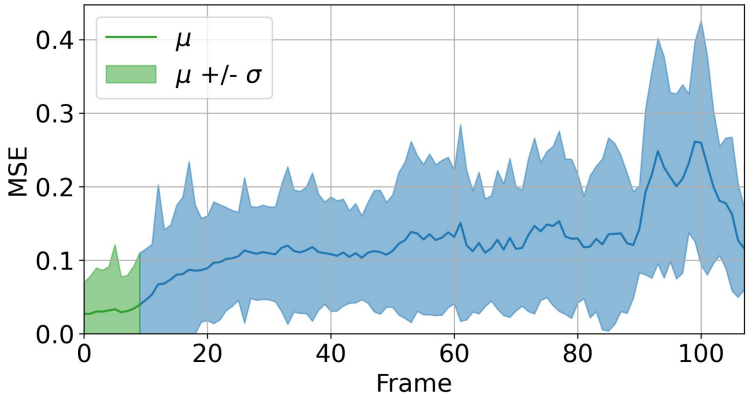


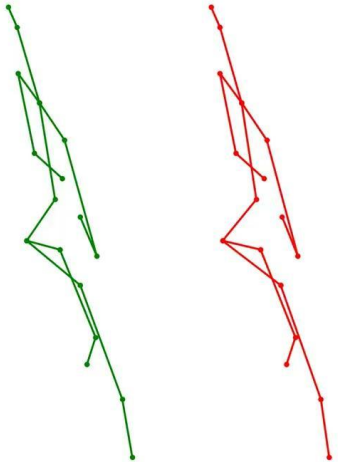
Model Pre-Training Results

Pretrained Model Evaluation

From the first 3D pose and 2D position, predict all subsequent 3D poses iteratively.

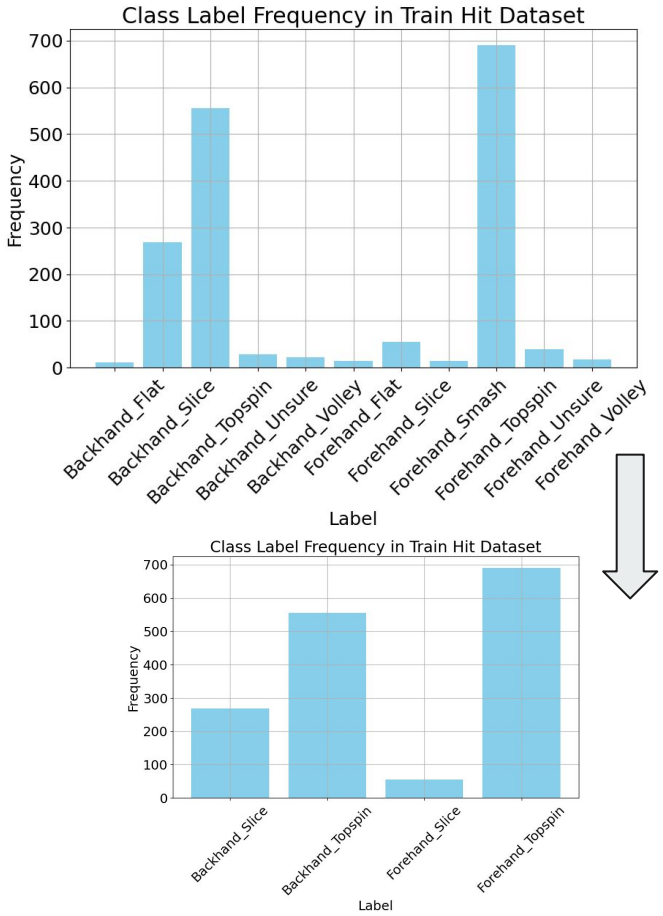
→ **Record quality drift over time**





Downstream Task - Shot Classification

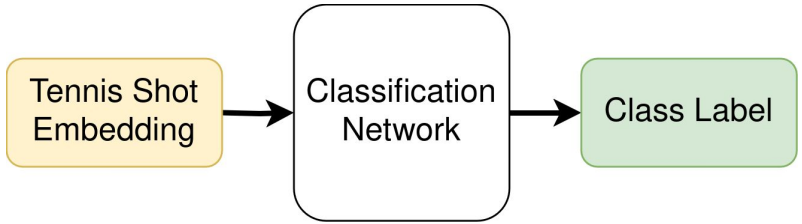
- Task: given an embedding of the shot, predict the shot type
- 11 classes in original test dataset
- Kept top 4 classes due to extreme shot imbalance



