Journal of KIIT. Vol. 23, No. 7, pp. 115-122, Jul. 31, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.7.115

객체 탐지와 텍스트 융합 기반 멀티모달 인터랙티브 시스템

송영훈*, 김남기**[1], 정경용**[2]

Object Detection and Text Fusion-based Multimodal Interactive System

Younghun Song*, Namgi Kim**[1], and Kyungyong Chung**[2]

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(RS-2020-NR049579).
Additionally, this work was also supported by Kyonggi University's Graduate Research Assistantship (2025)

요 약

본 논문은 이미지 기반 객체 탐지와 텍스트 모델을 결합하여, 시각 정보와 자연어 질의를 동시에 처리하는 멀티모달 인터랙티브 시스템을 제안한다. 이를 통해 단일 모달 접근법의 한계를 넘어, 공항 안내 로봇, 자율 에이전트 정보 시스템, 학습 보조 등 다양한 분야에서 풍부하고 정확한 상호작용이 가능함을 시나리오 분석으로 입증한다. 또한 실시간 처리, 데이터 불확실성, 오류 관리 등 현실적 과제를 논의하고, 사용자 피드백과 추가 모달리티 확장을 통한 향후 연구 방향을 제시함으로써 차세대 인터랙티브 플랫폼으로의 발전 가능성을 확인한다. 이러한 결론을 토대로, 본 연구는 차세대 멀티모달 시스템 구현을 위한 유의미한 방향성을 제시한다.

Abstract

This paper proposes a multimodal interactive system that combines image-based object detection and text models to simultaneously process visual information and natural language queries. By overcoming the limitations of a single-modal approach, it demonstrates, through scenario analysis, that rich and accurate interactions are possible in various fields such as airport guide robots, autonomous agent information systems, and learning assistance. The paper also discusses practical challenges such as real-time processing, data uncertainty, and error management, and suggests future research directions through user feedback and the expansion of additional modalities, confirming the potential for development into next-generation interactive platforms. Based on these conclusions, this study presents a meaningful direction for the implementation of next-generation multimodal systems.

Keywords

multimodal interaction, object detection, natural language processing, NLP, fusion strategies, interactive systems

* 경기대학교 컴퓨터과학과 석사과정
- ORCID: https://orcid.org/0009-0006-1470-0130
** 경기대학교 컴퓨터공학부 교수(**[2] 교신저자)
- ORCID[1]: https://orcid.org/0000-0002-0077-6576
- ORCID[2]: https://orcid.org/0000-0002-6439-9992

· Received: Mar. 11, 2025, Revised: Jul. 16, 2025, Accepted: Jul. 19, 2025
· Corresponding Author: Kyungyong Chung
Dept. of Division of Engineering, Kyonggi University, South Korea
Tel.: +82-32-249-1382, Email: dragonhci@gmail.com

Ⅰ. 서 론

인공지능 기술의 급속한 발전과 더불어 다양한 정보 처리 방법이 주목받고 있다. 그중에서도 멀티모달(Multimodal) 접근 방식은 이미지, 텍스트, 음성 등 여러 형태의 데이터를 결합하여 단일 모달리티(Single modality)의 한계를 극복하는 데 기여하고 있다[1]. 단일 모달리티로는 표현하기 어려운 복잡하고 다양한 맥락 정보를 멀티모달 융합을 통해 정확하게 해석하고 반영하는 기술이 요구되는 것이다. 특히 자율 에이전트나 안내 로봇과 같은 인터랙티브 시스템은 사용자의 다양한 요구와 질문에 실시간으로 응답하고, 자연스러운 상호작용을 유지하기 위해 멀티모달 융합 기술의 중요성이 날로 높아지고 있다.

이에 따라 시각적 정보와 텍스트 데이터를 결합한 멀티모달 시스템에 대한 연구는 기술적, 실용적 측면 모두에서 중요한 의미를 지니고 있다. 이러한 시스템은 객체 탐지(Object detection) 기술과 자연어 처리(NLP, Natural Language Processing) 기술을 결합하여, 기존의 단일 모달 시스템보다 더욱 직관적이고 효율적인 사용자 상호작용을 가능하게 한다[2].

본 연구의 목적은 객체 탐지 기술과 최신의 텍스트 모델을 통합하여 멀티모달 인터랙티브 시스템의 개념적 설계안을 제안하는 것이다. 이를 위해 다음과 같은 세부 목표를 설정하였다.

