【명세서】
【발명(고안)의 설명】
【발명(고안)의 명칭】
음성인식 기반 멀티모달 상호작용 플랫폼 및 그 운영 방법 {VOICE RECOGNITION-BASED MULTIMODAL INTERACTION PLATFORM AND OPERATING METHOD THEREOF}
【기술분야】
【0001】
본 발명은 인공지능(AI) 기술, 특히 음성 인식(Speech-to-Text, STT), 자연어 이해(Natural Language Understanding, NLU), 컴퓨터 비전(Computer Vision) 기반 이미지 분석, 위치 기반 서비스(Location-Based Service, LBS), 및 챗봇 자동화 기술이 융합된 멀티모달 사용자 인터페이스(UI) 플랫폼에 관한 것이다. 더욱 상세하게는, 사용자의 음성 명령과 주변 환경으로부터 획득되는 이미지, 위치 정보 등 다양한 맥락적 입력 정보를 복합적으로 분석 및 처리하여, 전자상거래, 스마트홈, 자동차 인포테인먼트, 웨어러블 디바이스 등 광범위한 디지털 환경에서 사용자가 직관적이고 효율적으로 서비스를 제어하고 개인화된 정보를 제공받을 수 있도록 지원하는 플랫폼 및 그 운영 방법에 대한 것이다.
【발명(고안)의 배경이 되는 기술】
【0003】
기존의 웹·모바일 서비스는 대부분 터치·키보드 입력 중심의 그래픽 사용자 인터페이스(Graphical User Interface, GUI)에 의존하고 있었다. 그러나 스마트 디바이스의 보급률 증가와 인공지능 기술의 급속한 발전으로 인해, 음성을 주 입력 수단으로 사용하는 음성 사용자 인터페이스(Voice User Interface, VUI)에 대한 수요가 다방면에서 증대되고 있다.
【0004】
하지만 종래의 음성 인터페이스 시스템은 여러 가지 본질적인 한계점을 내포하고 있다. 첫째, 초기 VUI 시스템은 "온도 24도로 설정해줘"와 같이 사전에 정의된 엄격한 문법 구조를 따르는 정형화된 특정 명령어 인식에 국한되었다. 이로 인해, "방이 좀 덥네"와 같이 일상적인 대화에서 사용되는 비정형적인 자연어 발화에 담긴 사용자의 복잡하고 미묘한 의도를 정확하게 파악하는 데 근본적인 어려움이 있었다. 이러한 경직성은 사용자와 시스템 간의 유연한 상호작용을 저해하는 주요 원인이 되었다.
【0005】
둘째, 종래 기술은 '맥락적 고립(Contextual Isolation)'이라는 문제점을 가진다. 즉, 각각의 입력 채널을 독립적으로 처리하여 사용자의 통합적인 상황을 이해하지 못했다. 예를 들어, 시스템은 사용자의 음성 명령을 처리하면서 그 사용자가 현재 무엇을 보고 있는지(시각 정보) 또는 어디에 있는지(위치 정보)를 동시에 고려하지 못했다. 이러한 데이터 처리의 단절은 사용자의 실제 상황이나 주변 환경에 최적화된 개인화되고 상황 인지적인(context-aware) 맞춤형 서비스를 제공하는 데 결정적인 제약으로 작용하였다. 사용자가 특정 물건의 사진을 찍으며 "이것과 비슷한 거 찾아줘"라고 말하더라도, 이미지 분석 기술이 음성 시스템에 유기적으로 통합되지 않으면 해당 요청을 처리할 수 없었다.
【0006】
셋째, 기술적으로 종래 시스템들은 음성으로부터 변환된 텍스트, 이미지로부터 추출된 고차원 특징 벡터, GPS 센서로부터 수신된 지리 공간 좌표와 같은 서로 다른 형태와 차원을 가진 이종 데이터 스트림을 실시간으로 융합(fusion)하여 단일한 의사결정 컨텍스트로 통합하는 강건

한(robust) 메커니즘이 부재하였다. 이러한 기술적 한계는 음성 인터페이스의 실질적인 활용 범위를 제한하고, 비효율적인 상호작용을 유발하여 전반적인 사용자 경험(User Experience, UX)을 저해하는 요인으로 작용하였다. 따라서 사용자 편의성을 극대화하고 다양한 실제 환경에 적용 가능한 차세대 인터페이스 기술의 필요성이 증대되고 있다.

【선행기술문헌】

【특허문헌】

【0007】

(특허문헌 1) 대한민국 공개특허 제10-20XX-XXXXXXX호 (명칭: 음성 명령을 이용한 정보 검색 장치 및 방법)

【비특허문헌】

【0008】

(비특허문헌 1) "A Survey of Multimodal Fusion Techniques," ACM Computing Surveys, Vol. X, No. Y, 20XX.

