인물 객체 간 상호작용 인식을 위한 물리접촉 검출

박승보*, 정의손O, 함동균*, 금용호*
*인하대학교 소프트웨어융합공학과,
O인하대학교 소프트웨어융합공학과
e-mail: molaal@inha.ac.kr*, {junsticehand1999O, hameast20*, go9149*}@inha.edu

Physical Contact Detection for Recognizing Interactions between Person Objects

Seung-bo Park*, Eui-son JungO, Dong-gyun Ham*, Yong-ho Keum*
*Dept. of Software Convergence Engineering, INHA University,
ODept. of Software Convergence Engineering, INHA University

● 요 약 ●

본 논문은 영화의 스토리 인식을 위해 인물 간 상호작용 중 물리적 상호작용 즉, 물리접촉을 검출하는 방법을 제안한다. YOLO를 사용해 영상에서 인간객체를 탐지하고, Mediapipe를 사용해 골격 감지를 진행함으로써 인물의 뼈대를 랜드마크화 하고 타 객체 간의 랜드마크가 일정값 이하로 내려오면 Threshold를 적용해 객체 간의 물리적 접촉을 판단한다, 실험 결과, 50개 17,741 frame의 영상에서 정확도 99.66%의 정밀도 77.27%, 재현율 62.38%로 모델의 전반적인 성능을 나타내는 F1점수는 69%로 나타났다.

키워드: 객체탐지(object detection), 골격추출(skeleton detection), 상호작용(interaction)

I. Introduction

영화의 스토리는 여러 인물 객체 간의 상호작용으로 이루어져 있다. 따라서 영화 속 스토리를 인식하기 위해선 인물의 행위나 타 객체 간의 상호작용을 검출할 필요가 있다. 상호작용은 크게 의사소통, 시선, 물리적 접촉 총 3가지의 형태로 나눌 수 있다. 본 논문에서는 물리접촉 즉, 인물들 간의 접촉을 영상 적으로 검출하는 것에 목표를 둔다. 본 논문에서는 인물 객체 간의 여러 상호작용 중 물리접촉을 검출하여 물리적 상호작용 여부를 판단하는 방법을 제안한다. 본 논문은 2장에서 사용된 기술 및 기존 연구와의 문제점과 보안점 3장에서 방법론에 따른 자세한 설명 4장에서 그에 따른 성능 평가 보안사항 5장에서 결론을 기재하였다.

II. Processing

2.1. Object Detection : YOLO

먼저 영상 데이터에서 사람 Class의 영역을 추출한다. YOLO (You Only Look Once) v5 기반의 객체 탐지 알고리즘을 사용하여 객체를 탐지하고 탐지된 위치를 기반으로 Skeleton Detection을 진행한다. 즉시 Skeleton Detection을 하지 않은 이유는 객체 간의 혼선 및 배경의 차이를 줄이기 위함이다[1].

2.2. Skeleton Detection : Mediapipe

Mediapipe는 구글에서 제공하는 프레임워크로 인간 객체의 골격을 추출하여 각 골격의 관절 별 위치를 랜드마크 형식을 제공한다[2]. 본 논문에서는 객체 간 뼈대의 위치를 기반으로 접촉 여부를 판단하였다.

III. Physical Contact Recognition

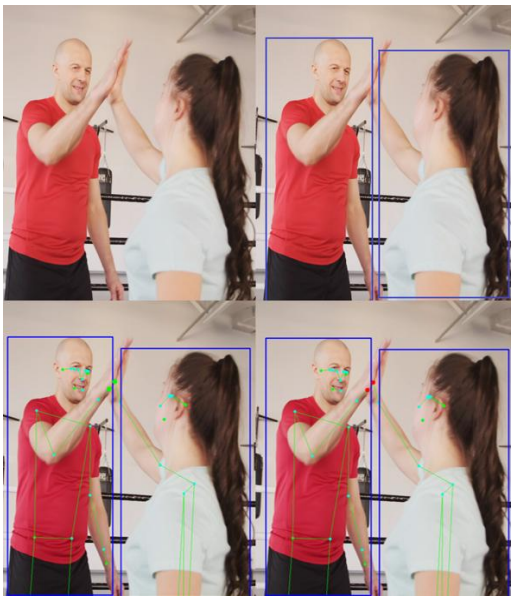

Fig. 1. Examples of Physical Contact

3.1. 물리접촉의 정의

본 논문에서 물리접촉(Physical Contact)은 객체와 객체끼리 서로 접촉을 통한 상호작용을 하는 것을 말한다. 예를 들어 악수, 포옹, 팔짱 끼기 등 두 객체가 물리적으로 접촉해있다면, 객체 간 상호작용을 한다고 볼 수 있다.