첫째, 객체 탐지와 텍스트 모델의 결합을 통해 사용자의 시각적 입력과 텍스트 기반 질의를 동시에 처리할 수 있는 시스템의 아키텍처를 제안한다. 둘째, 시스템 내 각 구성 요소의 명확한 역할을 정의하고, 구성 요소 간의 효율적인 정보 통합 전략을 도출한다. 셋째, 다양한 응용 시나리오를 설정하여 제안된 시스템의 이론적 장점과 실제 활용 가능성을 분석하고, 기대 효과를 도출한다.

이러한 접근법을 통해 본 연구는 멀티모달 인터랙티브 시스템의 실용성과 효율성을 높이는 데 기여할 것으로 기대된다.

본 논문의 구성은 다음과 같다. 제 Ⅱ장에서는 관련 연구 동향 및 문헌을 검토한다. 제 Ⅲ장에서는 시스템 구성 요소와 각 모듈의 역할을 상세히 설명하고, 모달 융합을 위한 통합 전략을 제안한다. 제 Ⅳ장에서는 다양한 활용 사례를 통해 시스템의 실질적인 응용 가능성을 분석한다. 제 Ⅴ장에서는 본 논문의 결과를 요약하고 연구의 한계점 및 향후 연구 방향을 논의한다.

Ⅱ. 관련 연구

2.1 객체 탐지 기술

객체 탐지는 이미지나 영상 내에 존재하는 물체의 위치와 클래스를 자동으로 식별하는 기술로, 최근 딥러닝 기반 알고리즘의 발전으로 처리 속도와 정확도 모두 크게 향상되고 있다. 대표적인 객체 탐지 모델 중 하나인 YOLO(You Only Look Once)는 단일 스테이지에서 전체 이미지를 대상으로 위치와 클래스를 동시에 예측함으로써, 실시간 처리가 필요한 분야에서 각광받고 있다[3]. 반면 Faster R-CNN은 영역 제안(Region proposal)과 분류 과정을 통합함으로써 높은 정확도를 보이는 모델 구조를 갖추고 있다[4].

최근에는 DETR(Detection Transformer)가 등장하여, CNN 기반 모델들과 달리 어텐션 메커니즘(Attention mechanism)을 통해 객체 간 관계나 맥락 정보를 더욱 풍부하게 반영하는 방식으로 주목받았다. 이와 같은 모델들은 지속적인 구조 개선과 최적화를 통해 객체 간의 상호작용까지 분석하는 고차원적 탐지, 저지연(Low-latency) 환경에서의 안정적인 성능 유지 등 다양한 연구 주제로 확장되고 있다[5.]. 따라서 객체 탐지 기술은 멀티모달 환경에서 시각 정보를 정확하게 획득하고 활용하기 위한 핵심 기제로 평가된다.

표 1은 대표적인 객체 탐지 알고리즘인 YOLO, Faster R-CNN, DETR를 비교하여 각 모델의 구조, 속도, 강점을 간략히 보여준다. YOLO는 단일 스테이지 구조(1-Stage)로 실시간 처리에 강점을 보이며, 구조가 단순하고 빠르지만 복잡하거나 작은 객체 탐지에는 한계가 있다. Faster R-CNN은 영역 제안 과정을 거치는 2-Stage 방식으로 높은 정확도를 제공하지만, 상대적으로 처리 속도가 느려 실시간 처리에는 부적합하다.

표 1. 객체 탐지 모델 비교
Table 1. Object detection model comparison

Model	Architecture type	Speed (FPS)	Advantages
YOLO	1-Stage	High	Real-time processing, simple
Faster R-CNN	2-Stage	Moderate	High accuracy, includes region proposals
DETR	Transformer	Varies with setup	Reflects contextual relationships

한편 DETR은 Transformer 기반 접근을 통해 객체 간의 맥락 정보를 더욱 풍부하게 반영하지만, 초기 학습 비용이 높고 하드웨어 환경이나 설정에 따라 처리 속도가 불안정할 수 있다. 따라서 실제 적용 시에는 정확도, 속도, 사용 환경 등을 종합적으로 고려해 모델을 선택해야 한다.

