ISSN 2288-4866(Print)
ISSN 2288-4882(Online)
https://www.jiisonline.org
J Intell Inform Syst 2026 June: 32(2): 273~287 http://dx.doi.org/10.13088/jiis.2026.32.2.273

TRMNet: 시간 헤드 기반 멀티모달 네트워크를 활용한 사용자 촬영 테니스 영상의 랠리 구간 자동 추출

정영훈
주식회사 바운드원
(younghun@boundone.co)

최학열
연세대학교 인공지능융합대학
(hychoe@yonsei.ac.kr)

본 연구는 실사용자가 직접 촬영한 미편집 테니스 경기 영상에서 랠리 구간을 자동 추출하는 멀티모달 네트워크 TRMNet(Tennis Rally Multimodal Network)을 제안한다. 사용자 촬영 영상은 실제 플레이 구간이 전체의 약 50%에 불과하며, 랠리 경계는 숙련된 라벨러 간에도 F1 약 0.88에 그치는 모호성을 갖는다. 이에 TRMNet은 경량 비디오 백본 causal MoViNet A0에 BiLSTM·Temporal Self-Attention 시간 헤드와 오디오 Late Fusion 브랜치를 결합하여 시각·청각 정보를 통합하고, Soft Label Ramp Mask·Focal-BCE 결합 손실·Modality Dropout·차등 학습률을 적용하고, Clean-to-Noisy 3단계 커리큘럼 러닝을 보조 전략으로 활용하여 소규모 이질 데이터에서의 파인튜닝 안정성과 일반화를 확보한다. Ablation 실험을 통해 시간 헤드의 기여가 AP_smooth 기준 0.1556으로 가장 크게 나타남을 확인하였다. 자체 구축한 테니스 경기 영상 118편(총 43.6시간)을 대상으로 한 실험에서 Test F1 0.8422, AP_smooth 0.9315를 달성하였으며, 실서비스(PerfectSwing, 바운드원)에 배포하여 비플레이 구간 약 50% 자동 제거의 실환경 적용 가능성을 확인하였다.

주제어 : 테니스 영상 분석, 랠리 세그멘테이션, 멀티모달 학습, 커리큘럼 러닝, TRMNet

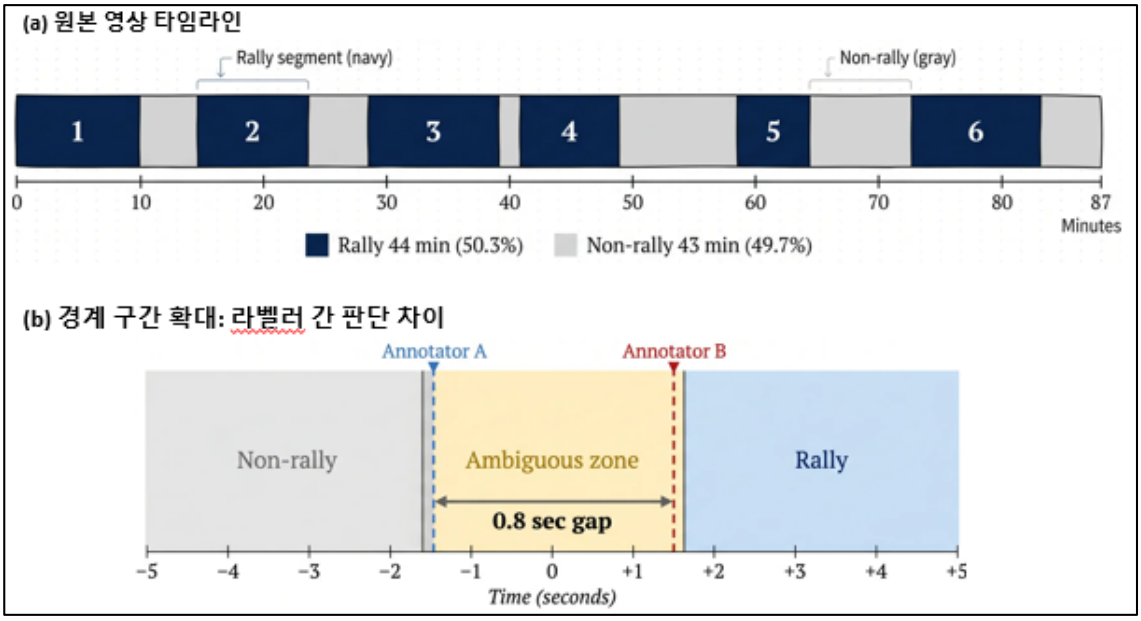
논문접수일 : 2026년 4월 27일 논문수정일 : 2026년 6월 1일 게재확정일 : 2026년 6월 2일
투고유형 : Fast track 교신저자 : 최학열

1. 서론

실사용자가 직접 촬영한 테니스 경기 영상에서 실제 플레이 구간은 전체의 약 50%에 불과하다. 그러나 액션 인식(Wu et al., 2022)과 이벤트 검출(Zhao et al., 2025) 연구들은 편집 완료된 방송 영상을 전제하며, 나머지 절반인 비플레이 구간 문제를 정면으로 다루지 않는다. 체인지오버, 서브 준비, 공 줍기 등이 차지하는 이 절반을 수작업으로 제거하려면 2시간 경기 영상에서 1시간의 추가 작업이 불가피하다.

랠리 구간 자동 추출은 다양한 실용적 가치를 지닌다. 사용자가 자신의 경기 영상에서 비랠리 구간을 자동 제거하여 하이라이트 영상을 손쉽게 제작하거나, 경기 분석 도구의 자동 입력으로 활용하여 개인의 기술 분석 효율을 높일 수 있다.

한편 본 연구가 일반 사용자가 촬영한 영상을 대상으로 하는 이유는 기존 상용 솔루션의 명확한 한계 때문이다. 대표적 상용 서비스인 SwingVision은 2026년 5월 기준 Apple Core ML 가속을 활용한 온디바이스 추론을 채택하여 iPhone 등 특정 기기에서만 동작하며, 1080p / 60fps 이상의 고해상

