Abstract

Predicting outcomes in competitive sports is a complex and high-impact task that benefits from recent advances in deep learning and multi-modal data analysis. However, fine-grained outcome prediction in sports like badminton remains underexplored, especially at the point-by-point level, where rapid player movements and subtle contextual cues make prediction challenging. This study aims to address this gap by developing a model that can accurately predict point outcomes using lightweight, scalable inputs derived from broadcast videos.

To this end, we propose a novel point-by-point winning prediction model for badminton, leveraging two complementary modalities: player posture and environmental audio. The Badminton Winning Prediction dataset, consisting of 12 full-length singles matches, was manually annotated and preprocessed to extract synchronized features from both video and audio streams.

Our proposed architecture consists of two key components. First, a Cross-Modal Fusion Module integrates Mel-Frequency Cepstral Coefficient (MFCC) and delta features extracted from environmental audio, with 2D body keypoints extracted using MMPose, through a bi-directional cross-modal attention mechanism, to form stage-level representations. Second, a Cross-Stage Fusion Module combines rally and non-rally segments via attention and gated mechanisms to capture temporal interactions across gameplay phases. The fused features are then aggregated over a sliding window of consecutive points and passed into a classification head (Linear, Long Short-Term Memory (LSTM), or Transformer) for point-wise outcome prediction.

Extensive experiments show that our multi-modal fusion strategy significantly outperforms single-modal baselines and simple fusion methods. Notably, the inclusion of non-rally segments and audio information contributes to improved accuracy and F1-score, highlighting the value of contextual cues beyond active play. Our best model achieves an accuracy of 88.75% in binary classification, demonstrating the practical feasibility of fine-grained prediction using lightweight and scalable inputs.

These results demonstrate that our multi-modal fusion approach effectively captures predictive information in badminton matches. In particular, even a simple linear classifier achieves excellent performance, highlighting the strength of the fused features. The significant improvements gained by incorporating audio cues and non-rally segments confirm the importance of contextual information beyond active play.

The proposed framework not only advances badminton analytics but also offers a generalizable approach for other racket sports such as tennis and table tennis, especially in scenarios with limited camera views or sensor inputs.

Keywords: Badminton, Multi-modal Fusion, Winning Prediction, Attention Mechanism, Sports Analytics