멀티모달 정보를 이용한 잡음에 강인한 야구 이벤트 시점 검출 방법

김영익 [a)†], 정현조 [b)†], 나민수 [a)], 이영현 [b)], 이준수 [b)]
Speech AI Lab[a)] & Vision AI Lab[b)], 엔씨소프트
{youngik, hyunjo7905, minsoona, younghyunlee, jslee509}@ncsoft.com,

Noise Robust Baseball Event Detection with Multimodal Information

Young-Ik Kim[a)†], Hyun Jo Jung[b)†], Minsoo Na[a)], Younghyun Lee[b)], Joonsoo Lee[b)]
Speech AI Lab[a)] & Vision AI Lab[b)], NCSOFT

요 약

스포츠 방송/미디어 데이터에서 특정 이벤트 시점을 효율적으로 검출하는 방법은 정보 검색이나 하이라이트, 요약 등을 위해 중요한 기술이다. 이 논문에서는, 야구 중계 방송 데이터에서 투구에 대한 타격 및 포구 이벤트 시점을 강인하게 검출하는 방법으로, 음향 및 영상 정보를 융합하는 방법에 대해 제안한다. 음향 정보에 기반한 이벤트 검출 방법은 계산이 용이하고 정확도가 높은 반면, 영상 정보의 도움 없이는 모호성을 해결하기 힘든 경우가 많이 발생한다. 특히 야구 중계 데이터의 경우, 투수의 투구 시점에 대한 영상 정보를 활용하여 타격 및 포구 이벤트 검출의 정확도를 보다 향상시킬 수 있다. 이 논문에서는 음향 기반의 딥러닝 이벤트 시점 검출 모델과 영상 기반의 보정 방법을 제안하고, 실제 KBO 야구 중계 방송 데이터에 적용한 사례와 실험 결과에 대해 기술한다.

1. 서론

TV 방송이나 뉴스, 영화 등 동영상 데이터의 기하급수적인 증가와 함께, 사용자들이 원하는 특정 이벤트를 찾거나 핵심 내용만 요약하는 연구가 지속적으로 요구되고 있으며, 최근에 와서는 음성 및 영상을 포함하는 멀티모달 정보를 활용한 이벤트 검출 방법이 다양하게 연구되고 있다 [1]. 스포츠 분야, 특히 야구 경기 중계 방송에서도 하이라이트 영상 생성을 위한 이벤트 검출이 중요한 기술 중 하나인데, 이벤트 종류나 필요에 따라 영상, 음향, 캡션 등 다양한 정보가 활용되고 있다 [2].

이 논문에서는, 야구 경기 중계 방송 동영상에서 투수의 투구 이후에 발생하는 타격 및 포구 이벤트의 정확한 시점을 검출하는 문제를 다룬다. 야구 하이라이트 영상 생성 기법에서 많이 활용되고 있는 영상 기반의 이벤트 검출 방법은 투구와 같은 상황정보를 얻는데 유리한 반면, 2 차원 영상 데이터 처리에 따른 계산량이 많고, 타격이나 포구 등 정교한 시점이 필요한 경우 검출 정확도 향상의 어려움이 있다. 반면 음향 기반의 이벤트 검출은 영상에 비해 계산량이 적고, 정확한 시점을 검출하기에 적합하지만, 야구 경기의 상황정보가 부족하여 잡음으로 인한 오검출 사례가 많이 발생한다. 우리는 이러한 음향과 영상 기반 이벤트 검출 방법을 효율적으로 융합하기 위해, (1) 계산량 및 정교함 측면에서 유리한 음향 기반의 모델을 통해 먼저 이벤트 검출 후보군을 얻고, (2) 상황정보가 필요한 경우에만 영상 기반 이벤트 검출을 계산하고, (3) 음향과 영상 검출 결과를 결합하여 최종 이벤트 시점을 결정한다.

2. 멀티모달 이벤트 시점 검출 방법

이 논문은 야구 경기 중계 방송 동영상에서 장면 전환과 장면 분류 모델을 통해 투수가 공을 던지는 투구 장면을 포함하는 클립을 입력 동영상으로 사용한다 [3].