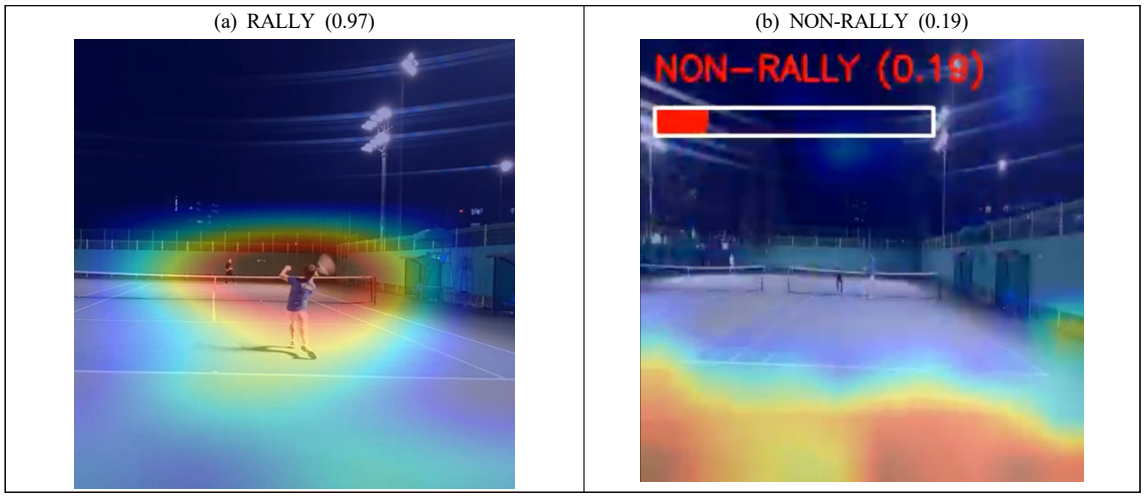


〈그림 1〉 사용자 촬영 테니스 경기 영상의 랠리 구간 분포(a)와 경계 모호성(b)

도·고프레임률 영상을 요구한다. 이를 클라우드 서버 형태로 확장하여 다양한 기기 환경에 대응하려면 고해상도·고프레임률 영상에 대한 추론 비용이 막대하여 대규모 일반 사용자 대상 서비스화가 현실적으로 어렵다. 따라서 저해상도 입력·저사양 환경에서도 동작하는 경량 모델의 필요성이 크며, 본 연구는 224×224 픽셀의 저해상도 입력에서 동작하는 멀티모달 네트워크 TRMNet을 제안하여 이러한 사각지대를 해소하고자 한다.

이 판별을 어렵게 하는 것은 분량 문제 이전에 경계 모호성의 문제이다. 테니스 랠리의 시작과 종료는 순간적으로 결정되지 않으며, 서브 준비 동작부터 실제 타구까지, 마지막 타구 후 포인트 확정까지 모호한 전환 구간이 존재한다. 이 모호성을 정량화하기 위해 본 연구에서는 숙련된 라벨러 2인이 동일 영상을 독립적으로 라벨링하는 예비 측정을 수행하였다. 그 결과 프레임 단위 F1이 약 0.88 (Cohen's Kappa $\kappa \approx 0.83$)에 그쳤으며, 이는 경계 판별이 인간 전문가에게도 비자명한 과제임을 정량적으로 보여준다. 이 수치는 동시에 AI 모델이 도달해야 할 인간 기준선으로 기능한다. <그림 1>은 실제 사용자 촬영 영상 1편의 랠리 구간 분포(a)와 동일 경계에 대한 두 라벨러의 판단이 어긋나는 경계 모호성(b)을 보여준다.

테니스 영상 분석에 특화된 선행 연구들, 예컨대 TrackNet(Huang et al., 2019)과 TTNet(Voeikov et al., 2020)은 전문 카메라 또는 방송 화질의 고정 앵글 영상을 전제로 설계되어 있어 편집자에 의해 경계가 이미 처리된 환경을 암묵적으로 가정한다. 상용 테니스 영상 분석 서비스 역시 삼각대 고정 촬영 환경을 전제로 설계되어 있어 카메라 흔들림·대각선 구도·역광이 포함된 일반 사용자 촬영 영상에서의 적용을 공식적으로 제한한다. 더불어 이들 연구는 시각 정보에만 의존하며, 타격음·심판



〈그림 2〉 causal MoViNet A0 Block 4의 Grad-CAM 활성화 시각화: (a) 랠리 구간에서 라켓 움직임과 타격 순간에 집중 활성화, (b) 비랠리 구간에서 코트 하단으로 활성화 분산

콜·관중 반응 등 랠리 경계의 강력한 청각 단서를 학습에 통합하지 않는다. TrackNet 계열의 공 추적 기반 접근은 카메라 고정과 선명한 화질을 강하게 전제하므로, 흔들림·대각선·역광이 불가피한 실사용 환경에서 추적 실패를 야기하고, 이는 곧바로 랠리 판정 오류로 전이된다. 이러한 시각 정보 단독 의존의 한계는 사용자 촬영 환경에서 랠리 경계 판별을 어렵게 만들며, 오디오와 같은 보완적 청각 단서를 통합할 필요성을 제기한다.

이 태스크를 딥러닝으로 접근할 때 두 번째 문제가 나타난다. 사용자 촬영 영상은 동일 도메인 내에서도 촬영 품질이 이질적으로 분포한다. 삼각대 고정·정면 구도의 고품질 영상(이하 Clean data)과 카메라 흔들림·대각선 구도·역광이 포함된 저품질 영상(이하 Noisy data)이 혼재하며, 본 연구 데이터셋 118편 중 Noisy data가 약 55%를 차지한다. 사용자 촬영 테니스 영상에 대한 대규모 공개 데이터셋이 존재하지 않아 소규모 자체 구축 데이터로 학습이 불가피한 조건에서, 영상 품질이 이질적인 사용자 환경 전반에 대한 일반화를 확보할 학습 전략이 필요하다. 이는 Noisy data를 고려하지 않은 학습 전략이 일반화에 구조적으로 불리함을 확인시켜 준다.

두 문제로부터 본 연구의 핵심 연구 문제가 도출된다. 연구 문제 1: 사용자 촬영 환경의 이질적 화질 조건과 모호한 경계 구간에서, 실사용 환경에서 품질이 불완전할 수 있는 오디오 신호를 시각 정보와 Late Fusion(Baltrušaitis et al., 2019)으로 결합하는 것이 랠리 경계 판별 정확도를 보완적으로 기여하는가? 연구 문제 2: 소규모 이질 데이터에서, Clean data 우선 난이도 순서로 학습하는 커리큘럼 전략이 단순 전체 데이터 학습 대비 일반화 성능을 향상시키는가?

