TinyLLaVA 기반 4 bit 양자화를 활용한 엣지 디바이스에서의 한우 승가 행위 실시간 검출 시스템

조영준*
*한국폴리텍대학 분당융합기술교육원 AI응용소프트웨어과
e-mail:samcho2017@kopo.ac.kr

Real-Time Hanwoo Mounting Behavior Detection on Edge Device Using 4-bit Quantized TinyLLaVA

Young-Joon Cho*
*Department of AI Application Software, Bundang Convergence Technology Campus of Korea Polytechnic

요 약

본 논문은 YOLOv4 기반 활동량 분석, YOLOv5 ARM 지능형 모니터링 시스템, VLM 하이브리드 파이프라인, SVLM 비교 분석으로 이어지는 한우 발정 검출에 대한 선행연구에서 GPT-4o Verifier를 활용한 VLM 기반 시스템이 높은 정확도(Precision 97.3%)를 달성하였으나, 클라우드 의존성으로 인해 네트워크가 불안정한 농촌 환경에서의 적용에 한계가 있었다. 이에 선행연구에서 권장된 TinyLLaVA를 4-bit AWQ 양자화로 경량화하여 Jetson AGX Orin에 탑재하고, YOLOv8n 전처리·맥락 인식 프롬프트·시간적 필터링·CLAHE를 결합한 완전 오프라인 실시간 검출 시스템을 제안한다. 본 연구에서 실시간이란 승가 행위가 발생하는 시점 내에 판정이 완료되어 알림이 발송됨을 의미하며, 판정 지연 6.9초는 승가 평균 지속 시간(7.2초) 이내로 이 조건을 충족한다. 평가 결과 Precision 96.8%, F1 91.8%, FPR 3.2%를 달성하였다.

1. 서론

저자는 한우 번식 관리의 핵심인 승가 기반 발정 검출 정확도 향상을 목표로 일련의 연구를 수행하여 왔다. 1단계[1]에서는 YOLOv4와 다중객체추적(MOT)을 이용하여 활동량 증가와 발정 발현의 상관성을 비침습 방식으로 검증하였다. 2단계[2]에서는 YOLOv5 기반 ARM(Augmented Recognition Model)을 도입하여 개별 한우 식별·활동 데이터 정량화·승가 감지를 통합하는 지능형 모니터링 시스템(IMS)을 구현하고 국제 저널에 게재하였다. 3단계[3]에서는 YOLOv8n Trigger와 GPT-4o Verifier로 구성된 2단계 하이브리드 파이프라인을 설계하여 Precision 97.3%를 달성하였으나, GPT-4o API 호출이 필수적이므로 네트워크 단절 시 시스템 전체가 중단되는 구조적 한계가 있었다.

이 한계를 해결하고자 4단계에서는 SVLM 프레임워크를 비교 분석하고 TinyLLaVA를 가장 적합한 후보로 제시하였다. 본 논문은 이를 실제로 구현하는 5단계 후속 연구이다. 본 연구에서 실시간 검출이란 처리 속도가 초당 수십 프레임에 달하는 것을 의미하지 않으며, 승가 행위가 지속되는 시간 내에 검출·알림이 완료되어 현장에서 유효하게 활용 가능함을 의미한다. 제안 시스템의 판정 지연(6.9초)은 해당 농가에서 측정된 승가 평균 지속 시간(7.2초) 이내이므로 이 기준을 충족하며, 완전 오프라인으로 동작한다.

2. 제안 시스템

전체 파이프라인은 선행연구[3]의 2단계 구조를 계승하되, Verifier를 클라우드 VLM에서 엣지 SVLM으로 대체한 것이 핵심이다. ① YOLOv8n이 30 FPS로 소 ROI를 실시간 추출하며, 두 개체 근접 시 해당 ROI를 TinyLLaVA-AWQ는 맥락 인식 프롬프트("앞발을 올리고 몸을 얹는 승가인지, 턱만 얹는 턱비비기인지, 일반 행동인지 판단. 반드시 'mounting / chin-resting / normal' 중 하나만 출력")로 행동을 분류한다. 프롬프트는 선행연구[3]의 페르소나 기법을 계승하되 엣지 추론에 적합하도록 출력 형식을 단순화하였다. ③ 연속 N=3회 'mounting' 판정 시(지연 6.9초 < 평균 지속 7.2초) 카카오톡 API로 알림을 발송한다. 이 구조는 승가가 진행 중인 시점에 알림이 도달하므로 실시간 검출로 정의한다.

TinyLLaVA-Phi2-SigLIP-3.1B에 AWQ 4-bit 양자화를 적용하여 GPU 메모리를 FP16(6.2 GB)에서 2.1 GB로 66% 감소시켰다. 야간 프레임(평균 휘도 μ<60)에는 CLAHE (clipLimit=2.0, YCrCb Y채널)를 자동 적용한다. 선행연구[3]에서 오탐의 주요 원인이었던 야간 저조도 환경의 개체 중첩 문제에 대응하기 위한 것이다.