2.2 텍스트 모델 및 자연어 처리

자연어 처리 분야에서는 BERT(Bidirectional Encoder Representations from Transformers)와 GPT (Generative Pre-trained Transformer) 계열 모델을 필두로 언어 이해와 생성 능력이 크게 발전하였다. BERT는 양방향 맥락 정보를 효율적으로 학습함으로써, 문서 분류, 질의 응답, 개체명 인식 등 다양한 다운스트림 과업에서 우수한 성능을 보이고 있다[6].

한편 GPT 계열은 단방향성을 기반으로 대규모 파라미터를 활용해 자연스러운 문장 생성 능력을 확보하였으며, GPT-3, GPT-4 등으로 확장되면서 텍스트 요약, 번역, 대화형 에이전트 등 광범위한 응용 분야에서 두각을 나타내었다. 특히 모델 경량화, 파인튜닝 기법 발전, 그리고 다양한 데이터셋 적용에 따른 성능 향상 연구가 진행되면서 실제 서비스 환경에서의 빠른 응답 속도와 높은 정확도를 동시에 달성하기 위한 방안이 적극 모색되고 있다[7].

표 2는 BERT와 GPT 계열 모델의 아키텍처, 파라미터 규모, 그리고 주요 강점을 비교·정리한 것이다. BERT는 양방향 Transformer 구조를 통해 문맥 이해 능력이 탁월하여, 문서 분류나 질의응답 등 여러 과업에서 높은 성능을 보인다. 그러나 BERT는 주로 이해 기반 모델로서 문장 생성 능력이 부족하고, 대규모 파라미터로 인해 응답 속도가 느릴 수 있다. 반면 GPT-3와 GPT-4는 단방향 구조와 대규모 파라미터를 통해 자연스럽고 일관된 텍스트 생성이 가능하며, 특히 GPT-4는 개선된 아키텍처로 맥락 이해와 생성 품질이 더욱 향상되어 보다 폭넓은 응용 분야에 효과적으로 활용될 수 있다. 다만 GPT 계열은 단방향성으로 인한 문맥 처리의 제한이 존재하고, 높은 연산 자원 요구와 함께 데이터 편향 문제도 여전히 과제로 남아 있다. 따라서 사용 목적과 환경에 따라 각 모델의 강점과 한계를 균형 있게 고려하여 적용해야 한다.

표 2. 텍스트 모델 비교
Table 2. Text model comparison

Model	Architecture	Parameter size	Advantages
BERT	Bidirectional transformer	340M	Excellent contextual understanding
GPT-3	Unidirectional transformer	175B	Generates natural, coherent text
GPT-4	Improved unidirectional transformer	1.76T	Enhanced context understanding and text generation

2.3 멀티모달 융합 전략

시각 정보(이미지, 영상)와 텍스트, 나아가 음성 등 다양한 데이터 소스를 결합하는 멀티모달 융합은 단일 모달리티로는 파악하기 어려운 복합적 맥락을 해석하는 데 필수적인 방법론이다[8]. 일반적으로 Early Fusion, Late Fusion, 그리고 Hybrid Fusion이 대표적인 융합 방식으로 제시된다.

Early Fusion은 각 모달리티 데이터를 전처리 단계에서 하나의 피처 벡터로 결합함으로써 풍부한 맥락 정보를 확보할 수 있다는 장점이 있으나, 모달리티별 특성을 세분화하여 반영하기 어렵다는 한계

가 있다. 반면 Late Fusion은 모달리티별로 독립적인 학습 과정을 거친 뒤 최종 결과 혹은 확률값을 통합하기 때문에 각 모달리티의 최적화를 용이하게 해주지만, 모달 간 상호작용이 제한적일 수 있다. Hybrid Fusion은 이러한 두 방식을 절충하여 중간 단계에서도 모달 간 피드백을 주고받고, 최종 단계에서도 합성 결과를 재검토함으로써 Early Fusion과 Late Fusion의 장점을 균형 있게 활용한다.

그림 1은 Early Fusion, Late Fusion, Hybrid Fusion의 세 가지 대표적인 멀티모달 융합 기법을 간략히 비교한 다이어그램이다. Early Fusion은 입력 단계에서 서로 다른 모달리티의 특징을 합쳐 풍부한 맥락 정보를 얻는 반면, Late Fusion은 각 모달리티를 독립적으로 처리한 뒤 최종적으로 합산된 결과를 사용한다. Hybrid Fusion은 이 두 방식을 절충하여 중간 단계에서도 교차 정보를 주고받고, 최종 단계에서도 재검토 과정을 거쳐 종합적인 의사결정을 내리는 방식이다. 그림을 통해 각 융합 기법의 특징과 데이터 흐름을 한눈에 비교할 수 있으며, 상황 및 목적에 따라 적절한 기법을 선택하는 데 도움을 준다[9].