본 연구는 이 두 연구 문제에 대응하여 멀티모달 네트워크 TRMNet(Tennis Rally Multimodal Network)을 제안한다. 기여는 세 가지이다. 첫째, 공 추적에 의존하지 않고 영상·오디오의 시간적 패턴을 직접 학습하는, 사용자 촬영 미편집 환경

특화 멀티모달 세그멘테이션 네트워크 TRMNet을 제안하였다. 둘째, 소규모 이질 데이터 환경에서의 파인튜닝 안정성과 일반화를 위한 네 가지 학습 전략과 보조 커리큘럼 학습을 체계화하였다. 네 가지 학습 전략은 라벨 경계 부드러운 처리(Soft Label Ramp Mask), 클래스 불균형 · 경계 혼동에 동시 대응하는 결합 손실 함수(Lin et al., 2017), 모달리티 누락 상황 대응 학습 기법, 사전학습 가중치 보존을 위한 차등 학습률로 구성되며, 보조 전략은 영상 품질 등급을 난이도 지표로 활용하는 3단계 커리큘럼 학습이다. 셋째, 사용자 촬영 테니스 경기 영상 118편(총 43.6시간)을 수집 · 라벨링하여 자체 데이터셋을 구축하고, 인간 기준선(F1 약 0.88) 대비 F1 0.8422를 달성하여 자동화의 실용적 한계를 실증하였으며, 실서비스(PerfectSwing, 바운드원)에 배포하여 비플레이 구간 약 50% 자동 제거의 실환경 적용 가능성을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 정리한다. 3장에서는 TRMNet의 구조와 학습 전략을 기술한다. 4장에서는 데이터셋과 실험 결과를 분석한다. 5장에서는 결론과 향후 연구 방향을 제시한다.

2. 관련연구

2.1 스포츠 영상 딥러닝 개관

스포츠 영상을 대상으로 한 딥러닝 연구는 지난 수년간 액션 인식, 객체 추적, 이벤트 검출, 하이라이트 요약 등의 하위 과제로 꾸준히 확장되어 왔다. Wu et al.(2022)은 스포츠 액션 인식 연구를 데이터셋 · 모델 · 응용 관점에서 포괄적으로 정리하였으며, Zhao et al.(2025)은 인식 · 이해 · 의사결정의 세 단계 관점에서 스포츠 딥러닝 연구 동향을 제시하였다. 두 서베이는 공통적으로 대부분의 연구가 짧은 클립 단위의 분류나 방송 영상 기반 이벤트 검출에 편중되어 있음을 지적한다.

본 연구가 다루는 랠리 구간 자동 추출(세그멘테이션)은 이러한 기존 과제와 구분된다. 수십 분 분량의 긴 미편집 영상 전체에 걸쳐 프레임 단위로 랠리 · 비랠리를 이진 분류하고, 시간적 경계를 일관성 있게 판별해야 하는 시계열 세그멘테이션(Temporal Action Segmentation) (Farha & Gall, 2019) 문제이기 때문이다. 따라서 짧은 클립 단위의 액션 인식 모델이나 요약형 하이라이트 검출 모델을 그대로 적용하기는 어렵다.

2.2 테니스 · 라켓 스포츠 영상 분석

라켓 스포츠 영상 분석 연구는 주로 공 · 선수 추적과 이벤트 검출을 중심으로 전개되어 왔다. Huang et al.(2019)이 제안한 TrackNet은 고속 · 소형 객체인 테니스공의 궤적을 딥러닝 기반으로 추적하기 위한 구조로, 이후 배드민턴 등 유사 라켓 스포츠 영상 분석 연구로 확장되었다(양준석 외, 2022). 이러한 접근은 공이나 선수의 위치 · 궤적 정보로부터 랠리 상태를 간접적으로 추정할 수 있다는 장점이 있다.

그러나 공 추적 기반 방법은 카메라가 고정되어 있고 영상 화질이 선명하다는 전제를 강하게 요구한다. 사용자가 직접 촬영한 테니스 영상은 삼각대나 스마트폰 스탠드를 활용하더라도 미세한 흔들림, 대각 촬영, 역광과 같은 요인을 포함하며, 이 경우 공 추적이 자주 실패한다. 추적 실패는 곧바로 랠리 판정 오류로 전이되므로, 실제 사용자 환경에서의 일반화에는 한계가 있다.

기존 스포츠 영상 자동 캐스팅 연구(김병조·최용석, 2019)는 본 연구가 대상으로 하는 사용자 촬영 영상과는 상이한 촬영·편성 조건을 전제하므로, 해당 접근을 본 연구의 입력 분포에 그대로 이전하기는 어렵다. 또한 위의 선행 연구들은 대체로 시각 정보에만 의존하며, 타격음이나 심판 콜과 같은 청각 단서를 학습에 활용하지 않는다.

2.3 비디오 백본과 시계열 모델링

긴 영상을 프레임 단위로 처리하기 위해서는 효율적인 비디오 백본이 필요하다. Kondratyuk et al.(2021)이 제안한 MoViNet 계열은 모바일 환경에서의 효율적 비디오 인식을 목표로 설계된 경량 3D 컨볼루션 네트워크이다. 특히 causal 변형과 Stream Buffer 구조는 미래 프레임에 의존하지 않는 스트리밍 추론을 지원하므로, 긴 영상을 슬라이딩 윈도우 방식으로 순차 처리하는 본 연구의 요구에 부합한다. 본 연구에서는 모바일과 서버 양쪽에서의 배포 가능성을 함께 고려하여 MoViNet 계열 중 가장 가벼운 A0를 백본으로 채택한다.

비디오 백본만으로는 수십 초에 걸친 랠리 구간의 시작·종료 경계를 안정적으로 판별하기 어렵다. 이를 보완하기 위해 시계열 학습에서 널리 활용되는 양방향 순환 모델과, 장거리 시간 의존성을 모델링하는 Self-Attention 기반 시간 맥락 모델링이 비디오 이해 연구 전반에서 폭넓게 사용되어 왔다(Vaswani et al., 2017). 특히 Vaswani et al.(2017)이 제안한 Transformer 구조는 Self-Attention을 중심으로 토큰 간 장거리 의존성을 직접 학습할 수 있는 일반적인 메커니즘을 제공하였으며, 이후 영상·오디오 시계열 모델링에도 광범위하게 확산되었다. 본 연구 역시 causal MoViNet이 생성한 프레임 단위 특징 위에 BiLSTM과 Temporal Self-Attention, 잔차 연결을 결합한 시간 헤드를 구성하여 경계 판정의 안정성을 확보한다.

2.4 멀티모달 · 오디오 · 커리큘럼 러닝

시각 정보만으로 판별이 어려운 구간을 보완하기 위해, 최근 영상 이해 연구에서는 오디오를 부가 모달리티로 결합하는 멀티모달 접근이 활발히 연구되고 있다. Baltrušaitis et al.(2019)은 멀티모달 기계학습의 개요에서 결합 시점에 따라 Early Fusion과 Late Fusion을 구분하고, 각 결합 방식의 장·단점과 설계상의 고려 사항을 정리한다. 테니스 영상의 경우 타격음, 심판의 점수 콜, 관중 반응과 같은 청각 신호가 랠리 경계를 유추하는 강력한 단서로 작용한다. 본 연구는 이러한 관찰에 기반하여 멜 스펙트로그램 기반 오디오 특징을 Late Fusion 방식으로 결합한다. Late Fusion은 영상·오디오 각 분기의 학습 안정성을 보존하면서 결합 지점에서 상호 보완 효과를 얻을 수 있어, 소규모 데이터 환경에서 상대적으로 유리하다(Baltrušaitis et al., 2019).