【발명(고안)의 내용】

【해결하려는 과제】

【0009】

본 발명은 상술한 종래 기술의 한계를 극복하는 것을 목적으로 한다. 구체적으로, 본 발명은 고도화된 자연어 이해(NLU) 기술을 통해 정형화된 명령뿐만 아니라 비정형적인 자연어 음성 발화로부터 사용자의 복합적인 의도를 정확하게 파악하는 것을 제1 과제로 한다.

【0010】

또한, 본 발명은 종래 기술의 '맥락적 고립' 문제를 해결하기 위해, 음성 정보 외에 사용자의 위치 정보, 시각 정보 등 다양한 멀티모달(Multimodal) 데이터를 실시간으로 융합 처리하여, 사용자의 현재 맥락에 최적화된 직관적인 맞춤형 서비스를 제공하는 통합 멀티모달 상호작용 플랫폼을 구축하는 것을 제2 과제로 한다.

【0011】

마지막으로, 본 발명은 이러한 이종 데이터 스트림을 효율적으로 처리하고 통합하여 복합적이고 상황 의존적인 사용자 요청을 신속하게 해결할 수 있는 구체적이고 실용적인 기술 아키텍처를 제공하는 것을 제3 과제로 한다.

【과제의 해결 수단】

【0012】

상기 과제를 해결하기 위해, 본 발명에 따른 음성인식 기반 멀티모달 상호작용 플랫폼은 도 1에 도시된 바와 같이, 사용자의 다양한 형태의 입력을 수신하는 사용자 인터페이스 계층, 수신된 이종 데이터를 분석하고 사용자의 통합적인 의도를 파악하는 핵심 처리 모듈, 파악된 의도에 따라 실제 서비스를 실행하고 사용자에게 결과를 제공하는 서비스 실행 및 피백 모듈, 그리고 시스템 운영에 필요한 데이터를 관리하고 보안을 담당하는 데이터 및 보안 계층을 포함하여 구성된다.

【0007】 사용자 인터페이스 계층은 마이크, 카메라, GPS 센서와 같은 다양한 입력 장치들을 통해 사용자의 음성, 이미지, 위치 데이터를 수집하여 각 모듈로 전달하는 역할을 한다. 이 계층은 사용자가 플랫폼과 상호작용하는 첫 번째 접점으로서, 다양한 입력 형태를 동시에 수용할 수 있다.

【0008】 핵심 처리 모듈 계층은 사용자 인터페이스 계층으로부터 수신된 원시 데이터를 가공

하고 사용자의 의도를 정확하게 파악하는 핵심적인 역할을 수행한다. 이 모듈은 음성 수집 모듈, STT 변환 모듈, NLU 처리 모듈, 멀티모달 정보 분석 모듈, 위치 기반 추천 모듈로 이루어져 있다.

【0009】 음성 수집 모듈은 마이크를 통해 사용자의 음성 발화를 수집하여 PCM(Pulse-Code Modulation) 형식의 디지털 음성 데이터로 변환한다. 이 과정에서 대화형 AI 모델을 통해 발화의 시작과 끝을 자동으로 인식하고, 주변 잡음 제거 및 음성 증폭과 같은 전처리 과정을 거쳐 음성 인식의 정확도를 높인 후 STT 변환 모듈로 전달한다.

【0010】 STT(Speech-to-Text) 변환 모듈은 음성 수집 모듈에서 전달된 디지털 음성 데이터를 텍스트 데이터로 변환한다. 이 모듈은 Google STT, Whisper API 등 고성능 음성 인식 엔진을 활용하여 높은 정확도로 음성-텍스트 변환을 수행한다.

【0011】 NLU(Natural Language Understanding) 처리 모듈은 도 2의 음성인식 및 NLU 처리 흐름도에 따라 STT 모듈에서 변환된 텍스트를 입력받아, 사용자의 **의도(intent)** 및 **핵심 엔티티(entity)**를 추출한다. 이 모듈은 언어 감지 기능으로 다국어 입력을 처리하고, 개체명 인식(Named Entity Recognition, NER) 및 텍스트 정규화를 통해 비정형적인 자연어를 정확하게 이해한 다음, 최종적으로 의도를 '회원가입', '상품 검색', '위치 기반 서비스', '이미지 분석 서비스' 등으로 분류하여 해당 서비스 로직으로 연결한다.