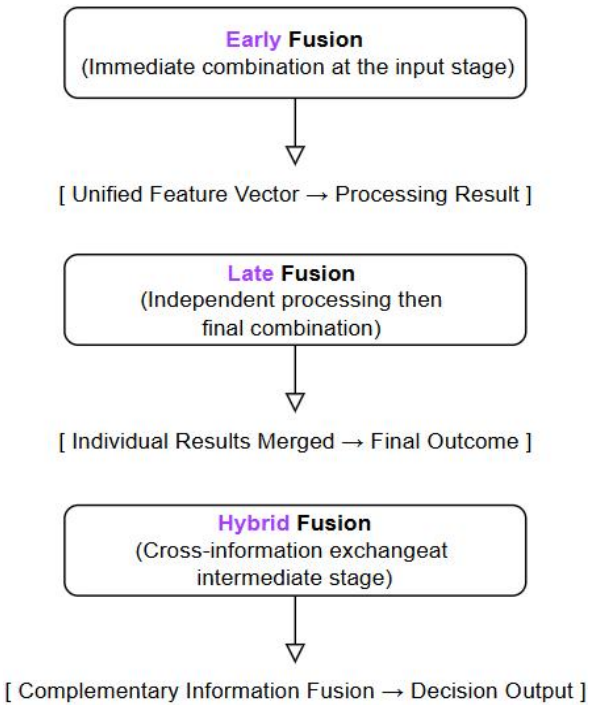

그림 1. 멀티모달 융합 기법 비교 다이어그램
Fig. 1. Comparison diagram of multimodal fusion techniques

최근에는 Transformer나 그래프 신경망(GNN)을 활용하여 모달리티 간 상호작용을 동적으로 학습하고, 상황에 따라 중요도를 가변적으로 조정하는 어텐션 기반 융합 기법이 각광받고 있다. 이러한 기술적 발전은 오류 전파를 최소화하고, 복합적인 상황에서도 정확하고 정교한 의사결정을 내릴 수 있도록 지원함으로써 멀티모달 융합의 활용 범위를 더욱 넓히고 있다.

2.4 대화형 인터렉티브 시스템

대화형 인터랙티브 시스템은 사용자와 에이전트가 실시간으로 상호작용하며, 사용자 요구와 의도를 정확히 파악해 자연스러운 반응을 제공하는 것을 목표로 한다. 초기에는 텍스트나 음성에만 초점을 둔 챗봇, 음성 비서 등이 일반적이었으나, 최근에는 시각적 정보를 결합하는 멀티모달 접근이 점차 확대되고 있다.

이를 통해 안내 로봇이나 자율 에이전트가 주변 환경의 객체를 실시간으로 인식하고, 사용자 질의와 결합하여 보다 직관적이고 상세한 설명을 제공할 수 있다. 예컨대, 객체 탐지를 기반으로 특정 사물이나 표지판을 식별한 뒤, 텍스트 모델을 통해 해당 사물의 기능이나 위치 정보를 자연어로 설명하는 방식이다. 또한 사용자 피드백을 즉각 반영해 대화 흐름을 유연하게 조정하고, 오류가 발생할 경우 신속하게 예외 처리를 수행할 수 있다는 점에서, 멀티모달 대화형 시스템은 사용자 경험(UX)을 혁신적으로 개선할 수 있는 잠재력을 갖춘다.

결과적으로 이러한 연구들은 사용자와 에이전트 간의 상호작용이 단순 정보를 주고받는 차원을 넘어, 인간과 유사한 수준의 이해와 맥락 인식을 달성하기 위한 필수 요소로 자리매김하고 있다.

III. 제안 시스템의 개념적 아키텍처 및 통합 전략

3.1 시스템 개요

본 시스템의 핵심 아이디어는 객체 탐지 모듈과 텍스트 모델을 상호보완적으로 결합하여, 사용자의

시각 정보와 텍스트 입력을 동시에 처리하는 데 있다. 이를 통해 단순히 시각적 정보만을 전달하거나 텍스트 질의에만 응답하는 시스템의 한계를 넘어서, 실제 사용자가 접하는 복합적인 맥락과 요구사항에 대한 맞춤형 대응이 가능해진다.