한편 랠리·비랠리는 프레임 단위에서 심한 클래스 불균형을 보이며, 특히 경계 근방의 프레임은 라벨 자체가 모호하다. 이러한 상황에서 하드 샘플에 가중치를 부여하는 Focal Loss(Lin et al., 2017)가 클래스 불균형 완화에 효과적임이 객체 검출 등 여러 분야에서 반복적으로 보고되어 왔다. 본 연구는 Focal Loss와 라벨 스무딩을 적용한 BCE를 선형 결합한 손실을 사용하여, 불균형과 경계 모호성을 동시에 다룬다.

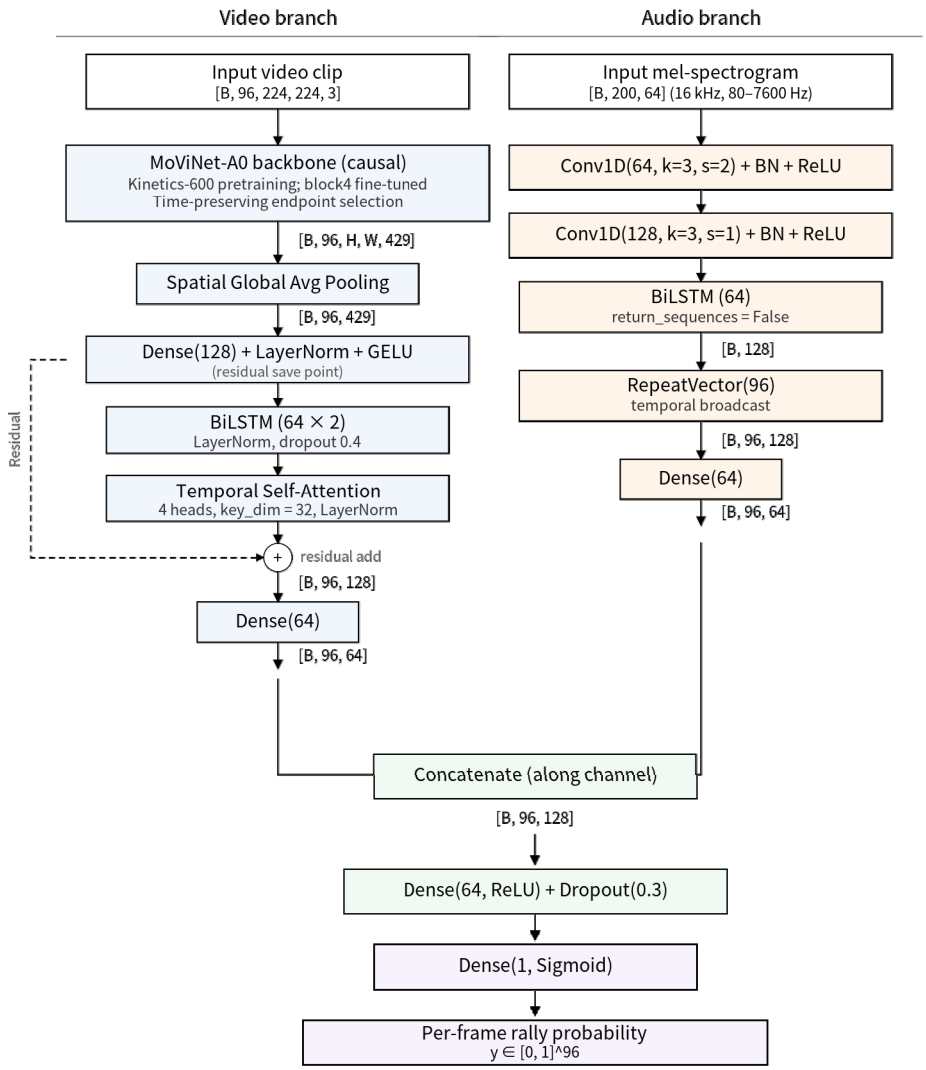
마지막으로 소규모 데이터 환경에서는 일반화 성능 확보가 핵심 난점이다. 커리큘럼 러닝은 학습 초기에 쉬운 샘플로 기본 패턴을 안정적으로 학습한 뒤, 점차 어려운 샘플로 확장하여 최종 성능을

끌어올리는 전략이다(Bengio et al., 2009). 본 연구는 이 원리를 비디오 세그멘테이션 과제에 적용하여, 영상 품질 등급(A/B/C)을 난이도 지표로 삼는 3단계 커리큘럼 러닝을 설계하였으며, 이 전략의 정량적 기여는 4장 Ablation 실험에서 확인된다.

3. 제안 방법

3.1. 전체 구조

제안 모델 TRMNet은 긴 테니스 영상을 입력받아 프레임별 랠리 확률 시퀀스를 출력하는 비디오-



〈그림 3〉 TRMNet 멀티모달 아키텍처 (비디오 브랜치 – 오디오 브랜치 Late Fusion 구조)

오디오 멀티모달 네트워크이다. 비디오 브랜치는 경량 3D CNN 백본 causal MoViNet A0 위에 순환·어텐션 시간 헤드를 얹고, 오디오 브랜치는 멜 스펙트로그램에서 Conv1D와 BiLSTM으로 시간 특징을 추출한다. 두 브랜치의 시간 정렬된 출력은 Late Fusion 헤드에서 결합되어 프레임별 랠리 확률로 변환된다.

학습·추론 단위는 96프레임 윈도우이며, 긴 영상은 스트라이드 48의 슬라이딩 윈도우로 분할해 순차 처리하고 중첩 구간은 평균 집계로 통합한다. 최종 출력 형상은 [B, 96, 1]이다.

3.2. 비디오 브랜치

비디오 백본으로는 Kondratyuk et al.(2021)의 causal MoViNet A0를 채택한다. causal 구조는 시간 의존성을 과거 프레임에 국한하고 Stream Buffer로 상태를 이월하므로, 긴 영상의 스트리밍 추론과 슬라이딩 윈도우 학습에 적합하다. A0는 MoViNet 계열 중 가장 경량이라 모바일과 서버 양쪽 배포를 고려한 선택이다. Kinetics-600 사전학습 가중치를 초기값으로 사용하며, 파인튜닝 시 block4만 학습 가능 상태로 두고 배치 정규화 계층은 전역적으로 동결하여 평가 모드로 동작시킨다. 백본 출력은 시간 헤드를 거쳐 경계 판정용 맥락으로 확장된다. 시간 헤드는 양방향 LSTM(각 방향 64 유닛, dropout 0.4)과 그 위의 멀티헤드 Temporal Self-Attention(헤드 4, key_dim 32)으로 구성되며, 두 층 사이에 잔차 연결을 두어 학습 안정성을 확보한다.