Downstream Task - Network Surgery and Fine-Tuning

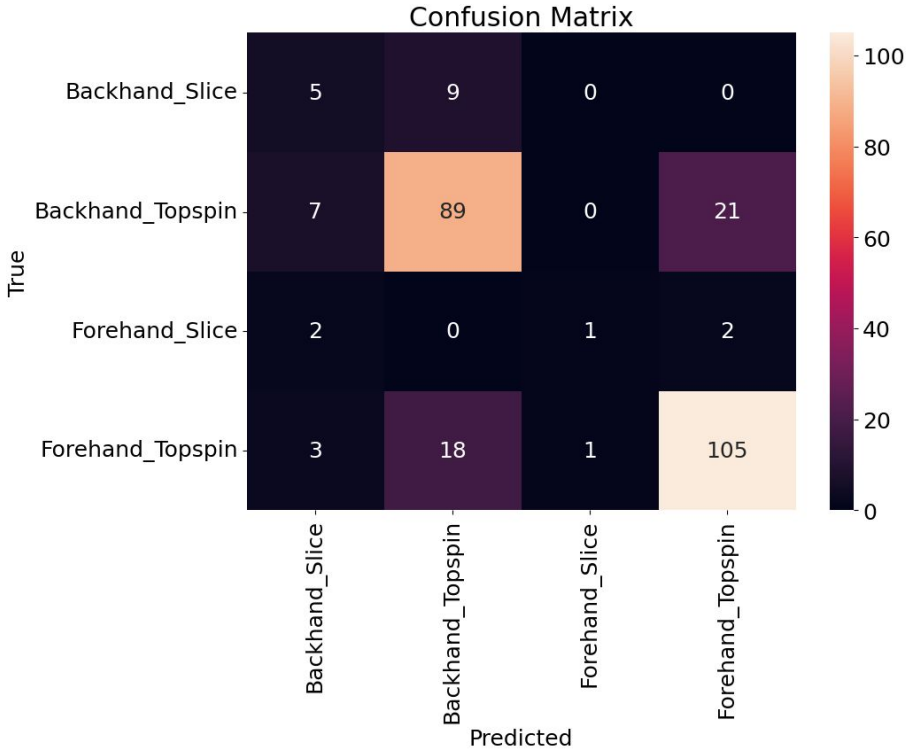
Procedure

- Add a simple MLP head to the model
- Freeze all weights except for the MLP
- Train for N epochs
- Unfreeze all weights
- Continue training



Downstream Task - Results

- Accuracy: 0.7605
- Precision: 0.7626
- Recall: 0.7605
- F1 Score: 0.7601



Conclusion

- Learning representations of sports shots through video -> 3d workspace -> graph construction is a promising direction for future research.

- Performing sequence modelling over continuous action spaces is a challenging task, and requires careful data curation and network architecting.

- The learnt representations are capable enough to serve as inputs to downstream tasks.

Future Work

Data Collection

- Quantity
  - Process more tennis matches
- Quality
  - Hand-keypoint extraction
  - Track ball and racket in 3d
- Modality
  - Collect { shot-video , caption } data samples to enable novel downstream tasks such as searching and coaching

Model Improvements

- Graph Construction
  - Use a temporal graph
  - Multiplayer graphs
- Encoder Variants
  - Try pre-training with BERT-style masking & reconstruction
- Multimodal
  - Contrastive learning on graph and text modalities
- Learning player style

References

- **Our Github:** https://github.com/ecbme6040/e6691-2024spring-project-TECO-as7092-gyt2107-fps2116
- Tenniset: https://github.com/HaydenFaulkner/Tennis
- TennisProject: https://github.com/yastrebksv/TennisProject
  - TrackNet: https://arxiv.org/abs/1907.03698
  - TennisCourtDetector: https://github.com/yastrebksv/TennisCourtDetector
  - Faster R-CNN: https://arxiv.org/abs/1506.01497
- MMPose: https://github.com/open-mmlab/mmpose
  - RTMDet: https://arxiv.org/abs/2212.07784
  - RTMPose: https://arxiv.org/abs/2303.07399
  - MotionBERT: https://arxiv.org/abs/2210.06551
- RNN Literature Review: https://arxiv.org/abs/1912.05911
- Graph Attention Network: https://arxiv.org/abs/1710.10903
- BERT: https://arxiv.org/abs/1810.04805
- CLIP multimodal embeddings: https://arxiv.org/abs/2103.00020
- V-JEPA:https://ai.meta.com/research/publications/revisiting-feature-prediction-for-learning-visual-representations-from-video/