3.2. 물리접촉 검출

물리접촉을 검출을 위한 순서도는 다음과 같다. 비디오 데이터가 들어오면 사람 class에 대한 영역을 추출 후 해당 위치를 기반으로 뼈대를 인식한다. 이는 다른 객체와의 혼동을 줄여 정확도를 향상시키기 위함이다. 추출된 뼈대를 기준 거리값을 기준으로 근거리에 존재하는 타 객체 중 가장 가까운 노드를 찾고 Threshold를 적용하여 일정 값 이하에 있을 시 해당 객체는 타 객체와 물리접촉을 한다고 정의할 수 있다.

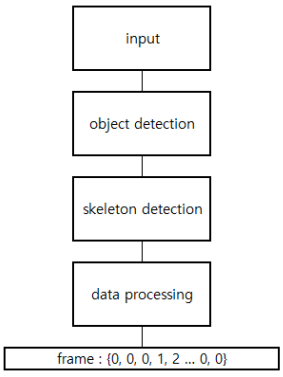

Fig. 2. Decision Flowchart for Physical Contact

3.2.1. Object Detection

$$F_t = \{o_0, o_1, o_2 \dots o_n\} \quad (1)$$

YOLO는 한 Frame 속에 있는 모든 객체의 집합을 F_t라 할 때, 식(1)과 같이 나타낼 수 있다.

$$o_\alpha = \{p_1, p_2 \mid p_1 = (x_1, y_1), p_2 = (x_2, y_2)\} \quad (2)$$

이 좌표값은 식(2)와 같이 바운드 박스의 좌표값을 표현하기 위한 형태로 구성되어있다. 여기서 p_1, p_2는 YOLO로 검출한 객체의 Bound Box의 좌표값이다.

3.2.2. Skeleton Detection

$$S(o_\alpha) = \{l_0, l_1, l_2 \dots l_k\} \quad (3)$$
$$where\ l_{i \in k} = (x_i, y_i)$$

객체 검출을 통해 얻어낸 좌표값을 기준으로 관심 영역을 생성 후 해당 좌표를 중점으로 Mediapipe를 통한 뼈대 추출을 식(3)과 같이 진행한다. 이는 배경, 타 객체와의 혼선과 같은 변수 들을 제거하기 위함이다.

3.2.3. Data Processing

$$D(l_i, l_j) = \underset{i,j}{\operatorname{argmin}} \sqrt{(l_i^2 - l_j^2)}$$
$$where\ l_i \in o_k, l_j \in o_c \quad (4)$$

한 객체의 뼈대를 추출하고 그의 관절에 해당되는 랜드마크 좌표값을 식(4)의 유클리드 거리 판별식을 이용하여 거리를 구한다.

$$f(l_i, l_j) = \begin{cases} 0, D(l_i, l_j) > threshold \\ 1, D(l_i, l_j) \leq threshold \end{cases} \quad (5)$$

구한 값을 Threshold 적용하여 해당 값을 벗어나면 객체 o_α와 o_β는 접촉하지 않는 것으로 판단한다.

IV. Results

4.1. Test Dataset

Table 1. Evaluation Metrics

성능 지표	값
TP	68
FP	20
TN	17612
FN	41
all frame	17741

4.2. Evaluation Metrics

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (6)$$

영상 50개 총 17741 Frame을 사용하여 성능 테스트를 진행하였다. 전체 데이터 중에서, 모델은 99.66%의 높은 정확도를 보였다. 그러나 정확도만으로는 모델의 성능이 완벽하다는 것은 아니다. 특히 불균형 데이터셋에서는 정확도가 과도하게 높게 나타날 수 있기 때문이다. 이 경우에는 양성 샘플(접촉 프레임)이 전체 데이터에서 차지하는 비율이 적기 때문에, 그 외의 다른 성능 지표를 고려하였다.

$$\mathrm{Precision} = \frac{TP}{TP + FP} \quad (7)$$

모델의 정밀도는 77.2%로, 이는 모델이 '접촉'으로 예측한 프레임 중 실제로 '접촉'인 프레임의 비율을 나타낸다. 이는 (FP)의 수가 상당히 있음을 의미하며, 이는 모델이 '접촉'이 없는 프레임을 '접촉'이 있는 것으로 잘못 분류하는 경향이 있음을 나타낸다.