3.3. 오디오 브랜치

오디오 입력은 비디오와 동일 구간에서 추출한 멜 스펙트로그램이다(80~7600Hz, hop size 256, 크기 [200, 64]). 이 특징은 두 층의 Conv1D로 지역 시간 패턴을 얻은 뒤 양방향 LSTM으로 시간 의존성이 모델링되고, RepeatVector로 96프레임으로 확장되어 비디오 브랜치와 시간 축이 정렬된다. 타격음, 심판의 점수 콜, 관중 반응과 같은 청각 신호는 시각적으로 모호한 경계 구간에서 유효한 보조 단서로 작용하므로, 본 브랜치는 단독 판정이 아니라 시각 브랜치의 보완을 목적으로 경량 구조로 설계하였다.

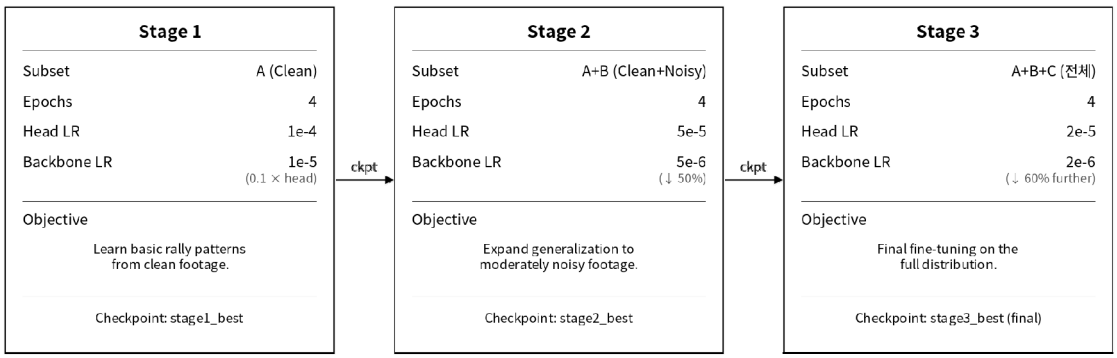
3.4. Late Fusion

두 브랜치의 시간 정렬된 출력은 특징 축에서 연결된 뒤 Dense(64)와 Dense(1)+Sigmoid 계층을 통과하여 프레임별 랠리 확률 [B, 96, 1]을 출력한다. Late Fusion을 채택한 이유는 초기 Fusion 대비 각 모달 분기의 학습 안정성을 보존하면서도, 소규모 데이터 환경에서 과적합 위험이 낮은 규모로 결합 표현을 학습할 수 있기 때문이다.

3.5. 학습 전략

학습 전략은 다음 네 가지 요소로 구성된다: 경계 모호성을 완화하는 부드러운 라벨 처리(Soft Label Ramp Mask), 클래스 불균형과 경계 혼동에 동시 대응하는 결합 손실 함수(Focal-BCE Combined Loss), 단일 모달리티 누락 상황에서의 강건성을 확보하는 학습 기법(Modality Dropout), 사전학습 가중치 보존을 위한 차등 학습률(Differential Learning Rate).

Soft Label Ramp Mask. 랠리 시작·종료 시점 각 3초(약 90프레임) 구간에 선형 램프로 완화된 타깃 확률을 부여하여, 경계 프레임에서의 과도한



〈그림 4〉 촬영 품질 기반 Clean–to–Noisy 3단계 커리큘럼 러닝 학습 절차

손실 증폭을 방지한다.

Combined Loss. Focal Loss(Lin et al., 2017)와 라벨 스무딩 BCE의 선형 결합으로 정의한다: L = 0.6 × Focal(α=0.7, γ=2.5) + 0.4 × BCE (label_smoothing=0.1). Focal 항은 경계 · 혼동 구간의 하드 샘플에 가중치를 두고, BCE 항은 소프트 라벨 및 라벨 스무딩과 맞물려 안정 수렴을 유도한다.

Modality Dropout. 학습 중 오디오 브랜치 출력을 50% 확률로 0 벡터로 대체하여 오디오 부재 · 잡음 상황에서의 견고성을 확보하고, Fusion 헤드가 시각 단서만으로도 합리적인 예측을 내도록 유도한다.

Differential Learning Rate. 시간 헤드 · 융합 헤드의 학습률은 커리큘럼 단계별로 1e-4, 5e-5, 2e-5로 감쇄시키고, 백본 학습률은 헤드의 1/10 수준으로 설정한다. 그 외 최적화 설정은 Adam, 가중치 감쇠 1e-5, 배치 크기 4, BF16 혼합 정밀도, 그라디언트 L2-노름 1.0 클리핑이다.

3.6. 커리큘럼 러닝

자체 구축 데이터셋은 촬영 품질에 따라 A · B · C 세 등급으로 분류되어 있다(4장 참조). 이 등급을 난이도 지표로 활용하여 3단계 커리큘럼을 수행한다. Stage 1은 A 등급 영상만으로 기반 파인튜닝, Stage 2는 A · B 등급 결합으로 학습 확장, Stage 3은 C 등급까지 포함한 전체 데이터로 최종 일반화를 수행한다. 학습률은 Differential Learning Rate 스케줄과 연동하여 단계별로 감쇄된다. 이러한 순차 확장은 초반에는 쉬운 분포로 기반 표현을 안정화하고 후반에 어려운 분포에 적응시켜, 소규모 데이터 환경에서의 일반화 성능을 강화한다.

4. 실험

4.1. 데이터셋

본 연구를 위해 총 118편의 사용자 촬영 테니스 경기 영상을 자체 수집 · 구축하였다. 전체 분량은 약 2,667분으로 43.6시간에 해당한다. 개별 영상의 길이는 최소 1.9분, 최대 85.8분, 평균 26.8분으로 분포한다. 모든 영상은 사용자가 직접 삼각대나 스마트폰 스탠드 등을 이용하여 촬영한 경기 영상이며, 방송 영상이나 전문 촬영 영상은 포함되지 않는다.

촬영 환경의 안정성과 주변 시각 요소의 영향을

기준으로 각 영상에 A·B·C의 세 등급을 부여하였다. A 등급은 53편(44.9%)으로, 카메라 흔들림이 거의 없고 주변 움직임 요소도 없이 선명하게 촬영된 영상이다. C 등급은 15편(12.7%)으로, 대각선 촬영이거나 주변에 지속적인 움직임 요소가 포함된 영상이다. B 등급은 A·C 어디에도 해당하지 않는 중간 품질 영상 50편(42.3%)이다. 본 등급은 엄격한 정량 기준이 아니라 라벨러의 주관적 판단에 기반한 서수적 난이도 분류이다. 이하 본 연구에서는 B등급과 C등급을 Noisy data로 통칭한다.