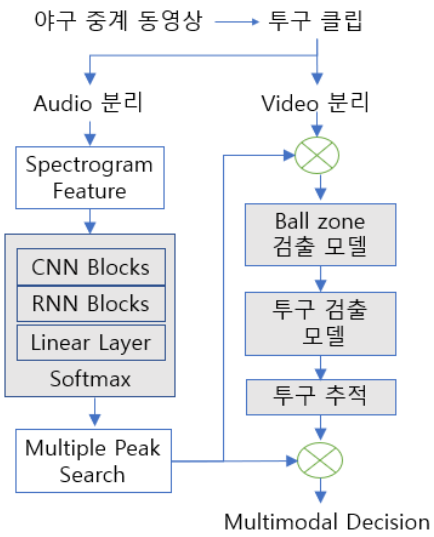

그림 1. 멀티모달 야구 이벤트 시점 검출 흐름도

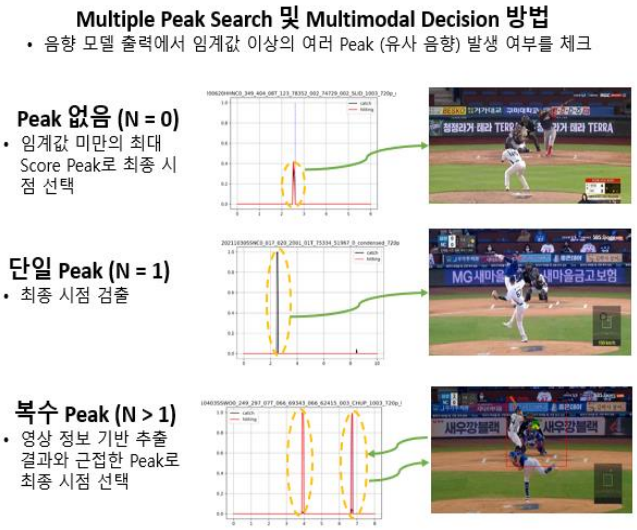

그림 2. 멀티모달 정보 융합 방법 및 사례

(1) **음향 기반 이벤트 시점 검출:** 정확한 타격/포구 이벤트 시점 검출을 위해 CNN 과 RNN 기반의 음향 모델로 DCASE 2019 챌린지에서 공개한 베이스라인의 기본 구조를 사용하였다 [4]. 음향 모델 훈련에서는, 전사된 이벤트 시점 앞/뒤로 τ_A 의 onset/offset 마진을 주었고, 타격과 포구의 2 가지 클래스를 구분하여 학습하였다. 이벤트 시점 검출 단계에서는, 투구 클립 내에서 모델 출력이 임계값 θ_A 를 넘는 모든 피크를 찾고, 선형 보간법을 통해 해당 피크의 onset/offset 시점을 추정하는 방법으로 정확도를 높였다. 그리고, 본 논문의 실험에서 τ_A 와 θ_A 값은 각각 50 ms, 0.5 를 사용했다.

(2) **영상 기반 이벤트 시점 검출:** 타격/포구 이벤트는 투수의 투구 이후 야구공이 홈 플레이트를 통과하는 시점에서 발생한다. 따라서 입력된 투구 동영상에서 전체 영역 보다는 홈 플레이트 부근 영역의 정보를 분석하는 것이 정확도 향상에 도움이 된다. 이를 위해 먼저 홈 플레이트 및 타자/포수 영역을 'Ball zone'이라 정의하고 해당 영역을 1 차적으로 검출한다. 그 이후 'Ball zone' 내부에서 야구공의 이동을 파악하기 위하여 야구공 검출을 수행한다. 야구공이 작고 빠른 객체이기 때문에 추적 기술을 사용하면 검출 성능의 저하를 보완할 수 있다. 'Ball zone' 및 야구공 검출에는 YoloR 모델[5]을 적용하였고, 야구공 추적에는 SORT 모델[6]을 적용하였다. 그리고, 영상 기반 이벤트 시점은 야구공이 검출 및 추적되다 사라지는 마지막 프레임 시점으로 추정한다.

(3) **멀티모달 정보 융합:** 음향 모델을 통한 이벤트 시점 검출 방법은 영상 기반의 방법에 비해 계산량 및 정확도 측면에서 유리한 반면, 야구 경기 상황 변화에 따른 잡음으로 복수개의 모델 출력이 발생하여 오검출 되는 경우가 빈번하게 발생한다. 이러한 경우, 그림 2 에서와 같이, 영상 기반의 이벤트 시점 검출을 통해 상황정보를 획득한 다음, 음향과 영상 이벤트 시점이 근접한 피크를 최종 멀티모달 이벤트 시점으로 결정한다.