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

모델의 재현율은 62.3%로, 실제 '접촉' 프레임 중에서 모델이 '접촉'으로 정확하게 예측한 프레임의 비율을 나타낸다. 이는 모델이 실제 '접촉' 프레임을 '접촉'이 없는 것으로 잘못 분류하는 경우(FN)가 다소 있다는 것을 나타낸다.

$$F1 = 2 \cdot \left(\frac{\mathrm{Precision} \times Recall}{\mathrm{Precision} + Recall} \right) \quad (9)$$

F1 점수는 69.0%로, 정밀도와 재현율의 조화 평균을 나타낸다. 이 점수는 모델의 전반적인 성능을 대표하며, 특히 불균형한 데이터셋에서 유용한 지표이다.

4.3. 보완점

위 와 같이 데이터셋의 분포가 일정하지 않아 성능지표를 제대로 표현할 수 없는 문제를 해결하기 위해서는 데이터셋의 추가확보나 모델 구조의 변경하는 등의 방법을 고려해야 한다. 아래는 현 모델에 문제점 및 보완점에 대해 언급한다.

4.3.1. 기존 모델 의존성

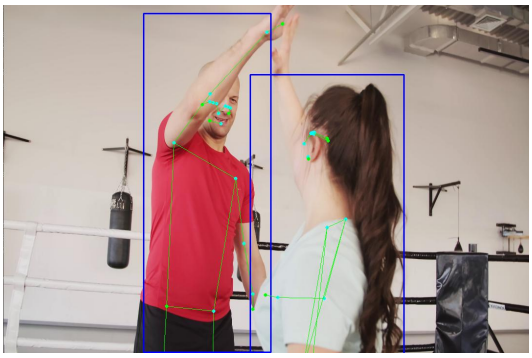

Fig. 3. Misrecognition of Mediapipe

본 논문의 방법론은 YOLO와 Mediapipe의 포즈 추정기의 출력에 의존하는데, 둘 중 하나의 출력이 잘못된 출력값이거나 출력을 내지 못 할 경우, 물리접촉 감지의 정확성도 저하된다. YOLO 모델을 더 큰 데이터셋으로 추가 훈련 시키거나, 더 정확한 모델을 사용하여 개선해야 한다. Fig. 3.과 같이 객체 간 접촉이 일어났음에도 Skeleton Detection이 올바르게 선행되지 않아 검출을 못 하는 것을 알 수 있다.

4.3.2. 예측성 모델

Fig. 4. Model close to Mediapipe predictions

MediaPipe의 Skeleton Detection은 검출뿐만 아니라 Prediction에 가까워 신체 부위가 검출되지 않을 시 예측하여 그려주기에 카메라 사각에 존재하는 접촉이나 잘못된 Skeleton Detection으로 인한 접촉 발생 등 문제가 있다.

V. Conclusions

본 논문은 영화 속 인물 간의 물리적 접촉을 감지하는 방법을 Object Detection으로 객체의 위치를 찾고 Skeleton Detection으로 인간객체의 뼈대를 추출하여 랜드마크를 뽑아내고, 타 객체 간의 랜드마크의 거리가 일정값 이하로 내려오면 Threshold를 적용해 물리접촉을 검출하는 방법을 제시하였다. 그러나, 성능 지표, 실험 결과를 고려할 때, 이 방법은 YOLO와 Mediapipe의 출력에 크게 의존하며, 이 두 기술의 출력 정확성이 물리접촉 감지의 정확성에 큰 영향을 미치는 것으로 나타났다. 따라서, 이를 보완하기 위해 데이터셋을 추가 보충하거나 더 정확한 기술을 사용하는 등의 작업이 필요하다. 이러한 개선 사항을 통해, 더욱 정확한 물리적 접촉 감지가 가능해질 것으로 기대된다.

ACKNOWLEDGEMENT

본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.

REFERENCES

[1] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[2] C. Lugaresi, et al. “Mediapipe: A framework for building perception pipelines,” in Proceedings of Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition, Long Beach, June 2019.