시스템의 전체 구성은 크게 네 가지 모듈로 나누어진다. 첫째, 사용자 또는 센서 디바이스로부터 입력되는 시각 정보를 처리하는 모듈이다. 둘째, 자연어 형태의 텍스트 입력을 분석하고 이해하는 텍스트 처리 모듈이다. 셋째, 두 모달리티에서 확보된 정보를 융합하여 필요한 의사결정을 내리는 모달 융합 모듈이고, 마지막으로 이 결과를 사용자에게 다양한 형태(텍스트, 시각, 음성 등)로 피드백하는 출력 및 피드백 모듈이다.

3.2 구성 요소 상세 분석

시각 정보 처리 모듈은 카메라, 영상 센서 등으로부터 획득한 이미지를 분석하여 객체 탐지, 분류, 공간 관계 파악 등을 수행한다. 구체적으로는 최신 객체 탐지 알고리즘(예: YOLO, Faster R-CNN, DETR 등)을 활용해 이미지 또는 영상 프레임 내의 객체를 실시간으로 식별하고, 해당 객체의 위치 정보(바운딩 박스), 클래스 정보, 객체 간 상대적 관계 등을 추출한다. 이렇게 획득된 객체 정보는 텍스트 모달리티와 결합될 때 중요한 단서가 되므로, 인식 정확도와 처리 속도를 균형 있게 유지할 수 있는 모델 선정이 핵심이다.

텍스트 입력 처리 모듈은 사용자 질의나 명령, 또는 시스템과의 대화 중 입력된 자연어 문장을 이해하고 의도를 파악한다. 이를 위해 BERT, GPT와 같은 사전 학습(Pre-trained) 모델을 기반으로 의도 분석, 키워드 추출, 문맥 이해 등의 과정이 수행된다. 예컨대 특정 객체에 대한 설명을 요청하거나, 현재 인지된 객체 목록과 관련된 질문을 하는 경우에도, 텍스트 모델은 해당 질의가 정확히 무엇을 요구하는지 해석하고, 답변에 필요한 관련 정보를 찾는 과정을 담당한다.

모달 융합 모듈은 시각 정보와 텍스트 임베딩 정보를 종합하여 최종 의사결정을 내리는 핵심 부분이다. 먼저 객체 탐지 결과에서 얻은 피처 벡터와 텍스트 모델에서 추출된 임베딩을 교차 어텐션(Cross-Attention) 기법을 통해 결합한다. 이때 Early Fusion, Late Fusion, 또는 Hybrid Fusion 방식 중 시스템 목적에 맞는 방법을 선택한다. 본 연구에서는 필요에 따라 두 모달리티 간 실시간 상호작용이 가능하도록, Hybrid Fusion 방식을 중심으로 설계하는 방안을 제안한다. 이 과정을 통해 객체 관련 맥락과 텍스트 내 의미 정보를 결합하여, 사용자 질문에 대한 답변 생성이나 시스템 내부 의사결정(예: 안내 경로 결정, 경고 메시지 표시 등)을 수행하게 된다.

최종 단계에서, 융합 모듈이 도출한 결과를 사용자에게 전달하고 사용자로부터의 피드백을 수집한다. 예컨대 대화형 응답 생성 시에는 텍스트 모델을 다시 활용해 자연스럽고 구체적인 문장을 생성하며, 시각적 피드백(디스플레이, AR 환경 표시 등)이나 음성 합성 기술(TTS)을 통해 멀티모달 형태의 결과물을 제공할 수 있다. 사용자 반응이 긍정적이면 해당 결과를 유지하고, 부정적 피드백이 들어오면 시스템 내부 파라미터를 조정하거나 예외 처리 루틴을 가동해 정정 절차를 진행한다. 이러한 순환 구조는 실시간 상호작용을 통해 시스템 성능을 점차 개선하는 중요한 피드백 루프를 형성한다.

3.3 통합 전략 및 알고리즘

본 시스템의 통합 전략은 멀티모달 입력 간의 비동기성을 해소하고, 상황 변화에 따른 동적 가중치 조정, 그리고 발생 가능한 오류에 대한 체계적 관리를 통해 신뢰성 있는 사용자 상호작용을 구현하는 것을 목표로 한다.