【0011】 멀티모달 정보 분석 모듈은 사용자의 음성 외에 이미지 및 위치 정보와 같은 비음성 데이터를 분석하여 요청에 대한 맥락적 이해를 돕는다. 이 모듈은 도 3의 GPS 및 이미지 분석 통합 처리 흐름도에 따라 동작한다. 이 모듈 안에는 AI 이미지 분석 모듈이 있어 카메라로부터 수신된 이미지를 딥러닝 기반의 합성곱 신경망(Convolutional Neural Network, CNN) 모델로 분석한다. 이 모듈은 이미지 내의 객체를 인식하고, 특징점을 추출하며, 객체를 분류(예: 신발, 옷, 가전제품)하여 시각적 정보를 텍스트 기반 NLU 결과와 융합할 수 있도록 **특징 벡터(Feature Vector)**를 생성한다.

【0012】 또한, 위치 기반 추천 모듈이 포함되어 GPS 센서로부터 획득한 GPS 좌표(위도, 경도)를 처리한다. 이 모듈은 공간 인덱싱(Spatial Indexing) 기법(예: R-tree, K-D tree)을 활용하여 대규모 지리 정보 데이터베이스에서 사용자의 현재 위치를 기반으로 주변의 특정 서비스나 상품을 효율적으로 검색하고 추천하는 역할을 한다.

【0013】 서비스 실행 및 피드백 모듈은 핵심 처리 모듈에서 파악된 사용자의 의도 및 통합된 맥락 정보를 바탕으로 실제 서비스 로직을 실행하고, 그 결과를 사용자에게 자연스럽게 전달하여 상호작용을 완료한다.

【0014】 자동화 실행 모듈은 NLU 및 멀티모달 정보 분석 모듈에서 얻은 결과를 종합하여 사전에 정의된 서비스 로직을 자동으로 실행한다. 이 모듈은 외부 REST API 또는 GraphQL API 연동을 통해 결제 시스템, 배송 시스템, 소셜 로그인 등 다양한 외부 서비스와 유기적으로 연동된다.

【0014】 대화형 피드백 모듈은 도 4에 도시된 서비스 실행 및 피드백 절차도에 따라 챗봇 기술 및 TTS(Text-to-Speech) 기능을 통해 사용자와 실시간으로 양방향 상호작용을 수행한다. 서비스 실행 결과를 자연스러운 음성 또는 텍스트 형태로 사용자에게 전달하고, 필요한 경우 추가 질문을 통해 대화를 이어나가며, 사용자의 피드백을 다시 입력으로 받아 재처리하는 순환 구조를 갖는다.

【0015】 데이터 및 보안 계층은 플랫폼의 안정성과 확장성을 보장하며, 모든 서비스 운영에 필요한 데이터를 안전하게 저장하고 관리한다. 이 계층에는 사용자 정보(프로필, 선호도), 상

품 정보, 위치 정보 등을 저장하고 관리하는 데이터베이스와 사용자 인증, 권한 관리, 데이터 암호화 등 전반적인 시스템 보안을 담당하는 보안 인증 모듈이 포함된다.

【0016】 또한, 다양한 언어의 음성 명령을 처리하고 응답할 수 있도록 다국어 NLU 모델 및 TTS 음성 데이터를 관리하는 다국어 지원 모듈이 있고, 사용자의 과거 이용 기록, 선호도, 개인 설정 등을 저장 및 학습하여 개인화된 서비스 제공의 기반을 마련하는 사용자 프로필 관리 모듈도 이 계층에 속한다.

【0014】

본 발명의 핵심적인 특징은 상기 NLU 모듈에서 추출된 의도 및 엔티티와 상기 멀티모달 정보 분석 모듈에서 분석된 시각 및 위치 정보를 유기적으로 '융합'하여 사용자의 단일한 요청 맥락(request context)을 재구성하고, 이를 기반으로 서비스 실행 모듈이 복합적인 서비스 로직을 수행하는 데 있다.

【발명(고안)의 효과】

【0015】

본 발명은 기존의 정형화된 명령어 인식 방식을 넘어선 고도화된 NLU 기술을 통해 사용자의 복잡하고 비정형적인 자연어 발화로부터 의도를 정확히 파악함으로써, 사용자의 인지적 부하를 현저히 감소시키고 서비스의 효율성과 전반적인 사용자 경험(UX)을 획기적으로 향상시키는 효과가 있다.

【0016】

또한 본 발명은 음성, 위치, 이미지 등 다양한 멀티모달 데이터를 실시간으로 융합하여 분석함으로써, 단순히 사용자가 '무엇을' 요청하는지를 넘어 '어디서', '무엇을 보면서' 요청하는지까지 종합적으로 이해하여 사용자의 현재 상황과 개인 선호도에 최적화된 하이퍼-개인화(hyper-personalized)된 추천 및 맞춤형 서비스 제공이 가능하다.