3. 실험

경기도 안성 소재 한우 농가에서 2025년 3~6월 수집한 CCTV 영상으로 데이터셋을 구축하였다. 선행연구[3]의 50클립 소규모 데이터셋과 달리, 본 연구는 승가 318·턱비비기 241·일반 행동 412클립(각 30초, 야간 40% 이상)으로 구성된 971클립 데이터셋을 신규 구축하였다. 훈련:검증:테스트를 6:2:2로 분할하였으며, YOLOv8n은 파인튜닝, TinyLLaVA는 Zero-shot으로 평가하였다. FPR† = FP/전체 비승가 × 100(%).

[표 1] 실험 환경 구성

구 분	내 용	비 고
엣 지 디바이스	NVIDIA Jetson AGX Orin 32GB	JetPack 6.0 / CUDA 12.2
모 델	TinyLLaVA-Phi2-SigLIP-3.1B	4-bit AWQ 양자화, llama.cpp
전 처 리	YOLOv8n (소 ROI, 30FPS)	야간: CLAHE (clipLimit=2.0, YCrCb Y채널)
데이터셋	한우 특화 971클립 (29,130 프레임)	승가 318 / 턱비비기 241 / 일반 412
알림 조건	연속 N=3 판정 → 카카오톡 API 발송	판정 지연 ≤ 6.9초 (승가 지속 시간 이내)
GPU 메모리	FP16: 6.2 GB → AWQ: 2.1 GB (66% ↓)	OOM 없이 안정 구동

[표 2] 모델별 성능 비교

모델 / 방법	Prec.	Rec.	F1	FPR	비고
선행연구 (GPT-4o)	97.3	94.0	95.6	2.7	클라우드 의존, 통신비 발생
YOLOv8n (단독)	91.4	89.2	90.3	8.6	턱비비기 혼동 多
TinyLLaVA FP16	87.6	84.1	85.8	12.4	OOM 불안정
TinyLLaVA-AWQ	93.7	88.5	91.0	6.3	4-bit 양자화
제안 시스템 (+CLAHE)	96.8	87.3	91.8	3.2	완전 오프라인 실시간 동작

표 2에서 선행연구[3]의 GPT-4o 기반 시스템은 Precision 97.3%로 가장 높은 정확도를 보이나 클라우드 통신이 필수적이다. 본 제안 시스템은 Precision 96.8%로 0.5%p 차이 내에서 근접한 성능을 유지하면서 완전 오프라인으로 동작하며 FPR을 YOLOv8n 단독(8.6%) 대비 5.4%p 감소시켰다. CLAHE 적용으로 야간 FP가 39% 감소하였으며, 판정 지연 6.9초는 승가 지속 시간(7.2초) 이내로 실시간 검출 조건을 충족하였다. FN의 63%는 지속 시간 3초 미만의 단기 시도성 승가(brief mounting attempts)로, 향후 YOLO 1차 미확정+VLM 2차 확정의 2단계 알림 체계로 개선할 수 있다.

4. 결론 및 향후 연구

본 연구는 저자의 선행연구 계보(YOLOv4 활동량 분석[1] → YOLOv5 ARM IMS[2] → VLM 하이브리드[3] → SVLM 비교분석[4])를 이어 받아, 네트워크 불안정 환경을 위한 완전 오프라인 실시간 한우 승가 검출 시스템을 실증하였다. 4-bit AWQ 양자화 TinyLLaVA와 YOLOv8n·시간적 필터링·CLAHE를 결합하여 선행연구[3](GPT-4o, Precision 97.3%) 대비 0.5%p 내의 정확도를 유지하면서 클라우드 의존성을 완전히 제거하고 FPR 3.2%를 달성하였다. 판정 지연 6.9초는 승가 평균 지속 시간(7.2초) 이내로 승가 발생 시점에 검출·알림이 완료되는 실시간 동작을 확인하였다.

향후 저널 연구에서는 (1) QLoRA 기반 한우 특화 파인튜닝으로 Zero-shot 한계 극복 및 GPT-4o 수준 정확도 달성; (2) YOLO 1차(미확정)+VLM 2차(확정) 2단계 알림으로 단기 승가 포착률 향상; (3) IR-RGB 열화상 융합을 통한 야간 오탐 근본 해소; (4) 선행연구[2]의 개별 한우 ID 추적 기능과 결합하여 발정 개체 특정 및 번식 이력 관리 플랫폼으로 발전시킬 예정이다.

참고문헌

[1] 조영준, 김종원, "ARM 모델을 활용한 한우의 활동량 데이터 수집과 발정 예측," 한국산학기술학회 논문지, Vol. 24, No. 3 pp. 551-556, 2023.

[2] Y. Cho and J. Kim, "AI-Based Intelligent Monitoring System for Estrus Prediction in the Livestock Industry," Applied Sciences, vol. 13, no. 4, p. 2442, Feb. 2023.

[3] 조영준, "Vision-Language Model(VLM) 기반의 하이브리드 파이프라인을 활용한 한우 승가 행위 검출 시스템 연구," 한국산학기술학회논문지, Vol. 26, No. 4, 2025.

[4] B. Zhou et al., "TinyLLaVA: A Framework of Small-scale Large Multimodal Models," arXiv:2402.14289, 2024.

[5] A. Gholami et al., "A Survey of Quantization Methods for Efficient Neural Network Inference," CRC Press, 2022.