먼저 정보 동기화 단계에서는 시각 정보와 텍스트 데이터가 서로 다른 시간축에서 수집되기 때문에, 이를 정렬하기 위한 구체적 메커니즘으로 슬라이딩 윈도우 기반의 타임스탬프 큐를 적용하였다. 사용자의 질문이 발생했을 경우, 시스템은 해당 질의 시점을 기준으로 최근 3초 이내의 객체 탐지 결과를 자동으로 선별하여 최신 정보에 기반한 정확한 답변을 유도한다. 이 3초 기준은 일반적인 사용자 반응 시간(2–4초) 및 실시간 센서 처리 주기(1Hz 이상)를 고려하여 설정된 경험적 기준이며, 짧은 시간 내 수집된 정보를 활용함으로써 반응성은

유지하면서도 과거 데이터의 노이즈를 최소화할 수 있다. 이를 통해 과거 탐지 결과가 질의에 잘못 연결되는 오류를 방지하고, 최신성이 높은 정보만을 질의 처리에 사용하도록 관리한다.

동적 업데이트 메커니즘은 두 모달리티의 신뢰도를 실시간 환경 변화에 따라 가변적으로 조정하기 위해 Cross-Attention Layer 내 동적 가중치 조정 수식을 적용한다. 예를 들어, 시각 입력 데이터의 품질이 낮거나 객체 탐지가 불완전한 경우, 시스템은 자동으로 텍스트 모델의 기여도(W_text)를 상승시키고, 반대로 시각 정보가 충분히 확보된 경우에는 객체 인식 결과의 비중(W_vision)을 높인다. 이 과정은 아래와 같은 수식을 기반으로 한다.

$$W_{total} = \alpha \times \mathrm{V} + (1-\alpha) \times \mathrm{T} \quad (1)$$

식 (1)에서 α는 동적 가중치 계수이며, 환경 평가 모듈이 사용자의 입력 상황, 시각 정보 품질, 센서 노이즈 수준 등을 실시간으로 평가하여 0.3에서 0.7 범위 내에서 자동으로 조정된다. 이를 통해 시스템은 시각 정보의 신뢰도가 낮은 경우 텍스트 모델의 기여도를 높이고, 반대로 시각 정보가 충분한 경우 해당 정보를 보다 적극적으로 활용할 수 있다.

마지막으로 오류 관리 및 예외 처리 단계에서는 발생 가능한 다양한 오류 상황에 대비하여 규칙기반 오류 처리 플로우를 적용한다[10]. 예를 들어, 객체 탐지 오류가 발생한 경우 시스템은 Top 3 후보 객체 리스트를 표시하여 사용자가 직접 선택하도록 유도한다. 텍스트 해석 과정에서 모호한 질의가 발생할 경우에는 "정확히 어떤 것을 원하시나요?"라는 메시지를 자동 생성하여 사용자의 명확한 피드백을 요청한다.

그림 2는 사용자의 입력으로부터 시각 정보와 텍스트 데이터를 수신하여 시간 기반 버퍼링 및 정렬을 수행하고, 멀티모달 융합 단계에서 가중치를 동적으로 업데이트한 후 결과를 생성하는 전체 과정을 나타낸다. 이 과정에서 오류가 감지되면 재확인 요청을 통해 결과를 보정하고, 문제가 없을 경우 정상 응답을 생성하여 최종적으로 사용자에게 응답을 전달하는 흐름으로 구성되어 있다.