【0017】

나아가 본 발명은 손을 사용하기 어려운 운전 중이나 요리 중에도 음성만으로 서비스를 완벽하게 제어할 수 있는 강력한 핸즈프리(Hands-free) 인터페이스를 제공하여 사용자 편의성을 극대화하며, 시각장애인이나 운동장애인 등 신체적 제약이 있는 사용자에게도 디지털 서비스 접근성을 크게 향상시키는 사회적 효과를 제공한다.

【0018】

뿐만 아니라 본 발명은 이미지 검색 시 고차원 특징 벡터 비교를 위해 코사인 유사도(Cosine Similarity)를, 위치 기반 검색 시 R-트리(R-tree) 공간 인덱싱을 사용하는 등 특정 과업에 최적화된 알고리즘을 채택함으로써, 복잡한 멀티모달 쿼리에 대해 신속하고 정확한 검색 결과를 보장하는 기술적 효율성을 달성한다.

【도면의 간단한 설명】

【0019】

【도 1】 본 발명의 일 실시예에 따른 음성인식 기반 멀티모달 상호작용 플랫폼의 전체 시스템 아키텍처를 도시하는 블록 다이어그램이다. 사용자의 입력부터 서비스 실행 및 피드백까지의 각 구성 요소와 계층 간의 관계를 보여준다.

【도 2】 본 발명에 따른 음성인식 및 자연어 이해(NLU) 처리 흐름을 상세하게 보여주는 흐름도이다. 음성 데이터를 텍스트로 변환하고, 의도와 엔티티를 추출하여 서비스 로직을 결정하는 일련의 과정을 나타낸다.

【도 3】 본 발명에 따른 GPS 및 이미지 분석 통합 처리 흐름을 상세하게 보여주는 흐름도이

다. 음성 입력과 함께 들어온 위치 정보 및 이미지 정보가 어떻게 NLU 결과와 융합되어 최종적인 서비스 요청을 완성하는지를 보여준다.

【도 4】 본 발명의 실시예 1에 따른 서비스 실행 및 피드백 절차를 나타내는 시퀀스 다이어그램이다. 시스템과 사용자가 주고받는 양방향 대화의 흐름을 단계별로 보여준다.

【발명(고안)을 실시하기 위한 구체적인 내용】

【0020】

이하, 본 발명의 바람직한 실시예들을 첨부된 도면을 참조하여 상세하게 설명한다. 하기의 설명에서는 본 발명이 속하는 기술 분야에 널리 알려져 있고 본 발명과 직접적으로 관련이 없는 기술 내용에 대해서는 설명을 생략한다. 이는 불필요한 설명을 생략함으로써 본 발명의 요지를 흐리지 않고 더욱 명확하게 전달하기 위함이다. 명세서 전체에서 동일한 참조 부호는 동일한 구성 요소를 나타낸다.

【0021】

도 1은 본 발명의 일 실시예에 따른 음성인식 기반 멀티모달 상호작용 플랫폼(100)의 전체 시스템 아키텍처를 도시하는 블록 다이어그램이다. 도 1을 참조하면, 본 발명에 따른 플랫폼(100)은 크게 사용자 인터페이스 계층(110), 핵심 처리 모듈(120), 서비스 실행 및 피드백 모듈(130), 그리고 데이터 및 보안 계층(140)의 4개 계층으로 구성될 수 있다.

【0022】

사용자 인터페이스 계층(110)은 사용자와 플랫폼(100) 간의 상호작용을 위한 접점 역할을 수행한다. 이 계층은 음성 입력부(111)(예: 마이크), 이미지 입력부(112)(예: 카메라), 및 위치 정보 수집부(113)(예: GPS 센서)와 같은 다양한 입력 장치들을 포함한다. 사용자의 음성, 이미지, 위치 데이터와 같은 원시 데이터(raw data)는 이 계층을 통해 수집되어 핵심 처리 모듈(120)로 전달된다.

【0023】

핵심 처리 모듈(120)은 사용자 인터페이스 계층(110)으로부터 수신된 원시 데이터를 가공하고 사용자의 통합적인 의도를 정확하게 파악하는 핵심적인 역할을 수행한다. 핵심 처리 모듈(120)은 STT 변환 모듈(124), 자연어 이해(NLU) 처리 모듈(123), AI 이미지 분석 모듈(121), 및 위치 기반 추천 모듈(122)을 포함할 수 있다.

【0024】

STT 변환 모듈(124)은 음성 입력부(111)를 통해 수집된 사용자의 아날로그 음성 발화를 PCM(Pulse-Code Modulation) 형식의 디지털 음성 데이터로 변환하고, 이를 다시 텍스트 데이터로 변환한다. 이 과정에서 배경 잡음 제거(noise cancellation), 음성 증폭(audio amplification) 등과 같은 전처리 과정을 거쳐 인식 정확도를 향상시킬 수 있다.