3. 실험 및 고찰

실험에서 사용한 데이터로는, KBO 2019-2021 시즌 동안 방송된 총 127 개 야구 경기 중계 방송 동영상에서 얻은 1355 개의 투구 클립을 훈련/검증/평가셋으로 80%/10%/10%의 비율로 나누어 사용했다. 그리고, 오디오 편집툴인 wavesurfer 를 이용한 스펙트럼 뷰와 전문가를 통한 음향 청취 방법으로 정확한 타격 및 포구 이벤트 시점을 정답 레이블링 했다.

표 1 에서는 음향 기반의 이벤트 시점 검출 방법을 사용하여 오차 허용범위에 따른 모델의 성능을 비교하였다. 타격 및 포구 이벤트를 활용하는 야구 하이라이트 영상 생성을 위해서는 이벤트 시점 예측에 대한 오차 허용범위가 명확하게 제시될 필요가 있는데, 일반적으로 많이 사용하는 30 FPS 동영상을 고려하였을 때 프레임 간격인 33 ms 보다 길지 않으면서 성능이 가장 우수한 30 ms 를 오차 허용범위로 정할 수 있다. 그리고 실험을 통해 오차 허용범위가 30 ms 인 경우, 시스템의 반응 속도와 오류율을 감안하여 모델의 출력 간격을 20 ms 수준으로 선정하는 것이 시스템 적용에 적절한 것임을 확인할 수 있었다.

오차 허용 범위 (ms)	모델의 출력 간격 (ms) - 오류율		
	10	20	30
10	26.0%	26.2%	25.4%
20	10.3%	10.3%	14.7%
30	6.9%	**6.2%**	9.4%
40	4.7%	4.9%	6.2%
50	2.4%	2.1%	3.1%

표 1. 음향 기반 방법의 오차 허용범위와 모델 출력간격에 따른 검출 오류율 비교

	오류율			
	피크 N=0	피크 N=1	피크 N>1	전체
음향	50.0%	4.8%	14.0%	6.2%
음향+영상	50.0%	4.8%	**11.5%**	**5.9%**

표 2. 음향 기반 방법과 멀티모달 방법의 검출 오류율 비교

멀티모달 정보 융합을 통한 성능 개선을 확인하기 위해, 음향 정보만 사용한 이벤트 시점 검출 방법과 영상 정보를 추가로 사용한 방법의 오류율을 표 2 에서 비교하였다. 실험 결과에서 영상 정보를 추가로 활용하는 멀티모달 방법으로, 피크가 없거나 1 개만 있는 경우의 성능에 영향을 주지 않으면서, 피크가 다수 존재할 경우의 오류율을 17.8% 줄여주는 효과가 있음을 알 수 있다. 표 2 에서 피크가 없는 경우는 관중 소음 등의 영향으로 임계값을 넘는 음향 모델 출력이 없는 경우에 해당하며, 향후 성능 개선이 필요한 부분이다.

4. 결론

이 논문에서는 음향 및 영상 정보를 융합하는 멀티모달 방법으로 야구 경기 중계 방송에서 투구에 대한 타격 및 포구 이벤트 시점 검출의 정확도를 개선하는 방법을 제안하였다. 홈런 shortcut 영상이나 삼진 모음 영상 등 다양한 야구 하이라이트 컨텐츠 제작을 위해 정교한 이벤트 시점 검출이 필요한 상황에서, 정보 융합을 통한 성능 개선은 향후 멀티모달 연구의 좋은 사례이다.

† These authors contributed equally to this work

※ 이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.171115245, 인명 구조용 드론을 위한 영상/음성 인지 기술 고도화)

참고문헌

[1] B. Duan, H. Tang, W. Wang, Z. Zong, G. Yang and Y. Yan, "Audio-visual event localization via recursive fusion by joint co-attention," WACV 2021

[2] C. C. Cheng and C. T. Hsu, "Fusion of audio and motion information on HMM-based highlight extraction," IEEE Trans. Multimedia, vol. 8, no. 3, pp. 585-599, 2006

[3] Y. Lee, H. Jung, C. Yang and J. Lee, "Highlight-video generation system for baseball games," ICCE-Asia, 2020

[4] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for DCASE 2019 Task 4," Detection and Classification of Acoustic Scenes and Events 2019

[5] C. Y. Wang, I. H. Yeh and H. Y. M.Liao, "You only learn one representation: Unified Network for Multiple Tasks." arXiv:2105.04206, 2021

[6] A. Bewley, G. Zongyuan, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," ICIP, pp. 3464-3468, 2016