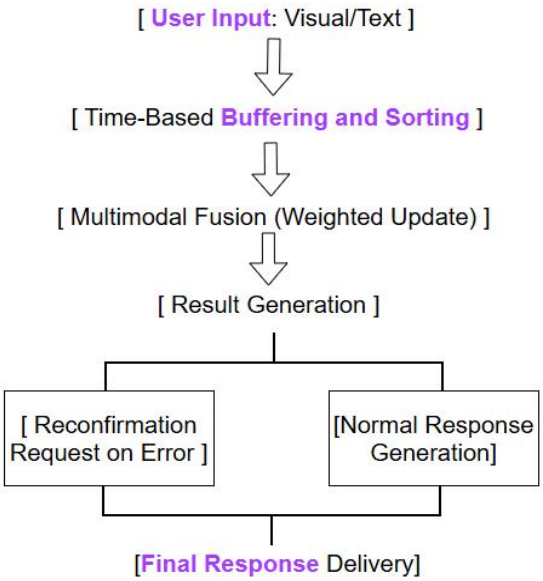

그림 2. 데이터 동기화 및 동적 업데이트 메커니즘 흐름도

Fig. 2. Data synchronization and dynamic update mechanism flowchart

Ⅳ. 활용 시나리오 및 예상 성능 평가

제안된 멀티모달 인터랙티브 시스템의 실질적인 응용 가능성을 확인하고, 각 모듈의 역할과 통합 전략이 실제 환경에서 효과적으로 작동하는지를 평가하기 위해, 본 연구에서는 도서관, 회의실, 출입구 등 다양한 실내 공간을 기반으로 한 세 가지 대표적 활용 시나리오를 설정하였다. 이러한 시나리오 설계는 사용자의 일상적 환경에서 발생할 수 있는 자연어 질의와 시각 정보의 조합에 시스템이 어떻게 반응하는지를 구체적으로 관찰하기 위한 목적을 지닌다.

각 시나리오에서는 실제 이미지 또는 영상 프레임을 기반으로 YOLOv5 객체 탐지 모델을 통해 주요 물체를 실시간으로 인식하였고, 사용자가 입력한 자연어 질의는 GPT-4 언어 모델을 활용하여 의미를 해석하고 적절한 응답을 생성하였다. 이 과정에서 시각 및 텍스트 모듈은 독립적으로 동작하면서도 최종 융합 단계에서 하이브리드 방식으로 상호 보완적으로 결합되어, 사용자 질의와 시각적 맥락 간의 의미 있는 상호작용이 가능하도록 설계되었다.

성능 평가는 정량적인 지표보다 실질적인 사용자 반응과 문맥 정합성에 초점을 둔 정성적 기준에 따

라 수행되었다. YOLOv5는 다양한 환경 조건에서도 주요 객체를 효과적으로 식별하였으며, GPT-4는 질의의 의도를 정확히 파악하여 자연스럽고 문맥에 맞는 응답을 생성하는 데 강점을 보였다.

표 3은 각 시나리오에서 사용된 입력 이미지, 사용자 질의, 탐지된 객체, 그리고 시스템이 생성한 응답을 정리한 것으로, 제안된 시스템의 개념적 타당성과 직관적인 상호작용 가능성을 입증하는 사례로 제시된다.

표 3. 다중 모달 상호작용 시나리오 결과
Table 3. Multimodal interaction scenario results

Input image	Query	Detection result	Response
Library	What is on the right side?	Desk, chair	There is a reading chair on the right
Meeting room	Is there any conference gear?	Monitor, remote control	A display for meetings is available
Exit door	What is that sign?	Exit sign	It is an emergency exit sign

본 결과를 통해 제안된 시스템이 멀티모달 데이터를 효과적으로 통합하고, 실시간 사용자 질의에 대해 의미 있는 응답을 생성할 수 있음을 확인하였다. 향후에는 실제 환경에서 다양한 사용자 집단을 대상으로 정량적 실험을 수행함으로써, 시스템 성능에 대한 보다 체계적인 평가를 진행할 계획이다.

V. 결 론

본 논문에서는 객체 탐지와 텍스트 모델을 융합하여 멀티모달 인터랙티브 시스템을 구현하기 위한 개념적 설계안을 제안하였다. 시각 정보와 텍스트 입력을 결합함으로써 보다 풍부하고 정확한 사용자 상호작용이 가능하다는 점을 강조하고, 이러한 시스템이 다양한 응용 분야에서 높은 효용성을 보일 수 있음을 사례 연구를 통해 제시하였다.

우선, 시각 정보 처리 모듈과 텍스트 모델을 상호보완적으로 결합하는 아키텍처를 구성함으로써, 단일 모달리티 접근법에 비해 한층 직관적이고 구체적인 피드백을 제공할 수 있음을 확인하였다. 객체 인식 결과와 사용자 질의를 결합하여 실시간으로 반응하고, 사용자의 피드백을 학습 과정에 반영함으로써 시스템은 반복적인 사용을 통해 점진적으로 개선될 수 있는 가능성을 보여주었다. 특히 지능형 안내 로봇, 자율 에이전트 기반 대화형 정보 시스템, 멀티모달 대화형 학습 시스템 등 각기 다른 도메인에서도 본 시스템이 가치를 발휘함을 사례 분석을 통해 도출하였다.