【0025】

NLU 처리 모듈(123)은 STT 변환 모듈(124)로부터 전달받은 텍스트를 분석하여 사용자의 핵심 의도(intent)와 그 의도를 구체화하는 핵심 개체(entity)를 추출한다. 예를 들어, "내 주변 축구화 찾아줘"라는 텍스트에서 의도는 '상품 검색', 엔티티는 '축구화', '내 주변'으로 식별된다.

【0026】

AI 이미지 분석 모듈(121)은 이미지 입력부(112)로부터 수신된 이미지를 분석하여 시각적 정보를 추출한다. 본 발명의 일 실시예에서, 이 모듈은 딥러닝 기반의 합성곱 신경망(Convolutional Neural Network, CNN)을 구현한다. CNN은 이미지의 픽셀 데이터로부터 의

미 있는 고차원 표현인 특징 벡터(Feature Vector)를 추출하는 데 매우 효과적이다. 이 과정은 다음과 같은 단계로 이루어진다. 첫째, 복수의 합성곱 계층(Convolutional Layers)이 이미지 위를 이동하며 에지, 질감, 색상 등과 같은 저수준 특징(low-level feature)을 감지한다. 이때, 동일한 필터(filter)의 가중치를 이미지 전체 영역에서 공유하는 '가중치 공유(weight sharing)' 메커니즘을 통해 객체의 위치 변화에 강인한 특징 추출이 가능하다. 둘째, 풀링 계층(Pooling Layers)은 특징 맵(feature map)의 차원을 축소하여 계산량을 줄이고, 미세한 위치 변화에 대한 불변성(invariance)을 확보한다. 마지막으로, 완전 연결 계층(Fully Connected Layer)을 통해 추출된 고수준 특징들을 조합하여 최종적으로 이미지를 대표하는 고차원의 수치 벡터, 즉 특징 벡터를 생성한다. 이 특징 벡터는 이미지의 의미적 내용을 압축적으로 표현하므로, 유사 이미지 검색의 기반이 된다.

【0027】

위치 기반 추천 모듈(122)은 위치 정보 수집부(113)로부터 획득한 GPS 좌표(위도, 경도)를 처리한다. "내 주변"과 같은 위치 기반 요청을 효율적으로 처리하기 위해, 본 발명의 일 실시예에서 이 모듈은 R-트리(R-tree)와 같은 공간 인덱스(spatial index)를 구현한다. R-트리는 다수의 지리적 위치 데이터를 계층적인 트리 구조로 구성하는 데이터 구조이다. 각 노드는 자신의 자식 노드들이 포함하는 모든 공간 객체들을 감싸는 최소 경계 사각형(Minimum Bounding Rectangle, MBR) 정보를 가진다. 특정 좌표 주변의 객체를 검색할 때, 알고리즘은 트리의 루트 노드부터 시작하여 쿼리 영역과 겹치는(overlap) MBR을 가진 자식 노드만을 재귀적으로 탐색한다. 쿼리 영역과 겹치지 않는 MBR에 해당하는 하위 트리는 탐색에서 제외(pruning)되므로, 전체 데이터를 선형적으로 스캔하는 방식에 비해 검색 속도를 획기적으로 향상시킬 수 있다.

【0028】

서비스 실행 및 피드백 모듈(130)은 핵심 처리 모듈(120)에서 파악된 사용자의 의도 및 융합된 맥락 정보를 바탕으로 실제 서비스 로직을 실행하고, 그 결과를 사용자에게 전달한다. 이 모듈은 자동화 실행 모듈(131)과 대화형 피드백 모듈(132)을 포함한다. 자동화 실행 모듈(131)은 외부 REST API 또는 GraphQL API 연동을 통해 결제, 배송, 소셜 로그인 등 다양한 외부 서비스와 연동하여 실제 기능을 수행한다. 대화형 피드백 모듈(132)은 챗봇 기술 및 TTS(Text-to-Speech) 기능을 통해 서비스 실행 결과를 자연스러운 음성 또는 텍스트 형태로 사용자에게 전달하고, 필요 시 추가 질문을 통해 대화를 이어간다.

【0029】

데이터 및 보안 계층(140)은 플랫폼의 안정성과 확장성을 보장하며, 서비스 운영에 필요한 데이터를 안전하게 저장하고 관리한다. 이 계층은 사용자 정보, 상품 정보, 위치 정보 등을 저장하는 데이터베이스(141)와 사용자 인증, 권한 관리, 데이터 암호화 등을 담당하는 보안 인증 모듈(142)을 포함한다.