랠리 구간 라벨링은 두 명의 숙련된 라벨러가 데이터셋 영상을 분담하여 프레임 단위로 수행하였다. 동일 영상을 두 라벨러가 중복으로 라벨링하지 않고 영상별로 한 명의 라벨러가 단독으로 라벨링하는 방식이며, 두 라벨러의 라벨 통합본을 데이터셋 118편의 정답 라벨로 사용하였다. 사용자 촬영 영상에서 랠리 구간이 차지하는 비율은 전체 길이 대비 약 50%로 관찰되었다. 라벨링 작업 이전에 두 라벨러는 모호한 경계 케이스에 대한 사전 룰을 합의하였다. 주요 합의 사항은 다음과 같다: (a) 경기 진행을 위한 공 주고받기가 아닌 경우(예: 상대편 코트로 공을 건네주는 동작)는 랠리로 간주하지 않는다. (b) 서브 폴트와 다음 서브 간 간격은 랠리로 간주하지 않는다. 두 라벨러는 동등한 위치에서 독립적으로 라벨링을 수행하였으며, 어느 한쪽을 절대 기준으로 두지 않았다. 라벨링 일관성을 정량적으로 확인하기 위해 A·B·C 각 등급에서 무작위 1편씩 총 3편을 층화 표집하여 두 라벨러가 독립적으로 라벨링하는 예비 측정을 별도 수행하였으며, 프레임 단위(30fps) 측정 결과 양방향 대칭 F1 $\approx$ 0.88, Cohen's Kappa $\kappa \approx 0.83$, raw agreement $\approx$ 0.92의 일치도가 산출되었다.

데이터 분할은 Train / Validation / Test = 83 / 23 / 12 (7:2:1)로 구성하였다. 분할 간 A·B·C 등급 분포가 크게 왜곡되지 않도록 조정하였다.

4.2. 평가 지표

평가 지표로는 프레임 단위의 F1-score와 두 가지 Average Precision(AP)을 사용한다. 모델은 각 프레임에서 랠리 여부를 확률로 출력하므로, 확률 시퀀스를 임계값으로 이진화하여 프레임별 정답 라벨과 비교해 F1, Precision, Recall을 산출한다.

AP는 랠리 클래스에 대한 Precision-Recall 곡선의 평균 정밀도로 정의한다. 본 연구는 AP_smooth를 주 보고 지표로 사용한다. AP_smooth는 예측 확률 시퀀스에 시간 축 이동평균을 적용한 뒤 계산한 AP로, 짧은 구간의 확률 변동을 완화하여 프레임 단위 국소 노이즈가 지표에 미치는 영향을 줄이는 역할을 한다. 참고로 이동평균 적용 전 원시 확률 시퀀스에 대한 AP(AP_raw)도 내부적으로 계산하였으나 표 1에는 포함하지 않는다.

추론 단계의 후처리로는 확률 시퀀스에 적용할 이동평균 윈도우 크기 W(프레임 단위)와 이진화 임계값 τ를 함께 최적화한다. 탐색 격자는 $W \in \{1, 9, 19, 31, 43, 55, 67, 75, 83, 91, 95\}$, $\tau \in \{0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75\}$로 구성하며, Validation Set에서 F1 최대화 기준으로 그리드서치한다. Validation에서 선택된 단일 조합 (W^*, τ^*)을 그대로 Test Set에 고정 적용하여 최종 성능을 보고하며, Test Set 기준 재탐색은 수행하지 않는다. 이 프로토콜은 Test Set에 대한 후처리 과적합을 배제하여 일반화 성능의 편향 없는 추정을 보장한다.

4.3. 구현 상세

학습은 NVIDIA GeForce RTX 4080 단일 GPU

환경에서 TensorFlow 2.x 기반으로 수행하였다. 입력 영상은 30fps로 샘플링하고 각 프레임을 224×224 픽셀로 정규화하였으며, 학습 · 추론의 기본 처리 단위는 96프레임(약 3.2초) 길이의 시간 윈도우, 슬라이딩 스트라이드는 48프레임이다. 모델 초기값으로는 Kinetics-600에서 사전학습된 causal MoViNet A0 가중치를 사용하였다.

최적화는 Adam 옵티마이저와 가중치 감쇠 1e-5를 사용하였으며, 배치 크기 4, BF16 혼합 정밀도, 그라디언트 L2-노름 1.0 클리핑을 적용하였다. Differential Learning Rate 설정에 따라 시간 헤드 · 융합 헤드의 학습률은 Stage 1에서 1e-4, Stage 2에서 5e-5, Stage 3에서 2e-5로 단계별로 감쇄시켰으며, 백본 학습률은 각 단계별 헤드 학습률의 1/10 수준으로 설정하였다. 커리큘럼 3단계는 각 단계당 4에폭씩 총 12에폭에 걸쳐 수행되었으며, 단일 GPU에서 단계당 약 10시간 내외가 소요되었다.

4.4. 실험 결과

전체 모델(Full)과 네 가지 Ablation 구성의 성능을 비교하였다. 모든 설정은 동일한 118편 데이터 분할(Train 83 / Val 23 / Test 12)과 동일한 평가 프로토콜을 사용하였다. 후처리 하이퍼파라미터(window, threshold)는 각 설정에서 Validation Set 기준으로 최적화된 값을 Test Set에 고정 적용하였다. 비교 결과는 <표 1>에 정리하며, 표 1에서 AP_s는 AP_smooth의 축약 표기이다.

각 설정의 Validation 최적 후처리 조합은 Full은 (window=95, threshold=0.65), w/o Audio는 (window=19, threshold=0.60), w/o Curriculum은 (95, 0.50), w/o Audio & Curriculum은 (95, 0.55), Baseline은 (window=19, threshold=0.55)이며, 동일 조합을 Test Set에 고정 적용하였다. 여기서 Baseline은 시간 헤드(BiLSTM + Temporal Self-Attention)를 모두 제거하고 MoViNet 백본 위에 단순 Dense 분류기만 둔 구성으로, 시간 모델링의 기여를 정량화하기 위한 하한 비교군이다.

각 구성 요소의 기여를 Test AP_smooth로 분해하면 다음과 같다. 첫째, 시간 헤드(BiLSTM + Temporal Self-Attention)의 기여가 압도적으로 가장 크다. 시간 헤드를 제거한 Baseline은 동일 입력(V)을 사용한 w/o Audio & Curr. 대비 Test AP_smooth가 0.9233에서 0.7677로 0.1556 하락하였으며, Test F1도 0.8368에서 0.6843으로 0.1525 하락하였다. 둘째, 오디오 Late Fusion과 커리큘럼 러닝의 독립 기여는 각각 노이즈 범위(±0.001) 수준이다. Test AP_smooth 기준으로 w/o Audio

〈표 1〉 제안 모델 TRMNet(Full)과 Ablation 구성의 성능 비교.