그럼에도 불구하고, 실시간 처리를 위한 고성능 하드웨어 요구와 데이터 불확실성, 오류 관리 문제 등은 향후 해결되어야 할 핵심 과제이다. 객체 인식의 오검출이나 텍스트 모델의 오인식처럼 예측이 어려운 오류가 발생할 수 있으며, 이를 최소화하기 위한 시스템 차원의 예외 처리와 사용자 피드백 반영 메커니즘이 추가적으로 보완되어야 한다. 또한 실제 사용자 환경에서의 장기간 운용 사례를 통해 체계적인 평가지표와 대규모 검증 과정을 거침으로써, 본 연구에서 제시한 개념적 설계안의 실효성을 보다 명확히 입증할 필요가 있다.

향후 연구에서는 제안된 프레임워크를 토대로 실제 프로토타입을 개발하고, 다양한 사용자 그룹과 실제 사용 환경에서 정량·정성적 평가를 수행함으로써 보다 현실적인 요구사항을 반영하는 후속 연구가 이루어져야 한다. 나아가 음성, 제스처, 생체신호 등 추가 모달리티를 결합하는 확장 연구를 통해 한층 폭넓은 응용 가능성을 모색할 수 있을 것이다. 이와 함께 멀티모달 융합 과정에서 발생하는 정보 손실이나 중복 문제를 줄이기 위한 고도화된 융합 알고리즘을 마련하고, 사용자 경험(UX) 관점에서 시스템의 인터랙션 흐름을 정교화함으로써 차세대 인터랙티브 플랫폼으로 발전시킬 수 있을 것으로 기대된다.

References

[1] S. E. Park, "Analysis of the Status of Natural Language Processing Technology Based on Deep Learning", The Journal of Bigdata, Vol. 6, No. 1, pp. 63-81, 2021. https://doi.org/10.36498/kbigdt.

2021.6.1.63.

[2] D. G. Lee, S. H. Ji, and B. Y. Park, "PointNet and RandLA-Net Algorithms for Object Detection Using 3D Point Clouds", Journal of the Society of Naval Architects of Korea, Vol. 59, No. 5, pp. 330-337, Oct. 2022. https://doi.org/10.3744/SNAK.2022.59.5.330.

[3] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger", 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 7263-7271, Jul. 2017. https://doi.org/10.1109/CVPR.2017.690.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", arXiv preprint, arXiv:1506.01497, Jun. 2015. https://doi.org/10.48550/arXiv.1506.01497.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers", arXiv preprint, arXiv:2005.12872, May 2020. https://doi.org/10.1007/978-3-030-58452-8_13.

[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv preprint, arXiv:1810.04805, Oct. 2018. https://doi.org/10.18653/v1/N19-1423.

[7] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, et al., "GPT-4 Technical Report", arXiv preprint, arXiv:2303.08774, Mar. 2023. https://doi.org/10.48550/arXiv.2303.08774.

[8] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 41, No. 2, pp. 423-443, Feb. 2019. https://doi.org/10.1109/TPAMI.2018.2798607.

[9] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences", arXiv preprint, arXiv:1906.00295, Jun. 2019. https://doi.org/10.48550/arXiv.1906.00295.

[10] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra, "Visual Dialog", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 326-335, Jul. 2017. https://doi.org/0.1109/CVPR.2017.121.

저자소개

송 영 훈 (Younghun Song)

2025년 3월 ~ 현재 : 경기대학교 컴퓨터과학과 석사과정
관심분야 : 데이터마이닝, 컴퓨터비전, 딥러닝, 빅데이터, 인공지능

김 남 기 (Namgi Kim)

1997년 2월 : 서강대학교 컴퓨터과학과(공학사)
2000년 3월 : KAIST 전산학과(공학석사)
2005년 3월 : KAIST 전산학과(공학박사)
2007년 2월 : 삼성전자 통신연구소 책임연구원
2007년 3월 ~ 현재 : 경기대학교 컴퓨터공학부 교수
관심분야 : 통신시스템, 네트워크

정 경 용 (Kyungyong Chung)

2000년 2월 : 인하대학교 전자계산공학과(공학사)
2002년 2월 : 인하대학교 전자계산공학과(공학석사)
2005년 8월 : 인하대학교 컴퓨터정보공학부(공학박사)
2006년 3월 ~ 2017년 2월 : 상지대학교 컴퓨터정보공학부 교수
2017년 3월 ~ 현재 : 경기대학교 컴퓨터공학부 교수
관심분야 : 데이터 마이닝, 빅데이터, HCI, 인공지능