【0030】

이하, 본 발명의 구체적인 실시예들을 도 2 내지 도 4를 참조하여 더욱 상세히 설명한다.

【0031】

【실시예 1】 음성 기반 회원가입 서비스 제공 방법

본 실시예는 사용자의 음성 발화로부터 필요한 회원 정보를 단계적으로 추출하여 회원가입을 완료하는 방법에 관한 것이다. 이 과정은 도 4에 제시된 시퀀스 다이어그램을 통해 설명될 수 있다. 먼저, 사용자가 "회원가입 할게"라고 음성 발화하면, 음성 데이터가 STT 변환 모듈(124)

로 전송되어 텍스트로 변환된다. NLU 처리 모듈(123)은 이로부터 의도를 '회원가입'으로 식별하고, 회원가입에 필요한 핵심 엔티티인 이메일, 이름, 비밀번호가 누락되었음을 파악한다. 이어서, 대화형 피드백 모듈(132)은 TTS를 통해 "이메일 주소를 말해주십시오"라고 사용자에게 음성으로 안내하고, 사용자는 자신의 이메일 주소를 발화한다. 사용자가 "jaeman 골뱅이 gmail 점 컴"이라고 발화하면, 시스템은 동일한 STT 및 NLU 과정을 거쳐 텍스트 "jaeman@gmail.com"으로 변환하고 이를 '이메일' 엔티티로 식별하여 데이터베이스(141)에 임시 저장한다. 이와 같은 대화형 과정을 반복하여 이름과 비밀번호를 순차적으로 수집한 후, 모든 필수 정보가 수집되면 자동화 실행 모듈(131)이 데이터베이스(141)에 최종 계정 생성을 요청한다. 마지막으로, 대화형 피드백 모듈(132)은 "회원가입이 완료되었습니다"와 같이 사용자에게 완료 메시지를 전달함으로써 절차가 마무리된다.

【0032】

【실시예 2】 위치 기반 상품 추천 서비스 제공 방법

본 실시예는 사용자의 음성 명령과 GPS 정보를 융합하여 주변 상품을 추천하는 방법에 관한 것이다. 사용자가 "내 주변 축구화 찾아줘"라고 음성 발화하면, NLU 처리 모듈(123)은 '상품 검색' 의도와 '축구화', '내 주변' 엔티티를 파악한다. '내 주변' 엔티티를 인식한 시스템은 위치 기반 추천 모듈(122)을 활성화하여 사용자 단말기의 GPS 좌표를 획득한다. 획득된 GPS 좌표는 위치 기반 추천 모듈(122) 내의 R-트리 공간 인덱스를 탐색하는 쿼리 포인트로 사용된다. R-트리 탐색 알고리즘은 사용자의 현재 위치를 중심으로 설정된 반경과 겹치는 MBR을 가진 노드들만을 따라 하향식으로 탐색을 진행하여, 해당 반경 내에 위치한 축구화 판매처 목록을 효율적으로 검색한다. 최종적으로 검색된 판매처 목록은 '거리', '평점', '재고 유무' 등에 가중치를 부여하는 알고리즘을 통해 순위가 결정되며, 대화형 피드백 모듈(132)은 "고객님의 위치에서 500미터 이내에 있는 'A스포츠 매장'에서 '나이키 팬텀' 축구화를 판매하고 있습니다"와 같이 자연스러운 음성으로 안내한다.

【0033】

【실시예 3】 이미지 기반 상품 검색 서비스 제공 방법

본 실시예는 사용자의 음성 명령과 이미지를 결합하여 유사 상품을 검색하는 방법에 관한 것이다. 사용자가 스마트폰 카메라로 특정 신발을 촬영하며 "이 신발이랑 비슷한 상품 찾아줘"라고 음성 명령을 내리면, NLU 처리 모듈(123)은 '유사 상품 검색' 의도를 파악한다. 동시에, 촬영된 신발 이미지는 AI 이미지 분석 모듈(121)로 전달된다. AI 이미지 분석 모듈(121)은 내장된 CNN 모델을 통해 이미지를 처리하여, 신발의 디자인, 색상, 형태 등 시각적 특징을 함축하는 고차원 특징 벡터를 추출한다.