Setting	Val F1	Val AP_s	Test F1	Test P	Test R	Test AP_s
Full (V+A+Curr.)	0.8198	0.9051	0.8422	0.8751	0.8117	0.9315
w/o Audio (V+Curr.)	0.8131	0.8988	0.8484	0.8664	0.8310	0.9327
w/o Curriculum (V+A)	0.8152	0.9060	0.8471	0.8492	0.8450	0.9305
w/o Audio & Curriculum	0.8187	0.9034	0.8368	0.8346	0.8389	0.9233
Baseline (V, no time head)	0.6844	0.7521	0.6843	0.6341	0.7431	0.7677

(0.9327)는 Full(0.9315)을 0.0012 상회하고, w/o Curriculum(0.9305)은 Full을 0.0010 하회한다. 셋째, 그러나 두 요소를 동시에 제거한 w/o Audio & Curriculum은 0.9233으로, Full 대비 0.0082 하락하여 두 요소 사이의 양(+) 방향 interaction이 관찰된다. 즉 오디오와 커리큘럼은 개별로는 한계 효용이 작으나 결합 시 모델 표현의 안정화에 기여한다. Validation 기준으로는 Full이 F1과 Precision에서 모두 최상위이며, Test에서 일부 ablation 구성이 F1·AP_smooth를 소폭 상회하는 현상은 소규모 Test Set(12편)에서의 통계적 변동 범위 내에 있다고 판단된다. Full 구성은 Validation 분포에서 안정적으로 최적점을 유지하며, Test Precision에서 +0.0087 ~ +0.0405의 우위를 일관되게 보여준다.

4.5. 논의

Test F1 0.8422는 A·B·C 각 등급에서 층화 표집한 3편에 대한 라벨러 간 F1 약 0.88(Cohen's Kappa $\kappa \approx 0.83$)의 예비측정치와 근접한 수준이다. 이는 본 모델이 사용자 촬영 환경의 다양한 품질 분포에 걸쳐 인간 라벨러의 일관성과 비교할 만한 수준의 랠리 구간 판정을 수행함을 시사한다.

본 연구는 모델의 작동 원리 해석을 위해 Grad-CAM(Selvaraju et al., 2017) 분석을 수행하였다. 그 결과 비디오 백본이 공의 위치보다 라켓 움직임과 타격 순간에 집중 활성화됨을 확인할 수 있었다(〈그림 2〉 참조). 이러한 활성화 패턴은 사용자 촬영 환경의 흔들림·역광 조건에서도 공 추적 의존도가 낮은 본 모델의 강건성을 시사하며, 동시에 오디오 통합의 효용에 대한 실증적 단서를 제공한다.

F1은 단일 임계값에 의존하여 후처리 설정에 민감한 반면, AP_smooth는 임계값에 무관하게 모델의 랭킹 능력을 평가하므로 구성 요소 기여 분석의 주 지표로 채택하였다. 연구 문제 1에 대해, 오디오 Late Fusion의 독립 기여는 소규모 Test Set의 통계적 변동 범위 내에 있어 단독 유효성을 단정하기 어렵다. 그러나 커리큘럼 러닝과의 결합 시 Test AP_smooth 기준 0.0082의 양(+) 방향 interaction이 관찰되어, 오디오는 단독보다 커리큘럼과의 결합에서 보완적 역할을 수행하는 것으로 해석된다. 본 실험의 한계는 Test Set 규모가 12편으로 제한적이라는 점이다. 일부 Ablation 구성에서 Test 지표가 Full을 소폭 상회하는 현상이 관찰되었으나, 이는 소규모 Test Set의 통계적 변동 범위 내에 있을 가능성이 높다. 후속 연구에서는 데이터셋 규모를 확대하여 오디오·커리큘럼 각 구성 요소의 한계 기여를 보다 정밀하게 정량화할 계획이다.

시간 헤드(BiLSTM + Temporal Self-Attention)의 기여는 Test AP_smooth 기준 0.1556으로 압도적으로 가장 크다. 한편 오디오 Late Fusion과 커리큘럼 러닝은 개별 제거 시 Test 지표 변화가 ±0.001 수준의 노이즈 범위에 있으나, 두 요소 동시 제거 시 Test AP_smooth 0.0082 하락이 관찰되어 두 요소 결합에 양(+) 방향의 interaction이 존재함을 확인하였다. 또한 본 연구의 대상 태스크인 테니스 랠리 세그멘테이션에는 공개된 외부 baseline 모델이 존재하지 않아, 본 실험에서는 구성요소 기여 검증에 초점을 두었으며, 외부 baseline과의 비교는 후속 연구로 계획한다.

5. 결론

본 연구는 사용자 촬영 테니스 경기 영상에서 랠리 구간을 프레임 단위로 자동 추출하는 멀티모달

네트워크 TRMNet을 제안하였다. TRMNet은 경량 비디오 백본 causal MoViNet A0 위에 BiLSTM과 Temporal Self-Attention, 잔차 연결을 결합한 시간 헤드를 얹고, 멜 스펙트로그램 기반 오디오 브랜치를 Late Fusion으로 결합한 구조이다. 라벨 경계 부드러운 처리, 결합 손실 함수, 모달리티 누락 강건성 확보 학습 기법, 차등 학습률을 학습 전략으로 도입하고, 영상 품질 등급에 따른 3단계 커리큘럼 학습을 보조 전략으로 활용하여 소규모 데이터 환경에서의 파인튜닝 안정성과 일반화 성능을 동시에 확보하였다.

자체 구축한 테니스 경기 영상 118편(총 43.6시간) 데이터셋에서의 실험 결과, TRMNet은 Validation Set에서 결정된 후처리 조합을 Test Set에 고정 적용하는 프로토콜 하에 Test F1 0.8422, AP_smooth 0.9315를 기록하였다. 5개 조건의 Ablation 분해(Full, w/o Audio, w/o Curriculum, w/o Audio & Curriculum, 시간 헤드 제거 Baseline)를 통해 각 구성 요소의 기여를 정량화하였으며, 시간 헤드(BiLSTM + Temporal Self-Attention)의 기여가 가장 크게 나타나 Test AP_smooth가 0.9233에서 0.7677로 0.1556 하락하는 것으로 확인되었다. 오디오 Late Fusion과 커리큘럼 러닝은 개별 제거 시 Test 지표 변화가 ±0.001 수준의 노이즈 범위에 있었으나, 두 요소 동시 제거 시 Test AP_smooth 0.0082 하락(0.9315 → 0.9233)이 관찰되어 두 요소 결합에 양(+) 방향의 interaction이 존재함을 확인하였다. Full 구성은 Test Precision 0.8751로 모든 설정 중 최상위였으며, 이는 오디오 단서가 비랠리 구간의 오검출을 억제하는 방향으로 작용한 결과로 해석된다. 이를 통해 연구 문제 1과 2에 대해, 오디오와 커리큘럼은 개별 기여보다 결합 시 상호보완적으로 모델 안정성에 기여함을 실증하였다.