서비스 실행 모듈(131)은 추출된 특징 벡터를 사용하여 데이터베이스(141) 내의 모든 상품 이미지 특징 벡터와 유사도를 비교한다. 고차원 벡터 공간에서는 유클리드 거리(Euclidean Distance)보다 방향성에 집중하는 코사인 유사도(Cosine Similarity)가 의미적 유사성을 더 잘 측정하는 경향이 있다. 코사인 유사도는 두 벡터 사이의 각도의 코사인 값으로 계산되며, 그 공식은 다음과 같다:

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|}$$

여기서 A와 B는 비교하고자 하는 두 특징 벡터이다. 유사도 값은 -1에서 1 사이의 값을 가지며, 1에 가까울수록 두 이미지가 의미적으로 유사함을 나타낸다. 서비스 실행 모듈(131)은

이 코사인 유사도를 계산하여 유사도가 가장 높은 상품들을 우선적으로 정렬하고, 그 결과를 대화형 피드백 모듈(132)을 통해 "찾고 계신 신발과 유사한 디자인의 'B브랜드' 'X모델'이 있습니다"와 같이 음성으로 제공하거나 시각적으로 제시한다.

【0034】

【실시예 4】 통합 멀티모달 정보를 활용한 복합 서비스 제공 방법

본 실시예는 음성, 이미지, 위치 정보를 모두 활용하여 복합적인 사용자 요청을 처리하는 방법에 관한 것이다. 사용자가 "이 사진 속 옷이랑 어울리는 신발을, 지금 내가 있는 곳 근처 매장에서 찾아줘"라고 발화하며 스마트폰으로 사진을 업로드하면, NLU 처리 모듈(123)은 음성으로부터 '상품 추천' 의도, '옷', '신발' 엔티티, '지금 내가 있는 곳 근처'라는 위치 맥락을 파악한다. 동시에 AI 이미지 분석 모듈(121)은 사진 속 옷의 스타일, 색상 등 특징을 추출하여 특징 벡터를 생성하고, 위치 기반 추천 모듈(122)은 사용자의 현재 GPS 위치를 파악한다. 이와 같이 음성, 이미지, 위치로부터 추출된 모든 데이터는 '멀티모달 융합' 단계를 거쳐 (의도: 상품 추천, 대상: 신발, 조건1: [옷 특징 벡터], 조건2:)와 같은 하나의 통합된 사용자 요청 맥락으로 재구성된다. 서비스 실행 모듈(131)은 이 재구성된 맥락 정보를 기반으로, 사진 속 옷의 스타일과 조화를 이루면서(조건1), 사용자의 현재 위치 근처에 있는 매장에서(조건2) 재고가 있는 신발을 검색하는 복합적인 로직을 실행한다. 최종적으로 대화형 피드백 모듈(132)은 "고객님의 현재 위치에서 2km 떨어진 'C백화점' 신발 코너에서 사진 속 옷과 어울리는 'Y 신발'이 판매 중입니다"와 같이 상세한 정보를 음성 및 시각적 이미지로 사용자에게 제공한다.

【부호의 설명】

【0035】

100: 음성인식 기반 멀티모달 상호작용 플랫폼
110: 사용자 인터페이스 계층
111: 음성 입력부
112: 이미지 입력부
113: 위치 정보 수집부
120: 핵심 처리 모듈
121: AI 이미지 분석 모듈
122: 위치 기반 추천 모듈
123: 자연어 이해(NLU) 처리 모듈
124: STT 변환 모듈
130: 서비스 실행 및 피드백 모듈
131: 자동화 실행 모듈
132: 대화형 피드백 모듈
140: 데이터 및 보안 계층
141: 데이터베이스
142: 보안 인증 모듈

【산업상 이용가능성】

【0036】

본 발명에 따른 음성인식 기반 멀티모달 상호작용 플랫폼 및 그 운영 방법은 다양한 산업 분야에서 광범위하게 적용 가능하며, 사용자 경험을 혁신하고 새로운 비즈니스 기회를 창출할 수 있다.

【0037】
전자상거래 분야에서는 음성으로 상품을 검색, 결제, 관리하고, 이미지와 위치 정보를 결합한 개인화된 상품 추천을 통해 쇼핑 편의성을 극대화할 수 있다. 자동차 인포테인먼트 시스템 분야에서는 운전 중 시선을 전방에 고정한 채 음성만으로 내비게이션, 미디어 재생, 통화 등을 안전하게 제어할 수 있다.
【0038】
스마트홈 및 사물인터넷(IoT) 분야에서는 음성 명령과 이미지 인식을 결합하여 조명, 에어컨, 로봇청소기 등 스마트 가전을 직관적으로 제어하고 상태를 확인할 수 있다. 웨어러블 디바이스 분야에서는 스마트 워치나 AR/VR 글래스와 같이 작은 화면의 제약을 극복하고 음성 명령으로 효율적인 조작과 정보 획득을 가능하게 한다.
【0039】
또한, 공공 서비스 및 관광 산업에서는 다국어 음성 안내 시스템을 구축하여 외국인 관광객이나 언어 소통이 어려운 사용자에게 맞춤형 정보를 제공할 수 있으며, 의료 및 헬스케어 분야에서는 의료 기록의 음성 기반 입력 및 조회, 약물 정보 안내 등에 활용하여 의료진의 업무 효율성을 높이고 환자의 접근성을 향상시킬 수 있다.
【청구범위】
【청구항 1】
사용자의 음성 발화를 수집하여 텍스트로 변환하는 음성 인식 모듈;
상기 텍스트로부터 사용자의 의도 및 엔티티를 추출하는 자연어 이해(NLU) 모듈;
사용자 단말기의 GPS 정보 또는 사용자가 제공하는 이미지를 분석하는 멀티모달 정보 분석 모듈;
상기 NLU 모듈에서 추출된 의도 및 엔티티와 상기 멀티모달 정보 분석 모듈에서 분석된 정보를 융합하여 서비스 로직을 실행하는 서비스 실행 모듈; 및
상기 실행된 서비스 로직의 결과를 사용자에게 자연어 음성 또는 텍스트로 전달하는 대화형 피드백 모듈;을 포함하는 것을 특징으로 하는 멀티모달 상호작용 플랫폼.
【청구항 2】
제1항에 있어서, 상기 NLU 모듈은 사용자 발화의 의도 분류 및 엔티티 추출을 수행하며, 언어 감지 및 텍스트 정규화 기능을 포함하는 것을 특징으로 하는 멀티모달 상호작용 플랫폼.
【청구항 3】
제1항에 있어서, 상기 멀티모달 정보 분석 모듈은, 입력된 이미지로부터 합성곱 신경망(Convolutional Neural Network, CNN)을 이용하여 의미적 특징을 포함하는 특징 벡터를 추출하는 것을 특징으로 하는 멀티모달 상호작용 플랫폼.
【청구항 4】
제1항에 있어서, 상기 멀티모달 정보 분석 모듈은, 수신된 GPS 좌표를 R-트리(R-tree)로 구현된 공간 인덱스에서 검색하여 상기 GPS 좌표에 인접한 지리 정보를 효율적으로 식별하는 것을 특징으로 하는 멀티모달 상호작용 플랫폼.
【청구항 5】
제1항에 있어서, 상기 서비스 실행 모듈은 상기 의도, 엔티티, 및 멀티모달 데이터에 따라 데이터베이스 조회, 외부 API 호출, 또는 시스템 제어 중 적어도 하나 이상의 동작을 수행하는 것을 특징으로 하는 멀티모달 상호작용 플랫폼.
【청구항 6】

제1항에 있어서, 상기 대화형 피드백 모듈은 텍스트-음성 변환(TTS) 기능과 챗봇 기능을 포함하여 사용자와 실시간 대화 상호작용을 수행하고, 필요한 경우 후속 질문을 통해 추가 정보를 획득하는 것을 특징으로 하는 멀티모달 상호작용 플랫폼.

【청구항 7】

컴퓨터에 의해 실행되는 멀티모달 상호작용 방법으로서,

(a) 사용자로부터 음성 및 이미지를 포함하는 복합 요청을 수신하는 단계;

(b) 상기 음성 정보를 텍스트로 변환하고, 자연어 이해(NLU)를 통해 상기 텍스트로부터 사용자의 의도 및 엔티티를 분석하는 단계;

(c) 상기 이미지 정보를 합성곱 신경망(CNN)을 사용하여 시각적 특징 벡터로 추출하는 단계;

(d) 상기 (b)단계의 의도 및 엔티티와 상기 (c)단계의 시각적 특징 벡터를 융합하여 통합 요청 맥락을 생성하는 단계;

(e) 상기 통합 요청 맥락을 바탕으로 서비스 로직을 실행하여 상품 데이터베이스에서 유사 상품을 검색하는 단계; 및

(f) 상기 검색된 유사 상품 정보를 사용자에게 음성 또는 시각적으로 제공하는 단계를 포함하는 것을 특징으로 하는 멀티모달 상호작용 방법.

【청구항 8】

제7항에 있어서, 상기 (e)단계의 유사 상품을 검색하는 단계는, 상기 추출된 시각적 특징 벡터와 상품 데이터베이스 내의 상품별 특징 벡터 간의 코사인 유사도(Cosine Similarity)를 산출하여 유사도가 기설정된 임계값 이상인 상품을 식별하는 것을 특징으로 하는 멀티모달 상호작용 방법.

【청구항 9】

제7항에 있어서, 상기 (a)단계의 복합 요청이 사용자의 위치 정보를 더 포함하는 경우, 상기 (e)단계의 서비스 로직 실행은 R-트리(R-tree) 공간 인덱스를 이용하여 사용자의 현재 위치 근처에 있는 매장을 검색하는 단계를 더 포함하는 것을 특징으로 하는 멀티모달 상호작용 방법.