본 연구에는 네 가지 한계가 있다. 첫째, Test Set 규모가 12편으로 제한적이어서 일부 Ablation 구성에서 관찰된 Test 지표의 미세한 역전이 통계적 유의성을 갖는지 단정하기 어렵다. 둘째, 본 연구가 대상으로 하는 사용자 촬영 테니스 경기 영상에 대한 공개 벤치마크·외부 베이스라인이 부재하여, 현 시점에서는 자체 데이터셋 내부 Ablation 비교에 한정된 검증만 가능하다. 셋째, 데이터셋의 총 분량은 43.6시간 수준으로, 실환경의 다양한 코트·조명·카메라 조건을 포괄하기에는 여전히 제한적이다. 넷째, 영상 품질 A·B·C 등급은 라벨러의 서수적 주관 판단에 의존하므로, 커리큘럼 난이도 정의의 객관성이 보장되지 않는다.

향후 연구에서는 다음 네 가지 방향을 추진하고자 한다. 첫째, 다중 시드 반복 실험을 통해 평균·표준편차를 함께 보고하여 오디오·커리큘럼의 한계 기여에 대한 통계적 신뢰성을 확보한다. 둘째, 데이터셋 규모 및 다양성을 확대하고 외부 라벨러 협력을 통한 교차 검증으로 라벨러 간 일치도 측정을 보강한다. 셋째, 영상 품질을 라벨러 주관이 아닌 정량 지표(카메라 흔들림, 조명 변화, 대비 등)로 재정의하여 커리큘럼 난이도의 객관성을 확보한다. 넷째, causal MoViNet의 Stream Buffer 구조를 활용한 실시간 모바일 추론과, 온디바이스 환경에서의 지연·정확도 트레이드오프를 체계적으로 보고한다.

참고문헌

[국내 문헌]

김병조, 최용석. (2019). 딥러닝 기반 스포츠 캐스팅. 정보과학회논문지, 46(10), 1020-1024. https://doi.org/10.5626/JOK.2019.46.10.1020

양준석, 이제욱, 박성제. (2022). 딥러닝 기반 배드민턴 선수 위치 추적시스템 개발. 한국스포츠학회지, 20(2), 851-868. https://doi.org/10.46669/kss.2022.20.2.072

[국외 문헌]

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443. https://doi.org/10.1109/TPAMI.2018.2798607

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. 26th Annual International Conference on Machine Learning (ICML), Montreal, Canada.

Farha, Y. A., & Gall, J. (2019). MS-TCN: Multi-stage temporal convolutional network for action segmentation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA.

Voeikov, R., Falaleev, N., & Baikulov, R. (2020). TTNet: Real-time temporal and spatial video analysis of table tennis. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA.

Huang, Y.-C., Liao, I.-N., Chen, C.-H., İk, T.-U., & Peng, W.-C. (2019). TrackNet: A deep learning network for tracking high-speed and tiny objects in sports applications. 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Taipei, Taiwan.

Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., & Gong, B. (2021). MoViNets: Mobile video networks for efficient video recognition. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. IEEE International Conference on Computer Vision (ICCV), Venice, Italy.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. IEEE International Conference on Computer Vision (ICCV), Venice, Italy.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. 31st Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA.

Wu, F., Wang, Q., Bian, J., Ding, N., Lu, F., Cheng, J., Dou, D., & Xiong, H. (2022). A survey on video action recognition in sports: Datasets, methods and applications. IEEE Transactions on Multimedia, 25, 7943-7966. https://doi.org/10.1109/TMM.2022.3232034

Zhao, Z., Chai, W., Hao, S., Hu, W., Wang, G., Cao, S., Song, M., Hwang, J.-N., & Wang, G. (2025). A survey of deep learning in sports applications: Perception, comprehension, and decision. IEEE Transactions on Visualization and Computer Graphics, 31(10), 9368-9386. https://doi.org/10.1109/TVCG.2025.3554801

Abstract

TRMNet: A Temporal-Head-Based Multimodal Network for Automatic Rally Segmentation in User-Recorded Tennis Videos

Younghun Jung* · Hakyoul Choe**

We propose TRMNet (Tennis Rally Multimodal Network), a multimodal network for automatic rally segmentation from long, unedited tennis match videos captured by amateur users. In user-recorded videos, actual play occupies only approximately 50% of the total duration, and rally boundaries exhibit inherent ambiguity-inter-annotator F1 reaches only approximately 0.88 even among skilled labelers. TRMNet integrates a lightweight causal MoViNet A0 backbone with a BiLSTM · Temporal Self-Attention temporal head and an audio Late Fusion branch, and applies Soft Label Ramp Mask, Focal-BCE combined loss, Modality Dropout, and differential learning rate, with Clean-to-Noisy three-stage curriculum learning as a supplementary strategy, to ensure fine-tuning stability and generalization under small-scale heterogeneous data conditions. Ablation experiments confirm that the temporal head (BiLSTM + Temporal Self-Attention) contributes most substantially, with AP_smooth improving by 0.1556 over the baseline without the temporal head. On a self-collected dataset of 118 tennis match videos (43.6 hours), TRMNet achieves a test-set F1 of 0.8422 and AP_smooth of 0.9315, with real-world applicability confirmed through deployment in a commercial tennis video service (PerfectSwing, Bound1 Inc.), where approximately 50% of non-rally footage is automatically removed.

Key Words : Tennis Video Analysis, Rally Segmentation, Multimodal Learning, Curriculum Learning, TRMNet

Received : April 26, 2026 Revised : June 1, 2026 Accepted : June 2, 2026
Corresponding Author : Choe, Hakyoul

* Bound1 Inc.
** Corresponding Author: Choe, Hakyoul
School of Computing, Yonsei University
50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea
Tel: +82-2-6000-2009, Fax: +82-2-6000-2084, E-mail: hychoe@yonsei.ac.kr

저 자 소 개

정 영 훈

주식회사 바운드원 대표이사. 테니스 영상 자동 편집 앱 PerfectSwing: Rally Highlights를 개발· 운영 중이다. 관심 연구 분야는 비디오 이해, 멀티모달 딥러닝, 실시간 모바일 추론이다.

최 학 열

현재 연세대학교 인공지능융합대학 겸임교수로 재직 중이다. 한국과학기술원(KAIST)에서 Database를 전공하여 석사학위를 취득하고, Carnegie Mellon University, School of Computer Science에서 Software Engineering으로 석사학위를 취득하였으며, 한국과학기술원(KAIST)에서 인공지능 분야로 박사학위를 취득하였다. 주요 관심분야는 자연어처리, 영상· 음향 분석, 스포츠 애플리케이션 등이다.